Big Data Analytics for Improving Customer Win-back Rate in Townhome segment

Mr. Warot Lilahajiva

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Engineering Management
(CU-Warwick)
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

การวิเคราะห์ข้อมูลขนาดใหญ่เพื่อเพิ่มอัตราการเรียกลูกค้าเก่ากลับมาในธุรกิจทาวน์โฮม

นายวรทย์ ลีฬหาชีวะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาการจัดการทางวิศวกรรม ศูนย์ระดับภูมิภาคทางวิศวกรรมระบบการผลิต
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Thesis Title	Big Data Analytics for Improving Customer Win-back Rate in Townhome segment
By	Mr. Warot Lilahajiva
Field of Study	Engineering Management
Thesis Advisor	Assistant Professor PISIT JARUMANEEROJ, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

.... Chairman
(Professor PARAMES CHUTIMA, Ph.D.)

.... Thesis Advisor
(Assistant Professor PISIT JARUMANEEROJ, Ph.D.)

.... Examiner
(Associate Professor JEERAPAT NGAOPRASERTWONG)

.... External Examiner
(Associate Professor Chuvej Chansa-ngavej, Ph.D.)

วรทย์ ลีพหาชีวะ : การวิเคราะห์ข้อมูลขนาดใหญ่เพื่อเพิ่มอัตราการเรียกลูกค้าเก่ากลับมาในธุรกิจทาวน์โฮม. (Big Data Analytics for Improving Customer Win-back Rate in Townhome segment) อ.ที่ปรึกษาหลัก : ผศ. ดร.พิศิษฎ์ จารุมณีโรจน์

อุตสาหกรรมอสังหาริมทรัพย์ไทยในปัจจุบันกำลังเผชิญหน้ากับการเติบโตที่ติดลบ (Negative growth) ส่งผลให้ผู้พัฒนาอสังหาริมทรัพย์ (Developers) ต่าง ๆ พบกับความยากลำบากในการสร้างรายให้ได้ตรงตามเป้าหมายที่ตั้งใจไว้ โดยเฉพาะจากกลุ่มลูกค้าใหม่ การกระตุ้นลูกค้าเก่า (Customer win-back) ที่มุ่งเน้นกลุ่มลูกค้าที่เคยเข้ามาในกรวยการขาย (Sales funnel) ผ่านข้อเสนอที่น่าดึงดูดมากขึ้นจึงนับเป็นหนึ่งในแนวทางการสร้างรายได้ที่น่าสนใจ โดยในการศึกษานี้ ผู้วิจัยจะมุ่งเน้นไปที่การพัฒนา Predictive analytics model ในการเพิ่มอัตรา Customer win-back ซึ่งการวิเคราะห์ต่าง ๆ จะอ้างอิงอยู่บนฐานข้อมูลของลูกค้าในปัจจุบัน และผลการดำเนินการขายที่ผ่านมา ทั้งนี้ผลลัพธ์ของตัวแบบที่ได้ หรือ Propensity model จะช่วยจัดลำดับความน่าจะเป็นของลูกค้าแต่ละรายในการกลับมาซื้ออสังหาริมทรัพย์เมื่อได้รับการติดต่อโดยพนักงานขาย ผู้วิจัยได้ทำการทดสอบวิธีการดังกล่าวผ่านวิธีการทดสอบแบบ A/B Test เป็นเวลา 3 สัปดาห์ จากนั้นจึงนำผลลัพธ์ที่ได้จากวิธีการดังกล่าวมาเปรียบเทียบกับวิธีการในปัจจุบัน ผ่านดัชนีการเปรียบเทียบ 2 ดัชนี ได้แก่ (1) จำนวนการส่งต่อลูกค้าไปยังฝ่ายขายโดยเฉลี่ย (2) อัตราการกลับมาของลูกค้าโดยเฉลี่ย จากการศึกษา ผู้วิจัยพบว่า Propensity model ที่ได้สามารถช่วยเพิ่มจำนวนการส่งต่อลูกค้าไปยังฝ่ายขายโดยเฉลี่ยประมาณ 11.8% ในขณะที่สามารถเพิ่มอัตราการกลับมาของลูกค้าได้ถึง 13.5% ซึ่งเทียบได้กับความสามารถในการเพิ่มรายได้ 20.9% หรือคิดเป็นมูลค่ากว่า 253.5 ล้านบาท

สาขาวิชา	การจัดการทางวิศวกรรม	ลายมือชื่อนิสิต ....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6071223021 : MAJOR ENGINEERING MANAGEMENT

KEYWORD: Big Data Analytics for Improving Customer Win-back Rate in Townhome segment

Warot Lilahajiva : Big Data Analytics for Improving Customer Win-back Rate in Townhome segment. Advisor: Asst. Prof. PISIT JARUMANEEROJ, Ph.D.

Thailand’s real-estate market is now facing with negative growth, where most developers are encountered with challenges in generating satisfactory revenue, particularly from new customers. In order to improve the revenue stream, customer win-back approach that focuses on current customers in the sales funnel and re-engages them by offering more attractive residential projects is therefore initiated. In particular, we focus on the development of predictive analytics model that potentially increases customer win-back rate by exploring the current databases, identifying significance for win-back (based on past performance), and testing the resulting predictions via several machine learning algorithms. The proposed method returns a propensity model that ranks customers based on likelihood to purchase if contacted. We test the proposed method by running an A/B test for 3 weeks and then compare (1) average number of refer cases to sales department and (2) average customer referral rate with the base case. We find that the propensity model impressively helps increase average number of refer cases to sales department by 11.8% and increase referral rate by 13.5%. In terms of finance, these significant improvements lead to a revenue uplift of 20.9% year-on-year, valued at THB 253.5 million.

Field of Study:	Engineering Management	Student's Signature ....
Academic Year:	2019	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincerest gratitude to the following of esteemed individuals, who have each contributed greatly to the success of this dissertation.

First and foremost, I would like to express my appreciation to Asst. Prof. Dr. Pisit Jarumaneeroj who provided support, guidance and showed understanding of my difficult situation in balancing a demanding career with a part-time postgraduate degree.

Secondly, I would like to express my utmost gratitude to Mr. Pochara Arayakarnkul, CEO of Bluebik Group, who has supported me endlessly into leadership at the company while providing me with all the support I needed to complete my postgraduate degree. Most importantly, thanks to the approval of my request to extend the company’s project into a dissertation and an academic study.

I would also like to provide a special thanks to the members of Ingenio team, a subsidiary of Bluebik Group, who provided technical knowledge and took time to help me understand more into Big data analytics.

Moreover, I would like to show my sincerest gratitude to the lecturers of WMG who have flown into Bangkok to provide comprehensive 5 full-day lectures. The knowledge and experiences shared during these sessions have been eye-opening and educative.

Furthermore, I would like to show great appreciation to the cohort members during my time at Chulalongkorn University and WMG joint-degree who have made the experience a joyful memory.

Lastly, I would like to express my utmost gratitude to my family and Ms. Yotsawadee Kittisutiphan, who have supported me throughout my postgraduate degree journey.

Warot Lilahajiva

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. Introduction

1.1 Background

1.1.1 Thailand's Residential Market Overview

The Residential market in Thailand has remained relatively stagnant between 2013 and 2017 with a slight negative compound annual growth rate (CAGR) of 0.54% (as shown in Figure 1). This is largely driven by rising housing debts which cause banks to impose stricter mortgage policies, thereby rendering consumer's affordability for homes (Pruksa, 2017).

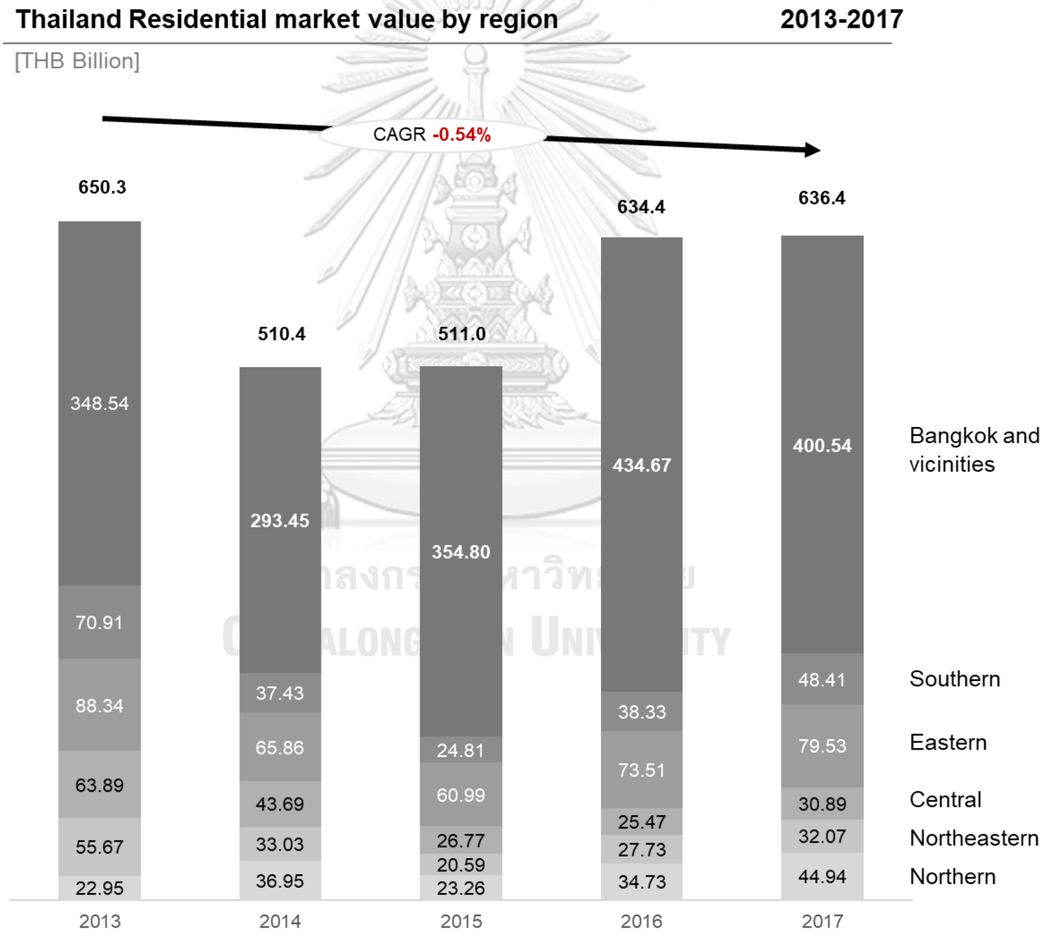


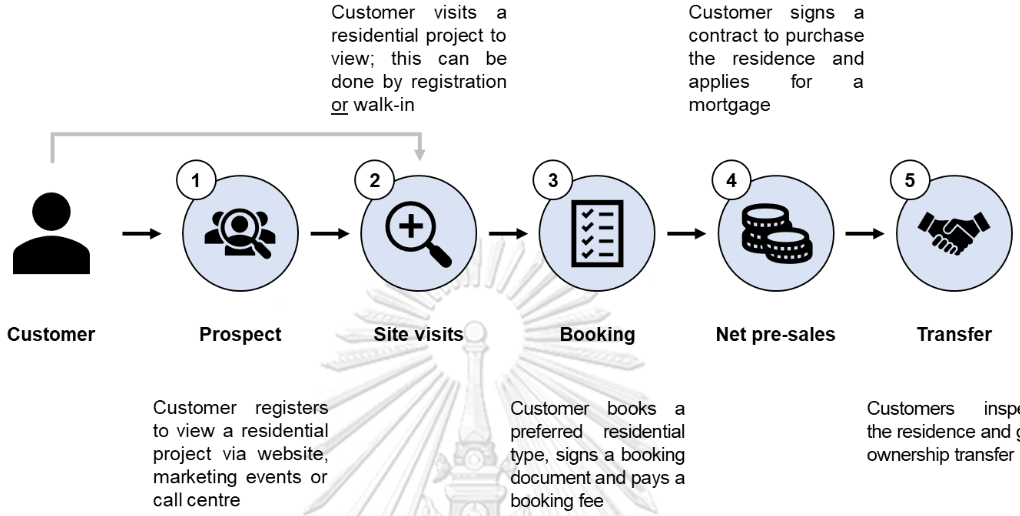
Figure 1: Residential market value in Thailand by region (2013–2017)

Source: Pruksa Annual report (2013 – 2017)

1. Introduction

Real-estate Sales Journey and Customer Win-back

The Real-estate sales journey encompasses 5 main steps, spreading across more than 30 days for the customer to complete the purchase, from registering to view the residences of interest through to inspect and transfer ownership (Figure 2).



Customer visits a residential project to view; this can be done by registration or walk-in

Customer signs a contract to purchase the residence and applies for a mortgage

Customer registers to view a residential project via website, marketing events or call centre

Customer books a preferred residential type, signs a booking document and pays a booking fee

Customers inspect the residence and get ownership transfer

Figure 2: Real-estate sales journey

1) Prospect – Customers show interests through registration to visit to view a given residential project(s)
2) Site visits – Customers visit a project(s) and fill in detailed demographic information, e.g. full name, income range, residential type preferences, family size, etc.
3) Booking – Customers book a residence and pay a booking fee; this often occurs at the end of the Site visits stage
4) Net pre-sales – Customers sign contract for the purchase of a residence and a pay a down payment fee. This stage usually sees an application for mortgages to real-estate developer’s partner banks.
5) Transfer – Customers inspect the residence and have ownership transferred from the real-estate developer

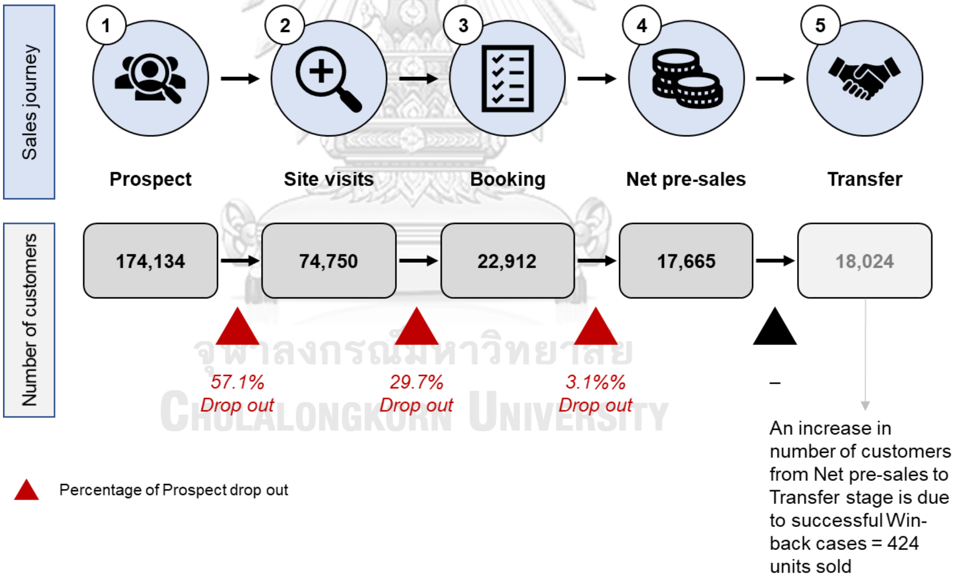
The real-estate developer, hereafter will be referred to as “The Company” experiences substantial customer drop out along the sales journey (as shown in Figure 3). To cope with the opportunity loss, the Company has established a specialised department called Customer win-back which has a similar function to Sales department. However, the sales are focused on customers who are tagged as

1. Introduction

“Opportunity lost” or customers who enter sales journey and tagged as defected, i.e. shown as “Win-back” in the Company customer relationship management system (CRM). The Win-back staff are tasked with outbound calls to customers who left the sales journey and gain insight into what the customers truly need or why the customers have not made any decisions. Afterward, the Win-back staff will provide an alternative offer which match the needs of the customers. If the customers are interested in re-entering the sales journey then Win-back staff will assign sales representatives from Sales department to service the customers further.

To summarise, Win-back department retargets customers who exit the sales journey and transfer them to Sales department for new offers. The main metric is Customer win-back rate (% refer):

$$Customer\ win - back\ rate = \frac{Number\ of\ cases\ referred\ to\ Sales\ department}{Number\ of\ outbound\ calls}$$



An increase in number of customers from Net pre-sales to Transfer stage is due to successful Win-back cases = 424 units sold

Figure 3: The Company’s customer drop-out rates along the sales journey

1. Introduction

1.2 Problem Statement

The Company is experiencing a relatively stagnant revenue between 2013 and 2017, in which it has experienced 2 consecutive years of falling revenue (Figure 4). This is driven from macroeconomic constraints and demand from consumers.

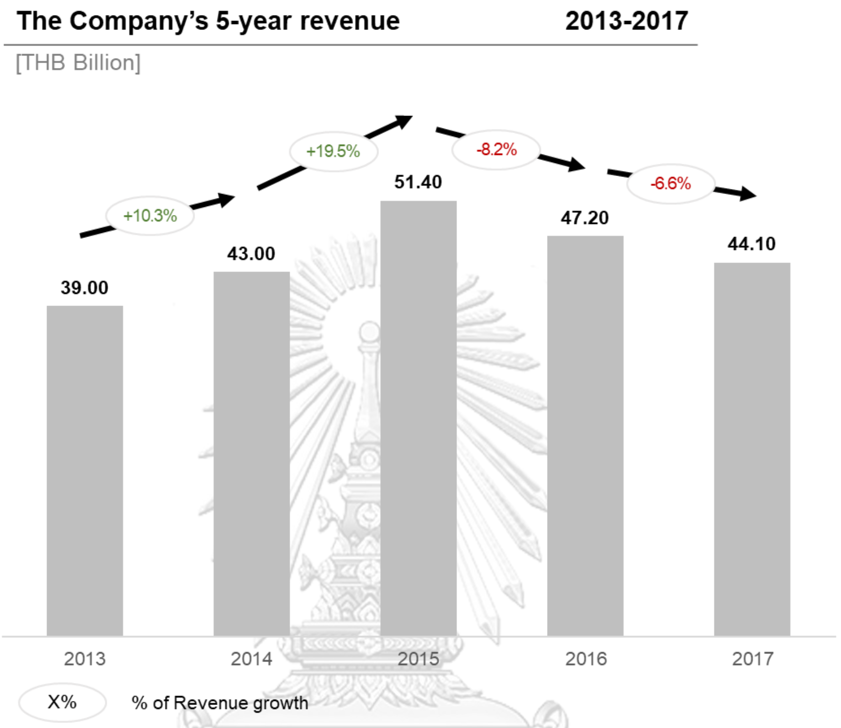


Figure 4: The Company's 5-year revenue trend (2013–2017)

A possible lever that the Company chose to explore is Customer win-back, a type of retention effort aimed at contacting back customers or leads that have entered the sales journey and left. The fundamental considers how these customers and leads have provided a set of data that can be leveraged to offer more attractive offers. However, this approach has not materialised.

Currently, the Company experiences approximately 10% of conversion from customers entering the sales journey through to transfer stage as shown in Figure 5. This represents a substantial opportunity loss.

The Company sees over 150,000 customers that leave the sales journey. All of which can be targeted for Win-back. However, Win-back department has limited capacity of 10 Win-back staff and can handle under 50,000 calls per year. Without a robust model for data utilisation, Customer win-back performance is likely to remain the same as previous year and the revenue is likely to decrease further.

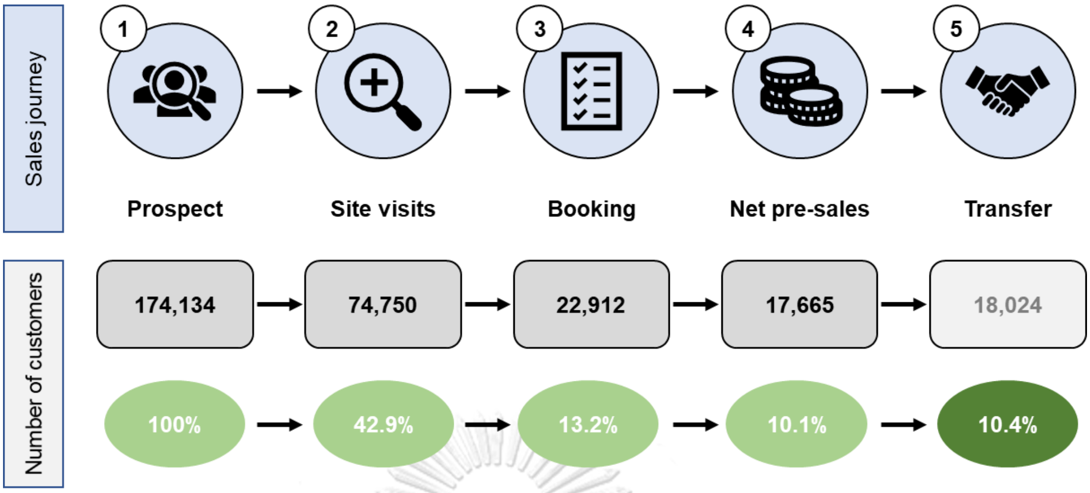


Figure 5: The Company's customer conversion

1.3 Research Objectives

The Company has engaged Blluebik Group Co., Ltd. to develop and implement Big data analytics model to increase Customer win-back rate which will ultimately increase revenue. The main objectives of this research are to:

1) Develop a Big data analytics model that prioritise customers with the highest propensity to win-back or to return and make a purchase
2) Deploy Big data analytics model in Customer win-back operation to improve Customer win-back rate
3) Measure the impact of the Big data analytics model in improving Customer win-back rate
4) Conduct revenue projection to measure the benefit of Big data analytics model development and implementation

1.4 Research Questions

Based on the problem state and research objectives, the fundamental questions that need to be addressed are the following:

1) Can Big data analytics increase effectiveness in terms of Number of calls referred (# refer) for Customer win-back?
2) Can Big data analytics increase efficiency in terms of Percentage of calls referred to outbound calls (% refer)?

3) Can Big data analytics increase productivity in terms of Percentage of calls that are picked up compared to total outbound calls (% reach)?
4) Can Big data analytics help increase revenue from Win-back operations for the Company?

1.5 Hypothesis Development

Based on the research questions proposed to address the problem statement and research objectives, hypotheses can be developed based on existing knowledge and research as follows:

- Customer win-back department can increase the number of calls made to customers using Big data analytics, i.e. increase in efficiency
- Customer win-back department can increase Customer win-back rate, i.e. percentage of number of calls made to customers that are successfully referred to Sales department to number of calls made to customers using Big data analytics model
- Customer win-back department can increase Percentage of calls that are picked up (% reach)
- Customer win-back department can substantially increase revenue projection from using Big data analytics modelc

1.6 Scope of Research

This research is focused on improving performance on Customer win-back for Prospect and Site visit stages on the sales journey only as there are limited capacity to serve a large pool of customers that dropped out from the sales journey.

The data collected from both stages which are stored across several databases will be transformed and tested for correlation using statistical model to identify significant features. These features will then be input for Adaptive machine learning (ML).

Big data analytics model will be,operationalised over a duration of 3 weeks. This will require 2 teams of 5 Win-back staff – A/B test. Team A will maintain as-is operation and Team B will be using the Big data analytics model and lead priority list to contact customers while. At the end of each day, 2 main metrics will be collected: (1) Number of outbound calls per day (2) Percentage Win-back on the day, and used as input for Adaptive ML.

At the end of the 2-week A/B test, the results will be compared and extrapolated to project revenue in 2020 for (1) Big data analytics model (2) Current operation which will be input for impact analysis, i.e. revenue increase from implementation.

Further discussion on the results will be made to explore viability in enhancing and fine-tuning the model for other use-cases within the Company or to productise and improve customer retention across other industries.

1.7 Outcome

Upon development and operationalisation of Big data for Customer win-back, efficiency and effectiveness for Customer win-back department is expected to improve, the following results are evident:

1) A significant increase in the Number of outbound calls per day performed by Customer win-back department over a period of 3 weeks
2) An evident increase in average percentage of Win-back cases to number of outbound calls over a period of 3 weeks
3) A substantial decrease in average number of calls that are duplicates, i.e. wrongfully marked as win-back
4) An increase in revenue projection for 2020 based on performance from operationalising Big data analytics model

2. Literature Review

2.1 Data and Relevancy

Mayer-Schönberger & Cukier (2013) state that data will become a very critical asset and will serve as vital inputs to help economic activities. Chase (2013) claim that a combination of data, statistical analysis and business know-how will lead to an increasingly precise demand forecasting. A notable commercialisation of data is in marketing. Huang (2013) propose an argument that using data in marketing will make advertising more effective. With increasing data points, advertisers ought to have a greater idea of what their target customers look like and what characteristics are shown for their non-targets, which helps with focussing advertising efforts on the aspects that matter, capturing the targets and increase return on investment for marketing activities.

2.2 Big Data: Definition and Characteristics

Traditionally, data from across various transactions are used to record activities without having any implications or commercial usages (Anshari, Almunawar, Lim and Al-Mudimigh, 2019). However, organisations across both public and private sectors have realised the importance and value of data in the digital age and have revolutionised the way data is viewed (Anshari et al., 2019). Anshari et al. (2019) further stated that a leap in processing capabilities have paved way for data analytics and Big data.

In its early conception, Big data was defined by Laney (2001) as having 3 characteristics (3 V’s): volume, velocity and variety.

- Volume refers to the mass volume of data that is beyond conventional processing capabilities
- Velocity refers to the increase in speed of interactions between data points that are used to interpret, analyse or summarise information
- Variety refers to amalgamation of various types of data, which are incompatible in terms of format, structure and logic, e.g. a collection of data tables, images and text files which are stored and used to provide a certain output

These characteristics and Big data are simplified by Ohlhorst (2013) as atypical large amount of data that cannot be processed using traditional analytics approaches. However, Jain (2016) propose additions to the 3 V’s: Variability and Value.

Variability, based on Jain (2016), is the way data points are captured and used which differ based on time and location. Value refers to the insights lying in within the amalgamation of data which can be used to improve the status quo. These characteristics are reinforced by Anshari et al. (2019), describing Big data as volume, velocity, variety, veracity and value of data. The common notion that is being discussed in literatures is that a vast amount of data both structured and unstructured have tremendous value if analysed comprehensively and this gave Big data analytics recognition in computer science field.

2.3 Big Data Analytics and Applications

2.3.1 Big Data Analytics Maturity

An integral part of Big data is Analytics. Gartner (2020) explain analytics as a “statistical and mathematical data analysis that clusters, segments, scores and predicts” the most likely outcome of something. It is further emphasised that analytics has been shifting more rapidly into business paradigm (Gartner, 2020).

De Jong (2019) summarise 4 levels in which Analytics evolve based on Gartner Analytic Ascendancy Model (Figure 6). This shows 4 main stages: Descriptive, Diagnostic, Predictive and Prescriptive analytics.

- Descriptive analytics is the simplest form of analytics. It provides a summary from data to support the event or outcome of a situation. De Jong (2019) emphasise on interesting insights that can be derived from descriptive analytics results simply by using arithmetic methods such as mean, median, mode, maxima and minima
- Diagnostic analytics moves up the level of advancement to look at historical relationship of data to understand why an outcome occurs. Correlations, probabilities and patterns are among the tools which are employed
- Predictive analytics is looking into the future based on historical occurrences. However, for predictive analytics to be accurate, data points need to be vast and with high quality. Common application is demand forecasting
- Prescriptive analytics is the most mature stage of analytics and answers the question of what needs to be done to exploit opportunities or to prevent damages from predictions. An interesting notion is the prescriptive

analytics leverages feedback and improve school, which ensure that the outcomes will be more favourable in each iteration

On the other hand, Srivastava (2015) argue that there are 5 stages to analytics development which slightly differ from Gartner's model: diagnostic, predictive, correlation identification between unknowns, prescriptive and monitoring. A further highlight is made on the application in the business world that diagnostic analytics, correlation identification and predictive analytics are most widely used.

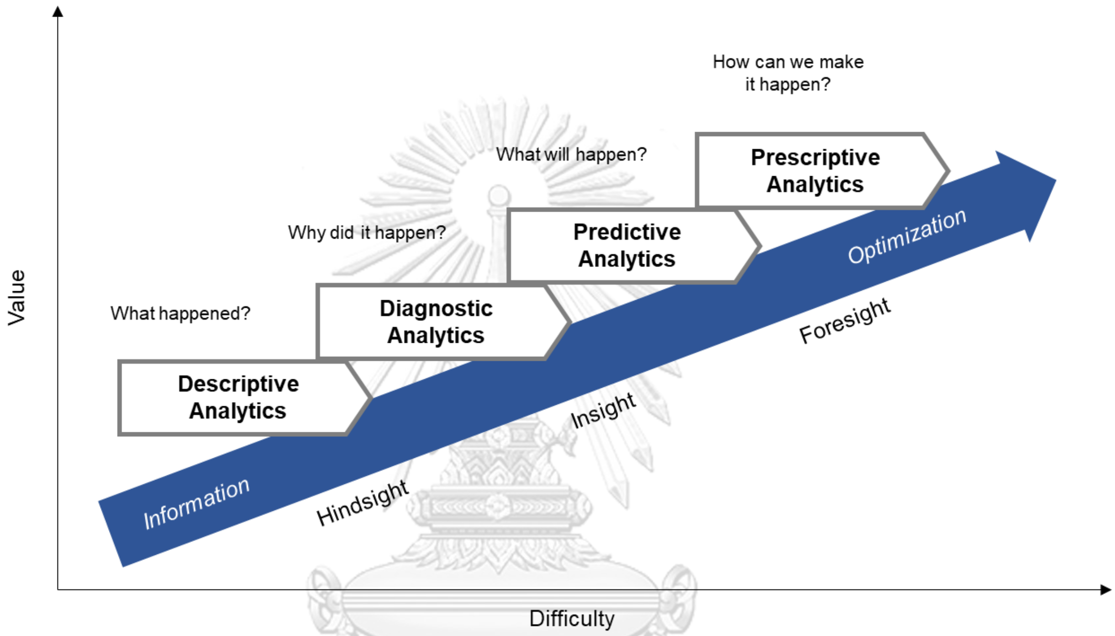


Figure 6: Gartner Analytic Ascendancy Model

Source: Adapted from De Jong (2019)

2.3.2 Application of Big Data Analytics in Marketing

While diagnostic analytics and correlation identification are among the most notable analytics deployment, predictive analytics provide a more impactful outlook. Marketing, as one of the most critical levers in business, is an area which has adopted Big data predominantly.

As customer data accumulate overtime to show characteristics of Big data, CRM requires new approaches to management and analyse (Anshari et al., 2019). The advent of Big data in CRM has enabled various approaches to improve and enhance customer interactions and relationships. The analytics power enables increasingly accurate assessment of customer's needs and trends. However, a cheaper and an effective approach is using Big data to drive Customer win-back

(Anshari et al., 2019). This is supported by a statement by Griffin and Lowenstein (2001) which highlight on a key success factor of Customer win-back being data.

2.4 Customer Relationship Management (CRM)

2.4.1 Definition of CRM

Customer relationship management (CRM) has since focused predominantly on customer retention and creation of loyalty (Thomas, Blattberg and Fox, 2004). Morgan and Hunt (1994) make and observation that the fundamental of CRM is for the organisation to know the customers “better”. To reinforce this view, Dowling (2002) state that developing a relationship with customers is the desirable approach to gain loyalty from customers and that loyal customers are likely to be more profitable that non-loyal customers. This is mainly due to customer acquisition cost. Loyal customers are prone to repeat purchases, warranting minimal customer acquisition activities. Therefore, the primary benefit and notion of CRM has been in help driving the company’s customer retention (Rust et al., 1996). However, Dodson (2000) state that companies have realised that customer retention is not the answer as customers leave eventually.

2.4.2 Emergence of Customer win-back

From realisation that customers still leave, the focus has begun to shift to reactivating customers who already left the sales funnel or recapturing lost customers, referred to as customer win-back. Customer win-back is defined as a process by which a company re-ignite its relationship with lost customers (Thomas et al., 2004). Kumar, Bhagwat and Zhang (2015) suggested that both product-based and service-based organisations are experiencing high customer attrition and finding it increasingly difficult to expand customer base. It is further suggested by Kumar et al. (2015) that recapturing lost customers benefits an organisation in revitalising profits and prevent competitors from acquiring its customers. Until recently, Thomas et al. (2004) claimed that Customer win-back had been neglected in comparison.

2.4.3 Customer win-back as a Value creator

Griffin and Lowenstein (2001) concluded that customer data within the firm is the key to increasing or maintaining customer retention. Further research has also shown that 20-40% of repeated sales can be achieved for lost customer segment using the right customer data (Griffin and Lowenstein, 2001). Thomas et al. (2004) assessed further to identify that engaging lost customers and trigger their re-entry into the sales funnel provides a return on invest (ROI) of 214% compared to a mere 23% ROI from acquiring a new customer. This approximately tenfold in value is adequate in convincing a company to focus on retaining and re-targeting current or lost customers. Travelocity, an American online travel agency has seen a 100% ROI compared from previous marketing campaigns from implementing a customer win-back campaign, whereby customers are made to feel exclusive within Travelocity's customer base (Kumar et al., 2015).

Stauss and Friege (1999) point out that customer win-back provide value beyond sales and profit addition, which are:

- Lower acquisition cost compared to new customer recruitment
- Product and service improvement opportunities identified from customer reasons for leaving
- Ability to identify at-risk customers from lost customers trend
- Ability to limit negative viral from defects and encourage positive word of mouth from win-back customers

Moreover, Tokman, Davis and Lemon (2007) suggest that companies are now realising that customer win-back is paramount to manage customer portfolio more effectively in the long run.

From the stated benefits, data inputs from lost customers are considered critical to realise the performance enhancement. Xu et al. (2005) conclude that the benefits of CRM lie in analytics, specifically in using customer data to extract patterns and identify behaviours that the organisation can better connect with customers, thereby strengthening customer relationship.

To contrast customer retention and customer win-back, Stauss and Friege (1999) state that traditional acquisition aims at customers who have no prior experience with the products or services offered by a company and ensuring that customers remain users once experience is established, but customer win-back is directed at customers who had prior experience and left the company elsewhere, i.e. defected or lost customers.

2.4.4 Customer Win-bank: A Case Study in Technology-based Payment Solution

An interesting case study is Paypal when it developed a churn analytics model to help reduce customer churn. H2o.ai (2017) state that the firm developed a predictive analytics model using a statistical method to uncover characteristics of existing churn reports. This model is then used to test across Paypal's entire customer base (H2o.ai, 2017). This resulted in a new set of KPIs used to monitor customer churn and deployment of more powerful data analytics which provide immediate call to actions. Commercially, this helps Paypal to create more effective campaigns to reduce customer churn.

2.5 Predictive Analytics and Model Development

Based on Figure 6, Predictive analytics is the future of business, employing the science of prediction to uncover "what will happen?" so that organisations can prepare to accommodate future occurrences. Kumar and L (2018) define predictive analytics as a branch of advanced analytics that "predict future events" through an amalgamation of techniques: statistical analysis, data mining, machine learning and artificial intelligence. Mayer-Schönberger & Cukier (2013) claim that the value of Big data lies within predictive analytics. This will ultimately help organisations into better serve predict what will happen so they can better prepare or create value from the expected. Siegel (2016) further emphasise on business implications, specifically pointing out that predictive analytics can help companies become more proactive and forward-looking into trends and behaviours that are likely to occur within the customer space, thereby gearing up profits.

Kumar and L (2018) point out that predictive analytics has a step-by-step process which data analyst need to follow in developing and deploying predictive model. This is illustrated in Figure 7. These 6 steps are vital to follow in order to allow real-world application to occur. Kumar and L (2018) detail these steps as follow:

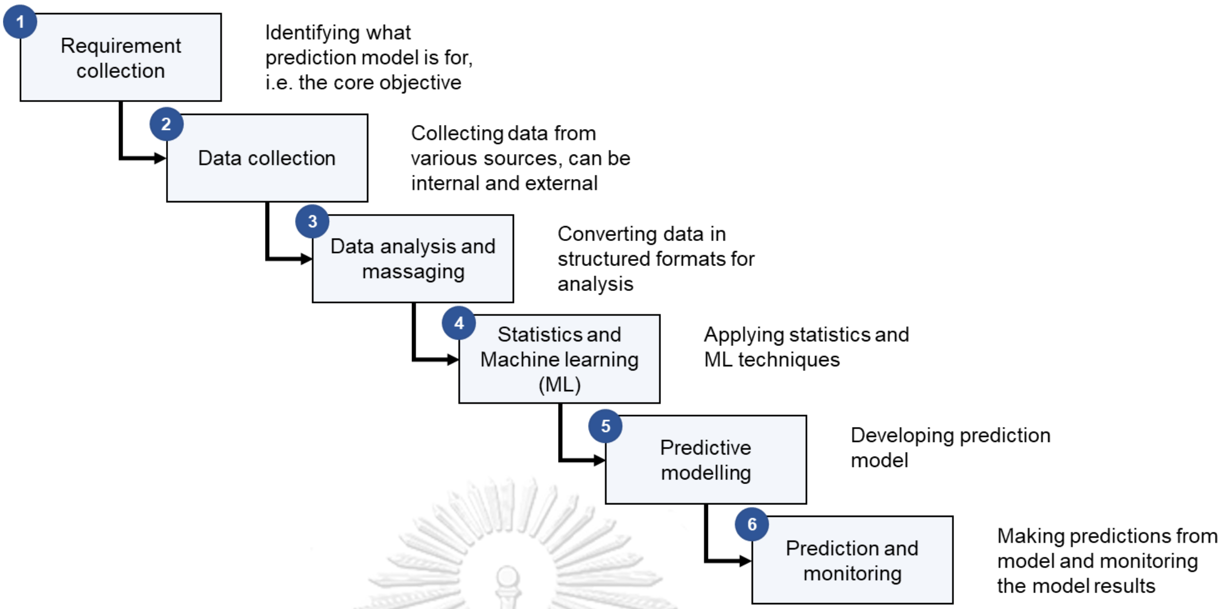

1
Requirement collection
Identifying what prediction model is for, i.e. the core objective
2
Data collection
Collecting data from various sources, can be internal and external
3
Data analysis and massaging
Converting data in structured formats for analysis
4
Statistics and Machine learning (ML)
Applying statistics and ML techniques
5
Predictive modelling
Developing prediction model
6
Prediction and monitoring
Making predictions from model and monitoring the model results

Figure 7: Predictive analytics process

Source: Adapted from Kumar and L (2018)

2.5.1 Requirement collection

The very first key action is to clearly define the objective of the prediction which will be the key framework for collection, analysing and predicting the outcome. Within the scope of this study, the objective has been clearly defined as Customer win-back prediction, i.e. to identify which customers will likely re-enter the sales funnel. Implementation approach to validate the model will be further discussed in the Research methodology section.

2.5.2 Data collection

After defining the objectives, the data analysts will need to identify the datasets needed. These datasets maybe structured or unstructured, i.e. in otherwise unusable formats (images, text files, etc.). Kumar and L (2018) provide an example as customer list who viewed the products on the website.

2.5.3 Data analysis and massaging

As mentioned above, the data collected maybe in various formats. This stage aims to prepare the data for further analysis; this will require converting all datasets in structured format. Moreover, there likely to be missing values in datasets and will need to be addressed. A possible solution is data imputation, i.e. using the pattern of existing data to estimate the likelihood of the missing values and randomise the inputs. Kumar and L (2018) stress that the impact of predictive analytics rests on the quality of data.

A fundamental stage of data analysis is data preparation, whereby data are transformed into a ready-to-use format. Zhang, Zhang and Yang (2003) suggest that companies are looking into cleaning up vast low-quality data which are available on the internet and within their databases. On the other hand, Abdallah, Du and Webb (2017) suggest that data preparation is an iterative process to organise data which are often unstructured into a structure that can be analysed.

Abdallah et al. (2017) suggest that though there are several approaches to prepare data for analysis, each would consist of key major tasks – from data profiling, cleansing, integration and transformation.

Data profiling as the first step concerns with reviewing the available data to identify its suitability and whether they are sufficient to feed into the analysis (Abdallah et al., 2017). Data quality is also a critical factor to consider when engage in data profiling. Abdallah et al. (2017) suggest 2 criteria to assess data quality: (1) Accuracy (2) Uniqueness.

1) Accuracy is described by Abdallah et al. (2017) as a function of 3 criteria: Integrity, Consistency and Density.
    a. Integrity concerns with the completeness of data, i.e. if there are any missing values, and validity of data, i.e. data holds true when constraints are lifted
    b. Consistency concerns whether the data contradict or whether there are anomalies within the datasets
    c. Density concerns with missing values in the datasets which equate to 'unknown'
2) Uniqueness is a criterion that is fulfilled when there are no duplicates.

Data cleansing is defined by Abdallah et al. (2017) as a process of removing inaccurate values or properties from databases. Data cleansing can simply be explained by a 4-stage process:

1) Define and identify errors and inaccuracies in the datasets
2) Clean the data by changing the values within the record
3) Record the errors
4) Measure the cleanliness of data by comparing it to user's requirements

Data integration is a critical step in data cleansing. Abdallah et al. (2017) suggest that integrating data from various sources is not a simple task but can bring unimaginable opportunities. Data across heterogeneous sources are in different formats – some are structured in tables while some are unstructured in texts. An example proposed by Abdallah et al. (2017) is combining socio-economic data of customers require data from other external sources, which can provide a more holistic view.

Data transformation is simply changing the format of a dataset from a source, integrated or not, into a different format that is desirable for users (Abdallah et al., 2017). This step occurs when a dataset is moved from one database to another, i.e. from databases into a data warehouse. Bonifati, Cattaneo, Ceri, Fuggetta and Paraboschi (2001) suggest that when a dataset needs to be analysed, they are moved from databases into a data warehouse, which is a database devoted to analytics processing.

2.5.4 Machine Learning and Predictive Analytics

Machine learning (ML) is becoming widely recognised as a newfound advanced application of technology and big data in modern society. Shalev-Shwartz and Ben-David (2017) define Machine learning as automated extraction of patterns from a large set of data (big data). Mohammed, Khan and Bashier (2016) explain that ML is a type of artificial intelligence that operates on intelligent software with adaptive statistical modelling that continuously build and enhance capabilities. Mohammed et al. (2016) summarise ML techniques into 4 branches: Supervised learning, Unsupervised learning, Semi-supervised learning and Reinforcement learning (Figure 8).

1) Supervised learning refers to an area of ML that input and output are somewhat pre-defined to allow ML to understand what inputs will lead of certain outputs and will enhance its analytics capabilities based on such functions (Mohammed et al., 2016).
2) Unsupervised learning concerns with a self-enhanced and learning method of ML whereby connection between inputs and outputs are not pre-defined and parameterised into the algorithms (Shalev-Shwartz and Ben-David, 2017).
3) Semi-supervised learning is a combination of Supervised and Unsupervised learnings. The technique aims to allow Unsupervised learning method to predict, enhance and improve the results so that it outperforms that of the Supervised learning (Mohammed et al., 2016).

4) Reinforcement learning is an approach that the software agents operate in a certain environment where the interpreter (often human controller) provides rewards to help them understand their actions. This is often cause be assigning reward to the pre-defined outputs. They aim to act in a way that maximises cumulative rewards they get (Mohammed et al., 2016).

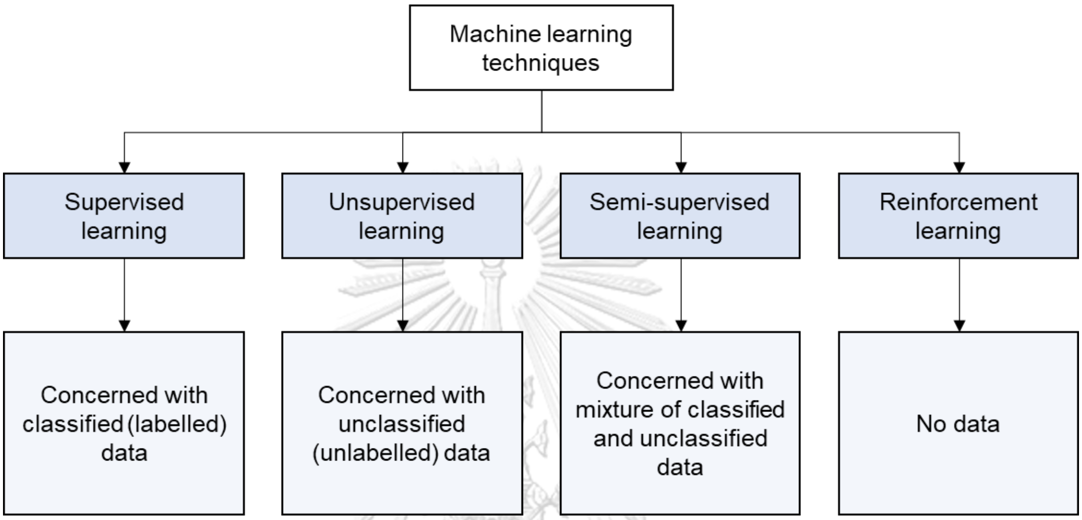

Machine learning techniques
Supervised learning
Unsupervised learning
Semi-supervised learning
Reinforcement learning
Concerned with classified (labelled) data
Concerned with unclassified (unlabelled) data
Concerned with mixture of classified and unclassified data
No data

Figure 8: Machine learning techniques

Source: Adapted from Mohammed et al. (2016)

Due to the datasets that have a large number of categorical inputs and expected outcome that requires insight interpretation, Supervised learning is a suitable ML technique to develop. There are several advantages to Supervised learning in practice. Joy (2019) summarised pros of Supervised learning as providing clarity for the data, training the algorithm is a simple process and detailing definitions for specific classifications can be achieved (in case of a unique or very specific definition). This research will explore 3 different Supervised learning algorithms: (1) Decision trees (2) Naïve Bayes (3) Random forest.

2.5.4.1 Decision Trees

Based on Mohammed et al. (2016), Decision tree algorithm is a statistical model widely used in classification. The fundamental considers classifying dataset into different classes and running through each class to reach a leaf. Mohammed et al. (2016) simply describe this algorithm as running a “query from each root to reach each leaf, representing classes”. Mohammed et al. (2016) summarise key steps in constructing Decision trees as (1) Assign training samples into a root (2)

Categorising training samples based on attributes (3) Select attributes using statistics and (4) Iteratively dividing training samples into classes until all the training samples are statistically evaluated or all the remaining samples belong to the classes generated. Chauhan (2019) distinguish Decision tree into 2 types: Categorical variable and Continuous variable.

To construct a decision tree, firstly each observation in the dataset must be represented as a point that lives in a high-dimensional abstract space. The dimensionality of this space depends on the number of features available (Bishop, 2006). The dataset that undergo imputation and feature selection has 16 dimensions. In order to explain decision trees, we shall first define the following notation, and assume p = 2 for simplicity. This assumption can be made without loss of generality, because in higher dimensions the process remains unaltered.

$$X = \{X_1, X_2, \dots, X_p\} \qquad \text{(Features in use)} \qquad [1]$$

There have been various applications of Decision trees in real-world data analytics cases. Al-Barrak and Al-Razgan (2016) developed a student's grade point average (GPA) prediction using Decision tree algorithm. The study considered a dataset of student's transcripts, looking at final GPAs and individual grade in each subject. The research also helped identify the most significant courses that students consider. Al-Barrak and Al-Razgan (2016) was able to help the university focus its resources on selected courses to help students in passing and therefore, gaining high GPAs. The research found that Software engineering-1 and Java2 are the most significant nodes in the Decision tree, i.e. these courses correlate highly with GPAs. Students with high grades in these courses are likely to obtain high GPAs.

2.5.4.2 Naïve Bayes

Another notable algorithm is Naïve Bayes classifier which is based on Bayes theorem (1). The fundamental considers that the features show independence between each other (Mohammed et al., 2016). Naïve Bayes is not a single algorithm but a family of algorithms that is based on independence among features. Budiyanto and Dwiasnati (2018) summarise Naïve Bayes algorithm as a statistical method which estimate probabilities by summing frequencies and combinations of values from a certain dataset.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad [2]$$

Where:
A and B are events
$P(A)$ *and* $P(B)$ are probabilities of A and B exclusive of each other
$P(A/B)$ = Posterior probability, i.e. probability of A given B holds true
$P(B/A)$ = Likelihood, i.e. probability of B given A holds true

Ray (2017) summarise Naïve Bayes algorithm as having assumed that a presence of one feature is independent of "any other feature". Although it is deemed simple, many occurrences have witnessed Naïve Bayes outperforming other sophisticated algorithms in predicting outcomes (Ray, 2017).

An example of Naïve Bayes application is a research by Budiyanto and Dwiasnati (2018) on identification and prediction of best-selling products for PT Putradabo Perkasa, a company that provides sales of CCTV system and Access control system. Budiyanto and Dwiasnati (2018) considered several factor inputs such as type, brand, quality and price of goods and target customers in order to predict best-selling product and help PT Putradabo Perkasa in supply preparation. The research deployed data mining and analysing sales transaction. Budiyanto and Dwiasnati (2018) found that (with 78.33% accuracy) IP camera product with type Infinity I-993V is the best-selling product. This has further implications in terms of managing stocks and order management.

2.5.4.3 Random Forest

With rapid development in ML, it is natural that an algorithm is developed from a combination of various classifiers. Random forest is a result of such development. Shalev-Shwartz and Ben-David (2017) define Random forest as a collection of Decision trees whereby each tree is applied an algorithm and a training dataset. The concept as proposed by Breiman (2001) can be summarised as correlation between predictions from individual Decision trees must be low. This is because each tree helps one another in mediating errors; while some trees may show errors, a majority will not and this notion helps maintain accuracy (Breiman, 2001). Gao, Wen and Zhang (2019) conclude a 3-step approach than a typical Random forest algorithm development must undertake: (1) Select the training set by using a Bootstrap random sampling approach (2) Establish a classification regression tree for each training set (3) Create a simple voting of the output of each tree.

Schott (2019) simplify understanding of random forest as a consolidation of several decision trees which “merge all answers from all trees” in order to identify the correct answer. Random forest operates on mean squared error (MSE) as its mathematics background (Schott, 2019).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \qquad [3]$$

Where:
N = number of data points
f_i = value outcome of the model
y_i = actual value for data point

This formula is used to measure the distance from each node within decision tree from its respective value. This will help determine which decision tree is a better fit for the random forest. Schott (2019) suggest that random forest uses Gini index to measure the likelihood of each decision tree and predict which tree is most likely to occur. Mathematically, the following formula is used:

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \qquad [4]$$

Where:
C is the number of classes
P_i is the class observed from the dataset

A research by Gao, Wen and Zhang (2019) adopted Random forest as an algorithm to help predict employee turnover. Employee turnover is a critical issue that threatens sustainability and planning of a company. Gao, Wen and Zhang (2019) developed a model that analyses data from a communications company in China. The research identified monthly income, age, overtime, distance from home, years of service and salary increase in terms of percentage as significant features. From this, Gao, Wen and Zhang (2019) developed a new analytics model which allows the company to better plan and manage employees to increase retention.

2.5.5 Predictive modelling

Testing the transformed data to ascertain the validity of datasets is of paramount importance as this is a decider to a win or a lose in terms of using Big data analytics in achieving an objective. To achieve this, hypothesis testing is required. Emmert-Streib and Dehmer (2019) define hypothesis testing as an approach to decide whether a data sample shows assumed characteristics compared to a population given that an assumption on the population holds true.

A typical statistical model suitable for running a hypothesis test in data science in Logistic regression. Katsaragakis, Koukouvinos, Stylianou, Theodoraki, and Theodoraki (2005) explain Logistic regression as a linear model which provides a prediction in terms of discrete output from continuous, discrete, dichotomous or a mixture of these inputs. Logistic regression provides no constraint on the sample, i.e. no assumption is made on the sample being normally distributed. Therefore, predictors and responses are not linearly related (Katsaragakis et al., 2005). In data science, Logistics regression is a suitable hypothesis testing method for categorical data analysis, i.e. a yes/no, true/false or 1/0 type of categories.

2.5.6 Prediction and monitoring

After testing for model's validity, the model is deployed for the organisation to use in predicting what will happen and decision-making to cope with the expected occurrences (Kumar and L, 2018). The result will be reported by the model, which would require close monitoring by users to ensure the accuracy and validity of the results.

The scope of this study also includes an empirical study to illustrate the impact of predictive analytics in helping organisations predict the future and act accordingly to improve performance.

2.6 Data Analytics Tools

2.6.1 Predictive Analytics Software and Resources

There are several tools which help with creating a tailored algorithm, running analytics and getting the results. This research will consider 3 tools that will be used throughout: (1) Scikit-learn (2) Apache Hadoop (3) Apache Spark^{TM}.

Scikit-learn is a Python module that integrates a large collection of ML algorithms for both supervised and unsupervised learning techniques (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot and Duchesnay 2011). Pedregosa et al. (2011) state that Scikit-learn package offers a non-specialist friendly setting for both academic and commercial access to browse a library of ML algorithms and develop their own codes based on the readily available tools.

Apache Hadoop is a framework software used in storing, managing and processing data. Dahiya, B and Kumari (2017) explain Apache Hadoop as having 2 main components: HDFS and MapReduce. HDFS is dedicated to storing large datasets while MapReduce is used for dataprocessing (Dahiya et al., 2017). MapReduce as a processing tool is simply and scalable, which is useful for data scientists to deploy across different use-cases (Dahiya et al., 2017). However, Hadoop is slow in terms of processing and is unsuitable for real-time analytics.

On the other hand, Dahiya et al. (2017) suggest that Apache Spark^{TM} is a cluster computing framework that couples and runs on top of Apache Hadoop, providing real-time data streaming and analytics. Dahiya et al. (2017) claim that Apache Spark^{TM} is used in place of MapReduce for lightning speed analytics, which MapReduce fails to deliver.

2.6.2 Operating Software and Tools

Apart from Apache Hadoop and Spark, Talend acts as an extraction, transformation and loading (ETL) software. Amazon web services (AWS) is another brand that is considered as a ready-for-use analytics tools.

ETL is critical in data integration. Talend (2020) highlight on the benefits of ETL as the actions which allow businesses to gather data across all databases and integrate them into a single source, thereby allowing different types of data to generate new values. ETL simply refines data and transfer the refined datasets into a data warehouse such as AWS Redshift and Microsoft Azure (Talend, 2020). ETL follows 3 main steps: (1) data extraction (2) data transformation (3) data loading.

Data are usually stored across different platforms and databases. However, Talend (2020) suggest that businesses usually use a variety of data analysis tools to product results for decision-making. To effectively product business intelligence, data must be able to “travel freely between systems and applications” (Talend, 2020). To achieve this, data across different databases are consolidated and stored in a single repository, both structured (in a structured format such as parameterised fields, categorical values, etc.) and unstructured (such as text files, images). These data can be extracted from CRM database, sales and marketing applications and analytics tools from each department that store results in different databases.

Secondly, data quality must be ensured otherwise consolidation would yield no benefit. Talend (2020) suggest that “rules and regulations” need to be applied to datasets in order to ascertain data quality and integrity which can be used to produce business impact. Talend (2020) further note 6 processes to follow when transforming data:

- Cleansing – this involves managing missing values and inconsistent data fields within the dataset, such as imputation, omitting certain data fields, etc.
- Standardisation – this creates a standard set of rules to be applied across all data fields to ensure that each selection has the same consistency and format ready for use.
- Deduplication – this omits any duplicate data fields or consolidate information within fields that fall under 1 single dataset, e.g. duplicated profiles of a single customer caused by spelling mistakes,
- Verification – this discards anomalies and data with unusable information
- Sorting – this categorises data which underwent transformation into groupings with shared characteristics.
- Additional tasks – further exploration into tasks that add validity or usefulness of data are carried out.

Transformation is a critical step as it determines whether data inputs are of use and can provide benefits after processed or analysed (Talend, 2020). If transformation is not carried out effectively then end benefits cannot be observed.

Data loading is the migration of newly transformed data into a new source. This can undergo full loading or incremental loading (Talend, 2020). Full loading occurs when all the transformed data from the transformation output enter the new data warehouse (Talend, 2020). However, if transformation aims at producing exponentially higher datasets then this could be more complicated to manage. On the other hand, incremental loading employs comparisons between existing and newly transformed data and consider the differences. These differences are then produced as new records and loaded to the datasets.

AWS tools consist of 3 main software which are critical for providing predictive analytics results. Firstly, Amazon Redshift which is a data warehouse supports the establishment of operations for data analytics (AWS, 2020a). This is where data that undergo ETL are loaded into for further extraction and analysis. Secondly, Amazon EMR extracts data from Redshift and processes data at scale with vast speed, operating on Hadoop framework (AWS, 2020b). Thirdly, AWS S3 is a storage space for output undergo EMR processing. AWS (2020c) state that S3 is an interface that is simplified for anytime, anywhere data retrieval.

2.7 Propensity Modelling

2.7.1 The Fundamental of Propensity Scoring and Model

A common application of predictive analytics is propensity scoring. A propensity model is a statistical approach to predict customer behaviours or decisions (Childs, 2002). HG Insights (2018) suggest that the fundamental of propensity model is to run various probabilities with various samples based on shared characteristics. This ought to provide an accurate prediction on future behaviours and is particularly beneficial in the real-world context such as economics, business, education and healthcare (HG Insights, 2018).

In terms of business, HG Insights (2018) suggest 3 notable models: (1) Propensity to buy (2) Propensity to churn (3) Propensity to unsubscribe. Propensity to buy considers customer’s willingness to purchase. Propensity to churn looks at customers who risk leaving the company. Propensity to unsubscribe concerns customers who have become intolerant to marketing campaigns and are ready to leave the service of the company.

2.7.2 Real-world Applications of Propensity Scoring: A Case of Sales and Marketing Across 160 Companies

Kucera and White (2012) survey on 160 senior executives which employ predictive analytics in sales and marketing confirms that adoption increases sales. The study focusses on the positive impact on marketing campaigns based on 2 key metrics: (1) Sales lift from a campaign and (2) Click-through rate (CTR) from a campaign. On average, sales lift increases from 4.8% to 8.3% after deployment of predictive analytics to target higher potential prospects (Kucera and White, 2012). Similarly, CTR is shown to increase from 4.5% to 7.9% after adoption of analytics (as illustrated in Figure 9).

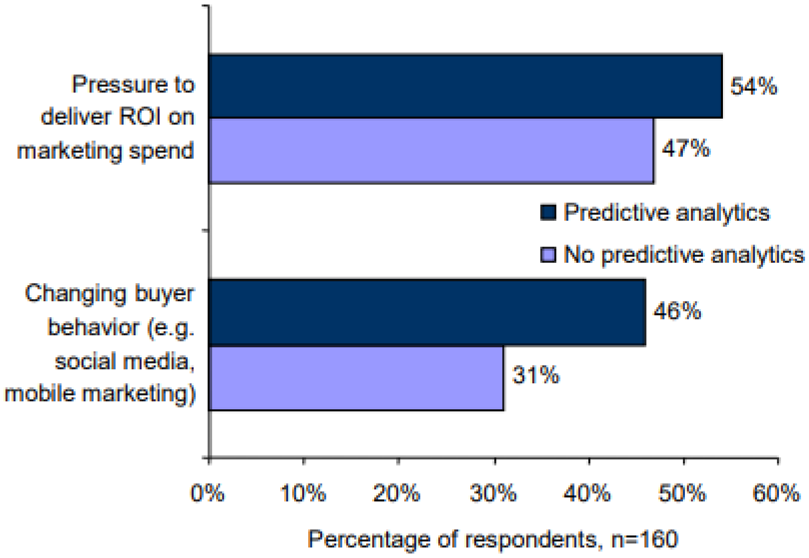
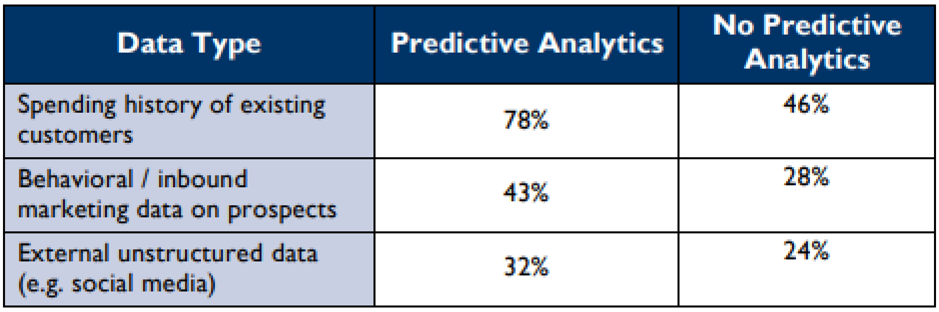


Figure 9: Impact of predictive analytics in marketing campaigns

Source: Kucera and White (2012)

Kucera and White (2012) also suggest that companies with predictive analytics adoption are likely to build business-data management competencies over peers. The basis for performance improvement in sales and marketing is largely behavioural data of existing customers and non-customers that have entered the sales funnel (Kucera and White, 2012). The study also shows that companies that utilise such tools have more access to data across various sources that peers that do not. This is expected as greater emphasis is placed on data-driven approach from companies that see improvement in sales and marketing metrics (Table 1).

Table 1: Data access between companies with and without predictive analytics



Data Type	Predictive Analytics	No Predictive Analytics
Spending history of existing customers	78%	46%
Behavioral / inbound marketing data on prospects	43%	28%
External unstructured data (e.g. social media)	32%	24%

This shows that predictive analytics is impactful in help gearing company’s performance, especially in marketing and sales aspects. This study will explore how predictive analytic can be developed for customer win-back in real-estate sector, particularly in the townhouse segment in Bangkok, Thailand. This empirical study is a real world Proof of concept (PoC) to help improve customer win-back for a leading real-estate developer in Thailand which.

2.8 Student’s T-test and Validation

Student’s t-test is a statistical used to measure “how changes affect small samples of a larger population” (Ruth, 2019). This approach is deemed most effective when the samples used in comparison are from separate datasets but from the same population (Ruth, 2019). To ensure that the samples used to run statistical analysis provide valid arguments, the following assumptions need to be made:

- The data used are normally distributed
- The data share the same standard deviation

Hole (2009) state that there are 2 variations of t-test: (1) dependent-mean t-test which considers samples undergoing both conditions tested and (2) independent-mean t-test which considers 2 sets of samples undergoing different conditions and comparing the results to identify whether the condition differentials have a significant impact.

Kim (2015) further suggest that when samples of *n* are analysed from a population of $N(\mu, \sigma^2)$, the distribution of mean x^2 should be normally distributed, with $N(\mu, \sigma^2/n)$. Furthermore, an independent-mean t-test should base on assumptions that variances under the 2 testing conditions are identical (Kim, 2015). Kim (2015) illustrate this further by extracting 2 independent samples from a normally distributed population and calculate the difference between the mean values of the 2 samples. Although the data is from the same population, the difference between the mean values are not always equal to 0. Figure 10 illustrates Kim (2015) experiment and compare this to theoretical mean difference. This would also mean that variances between the sample sets would also differ in practice.

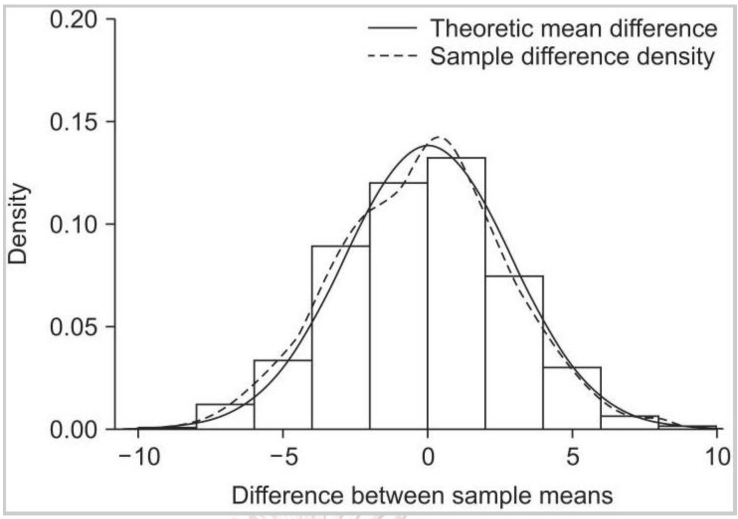


Figure 10: Simulation of different sample means

Source: Adapted from Kim (2015)

Independent-mean t-test or unpaired t-test follows 2 variations: equal variances and unequal variances. The unpaired equal variances assume that the sample sets from a population are equal (Armitage and Berry, 1994), the calculations are as follow:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad [5]$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2} \quad [6]$$

Where:

$\bar{x}_1$ and $\bar{x}_2$ are sample means

s^2 is the pooled sample variance

n_1 and n_2 are sample sizes

t is a Student quantile with $n_1 + n_2 – 2$ degree of freedom

Although unequal variances assumption assumes that 2 sample sets have different variances, Armitage and Berry (1994) suggest that this is unsuitable if the actual sample variances differ significantly. The calculations for unpaired unequal variances t-test are as follow:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad [7]$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \quad [8]$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2}{n_1 - 1} \quad [9]$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_2 - 1} \quad [10]$$

Where:
$\bar{x}_1$ and $\bar{x}_2$ are sample means
s^2 is the pooled sample variance
n_1 and n_2 are sample sizes
d is “Behrens-Welch test statistic evaluated as a Student t quantile with df freedom using Satterthwaite's approximation”

However, to determine or calculate the difference between sample sets from a single population is difficult (Kim, 2015). Therefore, certain assumptions should be made to in order to simplify the assessment, such as normal distribution and equal variances. Although Ruth (2019) argue that mean distribution is usually always not normally distributed, but an experiment-based statistical analysis should assume normal distribution for simplification. Moreover, equal variances should be as well.

3. Research and Methodology

The research will focus on 4 key steps (as shown in Figure 11) from (1) Data collection (2) Big data analytics model development (3) Implementation and (4) Impact analysis. Firstlly, Data collection will start with identifying the use case for the Company that warrants Big data analytics model. This will create a set of requirements for data input which will be identified through focus groups and operations observation. Secondly, the Company's data will be explored and prepared for Hypothesis testing to identify correlation with Win-back from historical performance. This stage will help select features that are significant for Win-back. These features will be inputs for Adaptive machine learning (ML). Thirdly, the developed model will be implemented through A/B testing. This will require a new operation setup for Win-back department to compare current process with operating model for Big data analytics. The results at the end of each day will be compared and fed back into feature selection to revise features in the Data preparation phase. This loop will repeat daily to refine the model. Lastly, the results over 3 weeks will be analysed for impact in terms of revenue projection for the Company. Further discussion will be made to identify underlying factors that may impact future results and viability in deploying similar model in other business functions within the Company. Furthermore, opportunities for the model to be customised for other use-cases in different industries will also be explored.

3. Research and Methodology

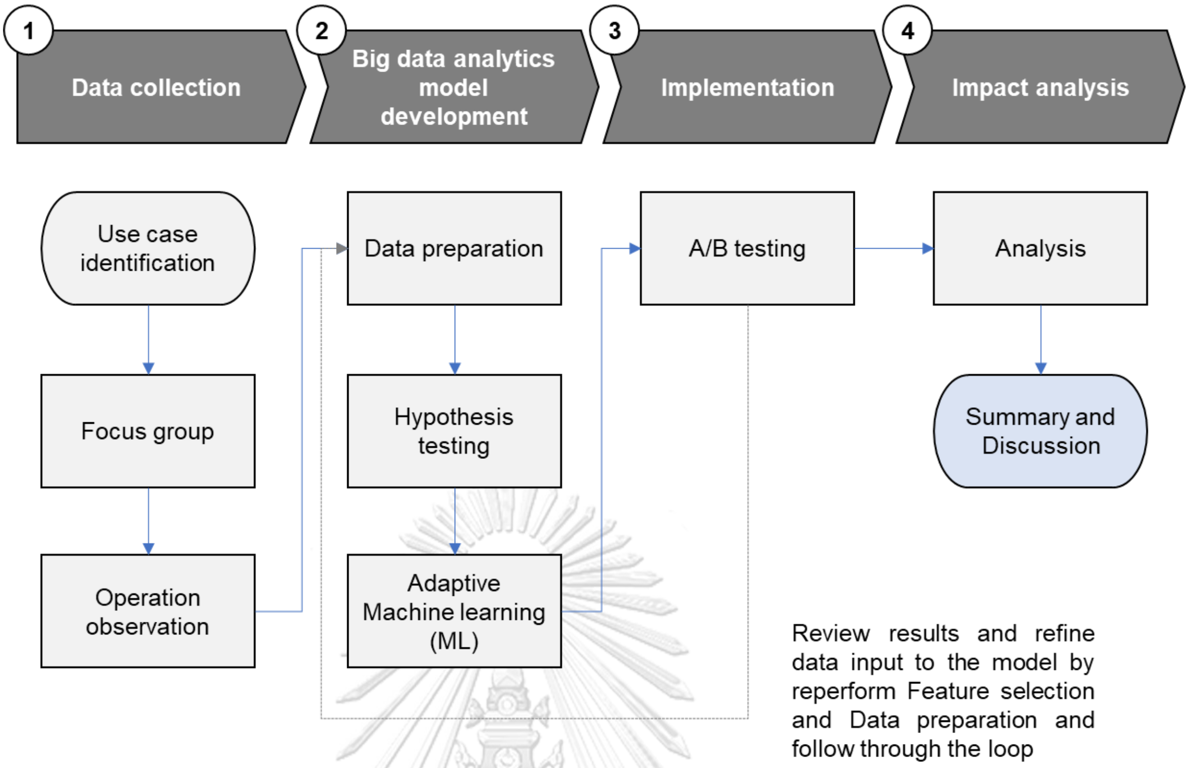

1
Data collection
2
Big data analytics model development
3
Implementation
4
Impact analysis
Use case identification
Focus group
Operation observation
Data preparation
Hypothesis testing
Adaptive Machine learning (ML)
A/B testing
Analysis
Summary and Discussion
Review results and refine data input to the model by reperform Feature selection and Data preparation and follow through the loop

Figure 11: Research methodology

3.1 Data Collection

3.1.1 Use Case Identification

From the problem statement, there are various approaches to improve Customer win-back using Big data analytics model. However, to create an impactful result, a clearly defined use case must be established.

The main business issue with customer win-back department is that there is limited capacity of 10 staff on call per day versus 220,265 customer records flagged as potential for customer win-back on CRM system to retarget. Therefore, the objective focusses on tackling this issue by using predictive analytics to help target high potential customers to win back. Hence, Lead prioritisation is the primary use-case in this study. This is an application of predictive analytics, specifically propensity scoring to help score which customers will be more likely to re-enter the sales funnel.

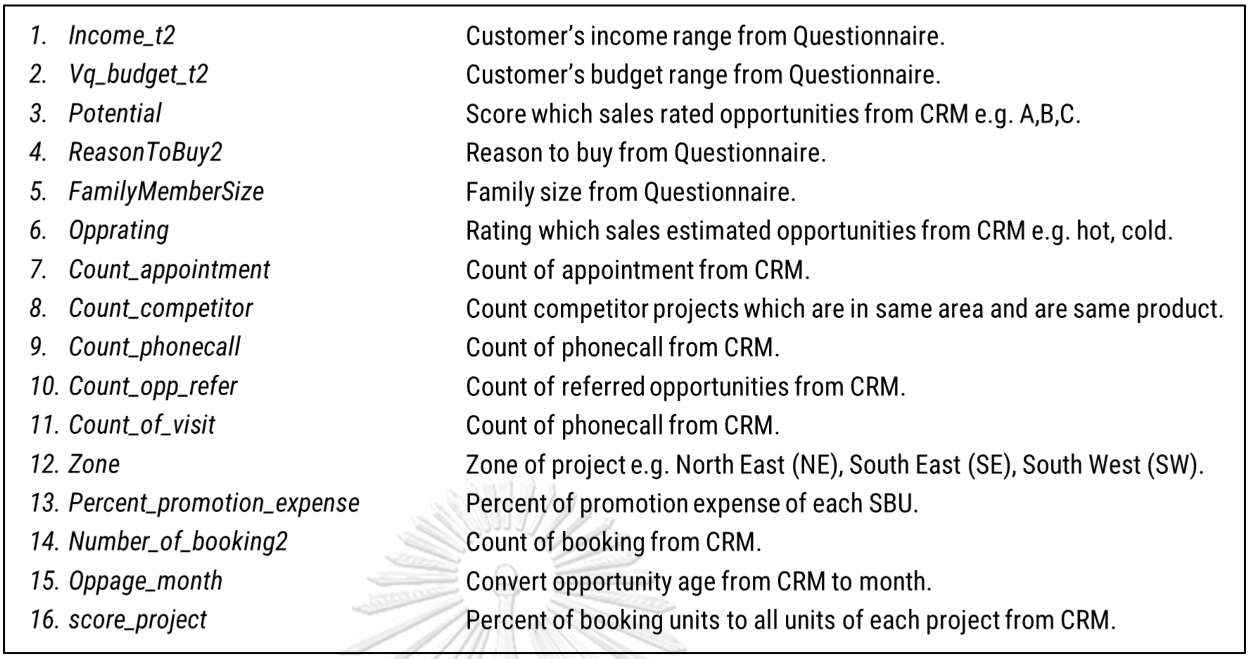
Therefore, this study will focus on developing and implementing a propensity model that helps prioritise leads for customer win-back.

3.1.2 Focus Group

To better understand factors that customers consider when deciding to purchase a residence, a focus group with Customer Win-back department will be conducted to understand more from customer engagement point of view. The sessions will help identify factors that are significant in deciding which customers to engage.

A set of interview guidelines will focus on the staff's experience in facing customers and what factors are visible for customers that are referred to sales staff, i.e. re-enter the sales funnel. For ease, these factors will be ranked from 1 to 16 in descending importance, with 1 being the most important feature and 16 being the least important (Figure 12). This is to help in selecting which features in CRM that may be of importance when prioritising customers to win-back. It is worth noting that this qualitative outcome is based on experience of staff which may be proven otherwise during hypothesis testing, i.e. feature selection. There are 4 key steps to understanding what factors are most significant:

1) Provide each customer win-back staff with a list of 16 factors that will be explored for significance (as shown in Figure 12) and allow each staff to independently rank the factor based on their personal opinion.
2) Consolidate the results and sum the score for each factor
3) Rank the scores with the lowest number being the most significant and the highest number being the least significant
4) Present the finding to the customer win-back department and demonstrate intent to cross-check the results with actual data from CRM



1. *Income_t2*	Customer's income range from Questionnaire.
2. *Vq_budget_t2*	Customer's budget range from Questionnaire.
3. *Potential*	Score which sales rated opportunities from CRM e.g. A,B,C.
4. *ReasonToBuy2*	Reason to buy from Questionnaire.
5. *FamilyMemberSize*	Family size from Questionnaire.
6. *Opprating*	Rating which sales estimated opportunities from CRM e.g. hot, cold.
7. *Count_appointment*	Count of appointment from CRM.
8. *Count_competitor*	Count competitor projects which are in same area and are same product.
9. *Count_phonecall*	Count of phonecall from CRM.
10. *Count_opp_refer*	Count of referred opportunities from CRM.
11. *Count_of_visit*	Count of phonecall from CRM.
12. *Zone*	Zone of project e.g. North East (NE), South East (SE), South West (SW).
13. *Percent_promotion_expense*	Percent of promotion expense of each SBU.
14. *Number_of_booking2*	Count of booking from CRM.
15. *Oppage_month*	Convert opportunity age from CRM to month.
16. *score_project*	Percent of booking units to all units of each project from CRM.

Figure 12: List of Features from CRM that are used to Rank by Customer win-back staff

Source: Extracted from company's CRM

3.1.3 Operations Immersion

To provide more understanding on as-is operation, immersion session will be deployed. This requires the team to observe how customer win-back department operates, what factors do the staff prioritise when retarget customers. This is to reconfirm if the features identified by customer win-back staff are truly significant, if feature selection testing states otherwise.

It is vital to understand how the staff contact customers and what factors each staff considers when choosing which customers to win back. The team must firstly identify (1) how the staff select projects to win-back (2) how the staff select customers from the list within each project. This will benefit the team in designing how A/B testing will be carried out. Prioritising customers and provide a long list alone will increase steps in switching between projects and will end up lower effectiveness and efficiency. Therefore, the output of this session will identify ways in which propensity model will be operationalised.

3.2 Big Data Analytics Model Development

3.2.1 Data Preparation

Data input must truly correlate with customer win-back to ensure predictive analytics robustness. This will require using data input or features from the Company and run a correlation test to identify significant features for customer win-back.

Afterward, data cleansing will be used to fill in ‘Null’ fields, i.e. some customers will have missing values in some data fields (features). This will require data imputation. For numerical features, (quantitative) median value from the data set will be used. On the other hand, categorical values (qualitative) will use mode value from the data set. However, imputing data will only consider customer profiles with >90% of the fields filled.

3.2.2 Hypothesis Testing

To further validate feature selection, Logistic regression (Chi-square test) will be used to plot the correlation between features using P-value and 95% confidence level. This will be used to test the significance of each feature and to validate the results from the focus group.

The method used is mean decrease in accuracy which measures the extent which accuracy will degrade should the values of the attributes are shuffled randomly. The idea is if a variable is truly significant, mixing up the variable's values will ruin the accuracy. On the other hand, if the variable is insignificant, then it does not matter whether the attribute is shuffled, because the original was not crucial to the prediction in the first place.

3.2.3 Adaptive Machine Learning (ML)

To predict and identify significant features, Adaptive machine learning (ML) will be built. Due to clarity of results and implications required, Supervised learning is a ML technique of choice. Adaptive ML will deploy 3 different algorithms which are suitable for the type of dataset the Company stores, i.e. large number of factors will be deployed simultaneously to identify the best performed algorithm to implement: Decision trees, Random forest and Naïve Bayes.

Adaptive ML will select the most viable algorithm to implement the propensity model. This will undergo 3 main stages as shown in Figure 13.

Data points obtained from data preparation stage will be fed into the system to allow machine learning to take place. This allows the system to run statistical analyses on the datasets to uncover any patterns and learn to predict the outcome based on given data. The 3 algorithms deployed for this study are Decision trees, Random forest and Naïve Bayes.

The second stage is to predict the outcome of a different dataset from CRM based on the datasets that are fed into the model for machine learning. The new dataset will omit the result of win-back data field in order to map the predictions with the actual results. Each algorithm performance is measured based accuracy. Accuracy measures the proportion of all response cases that were correctly predicted. However, to understand the prediction and actual results, we firstly need to establish understanding of outcomes from predictive analytics. Table 2 illustrates the predictions versus actual results. True positive means that the predicted win-back matches with the actual outcome, i.e. successful referral. On the other hand, false positive means that the model predicts a win-back but the actual outcome is the contrary. False negative indicates a predicted failure to win-back but the actual result is that the customer was referred to sales department. Lastly, True negative means that predicted failure match the actual no refer outcome.

Table 2: Predictions and actual outcomes

		Actual	
		Yes	No
Predict	Yes	True positive	False positive
	No	False negative	True negative

$$Accuracy = \frac{True\ positives + True\ negatives}{(True\ positives + True\ negatives + False\ positives + False\ negatives)} \quad [11]$$

$$= \frac{Number\ of\ predictions\ that\ match\ with\ actual\ results}{Number\ of\ data\ points\ entered\ for\ Adaptive\ ML} \quad [12]$$

The accuracy score for each algorithm will be compared and the one with the highest accuracy score will be deployed in the implementation phase.

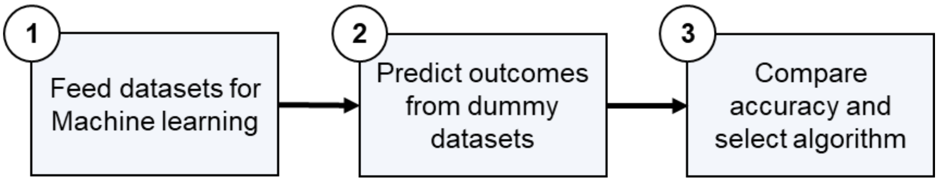

1
Feed datasets for Machine learning
2
Predict outcomes from dummy datasets
3
Compare accuracy and select algorithm

Figure 13: Algorithm selection process

3.3 Implementation

To prove the effectiveness and validate the impact of predictive analytics, the propensity model developed needs to be implemented and used to compare with the win-back performance without the model.

After identifying the suitable algorithm for propensity scoring, the model will be tested by customer win-back department.

3.3.1 A/B Testing

To measure this in comparable terms, A/B testing is proposed. This method requires 2 tests to run at the same time. Win-back staff will be separated into 2 teams: Team A and Team B, each with 5 staff. Team A will operate on the as-is process. Simultaneously, Team B will be assigned Big data analytics model as a Win-back tool.

At the end of each day, the results from both teams will be compared to one another. This will provide feedback for further iterative improvement in Data preparation and refinement of the Big data analytics model. Implementation will run over a duration of 3 weeks.

However, it is likely that there will be vacation or sick leaves during the implementation phase. Therefore, the performance measures will be adjusted to average or “per staff” basis to standardise and prevent potential bias.

3.4 Impact Analysis

3.4.1 Performance Measurement

Implementing propensity model will need to measure 2 aspects: effectiveness and efficiency. The following metrics will be used to measure the success of this study, which will serve the core basis for predictive analytics at scale:

1. **Effectiveness** measures how the model can help customer win-back department increase the average number of customers referred to sales

department per staff per day, i.e. *Average number of referrals*. This uses the following equation to compare the performance between as-is operation and model operation using predictive analytics:

$$Average\ number\ of\ referrals = \frac{Number\ of\ calls\ per\ day}{Number\ of\ staff\ per\ day} \qquad [13]$$

2. **Efficiency** measures how the model can help increase success rate or the percentage of customers successfully referred to the sales department for further product offering versus customers called, i.e. *Average referral rate*. This uses the following equation to compare the performance between the two modes of operation as mentioned above:

$$Average\ referral\ rate = \frac{Number\ of\ customers\ referred\ to\ sales\ department}{Number\ of\ calls} \qquad [14]$$

It is worth noting that within the limitation of project timeline, it is impossible to measure the success beyond the successful referral stage, i.e. to measure whether the customers referred to the sales department end up booking or transfer ownership of the townhouse or not.

3.4.2 Revenue Projection

The results at the end of the 3-week A/B test summarised into 2 metrics mentioned above will be used as key inputs to measure the. The difference in Number of referrals and Referral rates will be recorded over the 3-week implementation period. For the research to be determined a success, both metrics need to show improvements, i.e. predictive analytics and propensity scoring use-case can help improve effectiveness and efficiency statement is validated. The difference in number of refer cases will be used to estimate the impact on revenue by using the same conversion rate or sales success rate from previous year’s performance, which is 3.35% of sales closed from win-back refer cases. Ticket size is based on previous year’s revenue from win-back, with THB 3.25 million as the value input. From these parameters assumed, the following calculation is used to predict the impact of predictive analytics:

$$R_{Uplift} = N_{sales\_vol\_uplift} \times 3.25 \qquad [15]$$

$$N_{sales\_vol\_uplift} = N_{refer\_uplift} \times sales\ success\ rate \quad [16]$$

$$N_{refer\_uplift} = (N_{refer\_model} - N_{refer_{operation}}) \times 10\ [staff]\ \times 5\ [days] \times 52\ [weeks] \quad [17]$$

Where:

R_{Uplift} = Estimated increase in sales revenue from using predictive analytics

$N_{sales\_vol\_uplift}$ = Estimated increase in number of residential units sold from using predictive analytics

N_{refer_uplift} = Average increase in number of refer cases from win-back department to sales department per day

N_{refer_model} = Average number of refer cases per staff per day from team B with predictive analytics

$N_{refer\_operation}$ = Average number of refer cases per staff per day from team A with as-is operation

Sales success rate = Percentage of sales closed by sales department from win-back department refer cases at 3.35%

3.4.3 Test for Statistical Significance

Although the revenue projection might offer obvious improvement or outlook for using predictive analytics, a statistical method must be used to validate the hypothesis for improvement. To validate this, a Student’s t-test is used with 90% confidence interval, i.e. 9 out of 10 tests will find that predictive analytics would provide a superior performance.

The result of this study will help determine whether big data is the suitable for operationalisation. If there is a tangible improvement in performance, then predictive analytics will be replicated at scale and may expand to other use-cases beyond customer win-back.

4. Results and Analysis

4.1 Results from Focus Group

Conducting a focus group to identify CRM features that are significant to win back customers is a logical starting point. By doing this, the team can understand and see the difference between each staff and how each staff values each feature differently. Table 3 illustrates the consolidated ranking provided by 10 customer win-back staff. The dark green colour code represents the highest significance in each staff's opinion and experience.

This result is, to an extent, logical as when looking at customers to re-engage there must first present some sort of interests in the first place. Potential is a score given by sales staff based on their experience whether the line of questioning from the customers shows a pattern of someone who is interested in purchasing a residential property. Moreover, Oppage_month is also another logical thinking. If the customers came in over 3 months ago, it would be likely that they already purchased from a different project or from different real-estate developer. Score_project is from a sales point of view. It underlines the percentage of sales within a project. It makes sense to help drive a project which sees low sales rate to customers. This can easily be coupled with promotion, be it discount or other premiums, e.g. free air conditioners. Incidentally, Percent_promotion_expense which represents marketing expenses on a project is a factor that win-back staff views as a significant input. Buyers are prone to discounts or any privileges that come with a purchase and therefore, it would be more attractive to offer a product with such privileges.

Table 3: Consolidated feature significance from 10 customer win-back staff

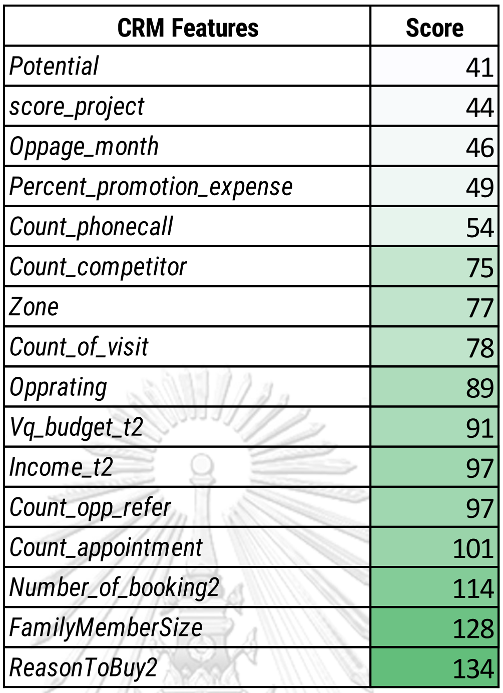
Staff CRM Features	1	2	3	4	5	6	7	8	9	10
Income_t2	10	13	10	13	9	2	14	8	7	11
Vq_budget_t2	11	14	11	12	14	1	5	7	6	10
Potential	1	3	2	3	5	9	10	6	1	1
ReasonToBuy2	9	15	13	15	15	13	12	15	14	13
FamilyMemberSize	8	16	12	14	16	16	15	16	15	14
Opprating	15	4	8	2	11	15	11	5	16	2
Count_appointment	6	12	14	10	10	14	13	14	8	15
Count_competitor	7	11	6	11	4	3	6	10	9	8
Count_phonecall	5	5	9	4	1	6	9	9	3	3
Count_opp_refer	12	9	5	16	12	12	8	13	10	16
Count_of_visit	13	10	15	7	2	4	7	12	4	4
Zone	14	6	4	6	13	11	2	2	12	7
Percent_promotion_expense	4	7	7	5	3	5	3	1	5	9
Number_of_booking2	16	8	16	8	7	7	16	11	13	12
Oppage_month	2	1	1	9	8	10	4	4	2	5
score_project	3	2	3	1	6	8	1	3	11	6

Table 4 summarises this by provide a final ranking based on ascending scores. This means that the feature with lowest score will provide the highest significance. The top 5 ranks are shown to be very similar in scores, with rank 1 and rank 5 merely 13 points apart. This indicates that all 5 features are comparably significant for customer win-back staff when looking to approach customers:

1) Potential
2) Score_project
3) Oppage_month
4) Percentage_promotion_expense
5) Count_phonecall

This result conforms with observations made from Table 3. While the top 5 are sensible, the lower rankings may present some counter intuitive traits. Vq_budget_t2 or customer's budget could have been placed with greater significance, as well as Income_t2 or income range of the customers. However, on a deeper level, these customer-based data can be easily recorded with falsified information. There is no evidence to support the claims from customers while the top 5 features are from internal data.

Table 4: Feature significance rank based on Focus group results



CRM Features	Score
Potential	41
score_project	44
Oppage_month	46
Percent_promotion_expense	49
Count_phonecall	54
Count_competitor	75
Zone	77
Count_of_visit	78
Opprating	89
Vq_budget_t2	91
Income_t2	97
Count_opp_refer	97
Count_appointment	101
Number_of_booking2	114
FamilyMemberSize	128
ReasonToBuy2	134

While it is simple to rank features with high significance, it is much more complicated to qualitatively pinpoint the less significant features with scores. The lower scores may present different significance scores entirely had the data been run through a statistical approach.

4.2 Results from Immersion Session

To reinforce the results from Focus group, an Immersion session is adopted to understand more on a-day-in-a-life of customer win-back department. From immersive session with 10 customer win-back staff, it is found that the win-back process undergoes 6 steps (as illustrated in Figure 14).

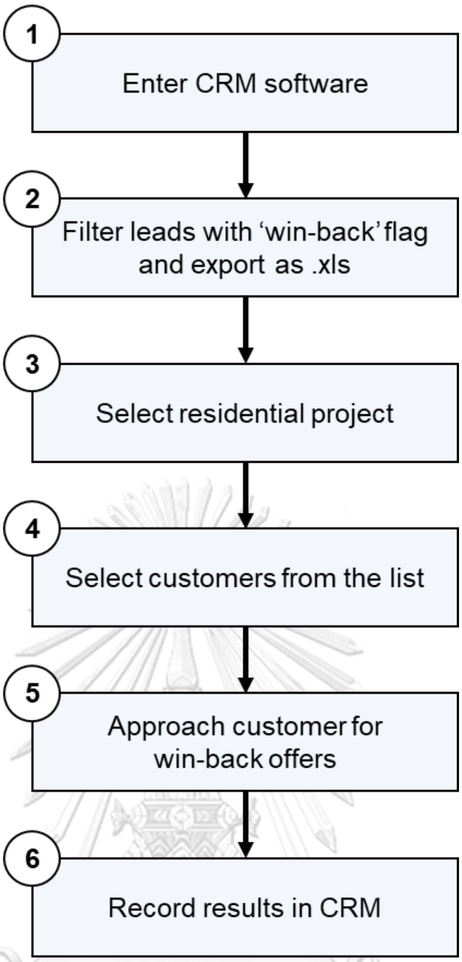

1
Enter CRM software
2
Filter leads with 'win-back' flag and export as .xls
3
Select residential project
4
Select customers from the list
5
Approach customer for win-back offers
6
Record results in CRM

Figure 14: Win-back as-is operation

Steps 3 and 4 are the main bottlenecks. Selecting residential projects is based on win-back staff preference. Moreover, customer list to win back is randomised based on 'win-back' filter from CRM only. This means that with merely 10 staff versus over 220,265 customer records, there are gaps for performance improvement using more targeted approach. This randomised operation is deemed to be ineffective in efficient.

For steps 3 and 4, each staff has different criteria that are considered when approaching customers. Step 3 for selecting residential projects, 40% of the staff use Zone as the key selector, i.e. they select projects to enter and approach customers based on the area in which they are comfortable with or have knowledge in, e.g. North, North East or South West zones. On the other hand, 60% select projects to win-back based on residential types, ranging from townhouses, single-detached houses and condominiums. For step 4, each staff considers different features from

CRM to help in approaching customers (as shown in Table 5). What immersion uncovers is that there is no standardised process in ensuring maximum performance or success from win-back attempts. However, these features reflect significance level based on results from Focus group. 70% of win-back staff consider Oppage_month (Rank 3 from Focus group) as one of the main criteria while 60% consider Potential as another (Rank 1 from Focus group).

These discrepancies may need to be formalised and data from CRM will need to be verified for patterns that can validate or void these significance claims.

Table 5: Customer win-back staff personal criteria when selecting projects and customers to win back

Staff	Project selection criteria	Customer selection criteria*
1	Zone	1. Potential 2. Oppage_month 3. Count_phonecall
2	Residential type	1. Oppage_month 2. Vq_budget_t2
3	Residential type	1. Number_of_booking2 2. Potential
4	Zone	1. ReasonToBuy2 2. Count_competitor
5	Zone	1. Oppage_month 2. Potential
6	Residential type	1. Opprating 2. Potential
7	Residential type	1. Opprating 2. Oppage_month
8	Residential type	1. Potential 2. Opprating 3. Oppage_month
9	Residential type	1. Oppage_month 2. Count_competitor
10	Zone	1. Potential 2. Oppage_month 3. Count_competitor

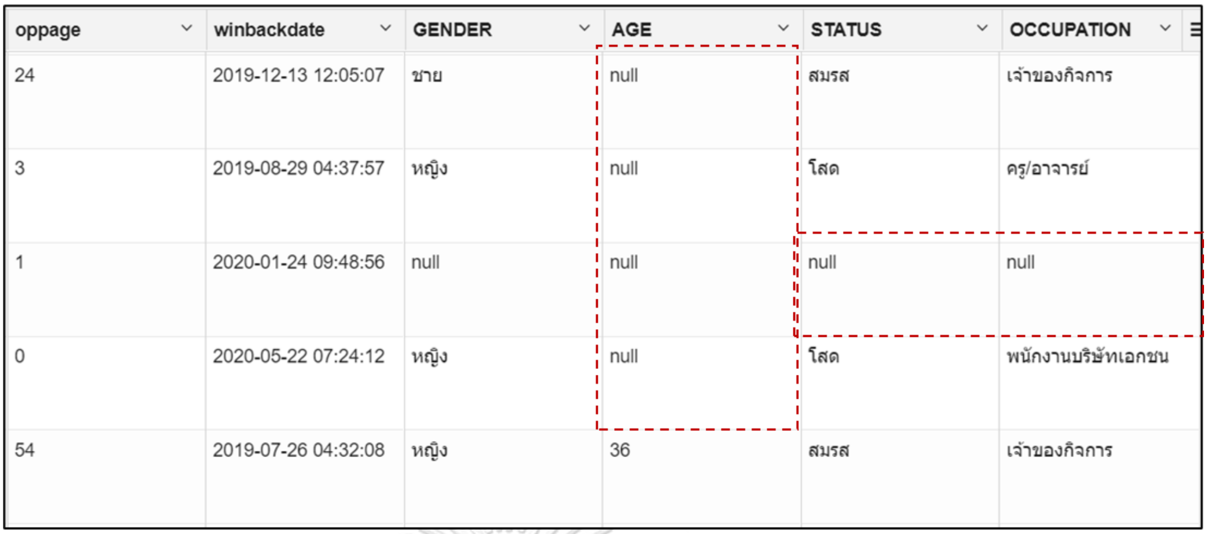
4.3 Data Preparation

To test whether the claims for feature significance are valid, the team needs to run a test on data from CRM to identify the significance level. However, the data is likely to be incomplete and some datasets may be invalid to use. Inclusion will lessen the effectiveness of the propensity model. Therefore, a data preparation known as data imputation needs to be employed.

It is found that there is a considerable amount of missing values. This is mainly caused by human error, i.e. sales staff did not make it mandatory for customers to fill certain data. More likely, this is caused by pre-sale days which see substantial customer visits and limit the average time spent per customer, thereby limiting sales staff to focus on a few key information needed to make the sales offer.

The team adopted 2 approaches for data preparation: (1) data imputation for datasets with 'null' fields or missing value <10% (2) discard datasets with 'null' field >10% as this categorical estimates will skew the accuracy of data inputs.

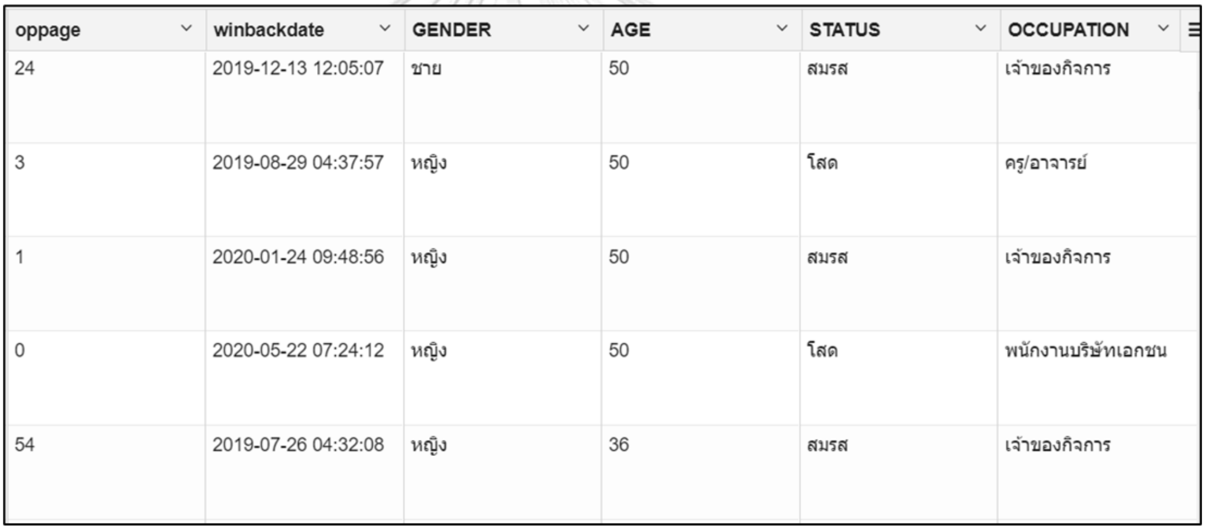
Data imputation fills missing values with estimates of what that missing values may be. This approach analyses the complete datasets and identify patterns from combination of data points and use the patterns to predict what the missing values for incomplete datasets may be. Figure 15 illustrates samples of missing data from CRM that are imputed (Figure 16). It is seen that the mode age is 50 years old for customers, which are imputed in the 4 sample missing values. Since this developer offers a price-sensitive range of residential projects, this can mean 3 scenarios: (1) Most customers (or those with available data from CRM) are from low-middle income groups which have established their wealth in their 40-50 year-old range (2) Customers might look to purchase for investment, i.e. to rent or anticipate price increase (3) Customers might purchase for family member, e.g. children looking to move out from their family home or children entering university and needing personal space. The imputed data status "สมรส" (married) and occupation "เจ้าของกิจการ" (business owner) in Figure 16 are likely to support scenario (2) and scenario (3).



oppage	winbackdate	GENDER	AGE	STATUS	OCCUPATION
24	2019-12-13 12:05:07	ชาย	null	สมรส	เจ้าของกิจการ
3	2019-08-29 04:37:57	หญิง	null	โสด	ครู/อาจารย์
1	2020-01-24 09:48:56	null	null	null	null
0	2020-05-22 07:24:12	หญิง	null	โสด	พนักงานบริษัทเอกชน
54	2019-07-26 04:32:08	หญิง	36	สมรส	เจ้าของกิจการ

Figure 15: Illustration of missing values

Source: Adapted from company''s CRM



oppage	winbackdate	GENDER	AGE	STATUS	OCCUPATION
24	2019-12-13 12:05:07	ชาย	50	สมรส	เจ้าของกิจการ
3	2019-08-29 04:37:57	หญิง	50	โสด	ครู/อาจารย์
1	2020-01-24 09:48:56	หญิง	50	สมรส	เจ้าของกิจการ
0	2020-05-22 07:24:12	หญิง	50	โสด	พนักงานบริษัทเอกชน
54	2019-07-26 04:32:08	หญิง	36	สมรส	เจ้าของกิจการ

Figure 16: Data imputed to legitimise missing values

Source: Adapted from company's CRM

4.4 Feature Significance from Hypothesis Test

A closer look at the variables used in predictive analytics should be explored further. In doing so, we will plot the variables importance for random forest, and give bar plots for likelihoods and posterior for Naïve Bayes method. Variable importance indicates the extent to which each variable affects the accuracy. The likelihood indicates the likely characteristics (for every feature) each response class has. Posterior probability, on the other hand, indicates the probability of each response class, given observations have some specified attributes (as given by the values of every feature). In brief, both likelihood and posterior indicate roughly the “properties” each response class has in terms of probabilities associated with each variable.

On the other hand, Decision tree model gave interpretable results, but more information can be garnered from bar plots of probability outputs attained from Naïve Bayes, and variable importance plots output from Random forest. This is because features that are deemed significant will dominate the statistical analysis and render other features’ importance. Random forest and Naïve Bayes focus on analysing characteristics on all features and so this study will focus exclusively on using Random forest and Naïve Bayes.

Mean decrease accuracy, as a measure of how the significance the feature is based on its impact from certain shuffles or changes, is used to highlight on the key features that are significant for customer win-back. As illustrated in Figure 17, it is found that Potential is evidently the most significant feature, with >100 mean decrease accuracy score. This is followed by score_project, count_phonecall and oppage_month with similar significant levels ranging between 75 and 85. However, these are still critical features as potential alone cannot determine the characteristics of a customer to approach for win-back offers. Another group of features with mean decrease accuracy between 40 and 60 are also included as significant. These are precent_promotion_expense, count_competitor, count_opp_refer and zone. These features do show similar significance as ranked by customer win-back staff from Focus group and Immersion session. Moreover, these features are logical when mapped with customer behaviour.

It is natural that having potential flagged on customers will be the first key to indicate if a customer is still interested. This usually derives from enquiries and behaviours that demonstrate eagerness to purchase a residence. Score_project is also logical as more options would mean there is a higher chance of a customer being impressed by an available plot. Count_phonecall and oppage_month are 2 features that can help indicate freshness of the customers. If a customer has been called several times, it is likely that this will spark unease and irritation, rendering the win-back effort. On the other hand, the longer the customers have been flagged as ‘win-back’ the more likely they will not be interested in purchasing. This is because they are likely to have purchased a property elsewhere.

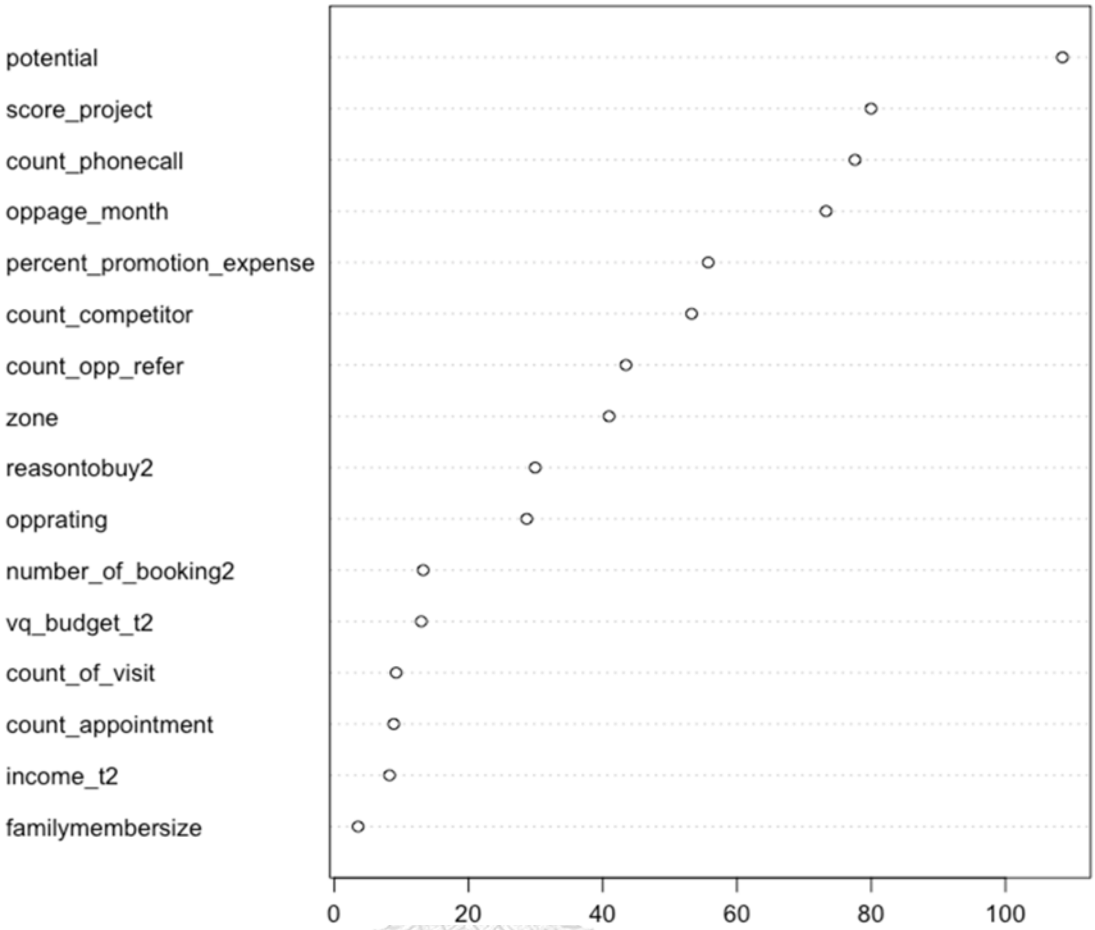


Figure 17: Feature significance from Hypothesis testing

Further exploration into features selected show conditions that are critical for win-back. With the measures being probability of win-back in percentage and mode for the highest win-back probability, Figure 18 and Figure 19 illustrate the results for each significant features.

Potential, measured in 5 increments: A=100, B=80, C=60, D=40 and E=20, surprisingly sees 67% chance of win-back if the customer is classed as C (as shown in Figure 18). This is something which may be counter-intuitive and that being classed as A would give the highest probability. However, there is an argument that observes how a customer or lead that enquires on the project with great interests may be a mystery shopper for competitors within the area. Alternatively, customers with potential classed A are likely to be from the more affluent background. This implies that there are greater options for them to consider and are likely to purchase from a more luxury real-estate developer – the company is a mass-market real-estate developer. This notion is supported by the income range of 68% win-back probability in the THB 20,000-50,000 region, i.e. the less affluent background. Further support is made from ~90% of win-back probability being placed on budget being < THB 3 million. However, this can be argued that most residential projects offered by the company are budget friendly.

On the other hand, a noteworthy result from Figure 19 is that most of high win-back rate are from project with score_project = 80. This would imply that projects with around 80% of the project sold are more likely to trigger customer win-back. This might seem counter intuitive as suggested earlier with higher vacancy might indicate more attractiveness. However, this does make logical sense in commercial terms. Projects with low vacancy can signal low interests from various reasons, e.g. unappealing location, unattractive pricing, better alternatives within area nearby. However, projects with around 80% sales can imply 2 scenarios: (1) the project is adequately attractive as there is a considerable booking proportion and (2) the project still has vacancies (20%) which can offer variety or choice for customers to consider.

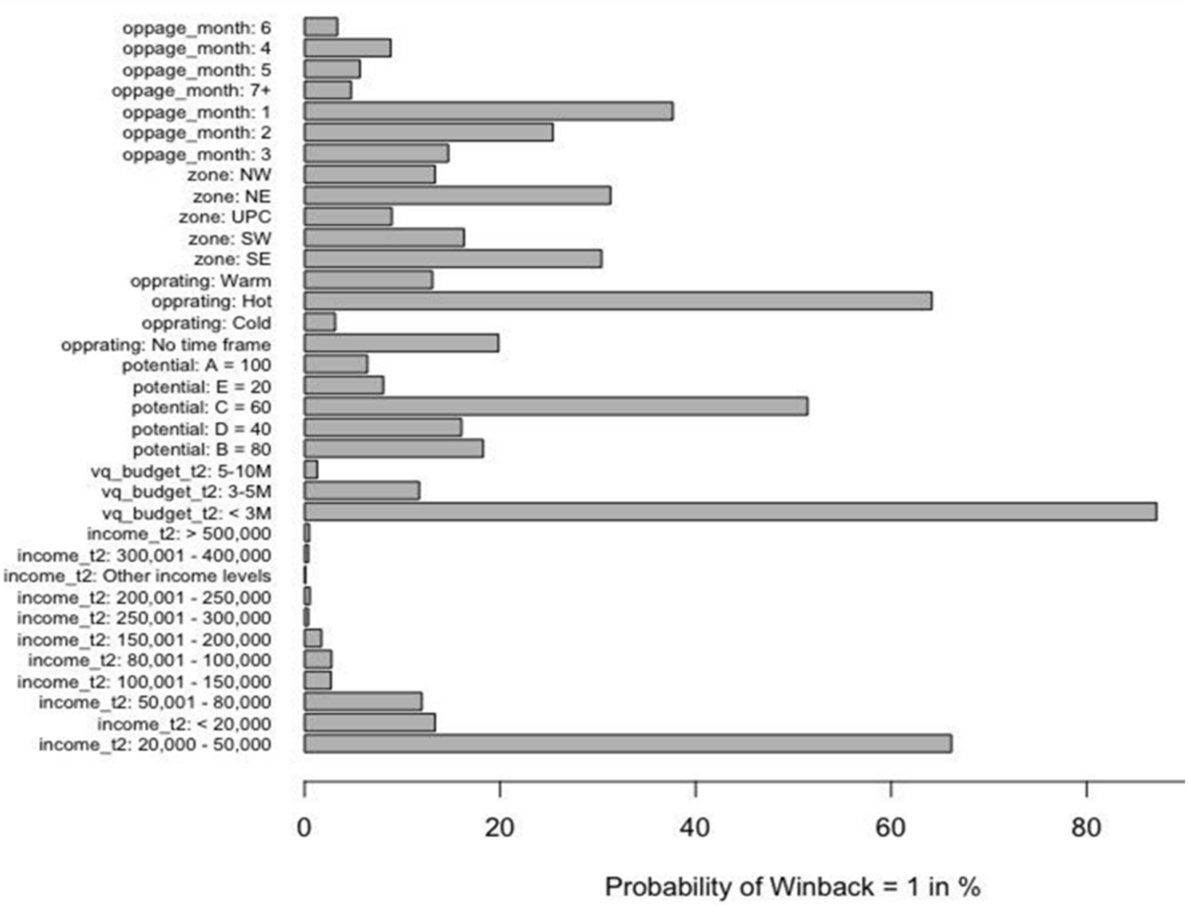


Figure 18: Insights from feature selection [1]

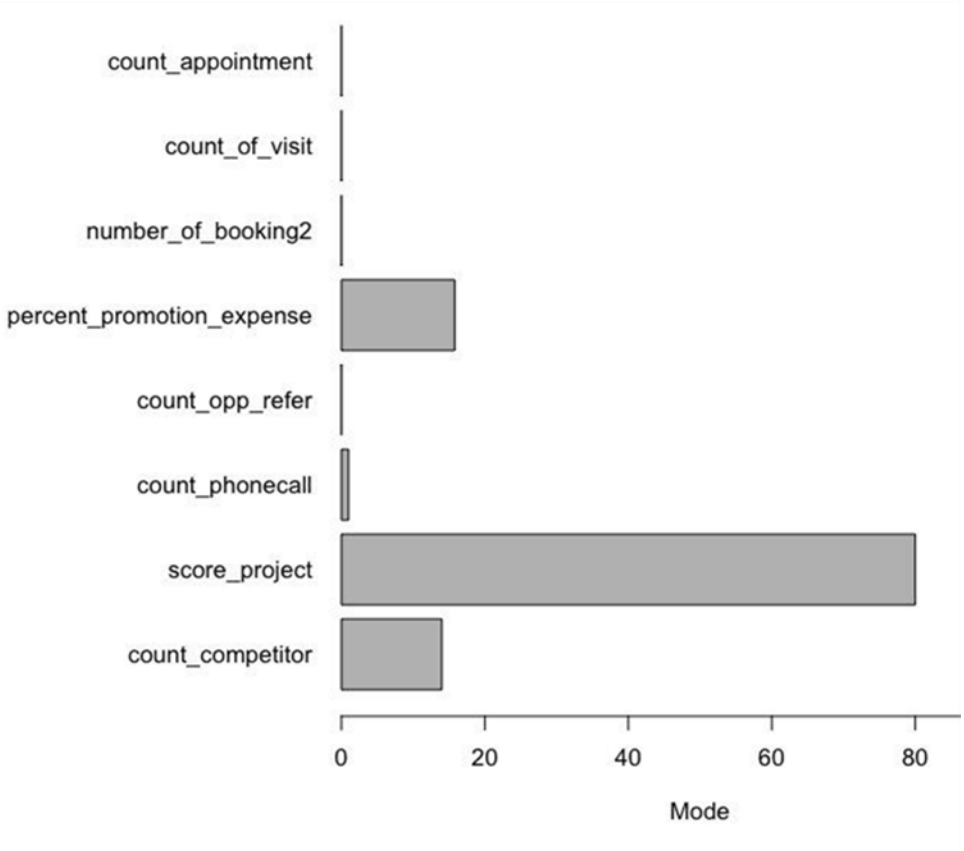


Figure 19: Insights from feature selection [2]

4.5 Algorithm Performance Comparison

From feature selection, we have identified features that are critical for win-back success, i.e. significant features that impact the performance. From the 3 algorithms explored, the most performed will have to be chosen for propensity model development. This will measure the accuracy based on how accurate they are in predicting the outcome of a win-back sample.

To test the performance of a model, the approach used is to randomly allocate 80% of the whole dataset to be the training sample and the remaining 20% as test sets (Pareto principle). The results from running the datasets on 3 algorithms are shown in Figure 20, with Random forest attaining the highest accuracy.

Feature significance has shown traits of interactions among features. This renders Decision tree and Naïve Bayes inapplicable for scenarios with interactions.

While Naïve Bayes is believed to be highly scalable, handling both discrete and continuous variables are simple, problems arise when it is used to train data with more complex characteristics. Its discretisation of continuous variables also disregard many possibilities within the values. Hence, it is evident in Figure 20 that Naïve Bayes presents the least accuracy. This realises Rennie, Shih and Karger (2003) argument against Naïve Bayes as offering too simplicity, discarding data with high skewness – a characteristic that is apparent in real-world situation.

On the other hand, Decision tree is not suitable for categorical values, i.e. fields with categories rather than numerical values. Moreover, the limitation to this algorithm is that the fundamental is biased in favour of features with more levels or more depth, rendering other features' significance. Thus, it is seen in Figure 20 that Decision tree shows less accuracy than Random forest.

Random forest provides the highest accuracy across all scenarios (as illustrated in Figure 20). The model is an extension of Decision tree. Random forest is a collection of trees running on the same training datasets. This does not simplify the prediction like decision tree but use a majority of outcomes across all decision trees under random forest to predict the most probable outcome. Therefore, Random forest excels at predictive analytics, with average accuracy score of 0.84. This means 84% of predictions on testing datasets based on machine learning from training datasets are true.

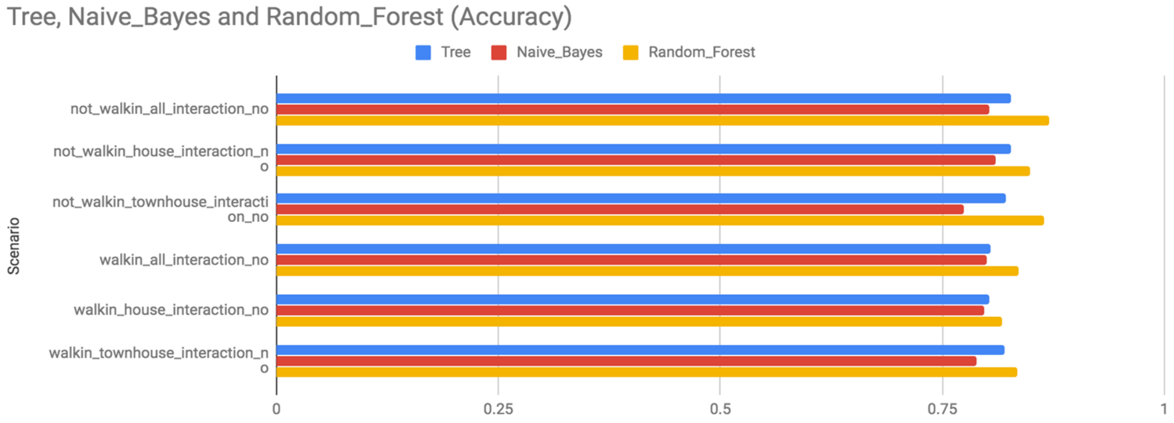


Figure 20: Machine learning algorithm accuracy comparison

4.6 Results from A/B Test

4.6.1 A/B Testing Process

As stated, to test the impact of predictive analytics, the customer win-back department consisting of 10 staff will be split into 2 teams: team A with as-is operation and team B with propensity model or predictive analytics tool. With this new way of working, the process by which team B will comply will need to change accordingly.

Based on as-is operation observed through Immersion session, an alternative approach to simplify propensity model was created, making it more optimal and closest to as-is operation as possible. Figure 21 illustrates the new approach for team B. The data will be run on a daily basis from CRM and extracted to prioritise customer lists and cluster in terms of residential projects and area in .xls file type. This is to minimise the time that win-back staff will need to select projects. The staff can then select from a priority list of projects with priority list of customers, each with a propensity score to indicate the likelihood that they can be won.

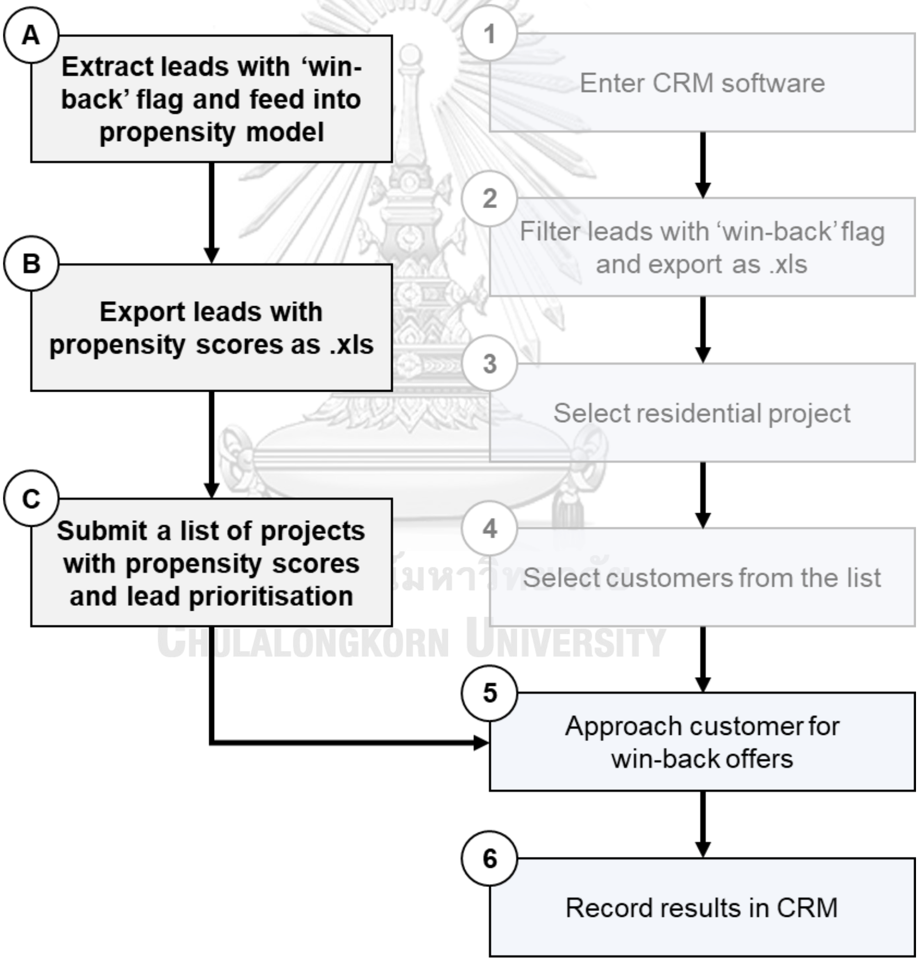

A
Extract leads with 'win-back' flag and feed into propensity model
B
Export leads with propensity scores as .xls
C
Submit a list of projects with propensity scores and lead prioritisation
1
Enter CRM software
2
Filter leads with 'win-back' flag and export as .xls
3
Select residential project
4
Select customers from the list
5
Approach customer for win-back offers
6
Record results in CRM

Figure 21: Predictive analytics for customer win-back approach

As A/B testing process is established, conditions need to be standardised in order ensure that there is no bias in favour of one or the other test stream. Initially, the test was said to be measured, in comparison, using total number of referrals and average referral rate.

- Total referral is the sum of customers being referred to sales department through the 3-week run.
- Average referral rate is the proportion of customers being referred to sales department versus total number of customers contacted (Denoted in (B))

However, the total number of customers referred might be subjected to bias, e.g. different capacity. Therefore, the metric used in standardised by using average number of referrals (Denoted in (A)).

Initially, the study also aimed to consider the quality of data improved through measuring the change in percentage of customers reached, i.e. number of calls that are picked up versus number of out-going calls. However, recording the number of calls picked up or not picked up take a considerable of time to record in practice, both during operations and after. This renders the capacity that the team has. Therefore, this objective was be eliminated from the test. However, the increase in quality of data in the model can inferred to by the improvement in number of duplicate calls. Duplicates are customers that have:

- Been contacted by the win-back team
- Already purchased from the company
- Already purchased from competitors

These behaviours or outcomes should have been eliminated and flagged as ‘lost’ in CRM which will ultimately not be in the win-back list. It poses time wastage and render efficiency and effectiveness. To conclude, the measure of data quality is changed from initial measure of percentage of reach to average number of duplicates discovered during testing (Denoted in (C).

$$Average\ number\ of\ referrals = \frac{Number\ of\ customers\ referral\ \ per\ day}{Number\ of\ staff\ per\ day} \tag{A}$$

$$Average\ referral\ rate = \frac{Number\ of\ customers\ referred\ to\ sales\ department}{Number\ of\ calls} \tag{B}$$

$$Average\ number\ of\ duplicates = \frac{Total\ number\ of\ duplicate\ cases}{Number\ of\ days\ tested} \tag{C}$$

4.6.2 Performance Comparison from A/B Testing

From the metrics mentioned in 4.6.1 the result from each day over the 3-week testing is recorded in Table 6. It is seen that day 1 (1 June) sees a mere 106 number of phone calls using the propensity model compared to 204 calls from as-is process. This is because team B is easing into the process set -up for propensity model. However, despite this, the referral rate in day 1 shows a successful 25.5% using the model compared to 20.1% from as-is operation. This marks a promising kick-off for testing. The following weeks show a promising trend, with model operation dominating as-is operation. However, there are 3 days where the team B performed less satisfactorily than team A in terms of referral rate (5, 11 and 17 June). This is mainly due to a mandate from management for team A to focus on offering special promotion for specific residential projects in Bangkok area. As seen from 4.4, promotion is a significant feature that affect customer win-back. Therefore, bias is created where team A does not have data that factor in such mandate. The marketing push from management does not immediately update into CRM which means that the data fed into propensity score do not have percentage_promotion_expense calibrated to match reality.

Despite bias being created, propensity model largely dominated throughout the 3-week A/B testing, with team B achieving a remarkable result of having reached 2,747 customers and referred 516 cases to the sales department. In contrast, team A which follows as-is process achieved a lesser outcome, with 2,366 customers reached and 401 customers referred to sales department.

To reinforce this outcome and validate the hypothesis whether predictive analytics can help increase effectiveness, we need to consider not the absolute number but at the average calls per staff. Firstly, the most obvious metric to measure is the number of calls a staff can handle per day. Using the following calculation, the first level efficiency can be measured before moving to the actual performance metric, i.e. average number of refer cases per staff per day (Figure 22):

$$Average\ number\ of\ customers\ per\ staff\ per\ day = \frac{Number\ of\ customer\ calls}{Number\ of\ staff} \qquad \text{(D)}$$

Team A achieved 40.8 or 40 calls per day while team B achieved 43.6 or 43 calls per day. In terms of absolute efficiency, predictive analytics shows a 6.8% increase in capacity. This is expected as the staff spend less time selecting residential projects from the database. Pre-selections based on priority help streamline the way of work noticeably.

Table 6: Daily performance comparison between as-is operation and propensity model

	Number of refer cases [value]		Total number of customers reached [value]		Referral rate [%]		Refer cases per staff [value]		Number of staff [value	
Date	*As-is process*	*Model*	*As-is process*	*Model*	*As-is process*	*Model*	*As-is process*	*Model*	*Team A*	*Team B*
01-Jun	41	27	204	106	20.1	25.5	8.2	6.8	5	4
02-Jun	38	40	181	183	21	21.9	7.6	10	5	4
04-Jun	46	57	186	222	24.7	25.7	9.2	11.4	5	5
05-Jun	36	39	188	254	19.1	15.4	9	7.8	4	5
08-Jun	24	38	125	176	19.2	21.6	8	9.5	3	4
09-Jun	28	29	173	165	16.2	17.6	7	7.3	4	4
10-Jun	16	28	160	147	10	19	4	5.6	4	5
11-Jun	20	24	96	149	22.4	16.1	8	6	4	4
12-Jun	20	33	172	233	11.6	14.2	6.7	6.6	3	5
15-Jun	29	59	144	232	20.1	25.4	9.7	11.8	3	5
16-Jun	26	39	166	249	15.7	15.7	6.5	7.8	4	5
17-Jun	44	48	258	291	17.1	16.5	8.8	9.6	5	5
18-Jun	12	20	155	134	7.7	14.9	2.4	5	5	4
19-Jun	21	35	158	206	13.3	17	5.3	8.8	4	4
	401	**516**	**2366**	**2747**	**17.0**	**19.0**	**7.2**	**8.1**	**58**	**63**

The former measure is a fundamental improvement which helps drive the key performance measures: (1) Average number of refer cases per staff per day [Effectiveness] and (2) Average referral rate [Efficiency]. A comparison between the average number of refer cases per staff per day is made and it is found that team A achieved a 7.2 or 7 refer cases while team B achieved 8.1 or 8 refer cases per staff per day. Taking team A’s performance as a representation of status quo, this indicates a major improvement of 13.5% in terms of effectiveness, i.e. getting more customers referred to sales department. Figure 22 illustrates the performance of team A and team B graphically.

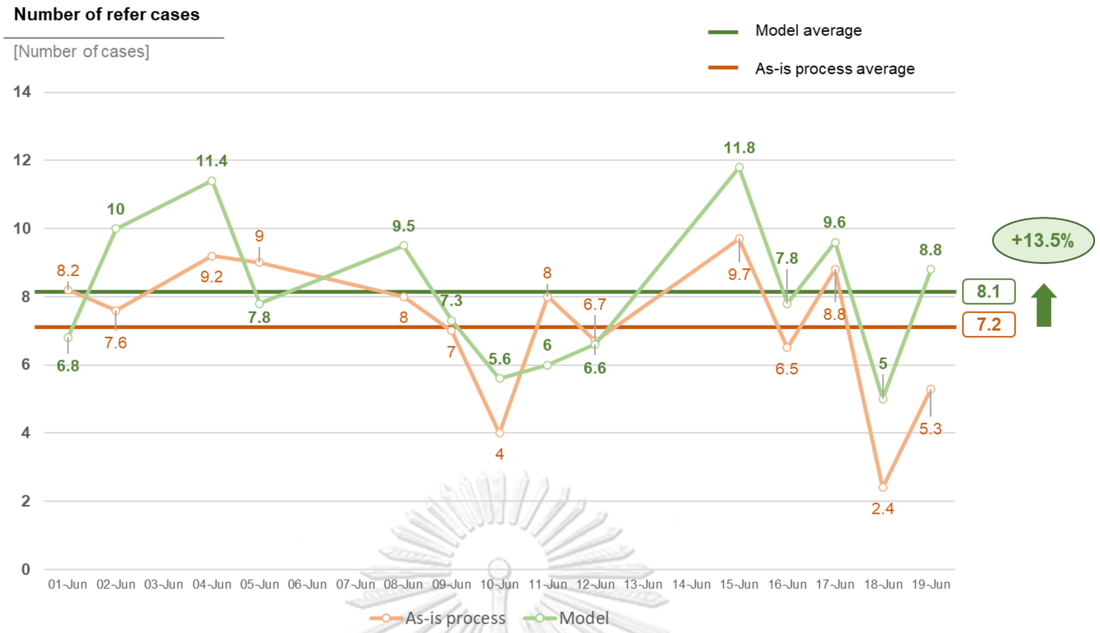


Figure 22: Number of refer cases per staff per day

While effectiveness shows an evident increase from status quo, efficiency is also another key metric that is measured. It is found that team A achieved a 17.0% while team B attained an 19.0% average referral rate. This means that for every 100 calls, as-is operation can create 17 refer cases while predictive analytics can help increase this to 19 refer cases. Similar to the effectiveness measure scenario, average referral rate for as-is operation can be taken as a representation of customer win-back average performance. This means that predictive analytics can help increase referral rate by 2.0%. Comparatively, this means that customer win-back sees an improvement of 11.8%.

Consider Figure 23, it can be inferred from performance graphs that referral rates see a minor downward trend. This is likely to be driven by 2 main reasons. Firstly, team A was subjected to direct mandate from top management regarding specific residential projects to engage customers. These, coupled with aggressive marketing promotion, will likely gain higher customer interests. As a result, higher referral rate is likely seen by sales department. However, as marketing push subsides, there is less incentive for customers to be referred to sales department. Secondly, marketing promotion trigger churned customers to re-enter and reconsider purchasing from the company. However, this attractiveness declines over time as promotional offers are taken up by potential customers.

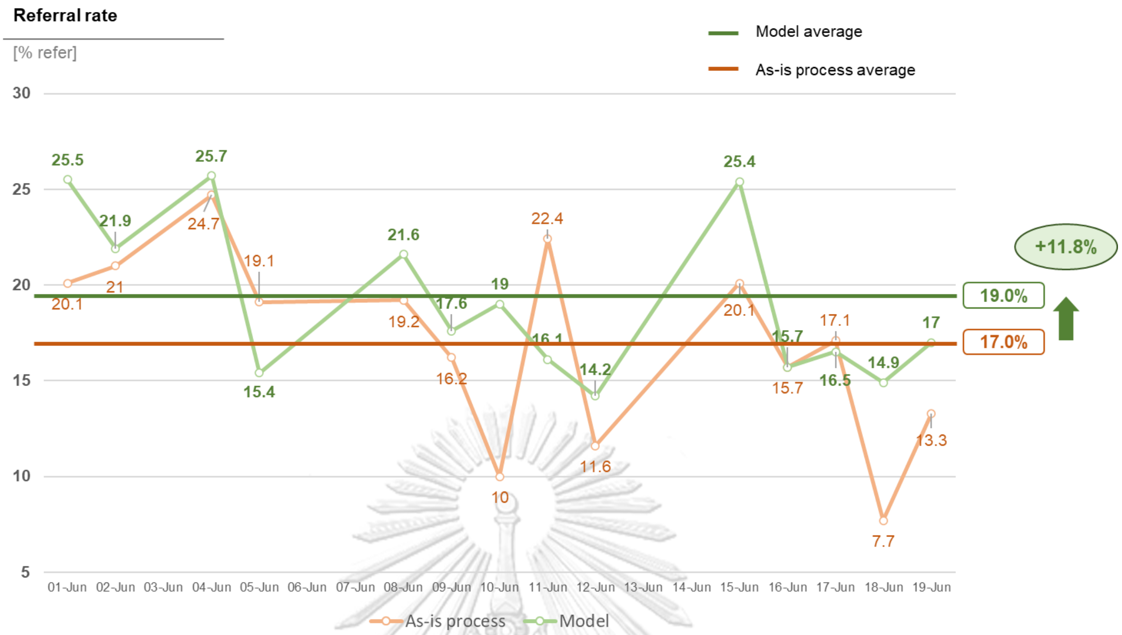


Figure 23: Percentage of refer cases to Sales department

As mentioned, number of duplicate cases – cases that have been contacted for win-back – is a measure that determines if the propensity model helps improve the quality of data. Despite the aim of the study is to prove that predictive analytics can increase efficiency and effectiveness, using data science automatically filters out errors or data that are duplicated. Figure 24 shows that team A has seen an average of 5.6 calls that are duplicates per day over the 3 weeks period. This means that out of the 40.8 calls per staff per day, approximately 35 calls are actual customers that can be offered incentives for referral. On the other hand, team B saw a mere 1.0 duplicate per day per staff out of 43.6 calls per day per staff. This means that team B can achieve approximately 42 calls per day, 7 calls more than team A. This 20% higher productivity indicates a substantial improvement and demonstrates a strong evidence for predictive analytics in increasing data quality.

This remarkable uplift in quality is primarily driven by analytics result that uncover the “winning” patterns and prioritise datasets that match with such characteristics. Therefore, the calls made by win-back staff are likely to be pre-screened by “winning” characteristics. However, if win-back staff were able to increase the capacity to match with the number of customers in the system, the difference in number of duplicates between as-is operation and propensity model will decrease, i.e. uplift in quality will lessen. This is because the staff would have reached lower potential profiles with characteristics that are less matched with the “winning” set.

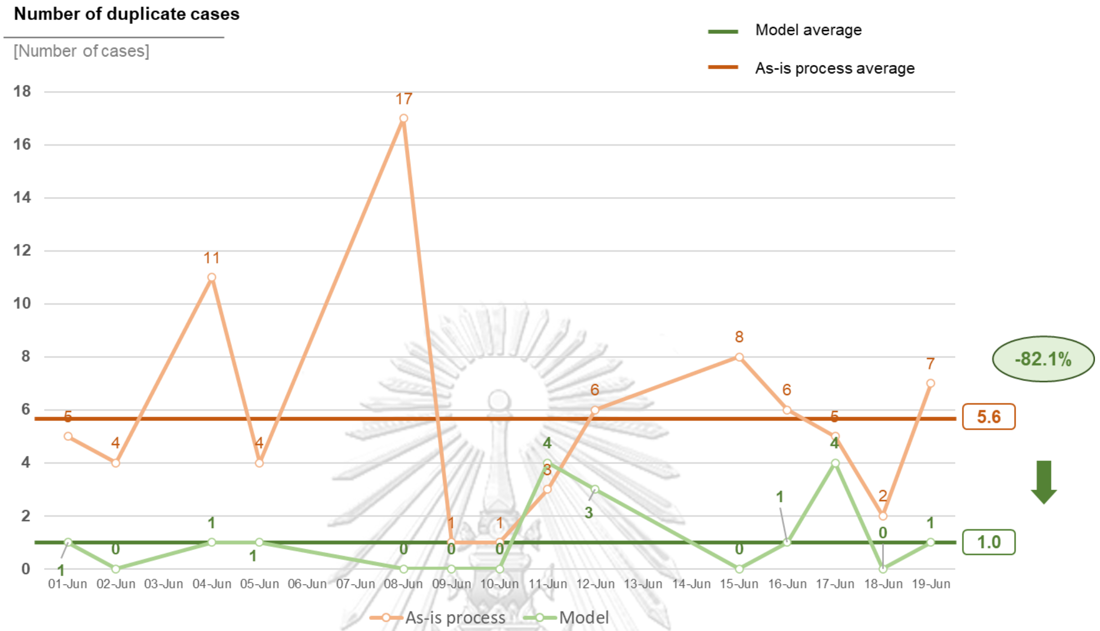


Figure 24: Number of duplicate cases discovered during Win-back calls

4.7 Impact Analysis and Revenue Projection

From results in the previous sections, initial hypotheses (1) (2) and (3) in 1.5 are validated. It is evident that predictive analytics show an increase in efficiency in terms of number of calls made in hypothesis (1). Moreover, efficiency in terms of referral rate is shown to increase as hypothesised in (2). Despite percentage of calls picked up by customers in hypothesis (3) is not validated directly, the core premise that predictive analytics helps improve data quality so win-back department can decrease wastage is addressed through decrease in number of duplicated calls. From these validations, it is methodical that hypothesis (4) that considers predictive analytics to increase revenue projection for the company will be validated.

This will require estimation based on calculations mentioned in 3.4.2. Firstly, the estimated increase in number of refer cases must be established from [17]:

$$N_{refer\_uplift} = (N_{refer\_model} - N_{refer_{operation}}) \times 10[staff] \times 5\ [days] \times 52\ [weeks]$$

$$N_{refer_{uplift}} = (8.1 - 7.2) \times 10 \times 5 \times 52 = 2{,}340\ units$$

Secondly, the estimated sales closed from this 2,340 units uplift will need to be determined in order to project revenue increase. Using [16] to estimate number of sales volume increase and 3.35% as sales success rate:

$$N_{sales\_vol\_uplift} = N_{refer\_uplift} \times sales\ success\ rate$$

$$N_{sales\_vol\_uplift} = 2{,}340 \times 0.0335 = 78.4\ units \approx 78\ units$$

Thirdly, this estimated number of sold units (using transfer to indicate a complete sales) will need to be estimated further for revenue uplift using average ticket size from past year of THB 3.25 million:

$$R_{Uplift} = 78 \times 3.25 = 253.5$$

This means that if the company decides to fully adopt predictive analytics with 10 staff, the revenue uplift will be THB 253.5 million per year. This marks a significant increase of 20.9% from THB 1,210 million from previous year. This evidently supports hypothesis (4) that predictive analytics can help increase revenue for customer win-back function.

4.8 Test for Statistical Significance

Although impact analysis illustrates evident business benefit, the results from 2 cases are compared using Student’s t-test to reinforce that the impact projection shows statistical significance.

Based on Table 6, each column comparing Operation and Model performance is tested for validation. These are (1) Number of refer cases (2) Total number of customers reached (3) Referral rate and (4) Refer cases per staff. The following results are shown from Student’s t-test with 90% confidence interval. For the purpose of validating that predictive analytics (Model) is superior to As-is process, a one-tailed test at 0.05 significance level is used.

Table 7: Student's t-test result for Refer cases

	As-is process	*Model*
Mean	28.64285714	36.85714286
Variance	115.7857143	134.2857143
Observations	14	14
Pooled Variance	125.0357143	
Hypothesized Mean Difference	0	
df	26	
t Stat	-1.943577152	
P(T<=t) one-tail	0.031422512	
t Critical one-tail	1.314971864	
P(T<=t) two-tail	0.062845025	
t Critical two-tail	1.70561792	

From Table 7, it is seen that the difference between Operation and Model performance for number of cases referred to sales department shows statistical difference and significance, with p-value = 0.03142 or less than 0.1. We further observe that mean value for Model is significantly higher than that of As-is process (29% higher), indicating an observable benefit for adoption of predictive analytics. However, since the experiment remains a prototype or a proof of concept, there exists key considerations for further developments. An observation can be made on the variability of outcomes. Result from Model has a slightly higher variance than As-is process, which reflects in higher standard deviation, but to a small extent. Moreover, the significantly higher mean value could offset this if the difference maintains throughout a longer period of testing.

Table 8: Student’s t-test result for Total number of customers reached

	As-is process	*Model*
Mean	169	196.2142857
Variance	1404.769231	2824.796703
Observations	14	14
Pooled Variance	2114.782967	
Hypothesized Mean Difference	0	
df	26	
t Stat	-1.565716247	
P(T<=t) one-tail	0.064753221	
t Critical one-tail	1.314971864	
P(T<=t) two-tail	0.129506442	
t Critical two-tail	1.70561792	

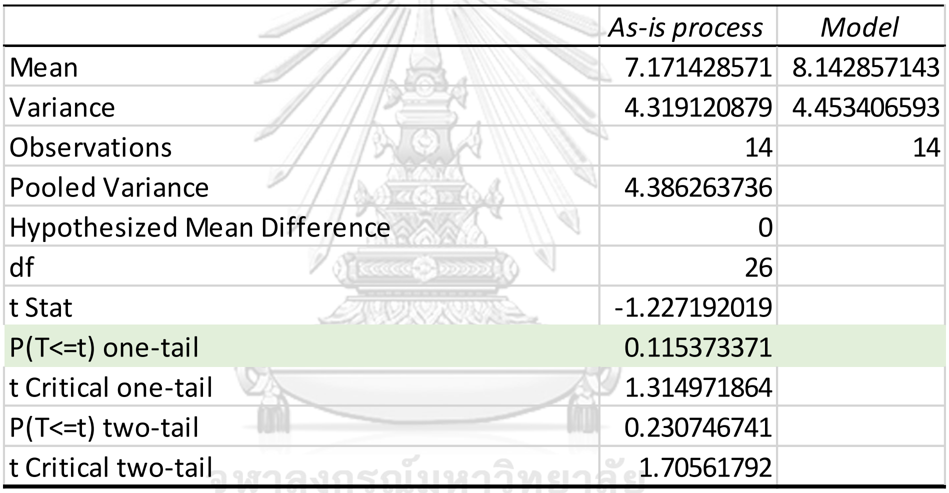
From Table 8, it is seen that the difference between Operation and Model performance for number of customers reached shows statistical difference and significance, with p-value = 0.06475 or less than 0.1. Although there is a considerable increase in mean value for Model, the result from Model shows a standard deviation of 53.15 calls while As-is process shows a value of 37.48. This difference could undermine the apparent mean difference. However, this does not factor in number of staff which means that unequal inputs (staff making calls) are not comparable. Therefore, Refer cases per staff will provide a more accurate and comparable variation of results.

Table 9: Student’s t-test result for Referral rate

	As-is process	*Model*
Mean	17.01428571	19.03571429
Variance	24.12285714	17.50554945
Observations	14	14
Pooled Variance	20.8142033	
Hypothesized Mean Difference	0	
df	26	
t Stat	-1.172269657	
P(T<=t) one-tail	0.125859257	
t Critical one-tail	1.314971864	
P(T<=t) two-tail	0.251718514	
t Critical two-tail	1.70561792	

From Table 9, it is seen that the difference between Operation and Model performance for Referral rate shows statistical difference and significance, with p-value = 0.1259 or more than 0.1. This does not indicate a statistical significance when compared the As-is process with Model results. Although the conversion rate shows a significant improvement with mean value of referral rate for Model 11.8% greater than that of As-is process, the test would require a longer duration to establish statistical significance. Despite this, the confidence level is set at 90%. Since the test is an experiment, this confidence level can be adjusted and a lower confidence level, i.e. less strict conditions, could have indicated statistical significance.

Table 10: Student’s t-test result for Refer cases per staff



	As-is process	*Model*
Mean	7.171428571	8.142857143
Variance	4.319120879	4.453406593
Observations	14	14
Pooled Variance	4.386263736	
Hypothesized Mean Difference	0	
df	26	
t Stat	-1.227192019	
P(T<=t) one-tail	0.115373371	
t Critical one-tail	1.314971864	
P(T<=t) two-tail	0.230746741	
t Critical two-tail	1.70561792	

From Table 10, it is seen that the difference between Operation and Model performance for Refer cases per staff shows statistical difference and significance, with p-value = 0.1154 or more than 0.1. This indicates that the mean value of 1 case higher per staff per day for Model which sets a 13.5% increase in productivity remains statistically insignificant. Although the business implication is apparent and acceptable, the statistical analysis suggests the contrary. However, extending the A/B test timeline could have established statistical significance. On the other hand, comparing the variances between Model and As-is process shows a negligible difference. With reference to Table 8, this shows (1) statistical validation for improvement in efficiency and (2) similar variation in terms of performance.

From the 4 comparisons made between Operation and Model (Team A and Team B), it can be concluded that, although the sample size is limited, predictive analytics offers superior performance in a business implication but would require an extended period to validate if this provides a statistical significance.

5. Discussion

The research objectives are to (1) develop a predictive analytics model that prioritise customers based on propensity to re-enter the sales funnel and make a purchase (2) deploy the model into customer win-back operation (3) measure the impact on customer win-back rate and (4) estimate impact in terms of revenue uplift. These objectives are thoroughly addressed. The propensity model was developed and implemented through an A/B testing approach. This fulfils objective (2) as well as create a comparative environment to highlight whether predictive analytics can improve performance. From this, customer win-back rate from both A and B tests were recorded and compared. Finally, revenue projection was drawn from performance improvement.

Despite evident improvements across all areas from conducting A/B tests, the study still has rooms for improvements and, on reflection, a number of limitations that may need to be thoroughly addressed if the company would like to scale up the use of predictive analytics to other use-cases. Furthermore, predictive analytics is not limited to customer win-back or customer retention-based use-cases. These limitations and critical evaluation of the study and how the results are generated will serve as a precedence for future developments and extensions into more novel use-cases, be it in real-estate or other industries. However, risks and mitigating approaches must be explored further as big data analytics adoption on a corporate scale poses a considerable challenge.

5.1 Implications and Reflective Evaluation

The outcome of the study shows a highly positive impact outlook for the company, with performance improvement across all metrics used. An increase in effectiveness by 13.5%, efficiency by 11.9% and productivity from decrease in duplicate cases by 82.1% all contribute to 20.9% uplift in revenue estimation for the following year. However, there were several obstacles faced throughout the study. These stem from data readiness point of view, analytical point of view and operational point of view. These 3 aspects are essentially data, people and process limitations.

In terms of data incompleteness, the significance levels of each features extracted from CRM were adequate in constructing a meaningful propensity model. However, there were substantial data incompleteness and scattering across enterprise databases; data inconsistency is apparent. There was substantial data exploration and an extent of data cleansing to make the data valid. Despite this effort, not all datasets were cleansed and a significant proportion were not incorporated into the study. In hindsight, the study should be phased out to accommodate data readiness. The initial phase should consist of data exploration and cleansing to ensure data are as complete as possible. This does not necessarily have to involve business use-case identification but to structure data in the company across all databases to be in a ready-to-use format. This will enable users from any department to get access and utilise vast data pool more meaningfully. The following phase then should be to identify business use-case to test as a pilot project, i.e. customer win-back analytics is still a valid use-case and a preferable case due to data availability, in spite of their consistency and completeness. However, the process of extracting what features are significant, customer-engaging activities should be introduced to incorporate qualitative analysis from customers. Moreover, mystery shopping should be conducted to explore the company's sales operation and compare this to competitors. Identification of data gap where collection can be introduced should be analysed as well to maximise data collection opportunity. This will benefit the company in terms of understanding what features are truly important to consider and so investment can be made to enhance it and to identify future improvement for data collection to make the enterprise data increasingly more robust.

People are the centre of the customer win-back – be it win-back staff or customers. As such, there is always emotional involvement when attempting to win back customers. The study mainly focusses on how data can improve customer win-back as a whole but people management is somewhat omitted. In hindsight, what could have enhanced the differential further is if customer win-back staff was ranked based on past performance. Using this rank, the team could listen to recordings from the top performing staff and the least performing staff to extract key words, phrases or how they interact with customers. The output would be a standardised sales script for different scenarios which would increase the likelihood of referral to sales department, thereby increase win-back rate. Moreover, training could have been conducted in a more comprehensive manner so that win-back staff can follow the predictive analytics process with greater robustness. The training sessions provided were how the operation would take place. To ascertain higher impact, the team should have been conducted a work shadowing process. This is when both the business and data consultants are deployed to observe and guide win-back staff in team B as well as address any concerns. Moreover, customer win-back staff are entitled to sales commission if sales department could successfully close the customers referred to them. Communicating explicitly on the benefits of using predictive analytics model to help increase the number of customers referred and referral rate could help decrease resistance. On the contrary, this could help encourage customer win-back staff to adopt the model more rapidly.

In terms of process, the study was designed immaculately for the development and testing of predictive analytics. The competing tests were standardised and constructed fairly. The metrics, average number of refer cases and referral rate were the key drivers and measured clearly. However, some may argue that a 3-week testing period may be inadequate. Had the time limitation been lifted, the test should be run further to consider other factors as well. It could be considered fortunate for team B that marketing promotions were launched during the testing period and so prioritising projects with “desirable” promotions do provide competitive edge to team B. In hindsight, the consulting project should be separated into 2 main streams and over a longer period, from 3 months to 5 months. The first month should be dedicated to data exploration, business case exploration and hypothesis formation. The following 2 months should be on model development and insight gathering. This should help identify key features and are truly significant to customer win-back. The last 2 months should be dedicated to A/B testing to measure the performance over an extended period. This would help refine and strengthen the argument for performance improvement and create a convincing case for revenue uplift estimation. The 2-month period is designed to (1) establish a statistical significance within the first month or approximately 4 weeks and (2) once significance is established, the following 4 weeks should be used to validate that such significance is valid. A current set of results can be concluded as showing a statistical significance on some metrics using one-tailed testing method. However, if the test is to be improved, a 95% confidence interval could be deployed which would deem the current dataset inadequate to established a statistical significance. Moreover, a 14-day test is insufficient to determine the claim that predictive analytics provides a superior performance when adopted is valid.

In terms of impact analysis, the 20.9% increase in revenue projection compared to the figure that would have been achieved without predictive analytics is likely to be an overestimation. This is because the model descends in terms of propensity score, i.e. the list will eventually see less impact on win-back results as higher potential customers have been contacted and the list might exhaust. This does not necessarily mean that there are no newer customers entering the win-back basket but the probability of high win-back potential customers will be limited. Furthermore, revenue is realised only when transfer of ownership occurs. This is beyond the responsibility of customer win-back department and involves sales and inspection departments. This uncontrollable factor is one which can impact revenue realised at the end of the year.

In hindsight, there are several changes that could have been made to create a more convincing case for predictive analytics. Three key changes are to (1) restructure the efforts with the company to layout a more methodical process from firstly standardise data, analyse standardised data for insights, design proof of concept and conduct an A/B test for impact illustration (2) design a comprehensive qualitative research practice to identify insights, limitations and to help design a process for team B that is closer to as-is operations and (3) extend A/B testing to increase the validity of impact.

5.2 Reassessment for Statistical Significance

To explore if this experiment is to be revisited in the future, the 14-day period has been extended to 20 days with conditions maintained and average values used as shown in Table 11. However, a 95% confidence interval is used to increase the accuracy of prediction and to increase the validity of argument for predictive analytics.

5. Discussion

Table 11: A/B test extension with current average values

	Number of refer cases [value]		Total number of customers reached [value]		Referral rate [%]		Refer cases per staff [value]		Number of staff [value	
Date	As-is process	Model	As-is process	Model	As-is process	Model	As-is process	Model	Team A	Team B
01-Jun	41	27	204	106	20.1	25.5	8.2	6.8	5	4
02-Jun	38	40	181	183	21	21.9	7.6	10	5	4
04-Jun	46	57	186	222	24.7	25.7	9.2	11.4	5	5
05-Jun	36	39	188	254	19.1	15.4	9	7.8	4	5
08-Jun	24	38	125	176	19.2	21.6	8	9.5	3	4
09-Jun	28	29	173	165	16.2	17.6	7	7.3	4	4
10-Jun	16	28	160	147	10	19	4	5.6	4	5
11-Jun	20	24	96	149	22.4	16.1	8	6	4	4
12-Jun	20	33	172	233	11.6	14.2	6.7	6.6	3	5
15-Jun	29	59	144	232	20.1	25.4	9.7	11.8	3	5
16-Jun	26	39	166	249	15.7	15.7	6.5	7.8	4	5
17-Jun	44	48	258	291	17.1	16.5	8.8	9.6	5	5
18-Jun	12	20	155	134	7.7	14.9	2.4	5	5	4
19-Jun	21	35	158	206	13.3	17	5.3	8.8	4	4
day 15	28.6	36.9	169.0	196.2	17.0	19.0	7.2	8.1	4.1	4.5
day 16	28.6	36.9	169.0	196.2	17.0	19.0	7.2	8.1	4.1	4.5
day 17	28.6	36.9	169.0	196.2	17.0	19.0	7.2	8.1	4.1	4.5
day 18	28.6	36.9	169.0	196.2	17.0	19.0	7.2	8.1	4.1	4.5
day 19	28.6	36.9	169.0	196.2	17.0	19.0	7.2	8.1	4.1	4.5
day 20	28.6	36.9	169.0	196.2	17.0	19.0	7.2	8.1	4.1	4.5

The results are shown in Table 12, Table 13, Table 14 and Table 15. It can be seen that all simulated cases provide statistical significance with all p-values less than 0.05. This implies that if the experiment were to carry on for an extended period of time, the statistical validity could have been established and that the business impact projected is supported statistically.

Table 12: Reassessed Student’s t-test for Refer cases

	As-is process	Model
Mean	28.64285714	36.85714286
Variance	79.22180451	91.87969925
Observations	20	20
Pooled Variance	85.55075188	
Hypothesized Mean Difference	0	
df	38	
t Stat	-2.808393048	
P(T<=t) one-tail	0.003909419	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.007818837	
t Critical two-tail	2.024394164	

5. Discussion

Table 13: Reassessed Student’s t-test for Total number of customers reached

	As-is process	*Model*
Mean	169	196.2142857
Variance	961.1578947	1932.755639
Observations	20	20
Pooled Variance	1446.956767	
Hypothesized Mean Difference	0	
df	38	
t Stat	-2.262398803	
P(T<=t) one-tail	0.014736546	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.029473092	
t Critical two-tail	2.024394164	

Table 14: Reassessed Student’s t-test for Referral rate

	As-is process	*Model*
Mean	17.01428571	19.03571429
Variance	16.50511278	11.9774812
Observations	20	20
Pooled Variance	14.24129699	
Hypothesized Mean Difference	0	
df	38	
t Stat	-1.693883853	
P(T<=t) one-tail	0.049235069	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.098470138	
t Critical two-tail	2.024394164	

Table 15: Reassessed Student’s t-test for Refer cases per staff

	As-is process	*Model*
Mean	7.171428571	8.142857143
Variance	2.95518797	3.047067669
Observations	20	20
Pooled Variance	3.00112782	
Hypothesized Mean Difference	0	
df	38	
t Stat	-1.77324452	
P(T<=t) one-tail	0.04210182	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.08420364	
t Critical two-tail	2.024394164	

5.3 Risk and Mitigation

Since big data analytics is a relatively novel application within the business domain, challenges are likely to arise. There are 2 primary areas that will pose greatest challenges: (1) People (2) Data and its utilisation. These risks associated with real-estate developer and respective contexts will be explored and mitigating approach will be proposed. However, these mitigations are not key actions but high-level step by step process that the company or any organisations looking to adopt big data analytics can explore.

5.3.1 Risks Associated with People

Despite big data analytics association with technology, databases and AI, the key success factor that consolidate these associations into meaningful business impact is people. Without alignment and understanding of what data can help ease their jobs, resistance will pertain and will intensify. On the other hand, adopting big data analytics will require specific expertise. Data professionals are critical resources that need to be prioritised and these professionals are becoming rarer due to wide adoption of big data.

Harvey (2017) cite that 85.5% of organisations surveyed by NewVantage Partners are endorsing data-driven development. However, a mere 37.1% were successful (Harvey, 2017). The main obstacle is people. People as a key risk driver in preventing big data adoption success consists of 3 issues. Firstly, there is a lack of understanding as to "why" the organisation needs to adopt big data and endorse transformation. Secondly, there is a lack of commitment from middle management who oversees cascading top management's mandate. Thirdly, there is a lack of know-how and expertise which are requisites to fulfilling the data-drive goal. Since people issue is a sensitive matter but a core driver for business, the first 2 key risk drivers need to be mitigated and addressed internally while expertise can be sought externally to help accelerate change.

One possible solution is to empower a strong leader of change – one that understands data's possibilities and one that has influence over key employees. A chief data officer (CDO) could be appointed to lead the business in a data-driven direction. This role will be required to work closely with other C-level officers to create business use-cases, educate others on how data will make their jobs simpler and help drive more impact, which will in turn uplift their performance. This will likely enable higher success potential through 3 levers: communication, empowerment and incentive. A CDO will take a central role in transforming perception of data-driven operation and convince employees throughout the company to encourage alignment and endorsement. The next step is to empower middle management with thorough understanding and tools that will help them make a difference. This empowerment will create a sense of belonging and belief that they can deliver impact which will lead to the last lever: incentive. Tangible incentives, be it financial or non-financial, are a great motivator. With these incentives bound to the success of data, the eagerness to understand and adopt will follow. These levers are vital and must be leveraged together.

Another risk is when there is evidence of success from big data, an organisation will likely scale its big data operations rapidly. This will require aggressive recruiting for data-based roles. While this is considered straightforward, there is limited qualified candidates in the market, especially during tech-boom with candidates entering start-ups or leading corporates such as True corporations and Central groups. The risk of (1) having insufficient manpower to handle data transformation and (2) underqualified employees pose a costly and damaging challenge. A straightforward and simple solution is to increase recruitment budget to obtain talents and to invest in comprehensive training to build internal capabilities. However, this approach is costly and relatively time-consuming. An alternative would be to outsource the business-data stream to a third party who has the manpower, expertise and experience. This Data management as a service (DmaaS) can be costly in the short-run but it creates a fast-paced development with specialist and dedicated resources. Moreover, this expertise can be obtained through on-the-job training. The company can easily negotiate knowledge-transfer and coaching as a part of deliverables that the outsource has to submit, thereby building internal capabilities. Despite this, risk of data leakage is a common argument against this accelerated track. In practice, this can be easily overcome with licenses and tools under the company's account. Moreover, access to data can be restricted to on-site operations as well as data breach tracking installed in licensed devices.

5.3.2 Risks Associated with Data and Its Utilisation

A common trap that the company could easily fall into is over-reliance in data. It is inarguable that data is the new oil, i.e. it has exceptional commercial value, and as it is compiled in 1 dataset, the value grows exponentially. However, one must not over-rely on the data and abandon the emotional and human judgement. A discussion that emerged during the study was that the evident impact of predictive analytics in customer win-back should be a strong case for the company to develop this notion further and extend it to product recommendation. This differs from using digital footprints as behavioural traits and predict what the customers might wish to view or buy in the digital space, e.g. Amazon offers product recommendation on their website based on historical data or searches. The use-case dictates sales staff on the front line to each have a mobile tablet with an in-house product recommendation engine to offer customers. This means that the recommendation will rely solely on the inputs from the customers that enter the sales gallery and the data input from sales staff. However, the key issue is that real-estate is a high ticket size product. The purchasing decision realistically involves emotional triggers. As such, using this product recommendation engine, which overrules the emotional aspects, would likely distract the sales staff from displaying their expertise in salesmanship to focussing on what the engine might offer the customers. In order to mitigate this, sales department and data analytics team must first collaborate to extract the core values of sales operation and determine what predictive analytics tools are realistically beneficial for sales staff when engaging customers. This should not override the personal connection activities that sales staff focus on when engaging customers but to aid in decision-making for customers.

Another risk that could arise from having successfully implemented predictive analytics is over-confidence in data. This could prove costly and time-consuming. A great threat of over-confidence is that the company may put in all the data from CRM into the analytics tools and let ML runs unsupervised to uncover hidden insights or imperceptible patterns. However, this is time-consuming and is likely to produce non-value addition, i.e. useless results. This is driven by numerous data points that have little connection or linkage, if any, to each other. This means that these data points are insignificant when combined and analysed together. Whatever the results from such analytics may show, the company would likely misuse that data analysis and create no value. Potential causes that damage over-confidence are: (1) the current data points are inadequate so there are needs for new data points to collect (2) the datasets are not inconsistent and subject to partial errors and duplications.

5.3.2.1 Data inadequacy and mitigating approach

Data inadequacy occurs when the current datasets lack the necessary traits or characteristics required to perform analytics that address the business case proposed. This is a common issue which stems from uncompelling business use-cases. Business use-cases are critical inputs to design data requirements and identify what and how which data are to be collected and used for commercial benefits. In addition, unclear business cases are not the only issue. Lack of knowledge and know-how in terms of analytics also fuel the issue. While data from the past 20 years may be useful in that time, they are likely to become obsolete and do not apply for business cases presently or in the future. This would warrant new data requirements in terms of what data to collect, how to collect them and, more importantly, how to analyse them and generate meaningful insights. Janoschek (2020) cite 38% of organisations in a big data analytics challenge survey identifying lack of compelling business cases as an issue for failed adoption of big data analytics. Furthermore, 53% of the survey participants claim that lack of analytical know-how and skills are a great challenge (Janoschek, 2020).

A potential coping mechanism in successfully incorporating big data analytics in an organization is to form "compelling business cases". Business cases or use-cases will determine what analyses are required which will serve as the fundamental for data requirements. Data requirements are essentially a list of data points that are needed to satisfy the analyses identified. However, this list will help create a "how-to" approach to obtain said data. This new set of data requirements must be standardised across the organisation to enable complete and consistent datasets. However, a recent development in Thailand regarding data is Personal data protection act (PDPA) which will be enforced within 2020. Apart from reprocessing data requirements, the approach by which an organisation need to undertake will inevitably have to comply with PDPA to prevent privacy issues.

Therefore, from the company's current data structure, there is a need for an initiative to identify compelling use-cases across all business functions, be it sales, customer win-back, marketing or land acquisition. The use-cases must be specific, measurable, achievable, relevant and time-bound. It is true that data holds a golden value but realistic application is vital and will save costs and time spent on something overpromising or "too good to be true" expected results. Afterward, an initiative to review current data fields across all databases should be implemented to (1) conduct comprehensive data cleansing to ensure meaningful and useful datasets (2) standardise data structure and (3) identify potential data points to collect. Lastly, cascading said data point requirements to collection approach will warrant a cross-function discussion to conduct sanity check and finalise prior to implementation.

5.3.2.2 Data inconsistency and mitigating approach

Data inconsistency is a situation where there are multiple datasets of the same type of data, i.e. a duplication scenario to an extent, which are a result of inputs from different departments. This means that duplications of data points which vary in formats or structures create discrepancies and accuracy. This inaccuracy in data can cause much less impact to the business if used.

There Many organisations do have an approach to mediate this effect of data quality through data integration – having a standardisation in-process consolidation of data into a hub or central space. An organisation should create a master reference store to centralise and standardise data across all databases. A master reference store is a centralised rule-based data storage that serve as a source-of-truth (Kovalenko, 2019). This will require the company to create a data mart – a centralised database with standardisation of data inputs from across all enterprise databases for analysis – to help turn inconsistent data into consistent data structures. These datasets will be extracted directly in a ready-to-use format from data mart by end users.

To mitigate this, data governance and data strategy must be established. Data governance serves a policy and framework for managing of enterprise data. This governs how data is to be managed, analysed and utilised. Thomas (2014) propose a data governance framework that focus on 6 key areas: (1) Policy, standards, strategy (2) Data quality (3) Privacy / compliance / security (4) Architecture / Integration (5) Data warehouses and Business intelligence (6) Management support. These 6 components combined will help accomplish better decision-making, mediate conflict between operational functions, protect the needs of stakeholders, develop management and staff to adopt identical data management approaches, standardise processes, reduce costs and increase effectiveness and ensure transparent processes.

The key takeaway is that data management must be a top-down approach. Creating a robust governance framework to manage, oversee and standardise data management and utilisation across the company will help minimise risks, drive alignment and create new capabilities.

5.4 Future Work and Big Data Analytics Opportunity

In the past, any business use-cases that highlight on intensive applications of big data may seem unrealistic and overpromising. However, with proven impact on customer win-back, there are opportunities for the company to extend beyond customer win-back and replicate to other department and other business functions. Moreover, there are also opportunities to enhance customer win-back analytics further with tools that can help increase productivity, thereby improving performance further. Furthermore, customer win-back analytics as a branch of customer retention can easily be replicated to other industries, especially retail-based sectors such as banking and retail.

On the other hand, it is evident that e-Commerce and digital sales and services have been popularised and became the norm in customer services from COVID-19 pandemic. SCB (2020) suggest that there are substantial increases in online purchase activities, especially in online shopping, food deliveries and home streaming (32%, 30% and 42% respectively). This shift in behavioural paradigm could impact industries with larger ticket sizes, e.g. automotive and real-estate. Digitising real-estate sales could emerge as a new norm. This would require robust omni-channel integration model, digital data logging and management. This means competition will become more data-driven, ranging from designing residential layouts to suit target customers’ preferences to bundling products and aftersales services. Therefore, predictive analytics building on digital footprints such as online inquiries and offline-to-online interactions can help uncover customer product preferences so that future of residential projects could be fully tailored and personalised. Moreover, predictive analytics can thoroughly track interactions between sales executives and customers, which help monitor and identify approaches to uplift each sales performance.

5.4.1 Replicating to Other Functions Within the Organisation

From customer win-back successes, there have been discussions as to what potential use-cases may be in the future. In business, the main objective is inevitably to improve financial performance: revenue and costs. Three opportunities that could leverage data are pricing optimisation for land acquisition, predictive maintenance for condominiums and sales recommendation.

Land acquisition analytics is the use of the company's historical data on land purchases such as acquisition costs and appraised value at the time. However, optimising this will require external sources of data to be fed into the model. Current market values of surrounding properties, appraised values based on treasury department's estimations, economics of specific areas (for example, a piece of land that was just bought by a major retail outlet chain will likely drive up the price and affect the value of the entire area). All these data input can help the company in maximising returns from potential land acquisitions. Moreover, this can help prioritise acquisition in the pipeline as well. As a result, land acquisition analytics can help minimise cost to purchase.

Another use-case is the notable predictive maintenance. With over 40 high-rise residential buildings, the company has substantial expenses of over THB 40 million per annum on maintenance and upkeep of these properties. The process currently deployed is a typical preventive maintenance, a scheduled health-check every week for critical facility function, every month and quarterly for less critical functions. However, these are unnecessary expenses and there are rarely issues that need to be solved. Predictive maintenance can help predict when a component needs to be maintained through fitting integrated microsensors. Maintenance cost can be minimised and the THB 40 million budget could be easily reallocated to Data management as a service on annual subscription.

While the previous 2 cases are cost optimising analytics, a revenue enhancement opportunity is sales recommendation. This considers historical data on customers, their characteristics and purchases to create customer profiles. These profiles can be used to create a look-alike model, which essentially targets leads with similar characteristics to that of respective customer profile. Such customer acquisition intelligence can leverage social media space – Facebook and Instagram, for examples – to provide product offers that are most attractive to each lead based on lookalike profiles. This is a calculated likelihood that a lead with such characteristics will most likely find the offers attractive.

There are ample opportunities for the company to explore business use-cases with big data analytics to create an impact on both revenue and cost aspects. However, critical success factors remain data robustness and compelling business use-cases that justify utilising such data.

5.4.2 Extending Business Cases to Other Industries

Big data analytics applications are not limited to impact in real-estate but can be pivoted to lower ticket sized industries as well, particularly retail-based such as banking and retail sectors.

Banking sector has witnessed grave competition and will face tougher competitive rivalry in digital era (Thai PBS World, 2019). This means that there is a grave need for competitive edge in winning against other players. Data analytics can provide ample opportunities in triggering higher spending per customers, similar to customer win-back but to upsell and cross-sell current customers. However, this would require substantial data and powerful analytics. A potential use-case is using credit card spending, categorising their spending behaviour to predict what customer’s needs will likely arise. An example would be a customer who frequents an auto shop might be interested in a more premium motor insurance as this may reflect their concerns for their car. Potential cross-selling of bancassurance product can be common. On the other hand, small businesses in Thailand are more digitised and leveraging platforms such as Lazada and Shopee. The e-Commerce boom in Thailand has revolutionised small businesses and consumer behaviour. Kbank, a leading commercial bank, has seized this opportunity and established a partnership with Lazada, a prominent e-Commerce platform facilitating digital shopping, to offer digital lending at “fingertips” (Kbank, 2019). This phenomenon is enabled by transactional data on platforms which provide evidence of affordability and wealth despite lack of credit scoring. To expand this further, commercial banks can use digital footprints – social media profiles and activities and payment history to form a customer 360 view and create an alternative and robust credit assessment. This could potentially replace traditional credit scoring which is time-consuming and shift toward instantaneous loan approval. However, this needs to comply with regulation and know your customer (KYC) policy for risk management and compliance.

In retail, there are various possibilities that big data analytics can drive performance. A potential use of big data is customer journey analytics. Each step should provide information on customer behaviour so that data can be integrated across each stage in the customer journey and predict the root-causes of customer churn, i.e. churn prevention analytics. This can help the retailer redesign customer journey to help maintain customers in the sales funnel. Alternatively, big data should be incorporated with loyalty programs. In the past, loyalty programs were a mere tool for customers to collect points and redeem for premiums. The benefit for the retailer was to encourage repeated sales. However, with big data analytics, loyalty programs can provide a full-view of customers: where they visit, what they consume, how frequent do they enter the application and at which time. These data are critical to predict future spending or the likelihood of future actions. To commercialise on this prediction, the retailer could provide product recommendations that are targeted specifically for a certain group of customers.

There are various possibilities that big data analytics can unlock. Adoption across all industries are evidently possible. Real-estate customer win-back analytics can be replicated across other industries to drive retention as well as replicated to other business use-cases beyond retention, e.g. customer acquisition, operations improvement, etc.

6. Conclusion

The aim of this study is to prove whether Big data analytics can be used to improve Customer win-back operations for a leading real-estate developer. The results demonstrate that Big data analytics shows to help improve Customer win-back operations in Real-estate sector, particularly in the Town-house segment.

However, a research question that aims to establish whether the data model can prioritise customers with potential to pick-up when phoned remains invalidated. This is due to agreement with the company that this metric is better deployed in different data models in the future due to operational difficulty for team A and team B.

Nevertheless, running an extended period of 3 weeks for an A/B test experiment, where team A consists of 5 customer win-back staff using the proof of concept data model and team B consists of 5 customer win-back staff following the standard protocol, shows that:

- Team A shows to be 13.5% more efficient that team B, i.e. using data model allows win-back staff to have higher productivity
- Team A shows to be 11.9% more successful in referring customers to the sales team, i.e. higher conversion
- Extrapolating this trend, the real-estate developer is expected to earn THB 253.5 million more in revenue from Customer win-back operations if data model is deployed fully

The results from this study conform with some of the statements mentioned in the literature review. A claim by Griffin and Lowenstein (2001) that data is the key to customer retention and win-back is proved to an extent in this study. From structuring the datasets into more reliable inputs, the performance for customer win-back improves drastically.

However, there are limitations in terms of data, people and process which can be made different if the study were to be redone. In hindsight, more methodical data preparation could help ease the work throughout the project and ensure more effective data utilisation, more comprehensive analyses on people’s point of view, both customers and staff to understand the motives and extract insights for more robust feature selection, more robust process planning and company management could help better the outcome.

6. Conclusion

The success of predictive analytics in customer win-back for the real-estate developer can be replicated to other business functions within the company to increase revenue and reduce costs. Pricing optimisation in land acquisition, predictive maintenance in facility management and sales recommendation are future opportunities that the company could explore from leveraging their data. Furthermore, predictive analytics success is not limited to real-estate but can be replicated to retail-intensive sectors with vast databases such as banking and retail as well.

7. Reference

Abdallah, Z., Du, L. and Webb, G., 2017. Data Preparation. *Encyclopedia of Machine Learning and Data Mining*, pp.318-327.

Al-Barrak, M. and Al-Razgan, M., 2016. Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, 6(7), pp.528-533.

Anshari, M., Almunawar, M., Lim, S. and Al-Mudimigh, A., 2019. Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, 15(2), pp.94-101.

Armitage, P., Berry, G. (1994). Statistical Methods in Medical Research (3rd edition). Blackwell 1994.

AWS. 2020a. Amazon Redshift Management Overview - Amazon Redshift. [Online] Available at: https://docs.aws.amazon.com/redshift/latest/mgmt/overview.html [Accessed 23 June 2020].

AWS. 2020b. Amazon Elastic MapReduce [Online] Available at: https://www.amazonaws.cn/en/elasticmapreduce/#:~:text=Amazon%20Elastic%20MapReduce%20(Amazon%20EMR,cluster%20of%20Amazon%20EC2%20instances. [Accessed 23 June 2020].

AWS. 2020c. What is Amazon S3? [Online] Available at: https://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html [Accessed 23 June 2020].

Bishop, C., 2006. Pattern recognition and machine learning . New York: Springer.

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A. and Paraboschi, S., 2001. Designing data marts for data warehouses*. ACM Transactions on Software Engineering and Methodology*, 10(4), pp.452-483.

7. Reference

Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1), pp.5-32.

Budiyanto, A. and Dwiasnati, S., 2018. The Prediction of Best-Selling Product Using Naïve Bayes Algorithm (A Case Study at PT Putradabo Perkasa). *International Journal of Computer Techniques*, 5(6), pp.68-74.

Chase, C. W., 2013. Demand-Driven Forecasting: A Structured Approach to Forecasting. 2nd ed. New Jersey: John Wiley & Sons,.

Chauhan, N. S., 2019. Decision Tree Algorithm — Explained. [Online] Available at: https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4 [Access 25 June 2020]

Childs, G., 2002. What is ... a propensity model?. [Online]
Available at: https://www.campaignlive.co.uk/article/propensity-model/165289 [Accessed 20 Nov 2019]

Dahiya, P., B, C. and Kumari, U., 2017. Survey on Big Data using Apache Hadoop and Spark*. International Journal of Computer Engineering In Research Trends*, 4(6), pp.195-201.

De Jong, Y., 2019. Levels of Data analytics. [Online]. Available at: http://www.ithappens.nu/levels-of-data-analytics/ [Accessed 10 December 2019]

Dodson, J. (2000). “Find Out Why Your Customers Leave,” *Internet Week*, (Issue 800). 33–34

Dowling, G. R. (2002). “Customer Relationship Management: In B2C Markets, Often Less Is More,” California Management Review, 44 (3) 87–104.

Emmert-Streib, F. and Dehmer, M., 2019. Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference. *Machine Learning and Knowledge Extraction*, 1(3), pp.945-961.

7. Reference

Gao, X., Wen, J. and Zhang, C., 2019. An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, 2019, pp.1-12.

Gartner, 2020. Gartner Glossary: Analytics. [Online] Available at: https://www.gartner.com/en/information-technology/glossary/analytics [Accessed 20 March 2020]

Griffin, J. and Lowenstein, M. W., 2001. Customer Winback: How to Recapture Lost Customers—And Keep Them Loyal. San Francisco: Jossey-Bass.

Harvey, C., 2017. Big data challenges. [Online]. Available at:
https://www.datamation.com/big-data/big-data-challenges.html [Accessed 15 March 2020].

HG Insights, 2018. Propensity Modelling for Business [White paper]. Data Science Foundation. [Online] Available at:
https://datascience.foundation/sciencewhitepaper/propensity-modelling-for-business [Accessed 15 Nov 2019]

Hole, G. 2009. Handout 2009. University of Sussex. [Online] Available at:
http://users.sussex.ac.uk/~grahamh/RM1web/t-testHandout2009.pdf [Accessed 17 June 2020]

Huang, L., 2013. Marketing value of big data. Journal of Wuhan College Commercial Service, 27(5), pp. 17-19.

H2o.ai., 2017. Solving Customer Churn with Machine Learning: Case study. [Online] Available at: https://www.h2o.ai/wp-content/uploads/2017/03/Case-Studies_PayPal.pdf [Accessed 24 January 2020].

Jain, A., 2016. The 5 Vs of Big Data. [Online]
Available at: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/
[Accessed 20 March 2020]

7. Reference

Janoschek, N., 2020. The Most Common Problems Companies Are Facing With Their Big Data Analytics. [Online] BI Survey. Available at: https://bi-survey.com/challenges-big-data-analytics [Accessed 10 March 2020].

Joy, A., 2019. Pros and Cons of Supervised Machine Learning. Pythonista Planet. [Online] Available at: https://pythonistaplanet.com/pros-and-cons-of-supervised-machine-learning/ [Accessed 10 Nov. 2019].

Katsaragakis, S., Koukouvinos, C., Stylianou, S., Theodoraki, E. and Theodoraki, E., 2005. Comparison Of Statistical Tests In Logistic Regression: The Case Of Hypernatreamia. *Journal of Modern Applied Statistical Methods*, 4(2), pp.514-521.

KBank, 2019. KBank partners Lazada to provide online lending for sellers. [Online] Available at: https://kasikornbank.com/en/News/Pages/Lazada-KBank-Online-Lending.aspx#:~:text=Under%20the%20%E2%80%9CBetter%20Together%E2%80%9D%20collaboration,be%20approved%20within%20one%20minute. [Accessed 5 April 2020]

Kim, T., 2015. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6), p.540.

Kovalenko, K., 2019. Data Redundancy And Data Inconsistency Hurts Your Business. [Online] Available at: https://www.bizdata.com.au/blogpost.php?p=costs-of-data-redundancy-and-data-inconsistency [Accessed 21 March 2020].

Kucera, T. and White, D., 2012. Predictive Analytics For Sales And Marketing: Seeing Around Corners. Aberdeen Group. [Online] Available at: https://www.tibco.com/sites/tibco/files/resources/aberdeen-sales-marketing-analytics.pdf [Accessed 18 November 2019].

Kumar, V., Bhagwat, Y. and Zhang, X., 2015. Regaining “Lost” Customers: The Predictive Power of First-Lifetime Behavior, the Reason for Defection, and the Nature of the Win-Back Offer. *Journal of Marketing*, 79(4), pp.34-55.

7. Reference

Kumar, V. and L., M., 2018. Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications, 182(1), pp.31-37.

Laney, D., 2001. Application Delivery Strategies. [Online] Available at: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-ManagementControlling-Data-Volume-Velocity-and-Variety.pdf [Accessed 1 March 2020].

Mayer-Schönberger, V. & Cukier, K., 2013. Big Data: A Revolution that Will Transform how We Live, Work and Think. 1st ed. New York: Houghton Mifflin Harcourt.

Mohammed, M., Khan, M. and Bashier, E., 2016. Machine Learning: Algorithms and Applications. CRC Press.

Morgan, R.M. & Hunt, S.D., 1994. The commitment-trust theory of relationship marketing. Journal of Marketing, 58(3), p. 20

Ohlhorst, F., 2013. Big data analytics. Hoboken, N.J.: Wiley.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher Perrot M. and Duchesnay E., 2011. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12, pp.2825-2830.

Ray, S., 2017. 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R [Online] Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [Accessed 25 June 2020]

Rennie, J.D., Shih, L., Teevan, J. and Karger, D.R., 2003. Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

Rust, R. T., Zahorik A. J., and Keiningham T. L., 1996. Service Marketing, New York: Harper Collins College Publishers.

7. Reference

SCB. 2020. Scb.co.th. 2020. Thailand After COVID-19 Part 2: Business Opportunities And Survival. [Online] Available at: https://www.scb.co.th/en/personal-banking/stories/business-maker/thailand-after-covid-ep2.html [Accessed 23 June 2020].

Schott, M., 2019. Random Forest Algorithm for Machine Learning. [Online] Available at: https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb#:~:text=When%20using%20the%20Random%20Forest,better%20decision%20for%20your%20forest. [Accessed 25 June 2020]

Shalev-Shwartz, S. and Ben-David, S., 2017. Understanding machine learning. Cambridge: Cambridge University Press.

Siegel, E., 2016. Predictive Analytics. Wiley.

Srivastava, T., 2015. Nobody Tells You – 5 things Big Data 'CAN' and 'Cannot' Do. [Online]
Available at: https://www.analyticsvidhya.com/blog/2015/11/5-big-data-can-cannot/
[Accessed 20 April 2020]

Talend, 2020. What is Extract, Transform, Load? Definition, Process, and Tools. [Online] Available at: https://www.talend.com/resources/what-is-etl/ [Accessed 22 June 2020].

Thai PBS World, 2019. Thai Banks Face Tough Competition In Digital World. [online] Available at: <https://www.thaipbsworld.com/thai-banks-face-tough-competition-in-digital-world/> [Accessed 5 April 2020].

Thomas, G., 2014. The DGI Data Governance Framework. [Online] Available at: http://www.datagovernance.com/wp-content/uploads/2014/11/dgi_framework.pdf [Accessed 24 March 2020].

Thomas, J., Blattberg, R. and Fox, E., 2004. Recapturing Lost Customers. Journal of Marketing Research, 41(1), pp.31-45.

7. Reference

Stauss, B. and Friege, C. (1999). “Regaining Service Customers,” *Journal of Service Research*, 1 (4) 347–361.

Tokman, M., Davis, L. and Lemon, K., 2007. The WOW factor: Creating value through win-back offers to reacquire lost customers. *Journal of Retailing*, 83(1), pp.47-64.

Zhang, S., Zhang, C. and Yang, Q., 2003. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6), pp.375-381.

REFERENCES

VITA

NAME	Warot Lilahajiva
DATE OF BIRTH	17 January 1994
PLACE OF BIRTH	Bangkok, Thailand