CHAPTER IV
EXPERIMENT RESULT

4.1 Experiment Processing Results

In predictive modeling, it is important to obtain an unbiased estimate of the MSE or mean square error. The MSE is the most commonly used statistic for measuring the accuracy of the model. It is the squared difference between the target values and the predicted values, averaged over the number of observations that the model is fitting. The best model is the one that has the lowest MSE from the validation data set. This is because the MSE estimate from the training data set that is used to fit the model will almost certainty be overly optimistic since the same data is used to fit the model. It is important to understand that just fitting the model to the training data set does not mean that the model is necessarily correct and that the model will fit well to new data.

In statistics, the mean squared error or MSE of an estimator is one of many ways to quantify the amount by which an estimator differs from the true value of the quantity being estimated. The MSE can be written as the sum of the variance and the squared bias of the estimator:

$$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left(Bias\left(\hat{\theta},\theta\right)\right)^2$$

where $\hat{\theta}$ is the estimator, and θ is the estimated parameter.

As with all networks in this study, the mean squared error (MSE) performance function was utilized. The initial network implemented with three layers having 10 neurons in the hidden layer. The number of hidden neurons was changed regularly in an effort to choose the most robust neural network model. Table 4.1 summarizes the various networks investigated for the research.

According to the experiment results in Table 4.1, neural network based on PCA preprocessed data sets has higher performance. From the comparison between the value of the fourth column and the fifth column, it is very clear that the one using DC (Direct Connection) or the skip layer gets the better assessing results. It proves that some kinds of linear relationships are between the input data sets and output target. Furthermore, compared with the value of MSE for the testing data sets, the result of

Double dogleg algorithm with early stopping will be a little better. But the MSE of validation data set in Quasi-Newton with 8 neurons is 0.001583124, which is much smaller than the corresponding MSE of validation data set in Double dogleg with 5 neurons (MSE=0.001962573).

Therefore, the robust MLP neural network with Quasi-Newton algorithm in 8 neurons was selected to be the final model to go on the forecasting process. Other models with different neurons units' numbers are not suitable for this research study. Figure 4.1 illustrates the output predicted curl value from the robust neural network modeling (Symbols stand for the predicted value. Solid lines are the plots for the actual output values), Figure 4.2 represents the predicted curl values with actual curl values in a regression plot.

Data Type	Sub-Model	NO. Neurons	MSE with no DC	MSE with DC
Step-wise	Quasi-Newton	10	0.020037498	0.017573829
		15	0.020039678	0.017579368
		8	0.020110945	0.017711609
		5	0.020113452	0.017955018
	Double Dogleg	10	0.020033423	0.017550287
		15	0.020110136	0.017708773
		8	0.020039889	0.017549509
		5	0.020110450	0.017651599
PCA	Quasi-Newton	10	0.007023142	0.003112573
		15	0.005023245	0.002760182
		8	0.004964774	0.00261968
		5	0.005143774	0.002684601
	Double Dogleg	10	0.006890653	0.003113429
		15	0.006064724	0.002608188
		8	0.007464903	0.004387864
		5	0.005653675	0.002579321

Table 4.1: Training results summaries.

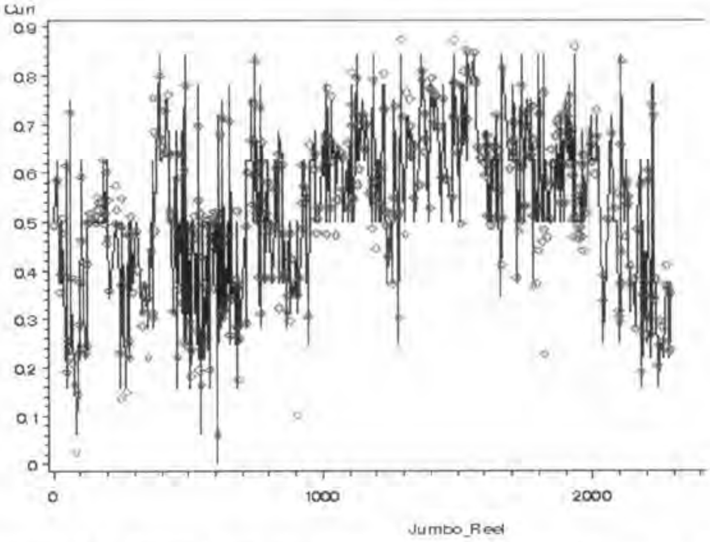


Figure 4.1: Actual curl vs. predicted curl in jumbo reel series.

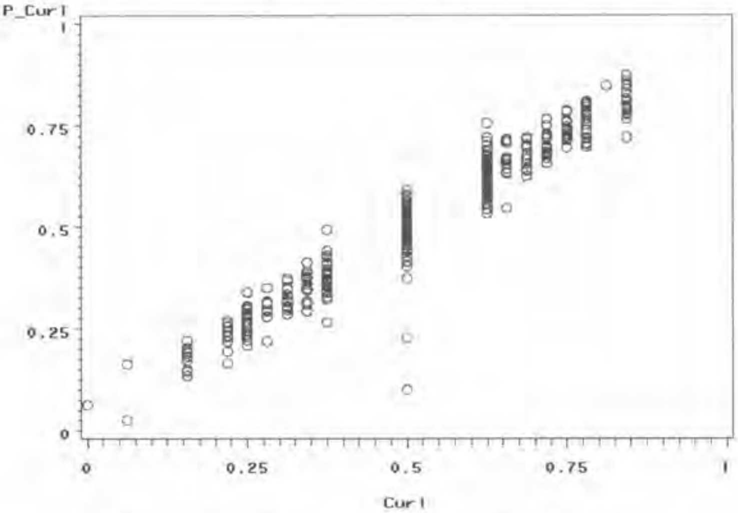


Figure 4.2: Actual curl vs. predicted curl.

4.2 Accuracy Evaluation

The MSE value of testing data set of the neural network model is 0.00261968. It is not an ideal value for our final target. Evaluating the accuracy of the model from a single validation data set can give inaccurate estimates at times. Therefore, re-sampling techniques, such as bootstrapping, should be considered in achieving stable and consistent parameter estimates in predictive models. The bootstrapping is performed k different times, each using a different partitioning of the input data set into training and validation data set with replacement. That is, placing the selected data point back into the population to be selected again, with the results the averaged by the k number of bootstrap parameter estimates. Therefore, bagging [12] technique based on bootstrap would be used to improve the accuracy of the final result.

Bagging stands for bootstrap aggregation. This modeling technique is analogous to bootstrapping where separate prediction estimates are created by re-sampling the data you want to fit. Aggregate bagging is performed, that is, bagging is based on bootstrap [16] sampling and aggregation by combing the prediction estimates. Bagging is analogous to bootstrap estimation where a fixed number of independent B samples are created by replacement. Refitting the statistical model to calculate the fitted values for each bootstrap sample of equal size, and then dividing by the number of bootstrap samples as follows to compute the bagged estimates of the predicted values:

$\hat{g}(x) = \frac{1}{B}\sum_{b=1}^{B} g(x)$, where g(x) is the bagging estimate from B bootstrap samples.

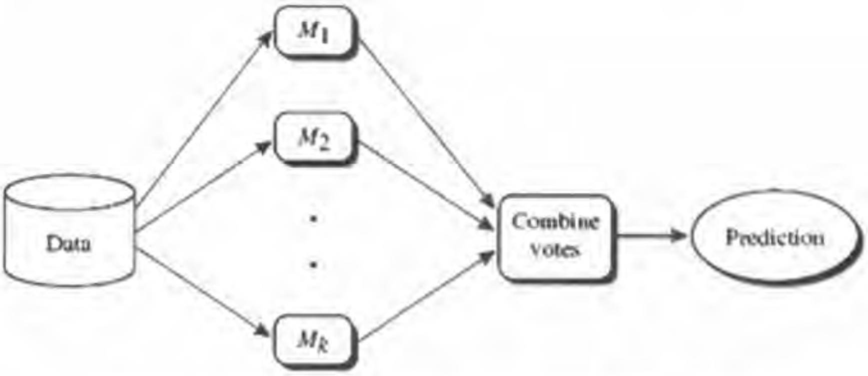


Figure 4.3: Increasing model accuracy design flow.

In this case, we divided the source data set into 10 bootstrap samples to train the model iteratively. The design idea shows in Figure 4.3. Bagging and boosting each generate a set of prediction models, $M_1, M_2, ..., M_3$. After bagging models step, the value of testing data set MSE decreased to 0.000017593 which displays a much higher performance shown in Figure 4.4 and Figure 4.5 than the former ones shown in Figure 4.1 and Figure 4.2. Most of predicted curls values are inosculate with the original output.

The number of testing data set is 684. The range of source target is [-40, 40], but after normalization, the source target has also been changed to [0, 1]. According to printing paper industry criteria, the jumbo reel of source target in the range of [-27.5, 27.5] can be accepted as the quality production. So the final normalized acceptance range is [0.15625, 0.84375]. When calculating the accuracy of final result, we set records in the acceptance range as "1", the part out of the acceptance range as

"0".

In statistics, false positive which also is named as Type I error and false negative which also is named as Type II error are used to describe possible errors made in a statistical decision process. The details can be got from Table 4.2.

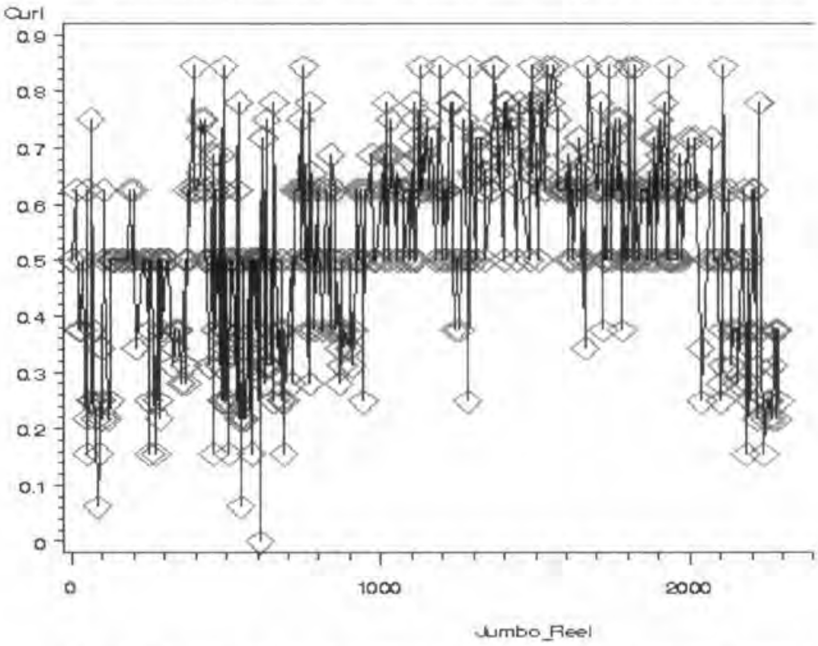


Figure 4.4: Actual curl vs. predicted curl in jumbo reel series (bagging).

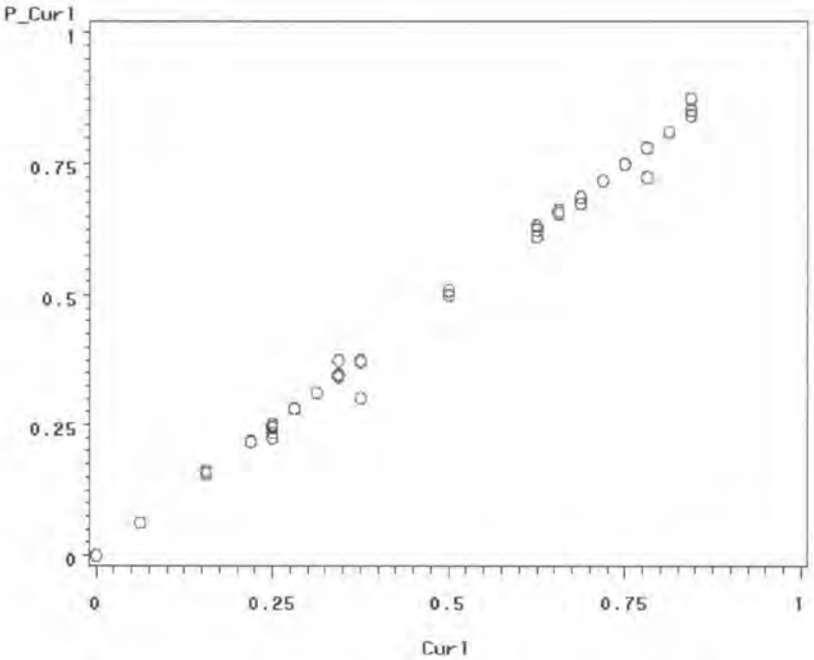


Figure 4.5: Actual curl vs. predicted curl (bagging).

Finally, there are 616 true positive (n_{TP}) records, 43 true negative (n_{TN})

records, 21 false positive (n_{FP}) records and 4 false negative (n_{FN}) records.

		Actual Condition	
		Accept	Recycle
Test result	Predict of "Accept"	True Positive (TP)	False Positive (FP)
	Predict of "Recycle"	False Negative (FN)	True Negative (TN)

Table 4.2: Test result analysis.

Sensitivity can be expressed as $\frac{n_{TP}}{n_{TP}+n_{FN}}=\frac{616}{616+4}$=99.35%

1) The proportion of accepted product which has a positive prediction testing result.
2) The ability of the test to detect accepted product condition, or the true positive rate.

Specificity can be expressed as $\frac{n_{TN}}{n_{FP}+n_{TN}}=\frac{43}{21+43}$=67.19%

1) The proportion of those accepted product which has a negative prediction testing result.
2) The ability of the test to identify accepted product condition fail, or the true negative rate.

Positive Predictive Value can be expressed as $\frac{n_{TP}}{n_{TP}+n_{FP}}=\frac{616}{616+21}$=96.7%

1) The likelihood that a positive test results indicates the existence of the accepted product condition.
2) The proportion of the population of testing positive which is accepted product.

Negative Predictive Value can be expressed as $\frac{n_{FN}}{n_{TN}+n_{FN}}=\frac{4}{43+4}$=8.51%

1) The likelihood that a negative test results indicates the existence of the recycled production condition.
2) The proportion of the population of testing negative which is recycled product.

Therefore, the **accuracy** is measured and the details of evaluation shown as Table 4.3:

$$\text{ACC}=\frac{n_{TP}+n_{FN}}{n_{TP}+n_{FP}+n_{TN}+n_{FN}}=\frac{616+21}{616+4+43+21}=93.13\%$$

	Predict	Target		Total
		Accept	Reject	
Frequency	Accept	616	21	**637**
Percent		90.06	3.07	**93.13**
Row Pct		96.7	3.3	
Col Pct		99.35	32.81	
Frequency	Reject	4	43	**47**
Percent		0.58	6.29	**6.87**
Row Pct		8.51	91.49	
Col Pct		0.65	67.19	
	Total	**620**	**64**	**684**
		90.64	**9.36**	**100**

Table 4.3: Cross-tabular frequency table.

4.3 ROC Curve

ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions. They are widely applicable, regardless of source of predictions. The Receiver Operating Characteristic (ROC) curve [17] is a plot of sensitivity versus (1- specificity). A model with no discriminatory power is represented by an 45° angle line through the origin (area = ½). A perfect model has 100% sensitivity and specificity for all cutoffs (area = 1).

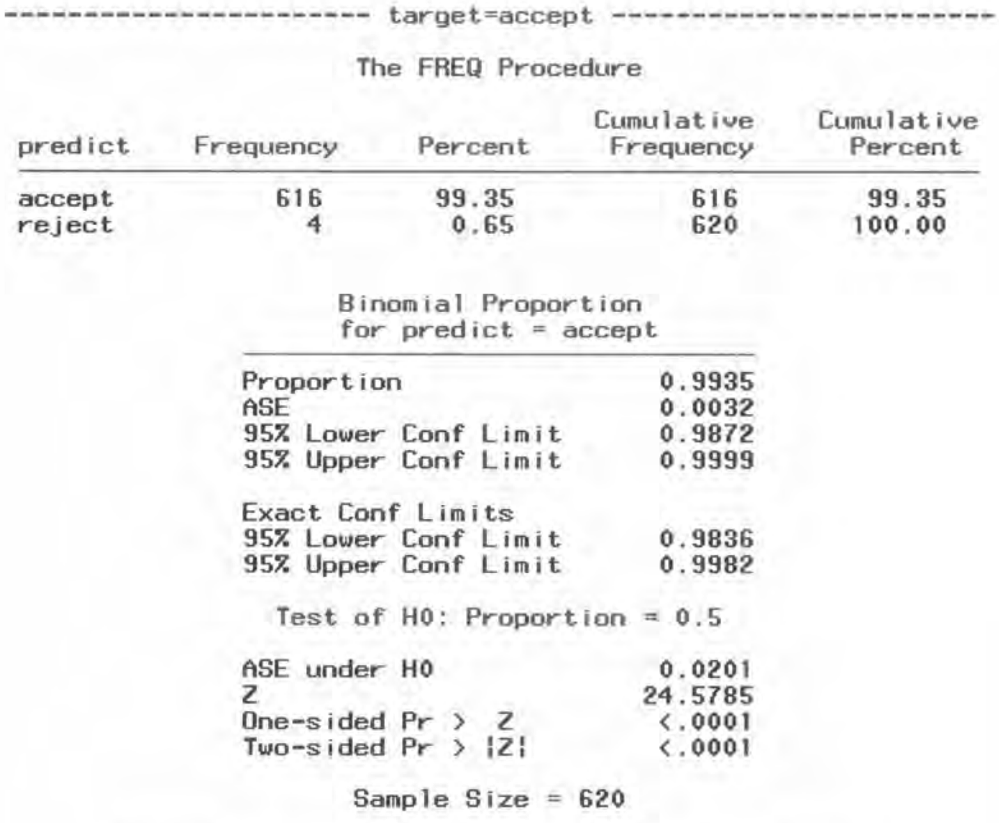
It is easy to recognize that these are all binomial proportions if one considers the corresponding denominators as fixed. This gives easy rise to the use of binomial functionality within PROC FREQ to compute interval estimates. The details of target as values of "accept" and "reject" are shown in the following Figure 4.6 and Figure 4.7.

The output under "Binomial Proportion" contains the key information.

Sensitivity was 99.35% and the asymptotic standard error (ASE) for the estimate of standard error is 0.32%. At this point, the asymptotic standard error for the estimate of binomial proportion p is give by

$$\text{ASE} = \sqrt{\frac{p(1-p)}{n}},$$

where n is the denominator of the binomial proportion. The confidence limits appearing under the ASE are based on asymptotic theory. If n is large, then the 95% confidence interval can be calculated using p± 1.96*ASE.



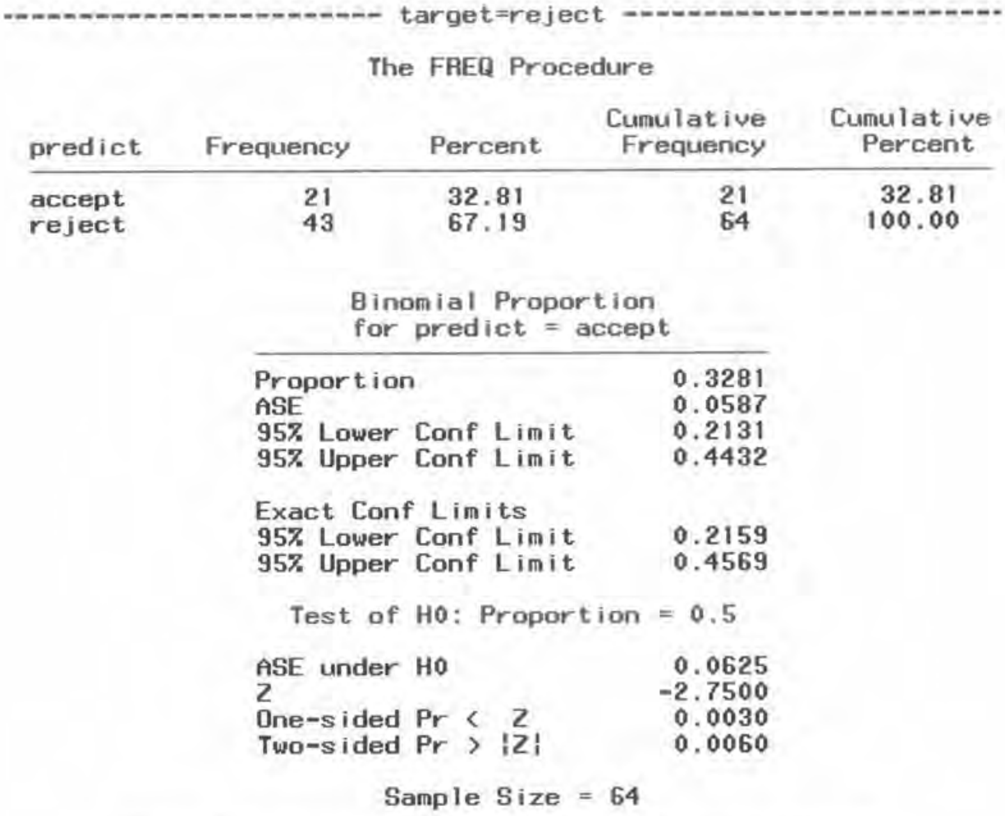
```
----------------------- target=accept -----------------------

                    The FREQ Procedure

                                              Cumulative    Cumulative
predict     Frequency      Percent            Frequency       Percent
-----------------------------------------------------------------------
accept           616         99.35                 616          99.35
reject             4          0.65                 620         100.00

                   Binomial Proportion
                   for predict = accept
               --------------------------------
               Proportion                0.9935
               ASE                       0.0032
               95% Lower Conf Limit      0.9872
               95% Upper Conf Limit      0.9999

               Exact Conf Limits
               95% Lower Conf Limit      0.9836
               95% Upper Conf Limit      0.9982

                 Test of H0: Proportion = 0.5

               ASE under H0              0.0201
               Z                        24.5785
               One-sided Pr >  Z         <.0001
               Two-sided Pr > |Z|        <.0001

                     Sample Size = 620
```

Figure 4.6: One-way frequencies for target value as "accept".

To compute the specificity, it only needs to subtract the estimates from 1. Therefore the specificity is 1-0.3281 = 67.19% and the confidence intervals are 1-0.4569 = 54.31% to 1-0.2159 = 78.41%. In terms of hypothesis testing the only change is in the sign of Z. Specifically, ASE and the p-values will be the same.

The ROC area is equivalent to the Mann-Whitney-Wilcoxon test statistic for comparing the distribution of the estimated posterior probabilities for the class 1 cases with that of the class 0 cases.

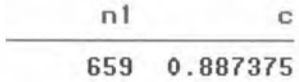
$$\text{ROC Area} = \frac{\sum_{y=1} rank\left(\hat{y}_i\right) - \frac{1}{2}\eta_1(n_1+1)}{\eta_1(n_1+n)}$$, where n_1 is the number of the predicted results who get the same value as the target result, $\sum_{y=1} rank\left(\hat{y}_i\right)$ is the sum of the target whose value is the 1 and the predicted value which already done by the process PROC RANK, η_1 is the calculated number of target values which equal to 1, and n is the total sample number.



```
------------------------ target=reject ------------------------

                        The FREQ Procedure

                                          Cumulative    Cumulative
predict      Frequency      Percent        Frequency       Percent
------------------------------------------------------------------
accept              21        32.81               21         32.81
reject              43        67.19               64        100.00

                       Binomial Proportion
                       for predict = accept
                   --------------------------------
                   Proportion                0.3281
                   ASE                       0.0587
                   95% Lower Conf Limit      0.2131
                   95% Upper Conf Limit      0.4432

                   Exact Conf Limits
                   95% Lower Conf Limit      0.2159
                   95% Upper Conf Limit      0.4569

                     Test of H0: Proportion = 0.5

                   ASE under H0              0.0625
                   Z                        -2.7500
                   One-sided Pr <  Z         0.0030
                   Two-sided Pr > |Z|        0.0060

                         Sample Size = 64
```

Figure 4.7: One-way frequencies for target value as "reject".

Thus, it can be interpreted as the probability that a randomly chosen class 1 case has a greater posterior probability than a randomly chosen class 0 case. And the result in this case is shown in Figure 4.8. In the Figure 4.8, the value of n_1 is 659, it means that in the whole 684 testing data, there are 659 testing data that is matched between the predicted value and the target value, in the groups of TP (True Positive) or FN (False Negative).



```
       n1          c
---------------------
      659   0.887375
```

Figure 4.8: Value of ROC areas.