Summary
Our team was tasked by Commercial Banking Corporation to create a retail credit scorecard as a part of the banking services to their department of Revolving Lines of Credit. The final scorecard was created by underlying a logistic regression model for the scorecard. The model has an AUC of 0.82, and an accuracy of 76 percent. We recommend granting credit to applicants with a score greater than 519. This score was chosen based on maximizing profit. Using this scorecard and cutoff for granting credit is expected to generate an average profit of $833 per applicant (up from $192) and an acceptance rate of approximately 75 percent. Also, the scorecard model is expected to reduce the current default rate in almost half to 1.64 percent.

Background
The Revolving Lines of Credit department of Commercial Banking Corporation has sought the help of our team for modifying their banking services. To achieve this purpose, we have created a scorecard for all retail credit applications for the bank. In this analysis, for precision in granting retail credit cards to applicants and regulatory compliance, we have considered the bank's past data on both accepted and rejected credit card applications. The current acceptance rate for granting credit is 75% with a default rate of 3.23%. The expected revenue from accepted customers that don't default is $2,000 and the cost of customers that default is $52,000. Based on the current data the estimated current profit per applicant is $192 and profit per accepted customer is $256. Our goal was to create a scorecard and find a cutoff score that would maximize profit.

Methodology and Results
The analysis was performed using two datasets: 1) the accepts dataset which had information about 3,000 accepted loan applicants and 26 variables with the target variable of whether a person defaulted on a loan; 2) the Rejects dataset contains the data on rejected applicants and isn't labelled (outcome variable is unknown). The overall process flow of scorecard creation is depicted in Figure 1.

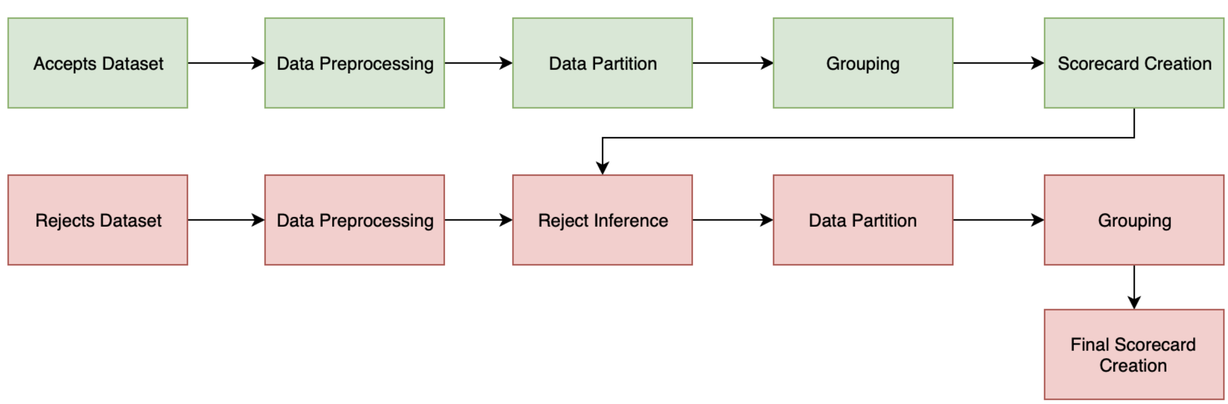

Figure 1: Process Flow

First, both accepts and rejects dataset was preprocessed which included imputation of missing and implausible values using RandomForest, and creation of new variables. Then the accepts dataset was partitioned into training and test set with a 70:30 partition. Once the training and tests sets were created, variable groupings were created on the training set using conditional inference tress and weight of evidence values were created for each grouping level. This binned numeric variables into categories and the categories were encoded as WOE values.

An initial logistic regression model was created using the WOE variables that were statistically significant. It had an AUC of 0.70 and accuracy of 0.68 on the test set. This initial model was used to predict labels for the rejects set and combined with the rest of the data (hard cut off augmentation). Weights of the observations in the rejects set was made proportional to the accepts set while maintaining the 3:1 ratio for accepts to rejects. This created a dataset that was representative of the entire applicant pool as required by the FDIC (reject inference requirement).

The combined data set was split into a train and test set and variable binning's and WOE value variables were created again. A second logistic regression model was created using the WOE variables that were statistically significant. This model had an AUC of 0.82 and an accuracy of 0.76 on the test set.

Scores for each variable and observation were calculated using the model coefficients and WOE values where a score of 500 was assigned to applicants with odds-ratio 20:1 and doubling the odds is associated with a change of 50 points in the scorecard. The variable scores for each observation were summed to create a total score for each observation. The scorecard model is displayed in table 1 below.

Score Card Model

Variable	Value	Points
Age	(0, 22]	21
Age	(22, 27]	56
Age	(27, 35]	93
Age	(35, 45]	109
Age	(45, inf]	149
Time at Address	(0, 21]	77
Time at Address	(21, inf]	91
Time at Job	(0, 18]	63
Time at Job	(18, 84]	86
Time at Job	(84, 158]	110
Time at Job	(158, inf]	129
Income to Cash Ratio	[0, 0]	95
Income to Cash Ratio	(0, 0.66]	90
Income to Cash Ratio	(0.66, inf]	84
Credit Card	Other	59
Credit Card	None	71
Credit Card	Visa	113
Credit Card	Cheque Card	136
Credit Card	Mastercard / Euroc	153
Adults in Household	One	66
Adults in Household	Two	104

Figure 2: Score Card Model

The score cutoff for granting a loan was chosen based on the predicted scores on the test set and actual outcomes on the test set (Figure 3). The test set was adjusted using the weights of each observation so that it would be a more representative of the population of all credit applicants. The possible profit for every score cutoff from 400 – 600 was calculated and the optimum cutoff that maximized profit was found to be 519. Using this cutoff, the expected average profit per applicant is $833. The graph shows that as the cutoff score increases the profit increase up until a score of 519. After that point the cost of customers that default start to outweigh the revenue from customers that don't default.

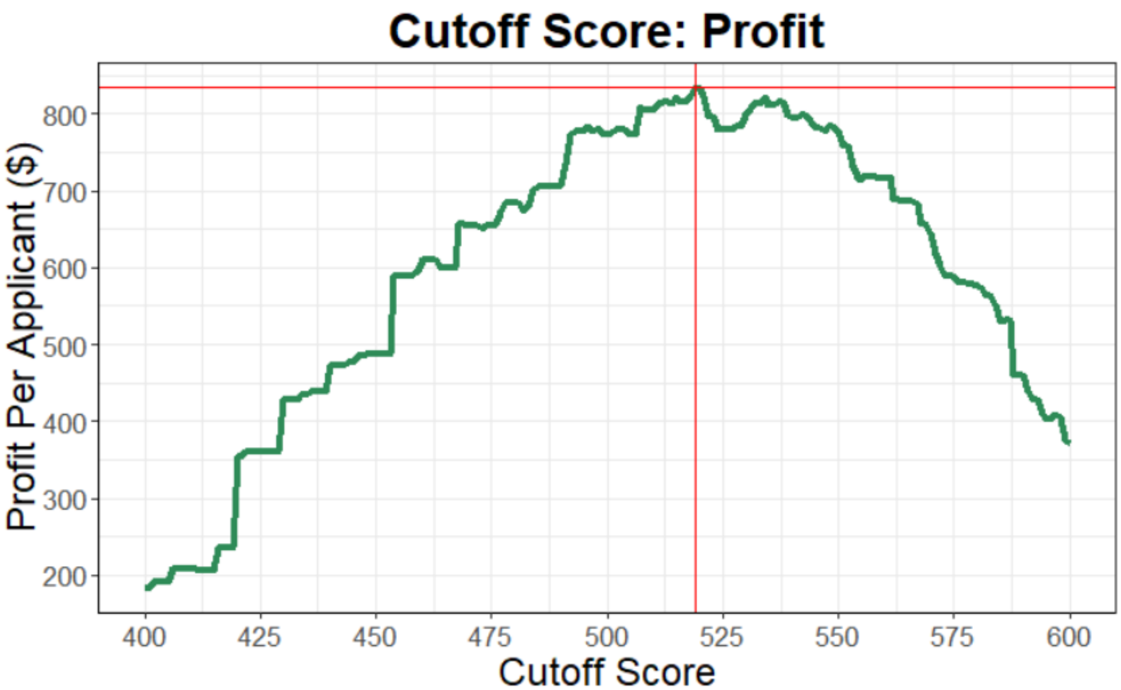

Figure 3: Cutoff Score – Profit

Figure 4: The cutoff score of >519 for granting credit is expected to result in an acceptance rate of ~75 percent and a default rate of 1.64 percent based on the adjusted test set. The plot shows that as the cutoff score decreases the acceptance percentage increases and the default percentage increases.

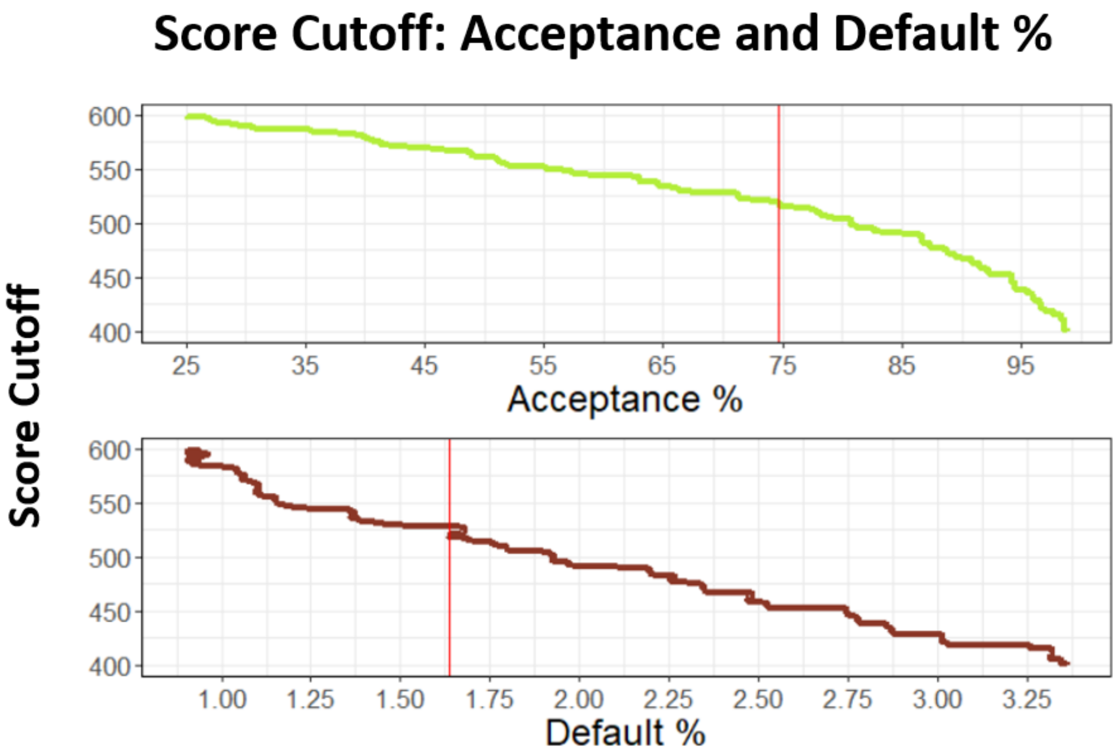

Figure 4: Cutoff Score - Accept and Default %

Figure 5: Scores for all samples were binned in intervals of 50. The predicted probability of default was averaged within each bin. The default rate was calculated by taking the number of people that defaulted within each bin divided by the number of people in the bin. Figure 4 shows that as the score increases the average predicted probability and the default rate decreases.

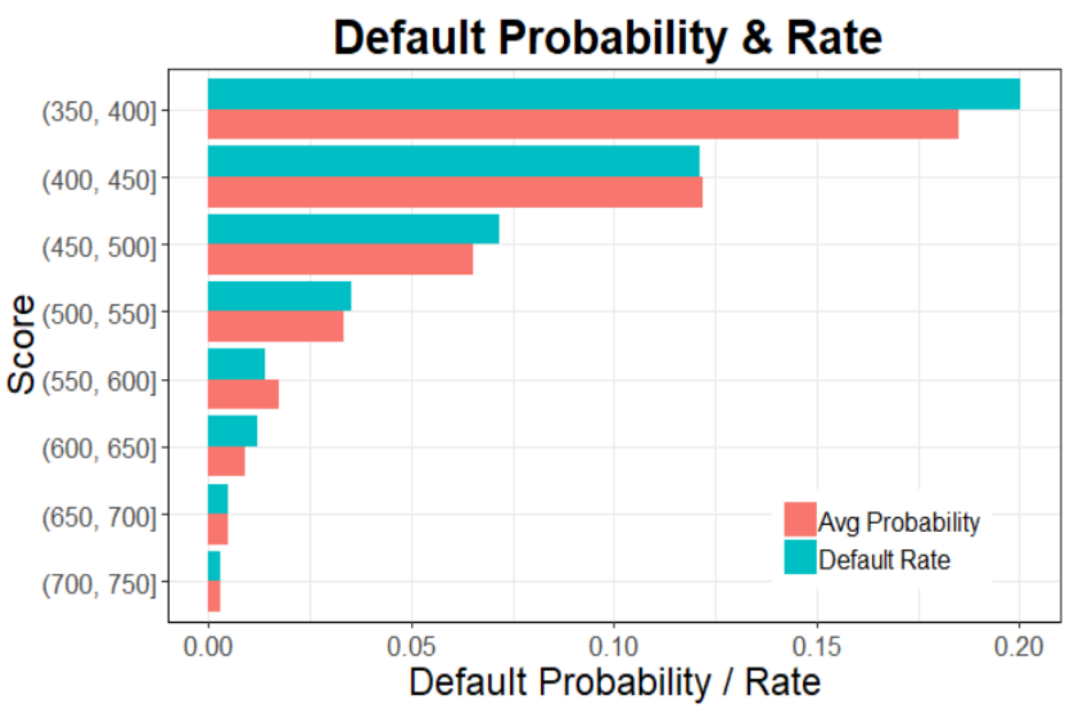

Figure 5: Default Probability and Rate

Figure 6: Confusion matrix for the adjusted test set and using a cutoff score of > 519 for granting credit. The test set as adjusted using the weights of each observation to create a sample that is more representative of the population of credit applicants. Using the scorecard model and a cutoff score of 519 for predicting whether a customer would default is expected to have an accuracy of 76 percent.

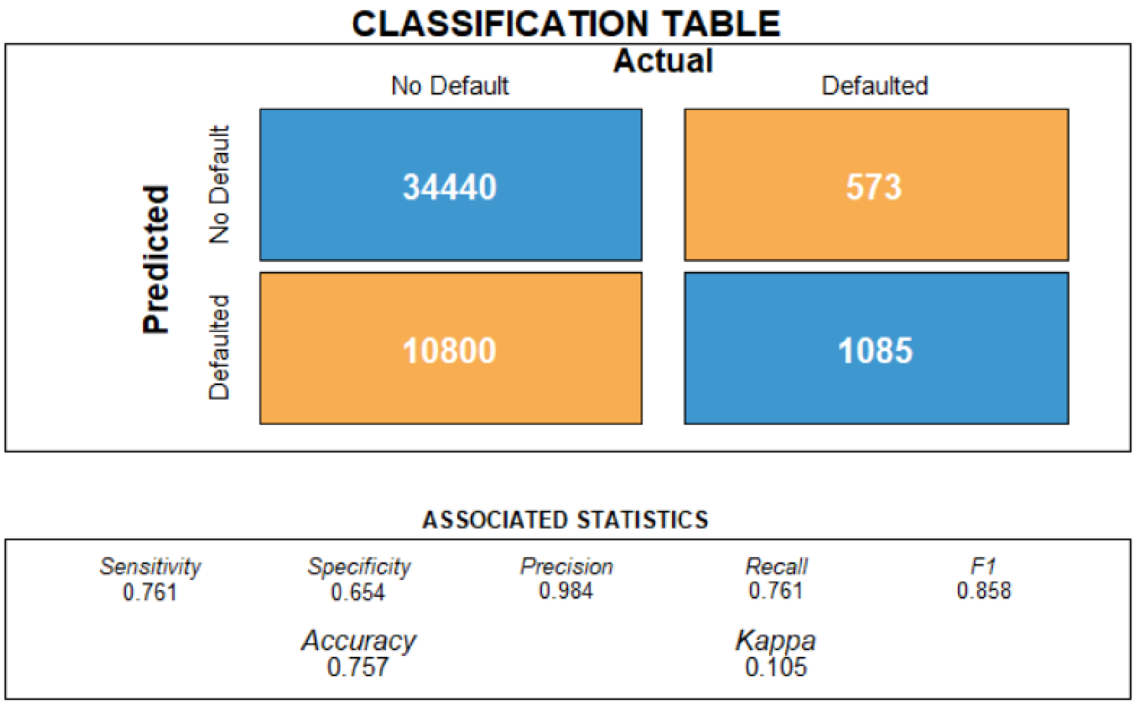

Figure 6: Confusion Matrix

Conclusion

There are inherent risks with lending money. In order to minimize those risks, while maximizing profit and return, we created a scorecard model for the Commercial Banking Corporation to use when reviewing customer's loan requests. The scorecard model we have created will make substantial improvements to profits and default rates. It is expected to maintain the current acceptance rate of 75%, reduce the default rate from 3.23% to 1.64%, increase the average profit per applicant from $192 to $833 and increase the average profit per accepted customer from $256 to $1,116.