|  | Retail Credit Scoring  Feb 1, 2021 |
| --- | --- |
|  |  |
|  |  |

# **Table of Contents**

[**Table of Contents 1**](#_heading=h.1fob9te)

[**Table of Figures 1**](#_heading=h.3znysh7)

[**Table of Tables 1**](#_heading=h.1t3h5sf)

[**Methodology 2**](#_heading=h.lnxbz9)

[*Data Used* 2](#_heading=h.35nkun2)

[*Scorecard Creation* 2](#_heading=h.1ksv4uv)

[*Reject Inference* 2](#_heading=h.m1wx5wfteh8c)

[*Behavioral Scorecard Evaluation* 3](#_heading=h.2jxsxqh)

[*Applicant Scorecard Evaluation* 4](#_heading=h.lplmh4fko96t)

[**results 4**](#_heading=h.3j2qqm3)

[*Tradeoff Plots* 5](#_heading=h.1y810tw)

[*Final Scorecard* 6](#_heading=h.kxvpujteq94u)

[**recommendations 6**](#_heading=h.2bn6wsx)

[**conclusion 6**](#_heading=h.qsh70q)

[**Appendix 7**](#_heading=h.bjb3lfokw4wf)

# **Table of Figures**

[Figure 1: SCORECARD DECILE PLOT 5](#_heading=h.4i7ojhp)

[Figure 2: DEFAULT VS. ACCEPTANCE RATE TRADEOFF PLOT 5](#_heading=h.2xcytpi)

[Figure 3: PROFIT VS. ACCEPTANCE RATE TRADEOFF PLOT 5](#_heading=h.1ci93xb)

[Figure 4: ROC CURVE AND AREA UNDER CURVE FOR ACCEPTED CUSTOMER 9](#_heading=h.2p2csry)

[Figure 5: ROC CURVE AND AREA UNDER CURVE FOR ACCEPTED AND REJECTED CUSTOMER 9](#_heading=h.147n2zr)

# **Table of Tables**

[Table 1: APPLICATION SCORECARD WEIGHT COMPUTATIONS/VALUES 3](#_heading=h.44sinio)

[Table 2: RETAINED VARIABLE INFORMATION VALUES 3](#_heading=h.z337ya)

[Table 3: FINAL SCORECARD 6](#_heading=h.3whwml4)

[Table 4: INPUT VARIABLE RANKED BY INFORMATION VALUE APPLICANT MODEL 7](#_heading=h.3as4poj)

[Table 5: INPUT VARIABLE RANKED BY INFORMATION VALUE BEHAVIORAL MODEL FULL 7](#_heading=h.1pxezwc)

[Table 6: VARIANCE INFLATION FACTORS FOR INPUT VARIABLES 8](#_heading=h.49x2ik5)

Retail Credit Scoring

**OVERVIEW**

The Commercial Banking Corporation (the “Bank”) is interested in the creation of a scorecard that follows FDIC regulations and can be used to evaluate its future retail credit applications, as well as the creation of a distribution to associate score buckets with default rate.

Our team created an applicant scorecard model based on logistic regression and reject inference, as well as a distribution of score buckets and their associated default rates. Both are depicted in the following report. We recommend that The Bank evaluate future applicants by accepting anyone with a score above 543, rejecting anyone below 487, and further investigating anyone with a score in between those two values. Using these cutoffs would allow the Bank to reach a profit of up to $56 million while increasing the current acceptance rate of 75% and decreasing the current default rate of 3.23%. By continuing to refine the scorecard’s underlying model with the collection of new variables, such as the reason for requesting a loan, we believe The Bank can optimize all three of these metrics even further in the future.

# **Methodology**

### *Data Used*

The team was provided with two data sets: one containing customers accepted for a loan and one having customers rejected for a loan. The “accepted customers” data set includes information on 3000 applicants, oversampled and split evenly between “good” applicants who sufficiently paid back their loans and “bad” applicants who went 90 days past due on a loan. The “rejected customers” data set contains information on 1500 customers who were denied a loan. In accordance with the Equal Credit Opportunity Act, we removed any variables related to race, color, religion, national origin, or sex to evaluate an applicant’s creditworthiness. We also decided to remove age, as it can be difficult to prove its viability in a model while disproving its discrimination.

### *Scorecard Creation*

To build our behavioral scorecard, we used only the acceptedcustomers dataset and created a 70:30 training to test split. We first conducted feature engineering to create two new variables: income per household member and percentage of life spent at current job. Next, we transformed all our binary variables with less than five levels into factors and binned all of our continuous variables. Binning these variables increases the interpretability for the final scorecard and helps standardize the variables in the modeling process. Moreover, binning can eliminate issues with outliers and missing values. Using the newly created bins, we calculated their respective weights of evidence.

We narrowed our variable pool down significantly by first selecting those with an IV value above 0.1 and then accounting for multicollinearity. We then created a logistic regression using our weight of evidence values as inputs and a weight of 30 to accurately reflect the population to account for bias from oversampling. This weight was validated by calculating the population ratio of good to bad customers in the accepted customer dataset (96.7% good : 3.23% bad). Finally, we layered a scorecard on top of the logistic regression model before performing reject inference.

### *Reject Inference*

As mentioned above, our dataset contains a group of granted loans and a group of denied loans. To include the denied loans in our full model, we created a behavioral scorecard model of only the accepted applicants and applied that model to the rejected applicants. This model's output provided us with a scores column that we then used to assign a good/bad target value. After assigning the values, we had to adjust the number of sampled rejects to reflect the number of rejects from the population accurately. To accommodate this difference, we used the following formulas:

**Table 1: APPLICATION SCORECARD WEIGHT COMPUTATIONS/VALUES**

| **Weight** | **Calculation** | **Value** |
| --- | --- | --- |
| Rejected Bad Weight | Always 1 | 1 |
| Rejected Good Weight | Population Sample Good/Bad Ratio | 30 |
| Accepted Bad Weight | Population Sample Accepted/Rejected Ratio | 1.5 |
| Accepted Good Weight | Population Sample Accepted/Rejected Ratio \* Population Sample Good/Bad Ratio | 44.94 |

Table 1 above shows the weightings for rejected and accepted cases applied in the applicant scorecard model. After using these weights, we selected a hard cutoff value of .0331 from the optimal K-S statistic and merged the two datasets back together to move forward with the applicant scorecard modeling process.

### *Behavioral Scorecard Evaluation*

After the accepted customer dataset was cleaned, information value was compared for the different variables to determine which would add value to the model. Two calculated variables were included in the analysis, Income per Household Member and Time at Job Per Year of Age. Table 2 below shows the output of this analysis ranked from highest information value to lowest. Every variable with at least an information value of .1 was kept in the model.

**Table 2: RETAINED VARIABLE INFORMATION VALUES**

| Variable | Information Value (IV) | Decision at IV Cutoff of .1 |
| --- | --- | --- |
| Time at Job | 0.2032 | **Keep** |
| Income per Household Member | 0.1987 | **Keep** |
| Credit Cards | 0.1569 | **Keep** |
| Persons per Household | 0.1528 | **Keep** |
| Time at Job per Year of Age | 0.1388 | **Keep** |
| EC Card Holder | 0.1349 | **Keep** |
| Income | 0.1146 | **Keep** |

Based on their information values above .1, the team selected Time at Job, Income per Household Member, Credit Cards, Persons per Household, Time at Job per Year of Age, EC Card Holder, and Income as the variables to move forward in the modeling process. After running the 1st logistic regression, we discovered that the Credit Card variable had semi quasi complete separation, meaning that one card type only had people who had defaulted. To solve this, we combined the problem category with the other card’s category. We then ran a logistic regression with all our significant variables and checked to see if there was any multicollinearity between our variables. Any variable with a Variable Inflation Factor (VIF) above 5.0 was considered a problem for modeling. The Credit Cards variable and the EC Card holder variable both had VIF over 5. The Income variable and the Income per Household Member also had VIF’s over 5. To solve for this multicollinearity, we removed the Credit Card variable, and the Income variable.

After removing those two variables we found we no longer had any multicollinearity problems. We then assessed our new logistic regression and found that Income per Household was not significant, so it was removed from the model. We then ran another logistic regression and found that Time at Job per Year of Age was also not significant, at a p-value of .8321, so it was removed from the model. We then ran a logistic regression on the remaining variables and found them all to be highly significant. The best performing model contained Persons per Household, Time at Job, and EC Card Holder as inputs. Figure 10 in the appendix shows the results from this initial model with an area under the curve (AUC) of 67.69%.

Once the underlying credit scoring model had been completed, the next step was to develop a scorecard that assigned points to every variable. The points were assigned based on the bank’s desire to have a score of 500 with a 20:1 odds of NOT defaulting. Additionally, we incorporated a change of score, 50 points, to correspond with a doubling in the odds of NOT defaulting.

After developing a model for accepted customers, the team underwent reject inferencing to include rejected applicants into the scorecard generation. With the model developed for accepted customers, a hard probability cutoff of .0331 was selected and used to infer whether a rejected customer would have been good or bad. Once calculated, the rejected and accepted customers were combined to develop the scorecard, including both groups.

### *Applicant Scorecard Evaluation*

Concluding creation of a behavioral scorecard, the team incorporated hard cutoff reject inference on customers who were rejected for a loan to adhere to the FDIC regulations. Moreover, incorporating reject inference allows the bank to have a more appropriate model to analyze new applicants. After rejected customers were incorporated into the dataset with a new designation of “good” or “bad” based on the behavioral model, similar steps were taken to create the behavioral scorecard. The variables selected for the final credit scoring model were based on IVs greater than 0.1 and can be seen in Table 4 in the appendix.

After the additional variable selection was complete using variable significance and accounting for multicollinearity, the final credit scoring model included persons per household, time at job, and EC cardholder. This model also had an AUC of 76.62%, as seen in Figure 5 in the appendix. This is similar to AUCs of other more complex models. By choosing a model with fewer variables, we can allow the bank to have good performance with an interpretable model. Once the underlying credit scoring model was created, we again assigned scores in accordance with the bank’s score and odds preferences as mentioned in the previous section. Each individual was then assigned an overall score. From the distribution of these scores, we were able to designate a cutoff point for accepting or rejecting a customer. The model’s results and implications for the bank's business will be provided in the next section.

# **results**

We used our final scorecard model to create several tradeoff plots that helped us determine cutoff points for the Bank’s applicants.

### *Tradeoff Plots*

Chart, bar chart

Description automatically generated

**Figure 1: SCORECARD DECILE PLOT**

Figure 1 above buckets the predicted scores from our scorecard into six equally sized buckets. Because our selected model only contains three predictor variables, there were not enough unique scores to create ten buckets. As the plot shows, the default rate steadily decreases over the first three buckets and then sharply declines from ~7% to ~1.5% between buckets 3 and 4 (at a score of 543). This information gave us a starting point for choosing a cutoff value.

| Chart, line chart  Description automatically generated  **Figure 2: DEFAULT VS. ACCEPTANCE RATE TRADEOFF PLOT** | **Diagram  Description automatically generated**  **Figure 3: PROFIT VS. ACCEPTANCE RATE TRADEOFF PLOT** |
| --- | --- |

Additionally, we used the default vs. acceptance rate tradeoff plot (Figure 2 above) to help determine our cutoff value. The average default rate is 1.57% and occurs around a score of 487, so any cutoff above this point would provide a default rate below average. At this point the acceptance rate is 86.6%, which is significantly higher than the current acceptance rate of 75%.

Overall, this plot proves that our model adds value to the Bank: At a 75% acceptance rate, our model brings the default rate down from 3.23% to nearly 1%, and at the current default rate of 3.23%, we can increase the acceptance rate to almost 100%.

We then analyzed the profit vs. acceptance rate tradeoff plot (Figure 3) to help us determine our final score range. Profit was calculated based on an expected revenue of $2000 from good customers and an expected cost of $52000 from bad customers. We determined that profit is maximized at a score of 543 and a value of approximately $56 million. At this score, the acceptance rate is 83.4%. The average profit made is 30.6 million, which is made at a cutoff of 452, so our cutoff should be above this average.

After considering all of the plots above, we chose a hard upper cutoff of 543. The default rate increases significantly beyond this point, profit is maximized at this point, and both the acceptance rate and default rate improve from their current status. We chose 487 as our hard lower cutoff. The profit and acceptance rates at this point are well above average ($54 million and 86.6%, respectively), and the default rate falls in the average range. Any scores between these two points should be inspected more closely.

### *Final Scorecard*

**Table 3: FINAL SCORECARD**

| **Variable** | **Level** | **Scorecard Points** |
| --- | --- | --- |
| Persons in Household | x = 1 person | 124.15 |
| Persons in Household | x > 1 person | 205.62 |
| Time at Job | x >= 6 months | 89.18 |
| Time at Job | 7 months < x <= 15 months | 123.27 |
| Time at Job | 15 months < x <= 84 months | 184.40 |
| Time at Job | x > 84 months | 228.76 |
| EC Card | No EC Card | 156.38 |
| EC Card | EC Card | 247.50 |

The scorecard displayed in Table 3 shows how each variable in the model affects an individual’s score. For example, an individual who lives alone (124.15 points), has been at their job for eight months (123.27 points) and has no EC Card (156.38 points) receives a total score of 403.8, which puts them below our hard lower cutoff of 487 and leads to immediate rejection. An individual who has the same qualities but has been at their job for over 84 months would receive a score of 509.3 and fall within our soft cutoff range, requiring a manual check. If this individual acquired an EC Card, that would boost them past our upper cutoff, causing them to be immediately accepted.

# **recommendations**

Taking into consideration all of the analysis discussed above, our team recommends that The Bank take the following actions to maximize profits while increasing the acceptance rate and reducing the default rate:

* Use the provided scorecard with a hard upper cutoff of 543 and a hard lower cutoff at 487 to make lending decisions; any applicants with a score between those values should be evaluated further.
* Collect more information on applicants, such as the reason they’re requesting a loan (Ex: small business loan) or their remaining balance in outstanding loans. These additional pieces of information could enhance model accuracy and subsequently increase profits.

# **conclusion**

Using our recommended applicant scorecard with a hard upper cutoff of 543 and hard lower cutoff of 487 will allow The Bank to make lending decisions that result in a profit of up to $56 million while increasing the current 75% acceptance rate and decreasing the current 3.23% default rate. It’s recommended that The Bank collect additional information on future applicants, as this could help build an even more accurate model and increase future profits. Our team has enjoyed working on this project and looks forward to an opportunity to work with The Bank again in the future.

# **Appendix**

**ADDITIONAL FIGURES AND TABLES**

**Table 4: INPUT VARIABLE RANKED BY INFORMATION VALUE APPLICANT MODEL**

| Variable | Information Value (IV) | Decision at IV Cutoff of .1 |
| --- | --- | --- |
| Time at Job | 0.5269 | Keep |
| Income | 0.4826 | Keep |
| Credit Cards | 0.4152 | Keep |
| Income per Household Member | 0.4014 | Keep |
| EC Card Holder | 0.3982 | Keep |
| Time at Job per Year of Age | 0.3849 | Keep |
| Persons per Household | 0.2791 | Keep |
| Number of Children | 0.1162 | Keep |
| Type of Vehicle | 0.1050 | Keep |

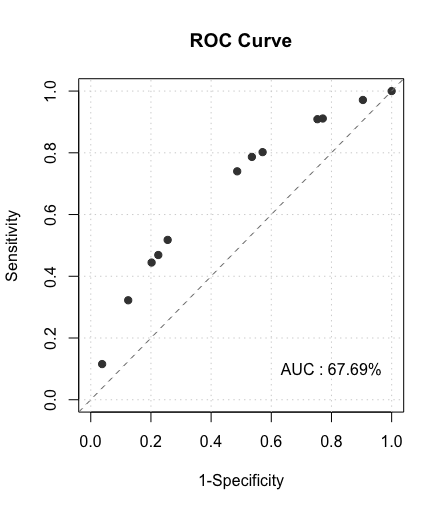
**Table 5: INPUT VARIABLE RANKED BY INFORMATION VALUE BEHAVIORAL MODEL FULL**

| Variable | Information Value (IV) | Decision at IV Cutoff of .1 |
| --- | --- | --- |
| Time at Job | 0.2032 | **Keep** |
| Income per Household Member | 0.1987 | **Keep** |
| Credit Cards | 0.1569 | **Keep** |
| Persons per Household | 0.1528 | **Keep** |
| Time at Job per Year of Age | 0.1388 | **Keep** |
| EC Card Holder | 0.1349 | **Keep** |
| Income | 0.1146 | **Keep** |
| Telephone | 0.0860 | Drop |
| Type of Vehicle | 0.0743 | Drop |
| Profession | 0.0662 | Drop |
| Number of Loans with Bank | 0.0513 | Drop |
| Number of Children | 0.0476 | Drop |
| Type of Credit Product | 0.0293 | Drop |
| Number of Loans Outside Bank | 0.0136 | Drop |
| Residence Type | 0.0127 | Drop |
| Finished Paying off Loans | 0.0106 | Drop |
| Credit Bureau Risk Class | 0.0067 | Drop |
| Region (Div) | 0.0055 | Drop |
| Location of Credit Bureau | 0.0004 | Drop |
| Time at Address | NA | Drop |
| Region | NA | Drop |
| Requested Cash | NA | Drop |
| Target (Good/Bad) | NA | Drop |

**Table 6: VARIANCE INFLATION FACTORS FOR INPUT VARIABLES**

| Variable | Variance Inflation Factor |
| --- | --- |
| Income per Household Member | 5.0250 |
| Income | 5.6981 |
| Persons per Household | 2.4101 |
| Time at Job | 3.6058 |
| EC Card Holder | 2.7628 |
| Time at Job per Year of Age | 3.5580 |

**Figure 4: ROC CURVE AND AREA UNDER CURVE FOR ACCEPTED CUSTOMER**



**Figure 5: ROC CURVE AND AREA UNDER CURVE FOR ACCEPTED AND REJECTED CUSTOMER**

