|  |  |
| --- | --- |
|  | **Retail Credit Scoring**    ORANGE HOMEWORK TEAM 8  Karthick Krishna Balaji  Camille Carter  Margeaux Johnson  Dillard McMichael  Nish Torane    January 30, 2023 |

**Table of Content**

[**Overview**](#_heading=h.17dp8vu) **1**

[**Methodology and Analysis**](#_heading=h.3rdcrjn) **1**

[**Data Used**](#_heading=h.haldq4efravl) **1**

[**Missing Observations and Outliers**](#_heading=h.ab8240tk7x6x) **1**

[**Separation Issues**](#_heading=h.g5cmkkuo4b3l) **1**

[**Binning Techniques**](#_heading=h.8llr5zil0q9q) **2**

[**Initial Model**](#_heading=h.h3701m3715um) **2**

[**Reject Inferencing**](#_heading=h.8e3c922ekz1t) **3**

[**Final Model**](#_heading=h.ktxbiyf65xmw) **3**

[**Results and Recommendations**](#_heading=h.cq2321jxrcvd) **3**

[**Final Scorecard**](#_heading=h.aphhouxcke3u) **3**

[**Final Cut-Off**](#_heading=h.wzjwp0q4l515) **4**

[**Recommendations**](#_heading=h.vns6zifpn2tr) **5**

[**Conclusion**](#_heading=h.esxwwxo6toq1) **5**

[**Appendix**](#_heading=h.exj2x5eo9xv8) **6**

**Retail Credit Scoring**

# **Overview**

The Commercial Banking Corporation (hereafter the “Bank”) has tasked Orange Homework Team 8 (hereafter the “Analysts”) to determine the retail credit eligibility of applicants and create a comprehensive scorecard for evaluating retail credit applications. The Bank provided two datasets detailing accepted and rejected applicants.

The Analysts created a credit scorecard by building a logistic regression on the accepted applicant dataset and obtained the optimal cut-off of 0.0321 using Kolmogorov–Smirnov (hereafter “KS”) statistic, which the Analysts used for hard cut-off reject inferencing. The scorecard was built on the following variables:

* Age of Applicant
* Credit Cards
* Number of Children
* EC Card Holder
* Number of Persons in Household
* Time at Job

These six significant variables resulted in 20 different levels, each with a unique point assignment as seen in **Table 1** in the Appendix. The Analysts trained the algorithm and evaluated its performance with an Area Under the Receiver Operating Characteristic curve (hereafter “ROC-AUC”) of 0.79. Using this scorecard, the Analysts recommend a final cut-off range in which the bank rejects any applicant below a score of 381, accepts any applicant above 529, and any applicant between 381 and 529 can be sent to a team for further investigation. The Analysts recommend the Bank improve its data entry process since there were many related issues. Additionally, the Analysts recommend that the Bank continue accepting 75% of applicants, as profit is optimized at this number. Since this outcome was noticed after increasing the acceptance percentage from the model, the Analysts would love to investigate with assistance from the Bank.

# **Methodology and Analysis**

## Data Used

The Bank provided two datasets for the proposal. The first dataset contained information regarding the accepted applicants. Out of 3,000 observations, half were categorized as “good”, meaning they did not default, and the other half were categorized as “bad”, meaning they were, at one point, 90 days past due. The actual default rate in the population was 3.23%, so over-sampling was corrected through a weight variable. The weights are assigned correctly: 30 for non-defaulters and 1 for defaulters. The second dataset contained information regarding the rejected applicants, and there were a total of 1,500 observations. Both datasets were checked for duplicates, and two observations were removed from the rejected data. To aid the prediction, the accepted applicant dataset contained 26 predictor variables and the rejected applicants dataset contained 24 predictor variables related to applicants’ personal information and existing accounts. For the initial and final model, a 70/30 split on the data for the training and testing sets was used, given the small sample size.

## Missing Observations and Outliers

The Analysts made many imputations to the data before modeling. Blank observations within Profession, Type of Credit Product, and Residence Type were imputed to missing. Seven variables had outliers: Requested Cash, Number of Children, Income, Number of Loans Outside the Bank, Number of Persons in Household, Time at Address, and Time at Job. The Analysts confirmed the highest value in each of these variables was an outlier with the Grubbs statistical test. The Analysts chose not to impute or remove these observations because, once binning (smbinning in R Language) was utilized, these outliers were grouped with the highest binned categories.

## Separation Issues

The Analysts identified five variables that presented separation issues: Credit Bureau Risk Class, Type of Credit Product, Profession, Telephone, and Credit Cards. The original Credit Bureau Risk Class categories (1, 2, and 3) were imputed into ‘0-1’ and ‘2’ bins. The newly imputed missing observations for Type of Credit Product and Profession were grouped with the ‘Other’ category for new bins called ‘Missing/Other’. The original Telephone categories (0, 1, and 2) were imputed into ‘0-1’ and ‘2’ bins. Lastly, any instances of ‘Visa’ within the Credit Cards variable were grouped into one ‘Visa’ bin, and the one instance of ‘American Express’ was grouped into the ‘Other Credit Cards’ bin.

The Analysts similarly imputed both the accepted and rejected datasets; however, other unique imputations were performed on the rejected data. First, the Profession and Type of Credit Product variables contained observations with commas, which caused issues in later Information Value analyses, so these were removed. There also seemed to be data entry errors within these two variables that the Analysts corrected for. Within the Profession variable, the following categories were grouped:

* ‘Civil Service M’ and ‘Civil Service’
* ‘FoodBuildingCa’ and ‘Food or Building’
* ‘Sea Voyage Gast’ and ‘Sea Vojage’
* ‘StateSteel Ind’ and ‘State or Steel Ind’

Within the Type of Credit Product variable, the following categories were also grouped:

* ‘Dept. StoreMail’ and ‘Dept. Store or Mail’
* ‘FurnitureCarpet’ and ‘Furniture or Carpet’
* ‘Radio TV Hifi’ and ‘Radio or TV or Hifi’

## Binning Techniques

The Analysts identified nine nominal predictor variables: Credit Bureau Risk Class, Type of Vehicle, Credit Cards, Nationality, Type of Credit Product, Number of Bank Loans, Profession, Region, and Telephone. There were also five binary predictor variables: Division, EC Card Holder, Finished Paying Off Previous Loans, Location of Credit Bureau, and Residence Type. The rest of the eight predictor variables were classified as interval or continuous: Age, Requested Cash, Number of Children, Income, Number of Loans Outside the Bank, Number of Persons in Household, Time at Address, and Time at Job.

## Initial Model

The Analysts used only the accepted applicant data to create the initial model. Through binning, the Analysts discovered no significant splits in the Number of Loans Outside the Bank. Given there was no statistical relationship with the target non-defaulter binary variable, it was not included in the initial model. The Weight of Evidence (WOE) plot for the Income variable caused some concern; **Figure 3** in the Appendix implies that applicants with no income are usually categorized as non-defaulters. Applicants with an income less than or equal to $2,500 are categorized as defaulters, while applicants with an income greater than $2,500 are categorized as non-defaulters. In the business sense, those with no income would likely default at least once, so it raised some questions about why the WOE was positive for applicants with zero income.

The Analysts created a variable importance plot from the accepted applicant training data with predictors ranked in descending order by their Information Value, as seen in **Figure 4** in the Appendix. The Analysts identified that the only predictor strongly related to the target, having the most predictive power (highest Information Value), was Age. The predictors classified as having ‘medium’ predictive power were Time at Job, Income, Number of Persons in Household, Credit Cards, and EC Card Holder. Other variables in the plot had Information Values less than 0.1 and were not included in the initial model.

The Analysts built the initial model using logistic regression. The target variable was a binary indicator with one corresponding to defaulters and zero corresponding to non-defaulters. The Analysts created WOE variables for each of the six significant variables in **Figure 4** in the Appendix and no other variable selection was performed because, in the banking industry, this is done only with Information Values. The overall performance of the accepted applicant training data in the initial model returned a ROC-AUC of 0.6994, as seen in **Figure 5** in the Appendix. This means that the initial model correctly ranked probabilities associated with defaulters over non-defaulters 69.94% of the time in the training data.

From the performance metrics, the Analysts also determined the optimal cut-off to be used for hard cut-off reject inferencing. The KS statistic of 0.3066, visualized in **Figure 6** in the Appendix by a black dashed line, correlates with an optimal cut-off of 0.0321. Lastly, the Analysts ran the initial model on the accepted applicant testing data, returning a ROC-AUC of 0.5378, as seen in **Figure 7** in the Appendix. This means that the initial model correctly ranked probabilities associated with defaulters over non-defaulters 53.78% of the time in the testing data.

## Reject Inferencing

The Analysts utilized hard cut-off reject inferencing. Given that the acceptance rate is 75%, the Analysts randomly chose 1,125 observations from the rejected applicant data (25% rejection rate) to accurately represent the population. It is given that the Bank assigns a score of 500 to applicants with an odds ratio of 20:1 and that doubling the odds is associated with a change of 50 points in the scorecard. The Analysts used this information to create an initial scorecard by assigning points to each level of the predictors used in the initial model. Rejected applicants were then scored from the initial model, and observations below the cut-off of 0.0321 were predicted to default, while those above the cut-off were predicted to not.

## Final Model

The Analysts created **Figure 8**,as seen in the Appendix, another variable importance plot from the combined applicant training data with predictors ranked in descending order by their Information Value. The Analysts identified three predictors strongly related to the target with the highest predictive power (Information Value): Age, Time at Job, and Number of Persons in Household. The predictors classified as having ‘medium’ predictive power were Credit Cards, EC Card Holder, and Number of Children. The analysis showed that Income, Number of Loans Outside the Bank, and Requested Cash had no significant splits, so there was no statistical relationship with the target variable. All other variables had Information Values less than 0.1 and were not included in the final model.

The final model produced different results from the initial model. Now, the previous Income variable concerns are no longer an issue, the Number of Persons in Household and Time at Job are strong predictors instead of medium predictors, and the Number of Children is a significant variable in the final model.

As with the initial model, the final model was created using logistic regression and WOE variables for each of the six significant variables were calculated. The Analysts made no other variable selections. However, the Analysts calculated the new weight variable using the proportion of defaulters to non-defaulters and the proportion of accepted applicants to rejected applicants. The overall performance of the combined applicant training data in the final model returned a ROC-AUC of 0.7882, as seen in **Figure 9** in the Appendix. This means that the final model correctly ranked probabilities associated with defaulters over non-defaulters 78.82% of the time in the combined training data. Next, the Analysts ran the model on the testing data, returning a ROC-AUC of 0.7904, as seen in **Figure 10** in the Appendix. This means that the final model correctly ranked probabilities associated with defaulters over non-defaulters 79.04% of the time in the combined testing data.

# **Results and Recommendations**

## Final Scorecard

The Analysts built the final scorecard considering six variables (Age, Time at Job, Number of Persons in Household, Credit Cards, EC Card Holder, and Number of Children) to determine applicants' retail credit eligibility. The scaling was determined using previously created point variables from the final scoring model. These six variables, the levels, and point allocations resulted in 20 levels with unique point assignments are shown in **Table 1** in the Appendix. The Bank can easily score new applicants by adding up the point values associated with their attributes. For example, an individual, who we will call John:

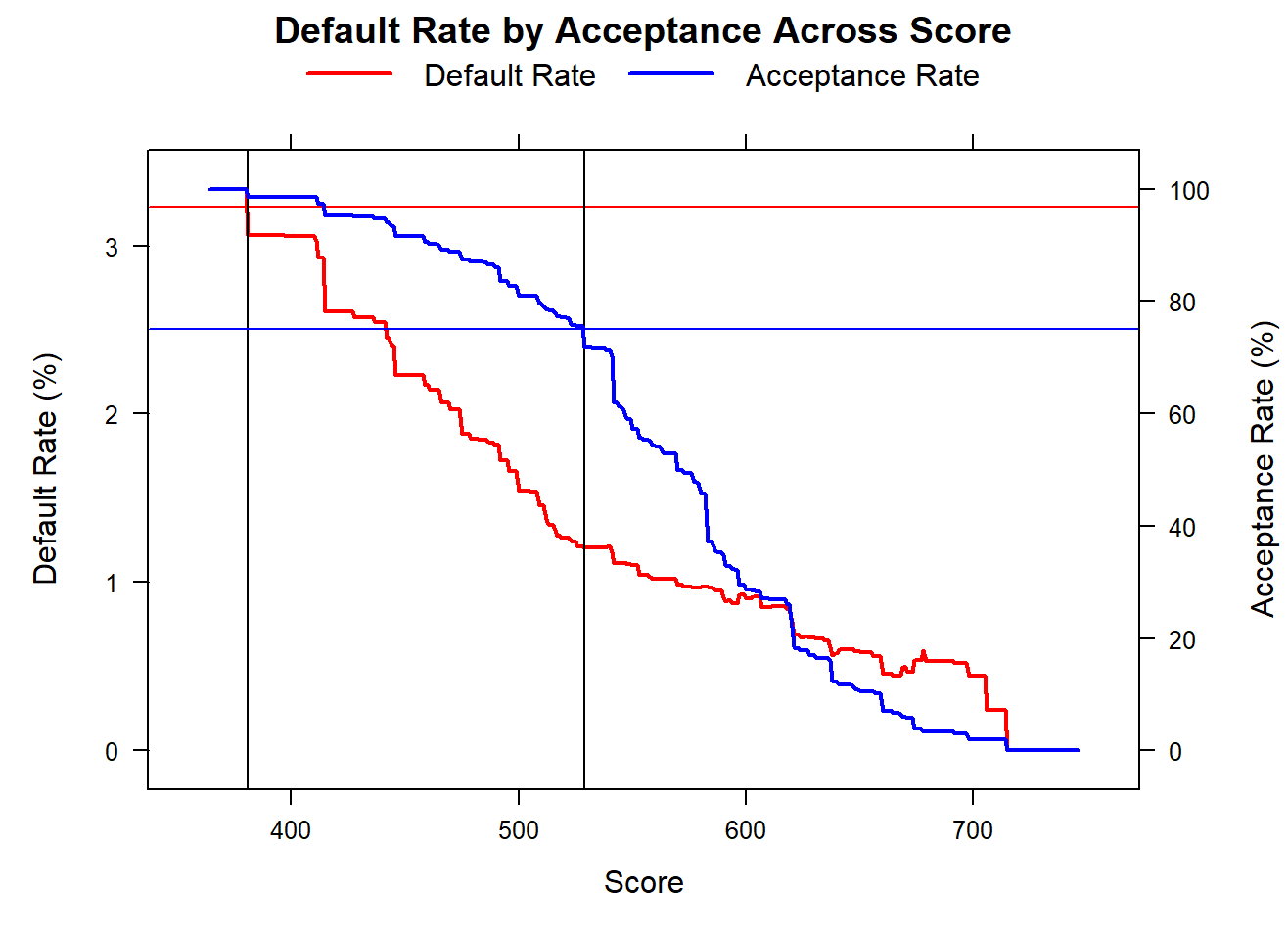
* Is 30 years old: 103 points
* Has a Mastercard: 175 points
* Has children: 78 points
* Is an EC Card holder: 83 points
* Has four people in the household: 104 points
* Has held the same job for three years: 52 points

These attributes sum to a total score of 595.

The model’s performance was evaluated by comparing the default rate across predicted scores. The plot of the deciles and the corresponding default rates are pictured in **Figure 11** in the Appendix. As the scores increase, the default rate consistently drops, indicating that the model can effectively distinguish between defaulters and non-defaulters. Individuals with a score above 529 will have a lower default rate than the current rate of 3.23%.

## Final Cut-Off

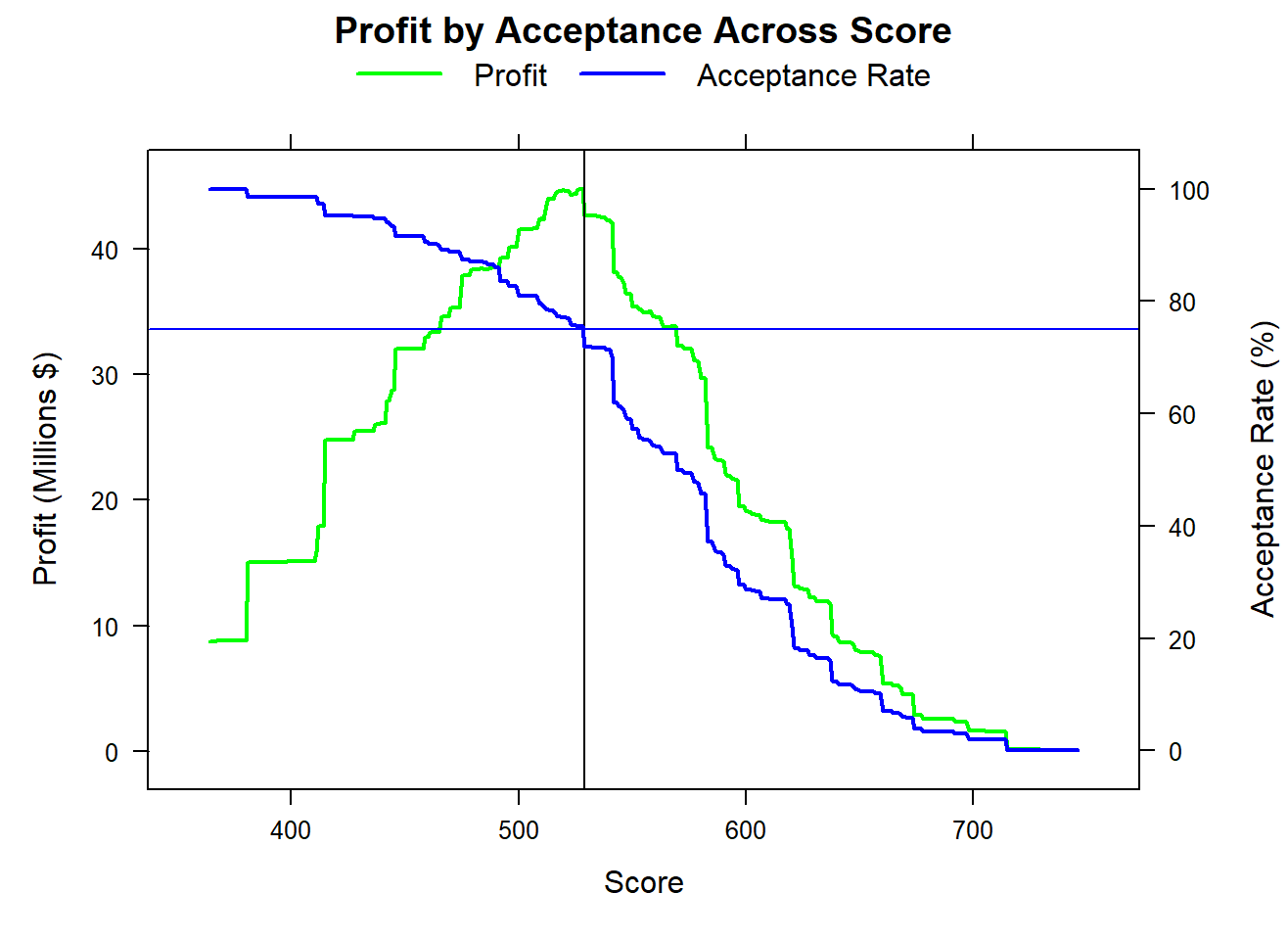
Based on the results of this final scorecard, the Analysts recommend that the Bank establish a final cut-off range corresponding to the current default rate (3.23%) and acceptance rate (75%). The current default rate is visualized in **Figure 1** by the red horizontal line, and the blue horizontal line represents the acceptance rate.



**Figure 1: Default and Acceptance Rate by Score**

The Bank can increase its acceptance rate to 97% while maintaining its current default rate by rejecting applicants below a score of 381, as seen by the first black vertical line in **Figure 1**. Similarly, they can maintain their acceptance rate and decrease the default rate from 3.23% to 2.5% by accepting applicants with a score above 529, as seen by the second black line in **Figure 1**. Using this information, the Bank can safely reject applicants who fall below the cut-off of 381 and accept those above 529 while sending applicants whose score fall between those cut-offs to an internal review process to decide their status.

Additionally, the Bank should continue accepting at least 75% of retail credit applicants to maximize profit. **Figure 2** on page 5 shows a blue horizontal line at the current acceptance rate of 75% and a vertical black line where profit is maximized. The profit, in millions, reaches its peak when the acceptance rate is at 75% and the score is at 529, corresponding to the recommended cut-off. This plot was made using an expected cost of $52,000 for accepting applicants who default and an expected profit of $2,000 for accepting applicants who do not default.



**Figure 2: Profit and Acceptance Rate by Score**

Since profit was already maximized at the current acceptance rate, this was not considered when creating the final cut-off. Instead, the Analysts opted to identify cut-off points to maintain the current acceptance rate while minimizing the default rate and maintain the current event rate while maximizing the acceptance rate. In the previous example, John received a score of 595, and given the final cut-off range, he would be accepted.

## Recommendations

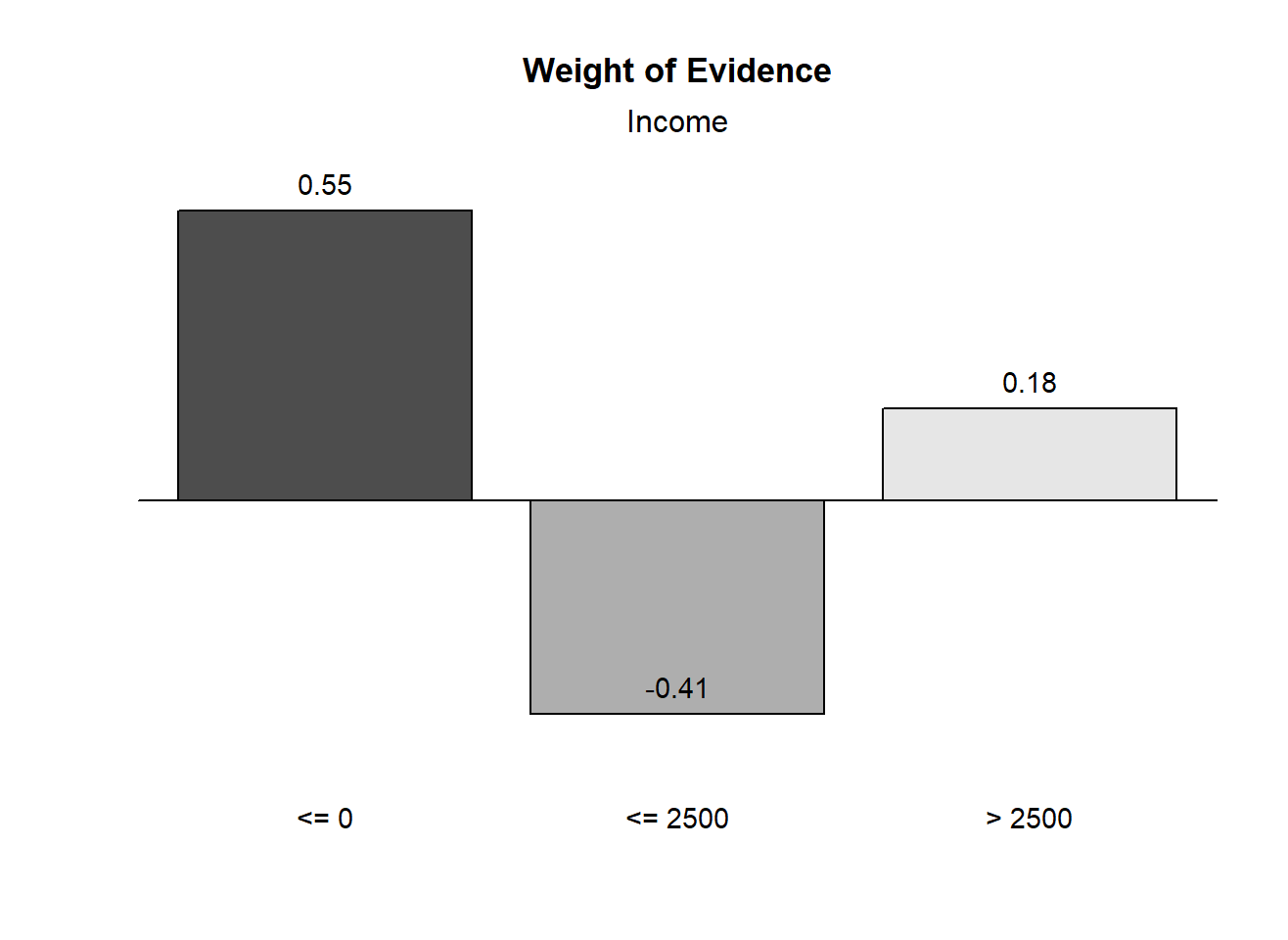
In addition to the recommendations of implementing a hard cut-off range and keeping the acceptance rate of at least 75%, the Analysts also recommend the Bank improve its data entry process going forward. Specifically, the Analysts encountered an unexpected finding when individuals with no income were categorized as less likely to default than applicants with an income less than or equal to $2,500. Additionally, the various imputations carried out by the Analysts showed that the collection and entry of data could be cleaned to improve further analyses. Going forward, the Analysts would also recommend a more thorough investigation into how altering the acceptance rate could increase profit as new information is added to the suggested model.

# **Conclusion**

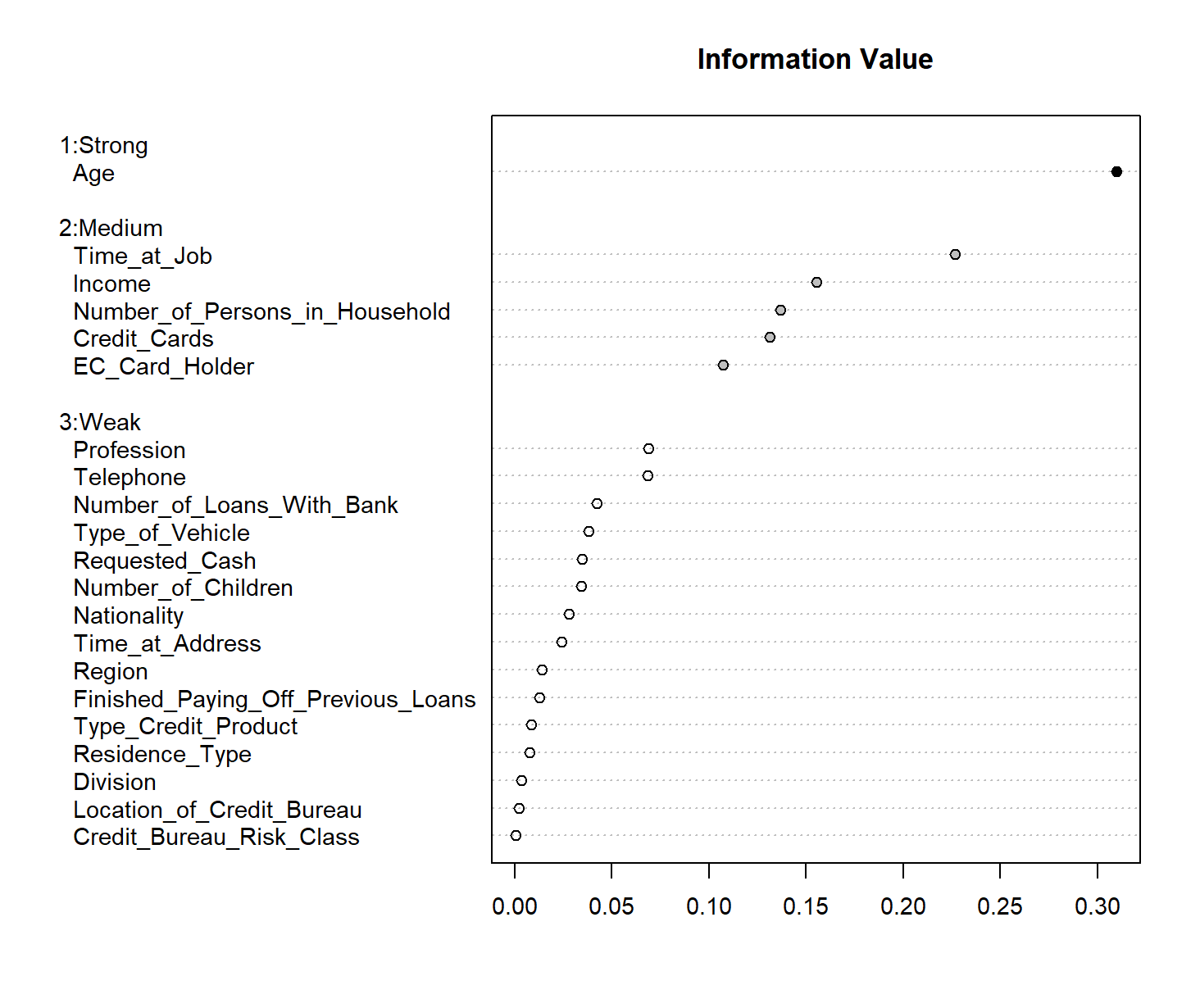
This report details the process of building a comprehensive scorecard and associated score buckets for evaluating all retail credit applications. During data processing, the Analysts handled separation issues and used binning on continuous variables, taking care of missing values and outlier issues. The Analysts created an initial model on the accepted applicant training dataset and used the KS statistic to determine the optimal cut-off of 0.0321. From the test dataset of accepted applicants, the Analysts obtained a ROC-AUC of 0.5378. The Analysts utilized hard cut-off rejected inferencing, built the final model on applicants from both the accepted and rejected data, and obtained a ROC-AUC of 0.7904 on the test dataset.

The Analysts created the final scorecard considering six variables to determine the retail credit eligibility of applicants. The resulting decile plot distinguished between defaulters and non-defaulters effectively. We concluded that individuals with a score above 529 would have a lower default rate than the Bank’s current default rate of 3.23%. The Analysts further recommend that the Bank improve its data entry process since there were many related issues. Additionally, we recommend that the Bank continue accepting at least 75% of applicants, as profit is maximized at this rate. Since this outcome was noticed after increasing the acceptance percentage from the model built, the Analysts would love to investigate with assistance from the Bank.

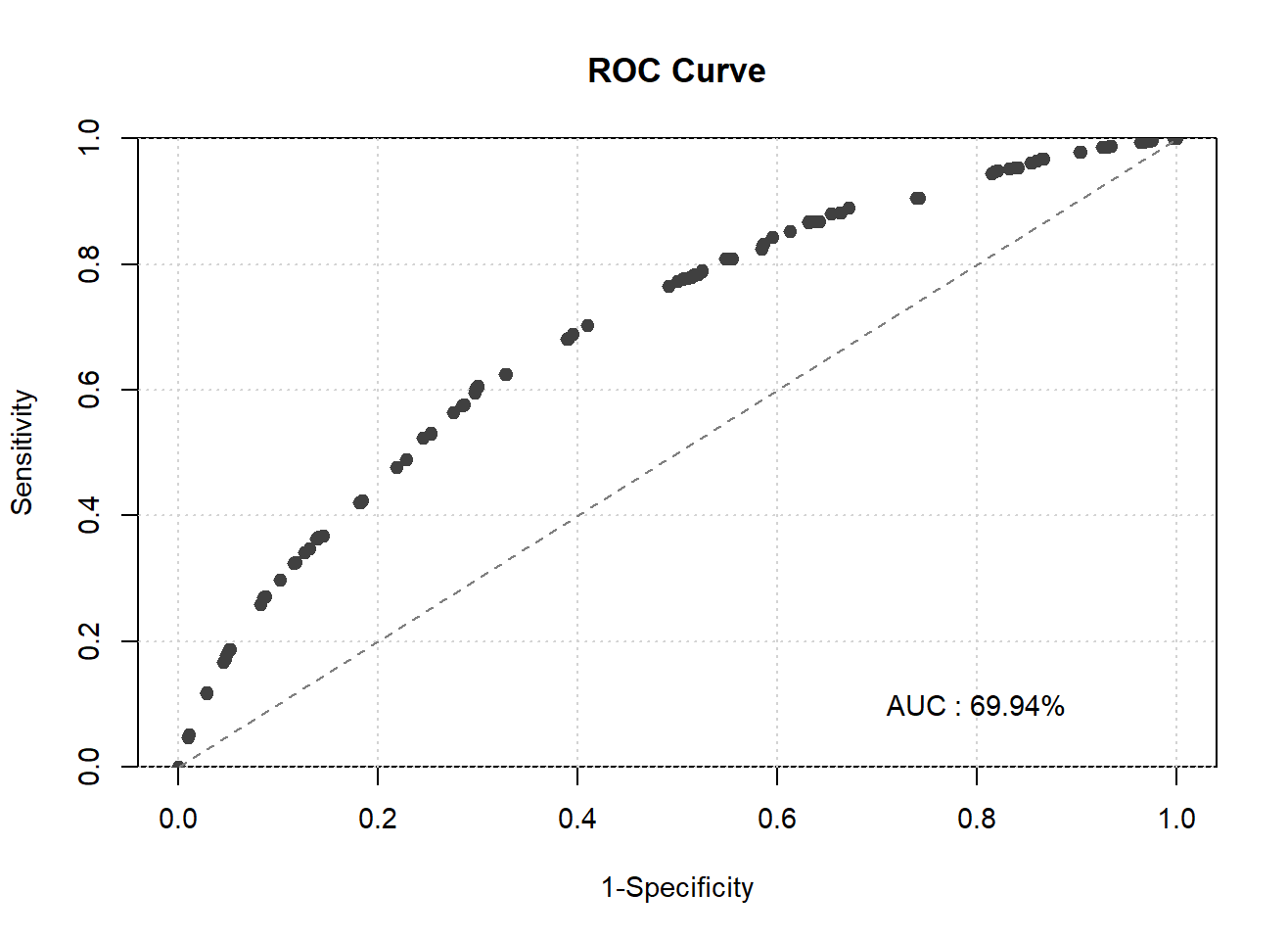
# **Appendix**



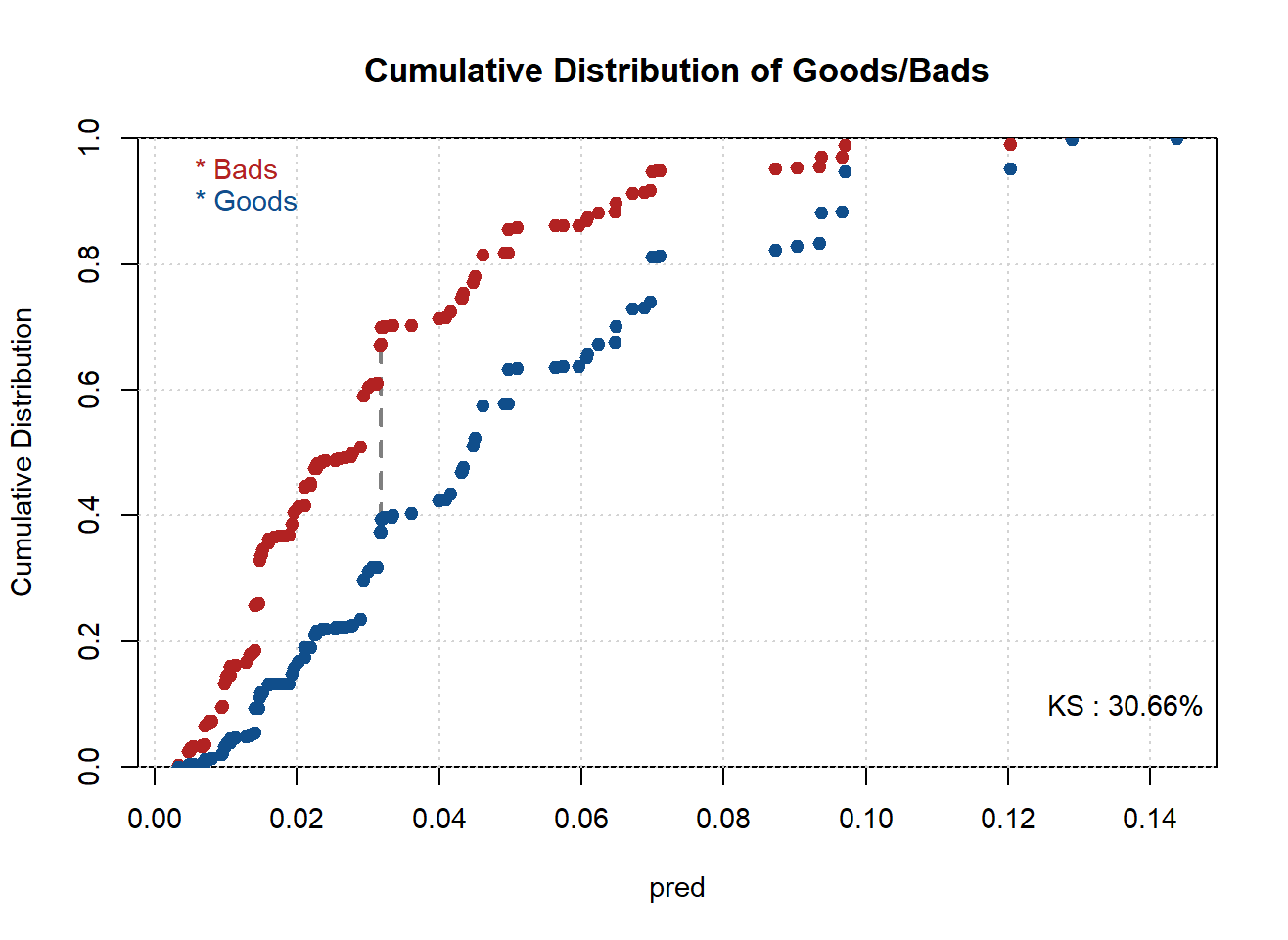
**Figure 3: Binned Income Weight of Evidence**



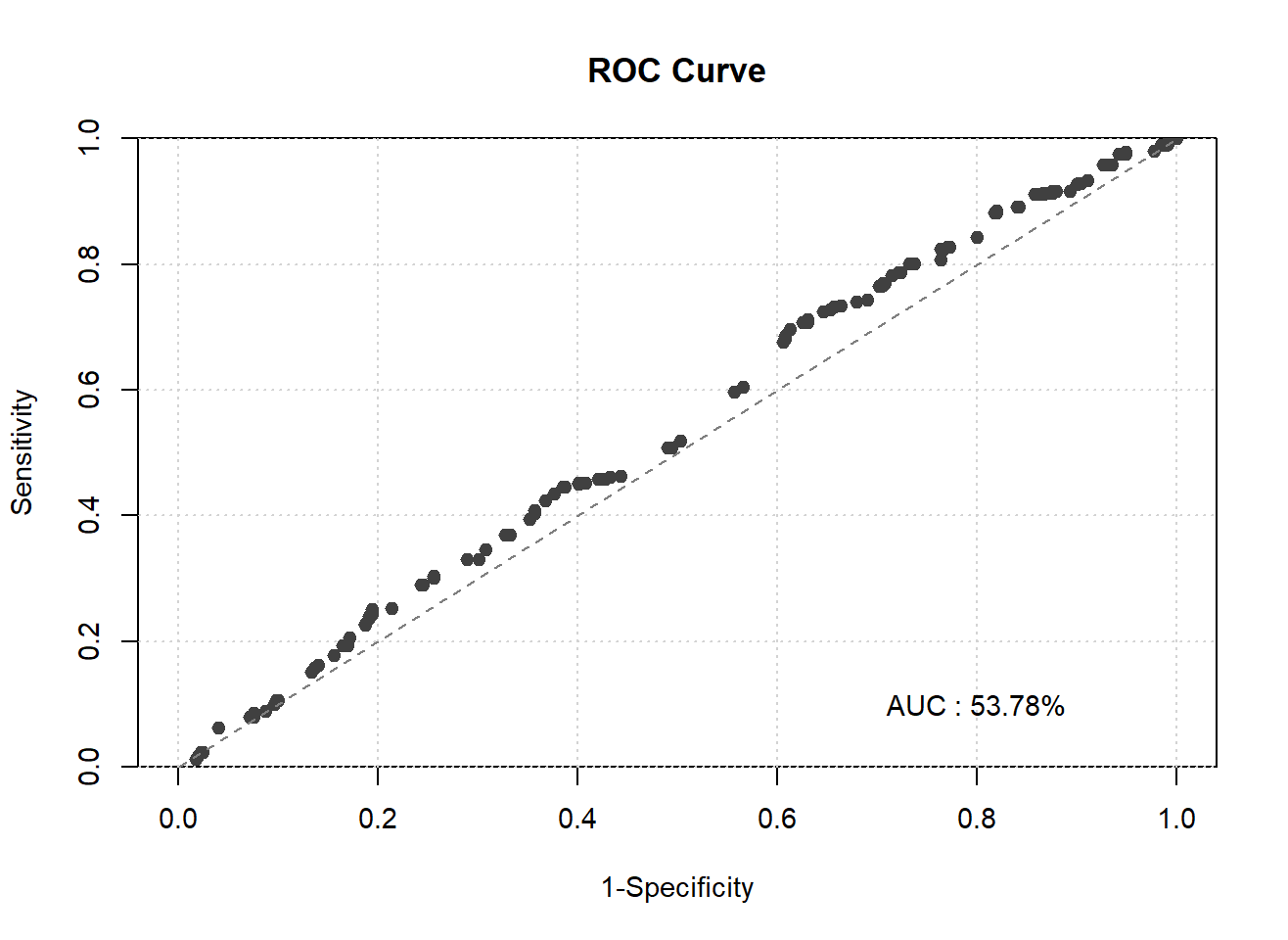
**Figure 4: Information Values for the Initial Model**



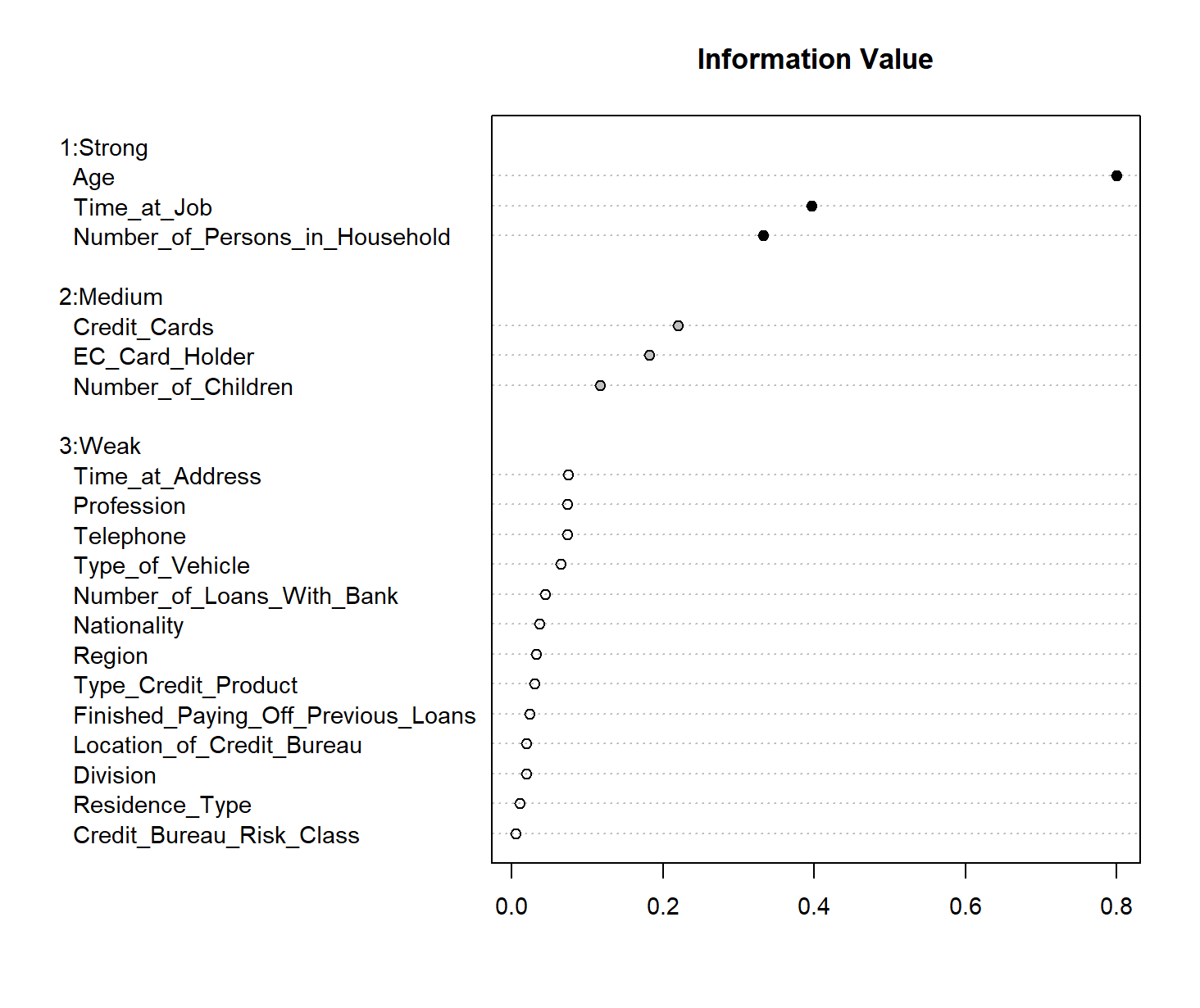
**Figure 5: ROC-AUC Curve from Initial Model on Accepted Training Data**



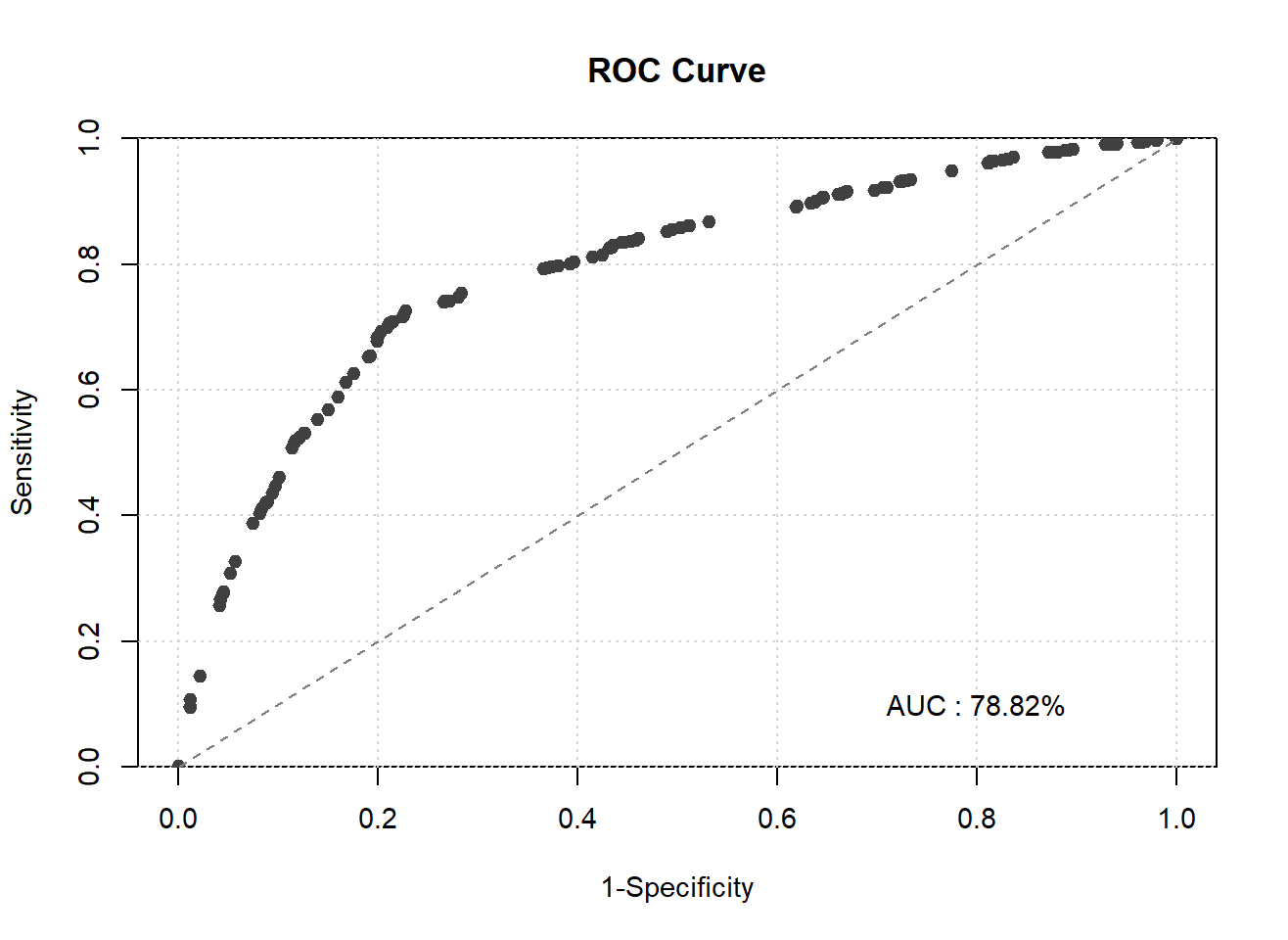
**Figure 6: Optimal cut-off from Initial Model on Accepted Training Data**



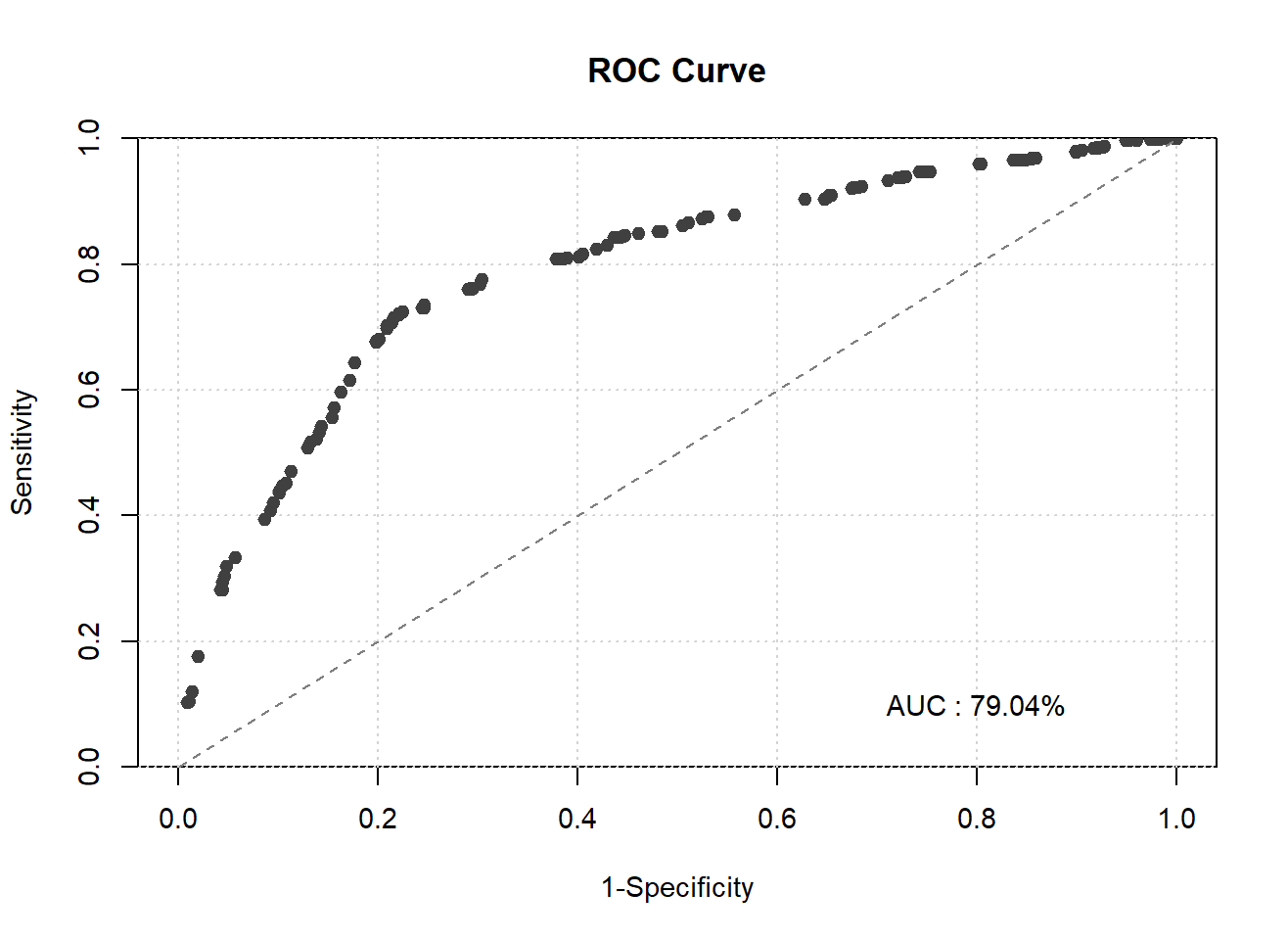
**Figure 7: ROC-AUC Curve from Initial Model on Accepted Testing Data**



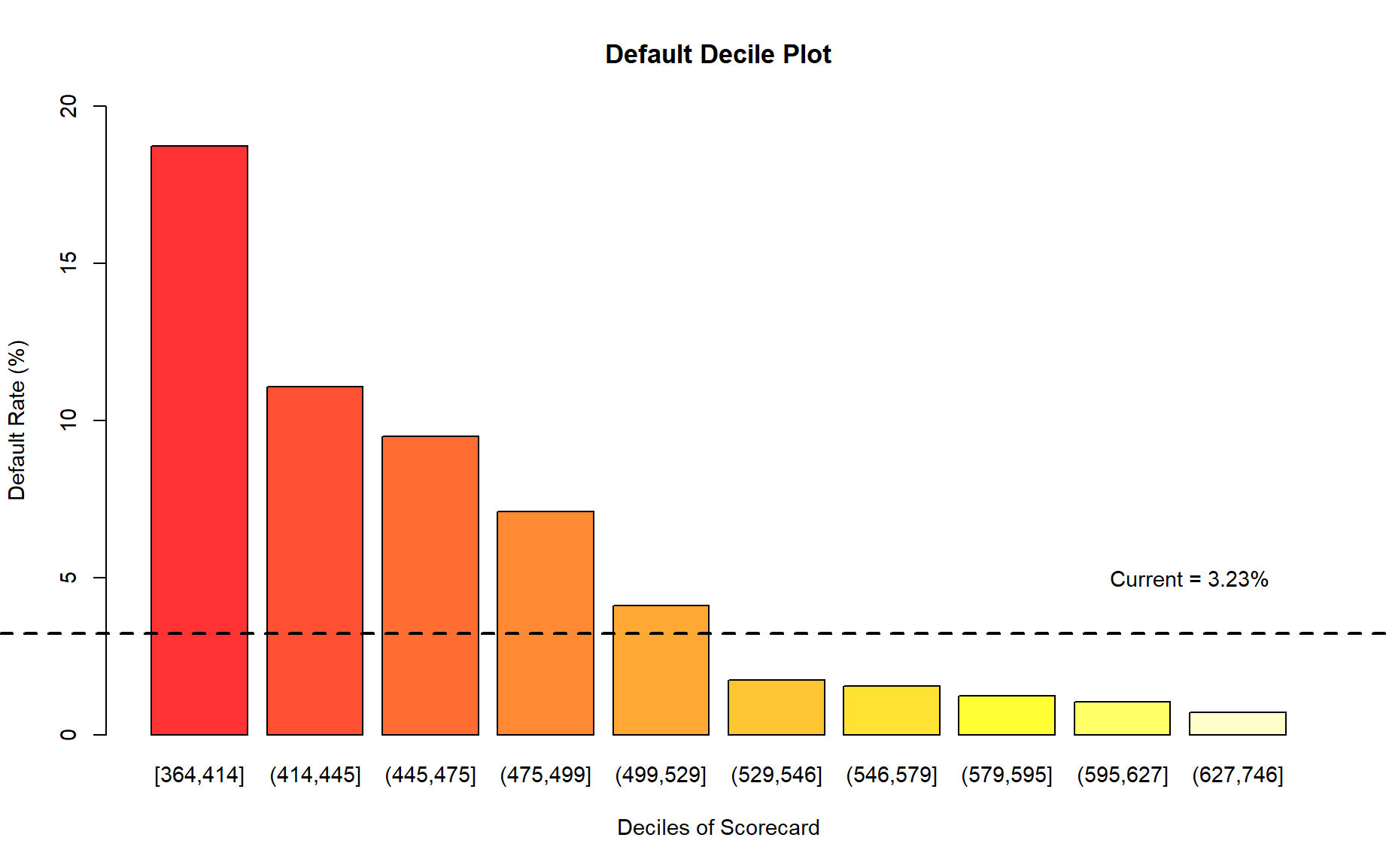
**Figure 8: Information Values for the Final Model**



**Figure 9: ROC-AUC Curve from Initial Model on Combined Training Data**



**Figure 10: ROC-AUC Curve from Initial Model on Combined Testing Data**



**Figure 11: Distribution to Associate Score Buckets**

**Table 1: Final Scorecard**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Level** | **Points** |
| Age | x ≤ 23 | 18 |
| Age | 23 < x ≤ 27 | 49 |
| Age | 27 < x ≤ 29 | 73 |
| Age | 29 < x ≤ 33 | 103 |
| Age | 33 < x ≤ 45 | 116 |
| Age | x > 45 | 153 |
| Credit Cards | Cheque Card | 149 |
| Credit Cards | MasterCard/Europe Card | 175 |
| Credit Cards | No Credit Cards | 66 |
| Credit Cards | Other Credit Card | 116 |
| Credit Cards | Visa Card | 19 |
| Number of Children | No Children | 95 |
| Number of Children | Children | 78 |
| EC Card Holder | No | 88 |
| EC Card Holder | Yes | 83 |
| Number of Persons in Household | x ≤ 1 | 58 |
| Number of Persons in Household | x > 1 | 104 |
| Time at Job | x ≤ 15 | 52 |
| Time at Job | 15 < x ≤ 84 | 86 |
| Time at Job | x > 84 | 127 |

**Homework Report Checklist**

The team member(s) responsible for checking each item should enter their initials in the field next to each question. All items should be addressed before submitting the assignment with the initial checklist attached.

**Sections & Structure - Nish & KK**

**Overview**

|  |  |
| --- | --- |
|  | Is the overview concise? |
|  | Does it provide context about the business problem? <Content> |
|  | Does it briefly address your team’s work, quantifiable results, and recommendations? <Action> |
|  | Does it offer audience-centered reasons for recommendations? <Context> |

**Body Sections**

|  |  |
| --- | --- |
|  | Does the report body include information on methods, analysis, quantifiable results, and  recommendations? |
|  | Is content grouped into appropriate sections (methodology, analysis, results, recommendations)? |

**Conclusion**

|  |  |
| --- | --- |
| NT | Does the report have a conclusion? |
| NT | Does the conclusion sum up the report and emphasize relevant takeaways? |

**Structure**

|  |  |
| --- | --- |
| NT | Does each major section have a heading? |
| NT | Are sections, subsections, and paragraphs organized logically for easy navigation? |

# 

**Visuals**

**Introduction, Discussion, and Captions**

|  |  |
| --- | --- |
| MJ | Is each visual introduced in the text before it appears? |
| MJ | Is each visual close to where it is introduced? |
| MJ | Does each visual include a title with the following information: type (table or figure), number, and a descriptive caption? |
| MJ | Is each visual discussed and interpreted in the text? |
| MJ | Are figures and tables numbered separately? |
| MJ | Are table captions above the table? Are figure captions below the figure? |

**Visual Design**

|  |  |
| --- | --- |
| MJ | Do figures/tables use audience-friendly labels rather than variable names? |
| MJ | Are the visuals easy to interpret? |
| MJ | Are the visuals appropriately sized? |
| MJ | Do tables appear on one page (not split between 2 pages)? |
| MJ | Are legends and axis labels included for figures? |
| MJ | Are numbers in tables right aligned? |
| MJ | Are the visuals designed well (ex: re-created in Word or Excel, not blurry or stretched,…)? |

**Document Design - Camille**

**Title Page Design**

|  |  |
| --- | --- |
|  | Does it include a descriptive title? |
|  | Does it state the team name, team members’ names, and the submission date? |

**Table of Contents Design**

|  |  |
| --- | --- |
|  | Does it list all the major sections of the report with corresponding page numbers? |
|  | Do the page numbers and sections in the Table of Contents match the report? |
|  |  |

**Document Design for Entire Report**

|  |  |
| --- | --- |
|  | Is a standard typeface (Calibri, Arial, etc.) used? |
|  | Is the size of the body text between 10-12 pt.? |
|  | Are headings and subheadings used to organize information? |
|  | Are distinctive text styles (bold, italic, etc.) used to distinguish between heading levels? |
|  | Are text styles for headings used consistently (ex: all level-one headings are bold)? |
|  | Are all paragraphs an appropriate length (fewer than 12 lines)? |
|  | Is white space used to indicate paragraph breaks? |
|  | Are bullet lists used for a series of items and numbered lists to show a hierarchy? |

# 

**Writing Style and Mechanics - Dillard**

**Spelling and Capitalization**

|  |  |
| --- | --- |
|  | Are spelling errors located and corrected? |
|  | Is spelling consistent throughout (no switching between acceptable spellings)? |
|  | Is capitalization used appropriately (proper nouns, etc.)? |
|  | Is capitalization of words consistent throughout the report? |

**Grammar and Punctuation**

|  |  |
| --- | --- |
|  | Are verb tenses used appropriately? |
|  | Are marks of punctuation used appropriately? |
|  | Is subject-verb agreement used in every sentence? |
|  | Is the grammar checker updated and are underlined grammar issues addressed? |

**Writing Style**

|  |  |
| --- | --- |
|  | Are all sentences in the report easy for your audience to understand quickly? |
|  | Are most sentences written in active voice? |
|  | Are idioms and vague words eliminated from the report? |
|  | Are acronyms introduced before being used? |
|  | Are well-written topic sentences included at the beginning of each paragraph? |
|  | Are lists parallel? |
|  | Is the appropriate point of view used when addressing your audience or describing team actions? |