**Breaking the Credit Barrier: Innovative Solutions for the First-Time Borrowers**

**Sebastian Gonzalez – 301336125**

**Centennial College**

**David Parent – Bilal Hasanzadah**

**Business Analytics Capstone**

**August 15, 2024**

**Table of Contents**

[**Executive Summary** 4](#_Toc174655724)

[**0.1.** **Executive Introduction** 4](#_Toc174655725)

[**0.2.** **Executive Objective** 4](#_Toc174655726)

[**0.3.** **Executive Model Description** 4](#_Toc174655727)

[**0.4. Executive Recommendations** 4](#_Toc174655728)

[**Introduction** 5](#_Toc174655729)

[**1.0.** **Background** 5](#_Toc174655730)

[**1.1. Problem Statement** 5](#_Toc174655731)

[**1.3. Objectives & Measurement** 6](#_Toc174655732)

[**1.4. Assumptions and Limitations** 6](#_Toc174655733)

[**Data Sources** 6](#_Toc174655734)

[**2.0. Data Set Introduction** 7](#_Toc174655735)

[**2.1. Exclusions** 8](#_Toc174655736)

[**3.0. Data Exploration Techniques** 9](#_Toc174655737)

[**3.1. Summary** 10](#_Toc174655738)

[**Data Preparation and Feature Engineering** 10](#_Toc174655739)

[**4.0. SMOTE** 10](#_Toc174655740)

[**4.1. Python Packages** 10](#_Toc174655741)

[**4.2. Data Split, Predictors and Outcomes** 11](#_Toc174655742)

[**Model Exploration** 11](#_Toc174655743)

[**5.0 Full Decision Tree** 12](#_Toc174655744)

[**5.0.1 Full Decision Tree Results** 12](#_Toc174655745)

[**5.1. Random Forest** 13](#_Toc174655746)

[**5.1.1 Random Forest Results** 14](#_Toc174655747)

[**5.2. Logistic Regression** 15](#_Toc174655748)

[**5.2.1 Logistic Regression Results** 15](#_Toc174655749)

[**5.3. Optional** 18](#_Toc174655750)

[**5.4. Model Comparison** 18](#_Toc174655751)

[**Model Recommendation** 19](#_Toc174655752)

[**6.0 Model Selection** 19](#_Toc174655753)

[**6.1 Model Theory** 19](#_Toc174655754)

[**6.2 Model Assumptions and Limitations** 20](#_Toc174655755)

[**6.3 Model Sensitivity to Key Drivers** 20](#_Toc174655756)

[**6.4 Additional Models to Address Business Objectives** 20](#_Toc174655757)

[**Conclusion and Recommendations** 21](#_Toc174655758)

[**7.0. Impacts on Business Problem** 21](#_Toc174655759)

[**7.1. Recommended Next Steps** 21](#_Toc174655760)

[**References** 23](#_Toc174655761)

# **Executive Summary**

## **Executive Introduction**

Those with no credit history at all, such as recent immigrants, have a difficult time getting credit cards and loans. While credit behavior was the main emphasis of the previous approval models, which are still relevant, these groups were frequently left out, which made it harder for them to establish credit. By addressing this issue, banks can expand their market and keep prospective customers from being financially excluded.

## **Executive Objective**

The main objective of this investigation is to identify the key variables that influence credit card approval decisions. Leveraging these insights, the goal is to develop strategies that can increase the approval rates for customers who do not have a strong credit history or have no credit history at all. The project aims for financial inclusion and expand access to credit for populations excluded.

## **Executive Model Description**

Multiple modeling approaches were employed in the analysis to determine the best fit for predicting the approval of credit card requests. After using analytics metrics like F1 score, ROC, and confusion matrix the model chosen for the investigation was logistic regression. This model provides clear, interpretable insights into the variables that influence credit card approval decisions and allows us to develop targeted strategies to improve approval rates for customers with limited or no credit history.

## 

## **0.4. Executive Recommendations**

The first recommendation for this project is to develop alternative methods that consider other factors than credit history. One possible approach is one model that incorporates payment history for rent, utilities, and other regular expenses. Implementing an alternative scoring model will enable the inclusion of diverse client segments and promote equitable access to credit.

Second, implement educational programs focused on credit scores for individuals with limited or no credit history. Providing guidance on improving credit profiles will enhance financial literacy and promote higher financial inclusion.

Third, develop a risk-based pricing model that adjusts interest rates and terms according to the credit risk profile of applicants. This approach offers flexibility in credit terms based on the client's overall financial situation, breaking barriers of financial isolation.

# **Introduction**

## **Background**

Credit Score has been used as a metric to approve or decline credit cards. This metric does not capture some of the financial information of a potential credit card client. This creates barriers for clients who do not have a credit score.

Also, there are newcomers into the country or into the credit score system who do not even have a credit history. These potential clients will face financial isolation. The purpose of this investigation is to validate that assumption and based on the results create strategies to reduce financial isolation.

## 

## **1.1. Problem Statement**

Persons with limited or no credit history struggle to get loans and credit cards, which creates barriers for those new to the credit system, including recent immigrants, which makes it more difficult to build and create their credit. Most financial institutions primarily focus on previous credit behavior to approve or decline credit requests.

## **1.3. Objectives & Measurement**

The main objective of this project is to identify key variables influencing credit card approval decisions to develop strategies that increase approval rates for customers with limited or no credit history, promoting financial inclusion and expanding access to credit.

Also, the problem will be approached by different models to identify which is the best for the project. To establish which is the best approach for the project these are the measures used to make the decision: Accuracy, F1 Score, and ROC-AUC.

## **1.4. Assumptions and Limitations**

This project does not make any assumptions about the dataset, but it does face a limitation regarding the sample. The dataset is unbalanced in the target variable, 'Status,' which has more approved values than declined ones. To address this imbalance, the SMOTE technique will be applied to achieve a better balance between approved and declined values.

# 

# **Data Sources**

The dataset used in this analysis was taken from Kaggle. This dataset contains 25128 clients from a bank who applied for a credit card. Each row has information about the customer like personal information, income, if it owns a car or property, and the credit card approval status, which will be explained in detail in the next section.

## **2.0. Data Set Introduction**

This dataset has information on the customers who applied for a credit card and includes the status of the application. The source did not mention which bank or institution was taken from. Below you will find a table describing the columns of the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Role** | **Level** |
| Applicant\_ID | ID number for each applicant | Rejected | Nominal |
| Applicant\_Gender | Gender | Rejected | Nominal |
| Owned\_Car | If the applicant owns a car | Input | Nominal |
| Owned\_Realty | If the applicant owns a property | Input | Nominal |
| Total\_Children | Number of children | Input | Nominal |
| Total\_Income | Total income | Input | Nominal |
| Income\_Type | Source of income | Input | Nominal |
| Education\_Type | Level of education | Input | Nominal |
| Family\_Status | Status of the marital situation | Rejected | Nominal |
| Housing\_Type | Type of house | Input | Nominal |
| Owned\_Mobile\_Phone | If the applicant has a mobile phone | Input | Nominal |
| Owned\_Work\_Phone | If the applicant has a work phone | Input | Nominal |
| Owned\_Phone | If the applicant owns a phone | Input | Nominal |
| Owned\_Email | if the applicant has an email | Input | Nominal |
| Job\_Title | The role that the applicant has in his job | Input | Nominal |
| Total\_Family\_Members | Number of members in his family | Input | Nominal |
| Applicant\_Age | Age | Rejected | Nominal |
| Years\_of\_Working | Number of years working | Input | Nominal |
| Total\_Bad\_Debt | Number of days of bad debt | Input | Nominal |
| Total\_Good\_Debt | Number of days of good debt | Input | Nominal |
| Status | Status of the credit card approval | Target | Nominal |

## **2.1. Exclusions**

The table above contains all the columns in the dataset. Additionally, two columns in the table show the variables role and type of variable. The rejected variables for this model can add biases to the analysis. In this case, it was determined that Applicant\_ID, Applicant\_Gender, Family\_Status, and Applicant\_Age were rejected as these variables are commonly removed from this type of analysis as they can create biases.

**2.2. Initial Data Cleansing or Preparation**

After choosing the variables that would make up part of our analysis. Analytics techniques were applied to prepare the dataset for Modeling. First, the dataset went through a null or NA check. There is no missing value for this dataset.

Second, a validation for unique values was made. The purpose of this was to understand the dimensions of the dataset. Some variables have many different values that can impact our analysis. In this case, it was decided to reduce the number of values for these variables. For the housing type variable were 5 different types of unique values. Some of these values were related. After transformation, the only two values for these variables are House and Apartment.

Also, a reduction was made for the Education Type as it was a value called ‘Secondary/secondary special’. This value was converted to ‘Lower Secondary’.

Finally, the type of variable was changed. As some categorical variables were assigned a different category. For this investigation, the categorical variables are Income Type, Education Type, Housing Type, and Job Title.

**Data Exploration**

As discussed in the previous sections, the dataset underwent several transformations. Following these adjustments, we began the data exploration phase, focusing on identifying patterns that could influence our analysis. The techniques employed for this exploration are detailed in the next section.

## **3.0. Data Exploration Techniques**

The data exploration techniques used for this analysis are correlation matrix and box plot charts to identify possible outliers. After running the correlation matrix, we found all clients owned a mobile phone, so this variable was discarded from the analysis. The box plot approach was used for income which no transformation was made as modeling will be rung first. If income is a potential variable to explain our model transformation this will be applied.

## **3.1. Summary**

The main changes in this section were the removal of the Owned Mobile variables and the outliers of the Total Income variable that will be fixed depending on the modeling results.

# **Data Preparation and Feature Engineering**

The dataset was prepared in this phase by dealing with outliers, filling in missing values, and transforming data types. The SMOTE technique was used to rectify the imbalance in the target variable ‘Status’ guaranteeing a balanced dataset for model training. These steps were essential to improve the dataset's quality.

### **4.0. SMOTE**

As mentioned above, a SMOTE (Synthetic Minority Over-sampling Technique) process was performed to address the data imbalance in the ‘Status’ variable. This technique involves creating synthetic samples for the underrepresented class, thereby balancing the dataset and enhancing model performance. To ensure consistency and reproducibility in the results, a random state of 1 was selected during the SMOTE process. This approach was essential for achieving more reliable and accurate predictions in our analysis.

## 

## **4.1. Python Packages**

In this section, I want to mention the packages required to run our project. The picture below includes all the packages.

A screenshot of a computer program

Description automatically generated

### 

### **4.2. Data Split, Predictors and Outcomes**

For our data split it was chosen a split of 60% for training and 40% for testing. Also, our outcome is status of the credit card approval, and our predictors are all variables after creating the dummies for the categorical variables.

**A screenshot of a computer program

Description automatically generated**

# **Model Exploration**

In our project, we explored three main different modeling approaches: Random Forest, Full Tree, and Logistic Regression. Every model was used to determine its effectiveness in predicting approval outcomes. The detailed results and comparative performance of these models will be presented in the following sections.

## **5.0 Full Decision Tree**

Decision Trees are an important model in our project because they can be used to visualize the main variables that impact credit card approval. They provide intuitive insights breaking down complex decision criteria into a series of simple binary choices. This is crucial for understanding the model’s behavior and explaining decisions to stakeholders. As it can be visualized the breakdown it helps to make decisions easier for strategies and allow us to interpret where the strategies should be focused.

### **5.0.1 Full Decision Tree Results**

From the full decision tree plot, we can observe that bad debt and good debt significantly influence the results. It is only after the third node breakdown that different variables start to emerge. This model supports our main hypothesis, demonstrating that individuals with no credit history or a poor credit history are at risk of financial isolation. Also, look at Confusion Matrix below for additional information on the accuracy.

In the picture below are the results after running the full decision tree.

**A diagram of a family tree

Description automatically generated**

A white background with black text

Description automatically generated

## **5.1. Random Forest**

Random Forest is part of the project as it is robust and versatile in handling complex datasets. This model builds multiple decision trees during training and adds their predictions to improve accuracy and control overfitting. This method is very useful because it can handle big datasets as this one which has entries and more than 40 variables after doing the dummies process. By incorporating Random Forest, it was expected to get a higher score than the Full Decision Tree.

### **5.1.1 Random Forest Results**

The Random Forest Results are focused on feature importance and standard deviation. The table below shows the importance of each feature. Also, below the table is the chart that shows the variables by importance in this model. Both show that the most significant variables of the model are Total Bad Debt and Total Good Debt with 5.10 and 2.25 correspondingly. This model also confirms that other variables do not have an important role in the approval for a credit card.

A screenshot of a computer

Description automatically generated

A graph with text overlay

Description automatically generated

## **5.2. Logistic Regression**

Logistic regression is recommended due to the effectiveness of binary classification tasks, where the objective is to predict the probability of the categorical outcome. In this case, approved or declined. The logistic regression shows the relationship between the predictor variables and the likelihood of approval. These results help us to create strategies based on the most important variables. In our analysis, we include the coefficients and odds ratio for each variable.

### **5.2.1 Logistic Regression Results**

Logistic regression coefficients and odds ratios were calculated after running the model. Below you will find the results. Our analysis is going to focus on the main variables that drive credit card approval.

**Positive influences for Approval**

Having a good debt increases your chances to get approval for a credit card. This might seem as obvious, but our main point is that clients without a good debt or low credit history will face a decline in their request.

Regarding the other variables that impact positively the credit card approval, we will only analyze the top 5.

* For every unit increase in good debt, the odds of credit card approval increase by 2.42 times.
* For every unit increase in total children, the odds of credit card approval increase by 1.75 times.
* For every unit increase in owned phone, the odds of credit card approval increase by 1.56 times.
* For every unit increase in owned car, the odds of credit card approval increase by 1.56 times.
* For every unit increase in higher education, the odds of credit card approval increase by 1.46 times.

**Negative influences for Approval**

The variable with the most significant impact on credit card approval is Bad Debt. I would like to highlight the top 3 variables that impact credit card approval.

* The coefficient for Total Bad Debt of -2.7105 and the odds ratio of 0.066 demonstrate that the presence of bad debt drastically reduces credit card approval by 93.3% (1 - 0.066 = 0.933)
* The coefficient for Total Family Members of -0.5771 and the odds ratio of 0.5645 demonstrate that the presence of bad debt drastically reduces credit card approval by 43.5% (1 – 0.5645 = 0.435)
* The coefficient for Owned Realty of -0.2951 and the odds ratio of 0.7441 demonstrate that the presence of bad debt drastically reduces credit card approval by 25.6% (1 - 0.7441 = 0.256)



The chart below shows the variable's coefficients and odds ratio. This picture shows the significant impact that have Total Good Debt and Total Bad Debt in the model. While other variables remain in the center between the ranges of [1,-1] for axis x and [2,0] for axis Y. This chart shows how clients can face rejection based on these only two variables.

**A graph with blue dots

Description automatically generated**

### **5.3. Optional**

In the modeling section of the code will be other models such as backward, forward, stepwise regression, and linear regression. These models were included in that section to find out if there were potential insights from them. Unfortunately, the best models for this investigation are Full Decision Tree, Random Forest, and Logistic Regression.

## 

## **5.4. Model Comparison**

For our model comparison, we used three different metrics. In the picture below is the comparison of the three models using ROC-AUC, F1 Score, and Precision. Using these scores, we determine the best model for our analysis is logistic regression.

A black and white text with numbers

Description automatically generated

# **Model Recommendation**

Logistic Regression was selected as the model for the analysis due to its effectiveness in handling binary classification. This model provides a clear interpretation through odds ratios and coefficients. Also, ensures a robust performance by identifying the variables that influence the credit decisions. That is the reason this model is ideal for developing strategies.

## **6.0 Model Selection**

As mentioned in the section above. The Logistic Regression model was chosen because has the highest score in the three metrics. This model helps us to understand better the situation for approval requests and create a strategy that addresses this issue.

## **6.1 Model Theory**

Credit card approval is typically based on traditional credit history models, which can disadvantage or reject applicants with poor or non-existent credit histories. While it may seem logical to deny credit to those without a strong credit history, some applicants with low scores may possess other financial characteristics that demonstrate their creditworthiness. As Equifax explains on their website, credit scores are primarily calculated based on credit history alone. However, this approach overlooks other critical factors that could provide a more comprehensive view of an individual's financial stability and ability to manage credit.

### **6.2 Model Assumptions and Limitations**

The model only has one limitation, that is the target variable is unbalanced, that is the reason we deployed a SMOTE on our data to have a more balanced number of values. After doing the SMOTE the results remained the same.

## **6.3 Model Sensitivity to Key Drivers**

The sensitivity of the model to key drivers is high. In this case, is considered as good as we want to focus on Total Good Debt and Total Bad Debt as these two variables determine the approval of a credit card. The main goal of this analysis is to decrease the significance of these variables in the model through our suggestions and business plan.

### 

### **6.4 Additional Models to Address Business Objectives**

The Random Forest model could be used as a secondary measure for the analysis of Total Good Debt and Total Bad Debt. As this model also determined these two variables play an important role in credit card approvals. This model can be another measure to decrease the importance of these variables, and a main goal could be to reduce the standard deviation in this model through the strategies proposed.

# **Conclusion and Recommendations**

The model proposed gives good insights for the investigation. Logistic Regression was able to prove clients can face financial issues for credit card approval if they have a credit card bad history or do not have a credit history at all. The main goal of this project is to reduce the financial isolation this problem can bring.

Another important conclusion of this project is that the other variables have a limited impact on credit decisions. In the model section of this project, we visualize the Coefficients vs Odds Ratio chart an we were able to see the gap that exists between these variables and the main key drives of credit card approvals.

## **7.0. Impacts on Business Problem**

The impact of the business is that the industry is losing some potential clients for using a rigid credit score method. This can impact on the business significantly. It is important to clarify that not all these clients deserve credit card approval. The focus of this project is to determine a strategy that allows both sides to win. In the case of the clients, get a credit card approval, and for the banks increase clients and growth.

## **7.1. Recommended Next Steps**

Establish an alternative credit score model. This will allow to get the financial information of the client to get a better picture of their creditworthiness. The current score method does not include payment history for rent, utilities, and other regular costs that can benefit the client when applying for a credit card or a loan.

Additionally, is important to educate clients about credit scores. This can help clients to improve their scores, understanding the ways they can leverage their scores. Also, for newcomers that do not have a good idea of this measure. They must receive the right information, so they can apply advice to their financial planning.

Finally, expand the portfolio for this type of client. Creating risk-based pricing and customized terms can help the clients to get access to financial products and avoid being isolated by financial institutions. These products can be dynamic as the client improves their scores.

# **References**

Caesar, M. (2022). *Kaggle - Credit Card Approval Prediction.* Retrieved from Kaggle: https://www.kaggle.com/datasets/caesarmario/application-data

Equifax. (n.d.). *Equifax Personal.* Retrieved from Equifax: https://www.equifax.com/personal/education/credit/score/articles/-/learn/what-is-a-credit-score/

TransUnion. (n.d.). *TransUnion Services.* Retrieved from TransUnion: https://www.transunion.com/credit-score