Retail Banking Customer Default Probability

Abstract

Background: This research focuses on predicting credit card default, specifically examining instances of more than 90 days of past due. Utilizing Exploratory Data Analysis (EDA) and Machine Learning in Python, complemented by graphical representations in Python and Tableau.

Objectives: Identify key factors influencing the probability of default, such as age, debt ratio, monthly income, open credit lines, dependents, and past due occurrences in the past 2 years.

Methods: Employing a Probit model, achieving a 93% accuracy, and an AUC-ROC curve of 0.7, indicating good explanatory power and potential adverse selection of credit card issues up to 7%.

Results: Significant variables include age, debt ratio, monthly income, open credit lines, dependents, and past due occurrences. The Probit model demonstrates a high accuracy of 93%, with the AUC-ROC curve confirming its robust explanatory power.

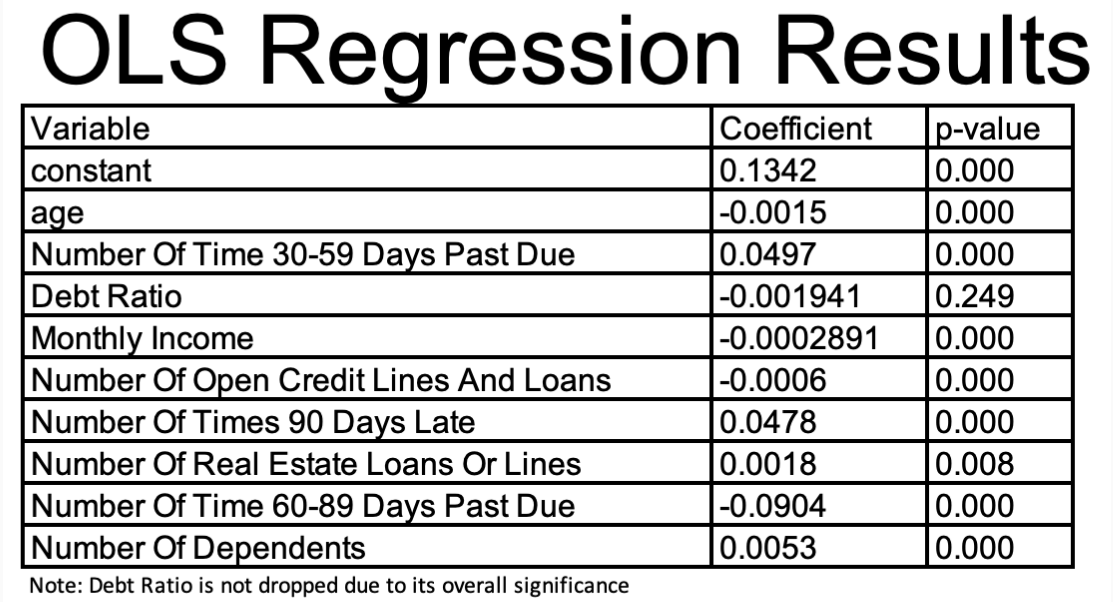
Conclusion: This research highlights demographic and financial factors crucial for predicting credit card default. The developed Probit model offers a reliable tool, with a 93% accuracy, potentially reducing adverse selection in credit card issuances.

Data

The dataset utilized in this research is the "Give Me Some Credit" dataset, accessible on Kaggle through the following link: https://www.kaggle.com/c/GiveMeSomeCredit/data. This dataset is a comprehensive compilation of credit-related information, offering a rich source for investigating the probability of credit card default. It encompasses various features crucial for predictive modeling, including demographic factors such as age and number of dependents, financial indicators like monthly income and debt ratio, and historical credit behavior, notably past due occurrences within the past 2 years. Leveraging this dataset, the research employs Exploratory Data Analysis (EDA) and machine learning algorithms to uncover patterns and develop a Probit model, contributing to a comprehensive understanding of the factors influencing credit card default probabilities.

Methodology & Results

The initial phase involved meticulous data cleaning, beginning with the removal of missing values to ensure dataset integrity. Subsequently, an Ordinary Least Squares (OLS) regression was implemented, targeting variables with a significance level below 95% (p-value < 0.050). This step aimed to refine the dataset by eliminating statistically insignificant variables.

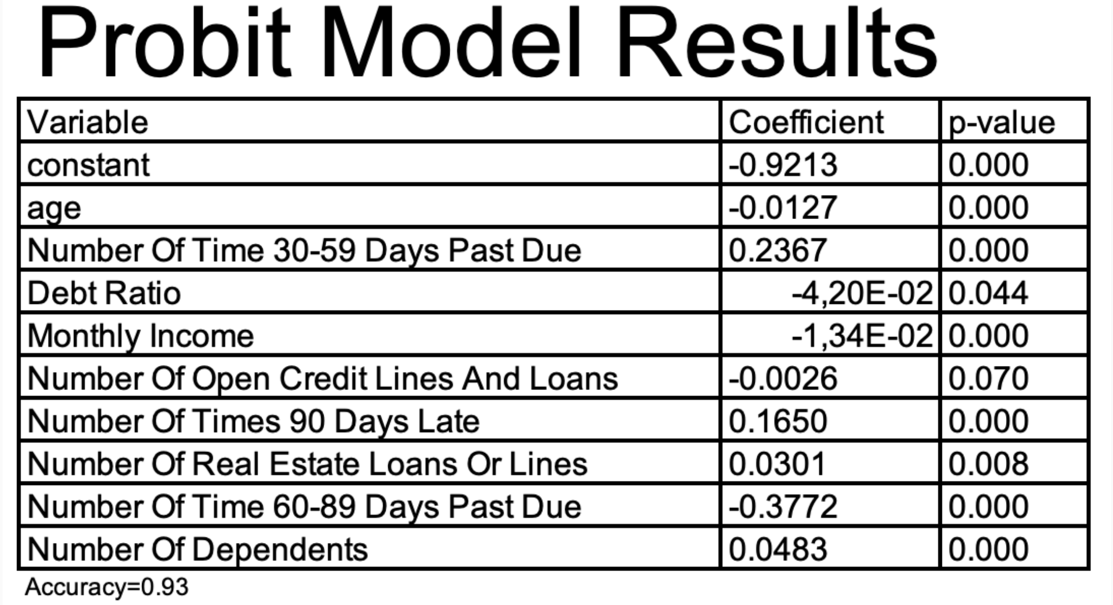


OLS Regression Results

Variable	Coefficient	p-value
constant	0.1342	0.000
age	-0.0015	0.000
Number Of Time 30-59 Days Past Due	0.0497	0.000
Debt Ratio	-0.001941	0.249
Monthly Income	-0.0002891	0.000
Number Of Open Credit Lines And Loans	-0.0006	0.000
Number Of Times 90 Days Late	0.0478	0.000
Number Of Real Estate Loans Or Lines	0.0018	0.008
Number Of Time 60-89 Days Past Due	-0.0904	0.000
Number Of Dependents	0.0053	0.000

Note: Debt Ratio is not dropped due to its overall significance

Following the data cleaning process, the refined dataset, free of dropped values, was employed to develop the Probit model. A test size of 20% (test_size = 0.2) was chosen to partition the data, considering the substantial size of the dataset, allowing for a robust evaluation with a sufficient number of observations.

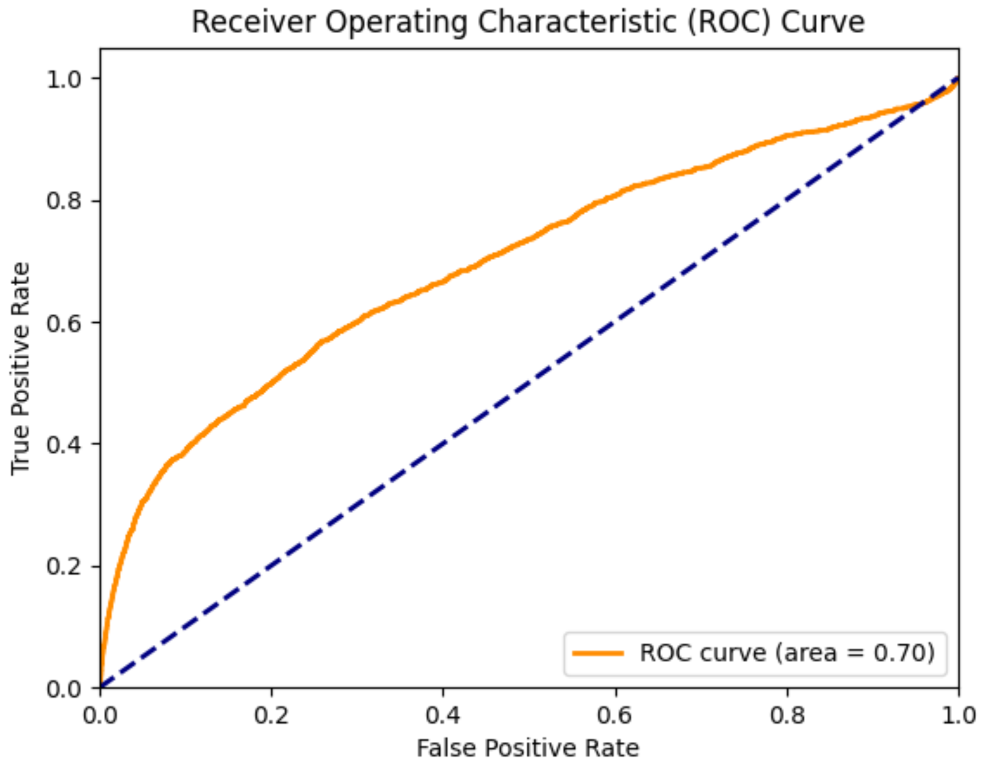


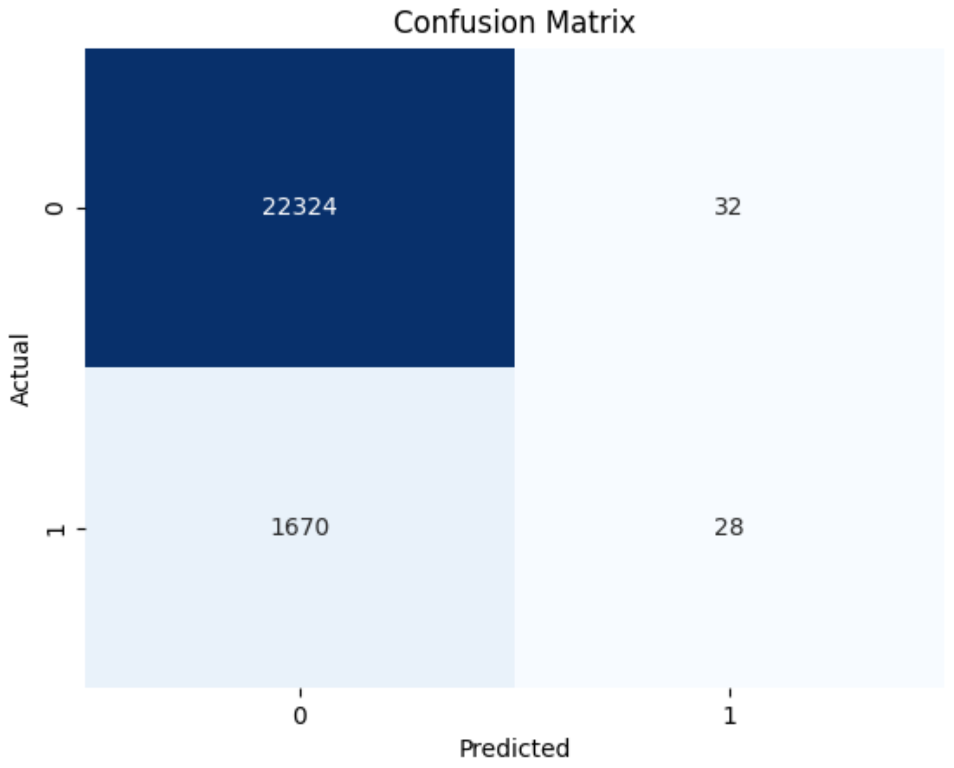
Probit Model Results

Variable	Coefficient	p-value
constant	-0.9213	0.000
age	-0.0127	0.000
Number Of Time 30-59 Days Past Due	0.2367	0.000
Debt Ratio	-4,20E-02	0.044
Monthly Income	-1,34E-02	0.000
Number Of Open Credit Lines And Loans	-0.0026	0.070
Number Of Times 90 Days Late	0.1650	0.000
Number Of Real Estate Loans Or Lines	0.0301	0.008
Number Of Time 60-89 Days Past Due	-0.3772	0.000
Number Of Dependents	0.0483	0.000

Accuracy=0.93

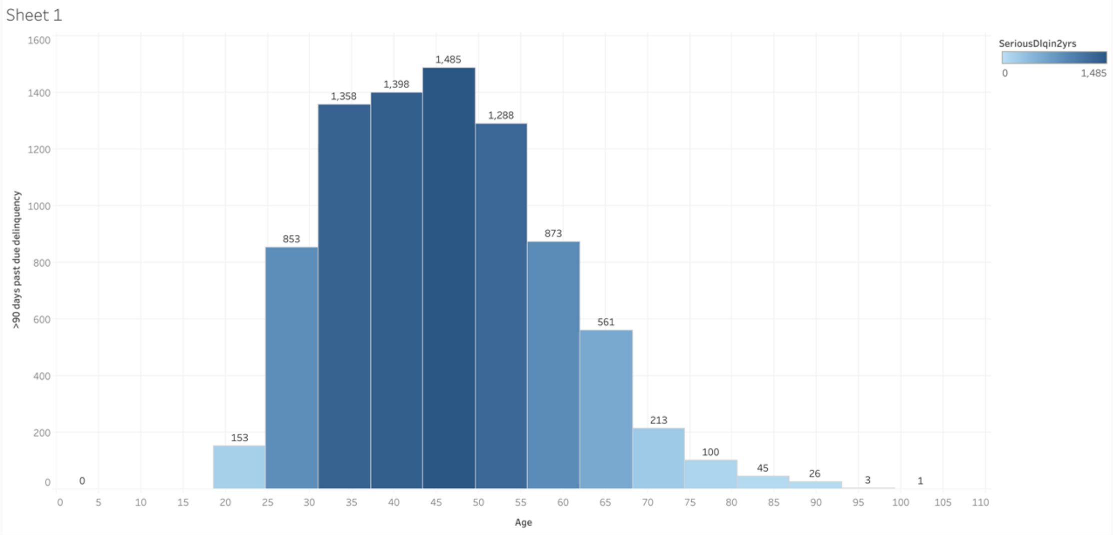
The evaluation of the Probit model yielded a Receiver Operating Characteristic (ROC) curve with an Area Under the Curve (AUC) of 0.7. The associated confusion matrix revealed that the 7% adverse selection is distributed as follows: 0.133% false positives, indicating instances where default was predicted but did not occur, and 6.943% false negatives, signifying cases where default occurred but was not predicted. These results provide insights into the predictive performance of the model and the potential occurrence of adverse selection in credit card issuances.



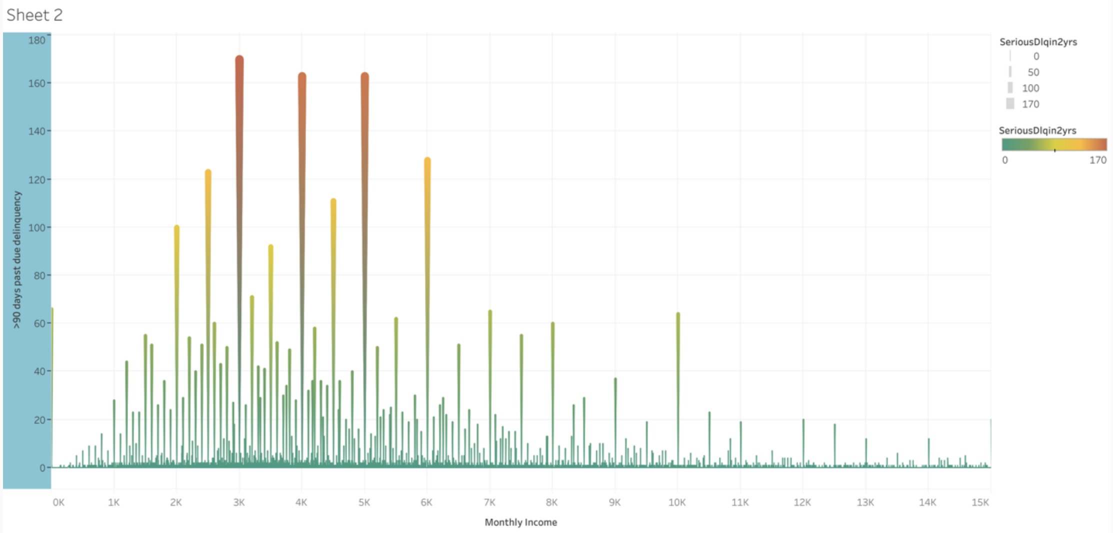


Decision - Conclusion

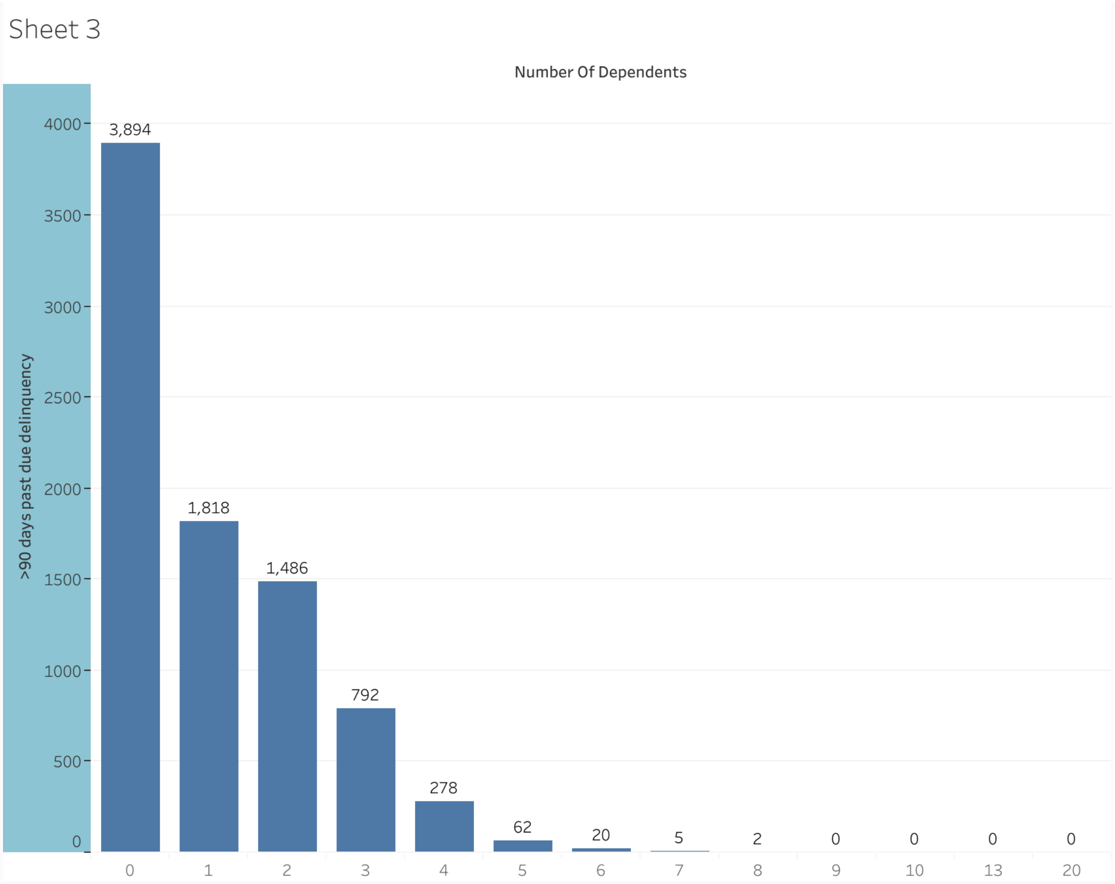
In conclusion, this study leverages the "Give Me Some Credit" dataset to explore the probability of credit card default, employing rigorous data cleaning and a Probit model. Through the elimination of missing values and statistically insignificant variables via an Ordinary Least Squares (OLS) regression, a refined dataset was prepared. The Probit model, developed on this cleaned dataset, exhibited an impressive 93% accuracy, underscoring its efficacy in predicting credit card default probabilities. The associated Receiver Operating Characteristic (ROC) curve revealed an Area Under the Curve (AUC) of 0.7, affirming the model's robust explanatory power. Adverse selection, quantified at 7%, was distributed as 0.133% false positives and 6.943% false negatives, emphasizing the model's potential in mitigating adverse selection in credit card issuances. These findings contribute valuable insights into credit risk assessment, offering a reliable tool for informed decision-making in the financial sector.



Those aged between 30 and 50 years old present higher probability of default.



There is higher probability of default for those who have a monthly income of 3.000 - 5.000$



Surprisingly, those without dependents present higher probability of default.