Credit risk modelling and Business implications: A case study of American Express

By

Temidayo Olowoyeye

Data and Business Analyst

Terms of Use

This credit risk analysis was prepared by Temidayo Olowoyeye, herein referred to as the "Author", to showcase his skills. The following terms and conditions apply to the use and interpretation of this document:

- Confidentiality: The information contained in this analysis is based on data provided on Kaggle.com. It demonstrates the author's analytical and reporting capabilities and does not disclose proprietary or confidential information.
- Opinions and Interpretations: All statements, findings, and views expressed in this document are solely those of the author and do not represent the views or opinions of American Express or any individual. The author takes no responsibility for the accuracy or applicability of the content to any actual business entity.
- Fictional Nature of Data: Any data, figures, or statistics presented in this report are fictional and created to illustrate data analysis and reporting techniques. They do not reflect any company or organisation's financial or operational performance.
- This document demonstrates the author's skills and is not intended to make business decisions or develop strategies for any company. It is important to note that it should not be relied upon for practical purposes.
- No Liability Assumed: The author assumes no liability for any consequences arising from the use of this report. The document is provided without warranties or guarantees of accuracy, completeness, or reliability.
- Unauthorised Use: This analysis is the author's intellectual property and is intended for personal use and demonstration only. Unauthorised reproduction, distribution, or use of any part of this document is strictly prohibited.
- Consulting the author or other qualified professionals is strongly recommended for business decisions or actions. Obtaining relevant, accurate, and up-to-date information from authoritative sources is essential for the responsible use of this document.

You agree to abide by these terms and conditions by accessing and using this document. If you do not agree with any part of these terms, you are not authorised to use or rely on the information presented in this report.

Supplementary Materials

- Data Sources: The data utilised in this analysis was sourced from Kaggle.com and is accessible via link.
- Python Script: The analysis was conducted using Python, and the corresponding code can be accessed here.

Conflicts of Interest

- The author declares no conflicts of interest.

Table of Contents

1. Introduction

1.1. Overview

In the Financial sector, granting credit and loans to customers has long been a fundamental practice. As time progresses, human needs continue to grow, illustrating the economic principle that human desires are insatiable. The increasing demand for financial resources presents an opportunity for the financial industry to address these challenges by providing credit to their customers. However, this opportunity is accompanied by significant challenges, particularly the discrepancy between customers' willingness to request credit and the willingness to repay.

Financial institutions have developed various strategies to address this behavioural difference to ensure timely and appropriate loan recovery. One practical approach involves using statistical and machine learning algorithms to predict customers' repayment behaviour. Analysing specific customer information can provide valuable insights into the likelihood of loan repayment, thus helping to mitigate potential losses and inform decision-making processes.

This case study applies predictive modeling methods to American Express (Amex) credit card data to evaluate their effectiveness in managing credit risk.

1.2. Business Problem

American Express (Amex) faces significant business challenges within its credit and loan department. A primary concern is approving credit requests while retaining its customer base and maintaining profitability. The organisation must ensure that its business priorities align with its decision-making processes.

To address this, Amex has gathered historical customer data to gain insights into credit risk. They aim to identify which customers are likely to default on their credit, determine the percentage of their customer base that falls into this high-risk category, and estimate the potential impact on profitability.

1.3. Objective

The primary objective of this study is to leverage historical customer data to enhance the credit risk assessment process at American Express (Amex). Specifically, the analysis aims to:

1. Develop a predictive model to identify customers likely not to default on their credit.
2. Determine the percentage of the customer base that falls into the high-risk category.
3. Assess the impact of these predictions on business operations and profitability.

2. Data Management

2.1. Data Source and Structure

The data for this analysis was sourced from Kaggle.com and is provided in two distinct datasets:

1. Training Dataset: This dataset, which contains 45,528 rows and 19 columns, is used for model development and will train the predictive model. It includes various features relevant to customers' credit and loan behaviours.
2. Testing Dataset: This dataset is used for decision-making, applying the developed model to real-world business scenarios. It contains 11,383 rows and 18 columns, omitting the target variable to evaluate the model's performance and effectiveness in predicting defaulting and non-defaulting customers.

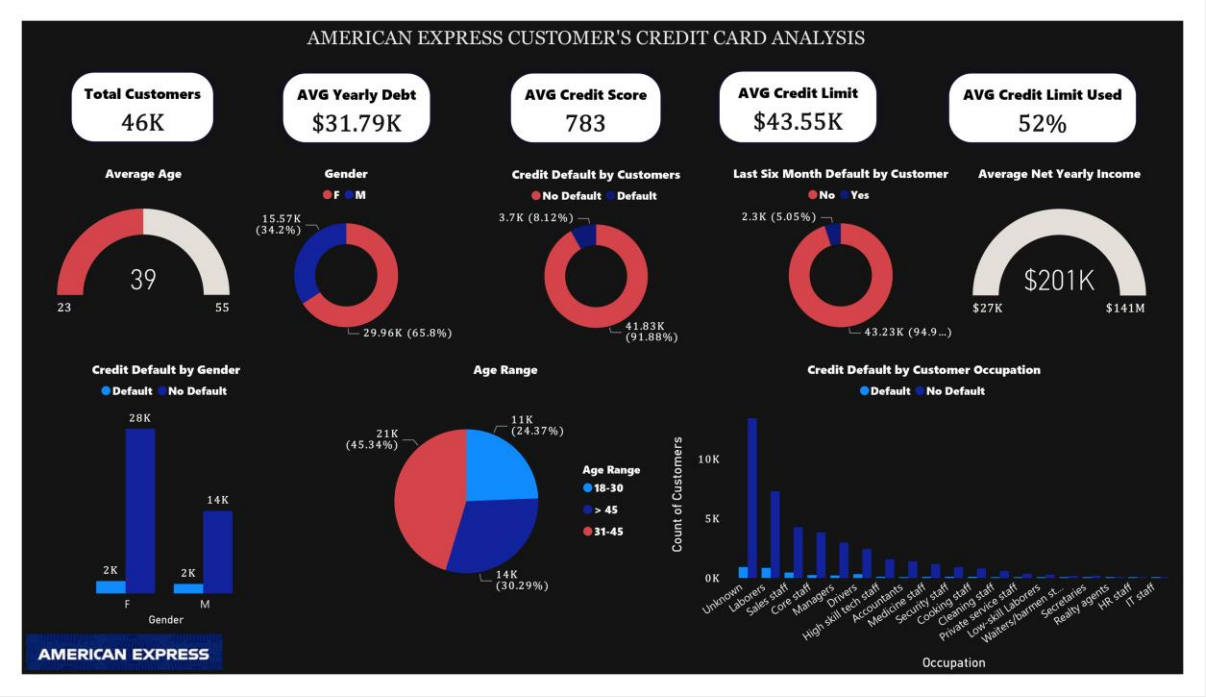
2.2. Data Quality Check and Processing

Ensuring the dataset meets the proper standard and quality is significant to effective data modeling. Several practices were undertaken to assess and improve the quality of the data, including:

- Feature Identification: The dataset was examined to identify categorical and numerical features essential for selecting appropriate modeling techniques and preprocessing steps.
- Missing Values Handling: The dataset was checked for missing values representing less than 1% (0.24%) of the data. Missing values were imputed using the following methods:
    - Categorical features: Imputed with the mode (most frequent value) to maintain the distribution of categorical variables.
    - Numerical features: Imputed with the median to mitigate the impact of outliers and preserve the central tendency of the data.
    - Substitution Method: Some numerical features were used to fill in missing values for other numerical features based on their similarity or relationship. For example, missing values for the number of days employed were filled with the corresponding occupational type.
- Error Value Replacement: To ensure data integrity and accuracy, further modifications were made to replace error values in the dataset.
- Data Type Conversion: Columns initially stored as float were converted to integers to streamline data representation and improve computational efficiency.

2.3. EDA

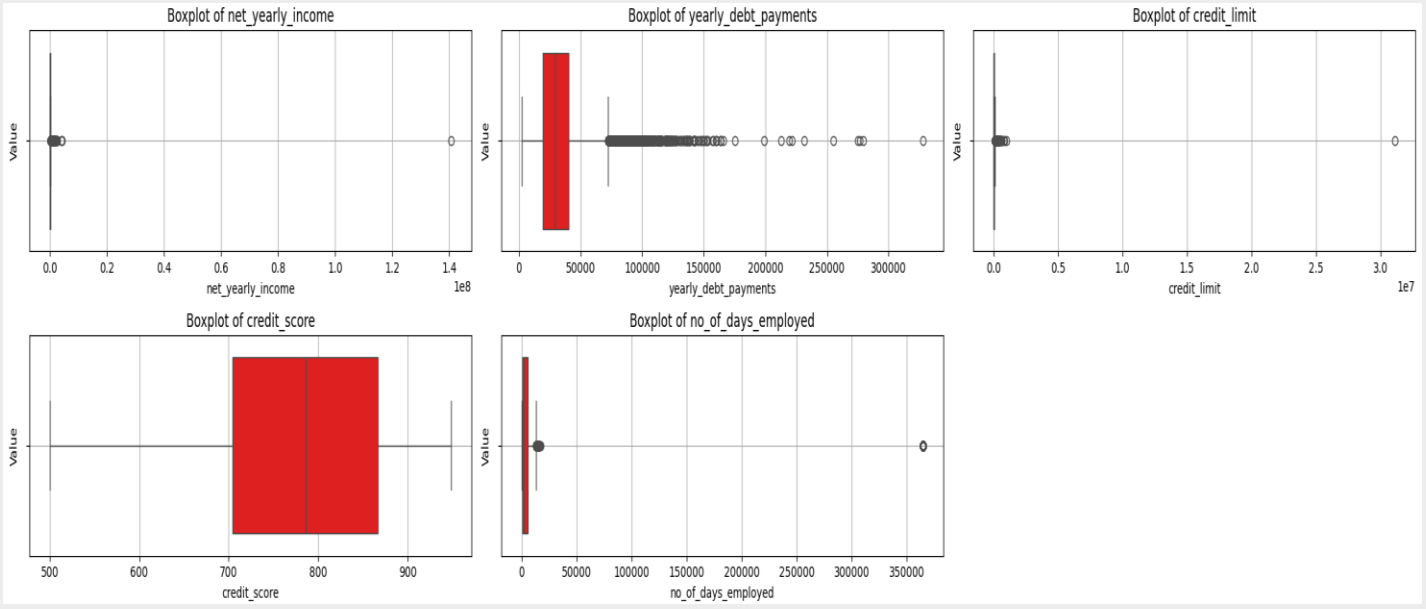
Exploratory Data Analysis (EDA) was carried out using PowerBI. Key insights on the customer's information were generated, as shown in the image below; the insight can be accessed via this link.



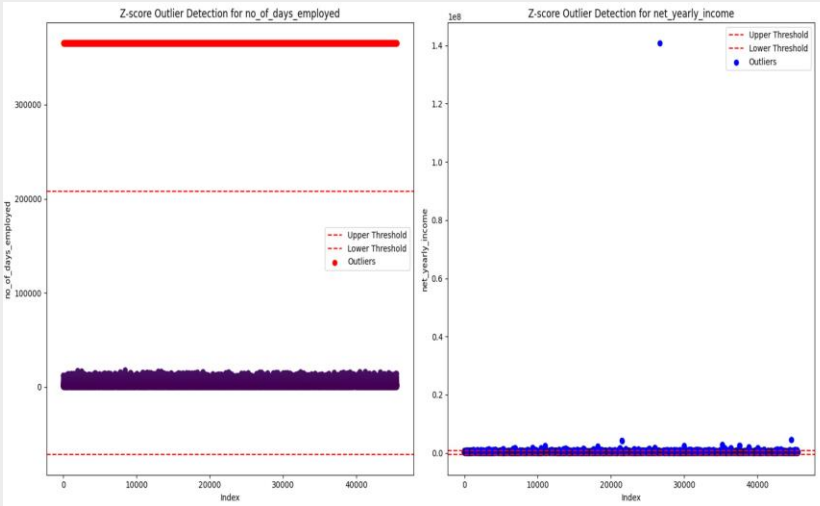
2.4. Statistical Preprocessing

Statistical preprocessing is crucial before modelling, as it helps eliminate discrepancies and ensures the data is appropriately aligned and fine-tuned. In this analysis, the following preprocessing steps were carried out:

- Outlier Detection and Removal: A Box plot was used to visualise the numerical variables and identify outliers. This visualisation confirmed the presence of outliers. The analysis identified two variables with prevalent outliers: Net Yearly Income (NIT) and Number of Days Employed (NDE). These variables were subjected to further statistical analysis using the Z-score outlier detection technique.



The Z-score technique involves subtracting the mean value from the original values of the variables and dividing the result by the standard deviation. Outliers were identified by setting a threshold of 1, meaning any Z value above 1 or below -1 was classified as an outlier. The results showed that some data indicated customers had been employed for over 300,000 days, which is erroneous. These outliers were removed from the analysis. A summary of the results and visualisations is presented below.



SUMMARY OF Z-Score OUTLIER DETECTION

Total rows corresponding to outliers for 'net_yearly_income': 107

Total rows corresponding to outliers for 'no_of_days_employed': 8229

Sum of rows that are outliers in df: 8336

Percentage of outliers in the dataset: 18.309611667545248

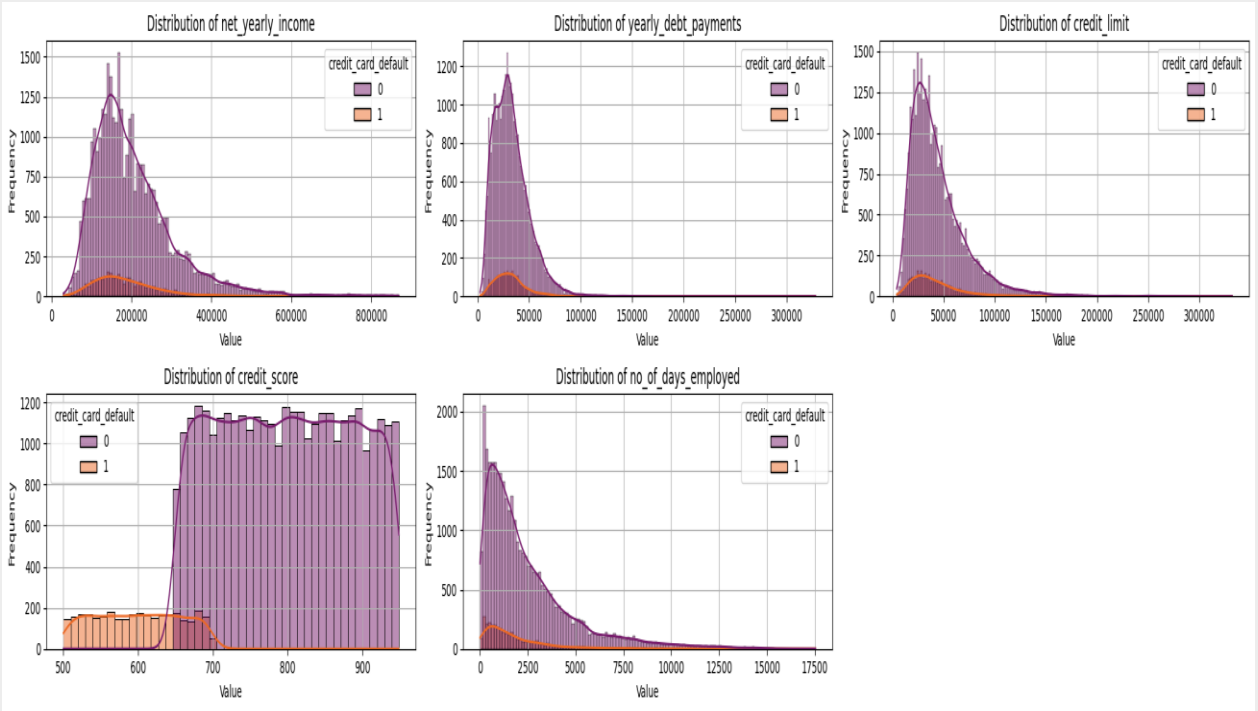
Min Outlier value for no_of_days_employed: 365240

Max Outlier value for no_of_days_employed: 365252

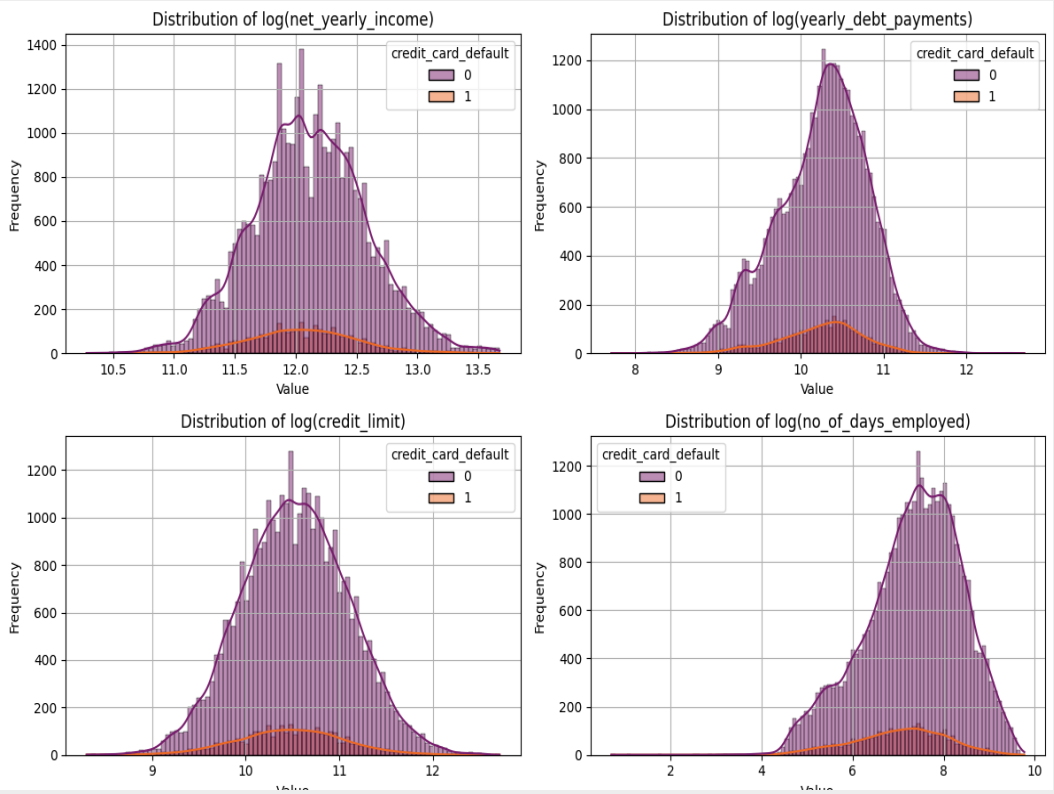
Min Outlier value for net_yearly_income: 869738.67

Max Outlier value for net_yearly_income: 140759012.73

- Observing Dataset Distribution: We used histograms and Kernel Density Estimates (KDE) to understand the dataset's distribution and skewness. These visualisations showed that most numerical/continuous variables are not normally distributed but are right-skewed. This means more values are below the mean, with a long tail extending to the right, as shown in the image below.

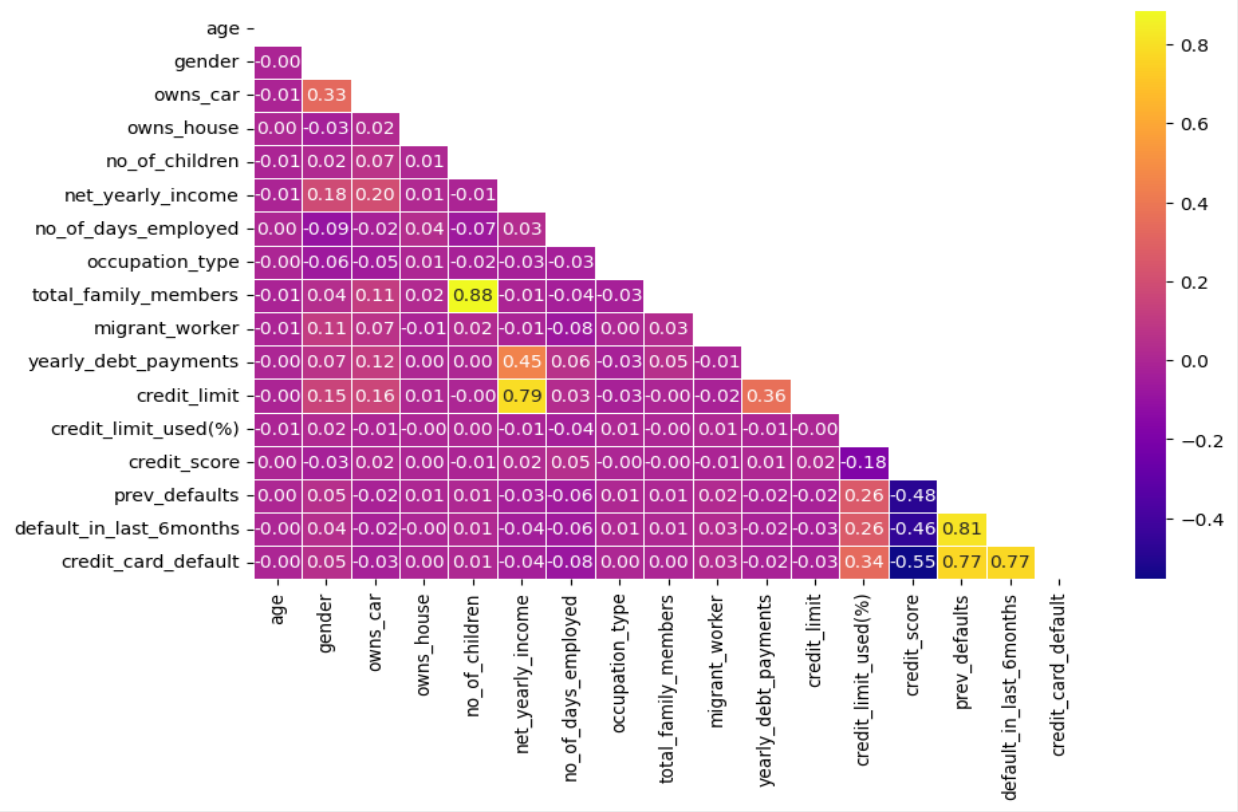


Since the numerical data is non-normally distributed and skewed to the right, it must be transformed to achieve a more centred distribution. This transformation was done using the logarithm function. The results after applying this transformation are shown below.



- Encoding categorical variables: This is another essential preprocessing step in modeling. Since categorical (non-numeric) variables cannot be directly used in modeling algorithms, converting them into numerical form is necessary. Different methods can be used for this purpose; however, Ordinal Encoding was adopted for this analysis. This method is suitable because some variables do not have binary characteristics but still require a meaningful numerical representation. Ordinal Encoding assigns a unique integer to each category of the categorical variable, maintaining any potential ordinal relationships among the categories.

- Check the Relationship Between Variables: A correlation matrix was used to examine the relationships between the features and the dependent variable as a statistical preprocessing step to identify potential multicollinearity. Despite some variables showing strong correlations with the dependent variable, all variables were included in the modeling process. This approach was chosen to ensure comprehensive representation, enhance predictive power, and capture potential feature interactions.



- Scaling: The final step in the preprocessing stage for this analysis involves scaling the variable values. Scaling is essential to ensure that all variables contribute equally to the model, especially those measured on different scales. This process helps improve the model's efficiency and accuracy by standardising the range of the independent variables. For this analysis, we applied standard scaling, transforming the data to have a mean of zero and a standard deviation of one. This technique is particularly useful for algorithms sensitive to the data's scale, such as logistic regression, support vector machines, and k-nearest neighbours. Standardising the variables enhances the model's performance and ensures that larger-scale variables do not unduly influence the results.

3. Model Methodology

With the data thoroughly preprocessed and deemed suitable for analysis, the next step is to apply modeling techniques to derive our predictions.

3.1. Model Preparation and Validation

The model preparation and validation process involves the following steps:

- Splitting Data into Features and Target: The dataset is divided into features (X) and the dependent variable (y). The features (X) include all the independent variables, while the dependent variable (y) represents whether a customer will default on their credit card (Credit card default).
- Train-Test Split: The data is split into training (X_train, y_train) and testing (X_test, y_test) sets using an 80-20 split. This implies that 80% of the data is used for training the model, and 20% is used for testing the model. This ensures that the model is trained on a substantial portion of the data and validated on a separate set to evaluate its performance.
- Model Training and Validation: Train the models on the training set (X_train, y_train) and evaluate their performance on the testing set (X_test, y_test), enabling the model's performance to be assessed on unseen data to validate its generalisation capability.

3.2. Choice of Model

The primary objective of our case study is to predict whether a customer will default on their loan, which is a binary classification problem (yes or no). This objective dictates our choice of models.

Based on this binary classification requirement, five (5) different models were adopted, which are:

- Logistic Regression
- K-Nearest Neighbors Classifier
- XGBoost Model
- Decision Tree Classifier
- Gaussian Naive Bayes

Each model will be evaluated based on predefined statistical metrics to determine the best fit, and according to these metrics, the chosen model will provide the most accurate and reliable predictions.

3.3. Statistical Evaluation

To evaluate our model performance, the following metrics were adopted:

- **Accuracy Score**: Measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances.
- **Precision Score**: Assesses the accuracy of positive predictions by calculating the ratio of true positive predictions to the total positive predictions (both true and false).
- **Recall Score**: This score measures the model's ability to correctly identify positive instances by calculating the ratio of true positive predictions to the total actual positives (true positives and false negatives).
- **F1 Score**: The harmonic mean of precision and recall provides a metric that balances both concerns.
- **Confusion Matrix**: This matrix provides a detailed breakdown of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP), which helps in understanding the model's performance in detail.
- **ROC Curve and AUC-ROC**: The Receiver Operating Characteristic (ROC) curve is visualised, and the Area Under the Curve (AUC) is measured. A model with an AUC value closer to 1 is preferred, as it indicates better performance in distinguishing between classes.

The models with the highest values for these metrics will be adopted. Specifically, we will look for the models with the highest accuracy, precision, recall, F1 score, and ROC value closest to 1. This comprehensive evaluation ensures that the selected model is robust and performs well across different aspects of classification performance.

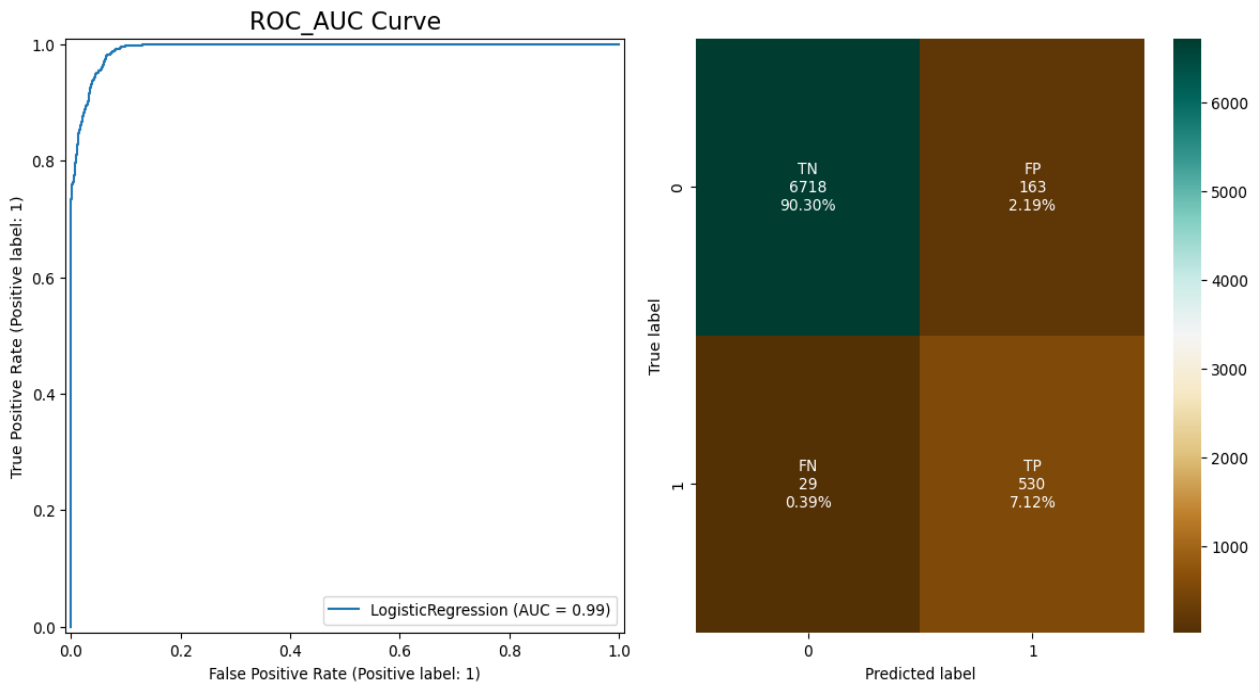
4. Result and Application

4.1. Model Result

The chosen models were trained on the training set and evaluated on the testing set. The results were assessed based on the statistical metrics initially set out. Although most models performed excellently, Logistic Regression outperformed the other four models with the following results:

- Accuracy Score: 0.97
- Precision: 0.76
- Recall: 0.95
- F1 Score: 0.85

As shown in the image below, these results indicate that the Logistic Regression model is highly accurate and robust, making it the best fit for predicting customer defaults in this analysis. The model's high AUC value (approximately 1) signifies excellent performance in distinguishing between defaulting and non-defaulting customers.



4.2. Model Application

The logistic regression model was selected and applied to the business (testing) dataset. The results are presented below, showing the Customer ID, the Probability of Class 0 (Prob_0), and the Probability of Class 1 (Prob_1). These probabilities indicate the likelihood of each customer defaulting on their credit card, with 0 representing non-default and 1 representing default, respectively.

Selected Model

```
model = LogisticRegression(random_state=42)
model.fit(X_train, y_train)
```

✓ 0.0s

```
LogisticRegression
LogisticRegression(random_state=42)
```

Predicted output

	customer_id	Prob_0	Prob_1	credit_card_default
0	CST_142525	1.0000	0.0000	0
1	CST_129215	0.9996	0.0004	0
2	CST_138443	0.0000	1.0000	1
3	CST_123812	1.0000	0.0000	0
4	CST_144450	0.0000	1.0000	1
5	CST_107341	0.9844	0.0156	0
6	CST_147879	1.0000	0.0000	0
7	CST_156027	1.0000	0.0000	0

5. Business Evaluation

5.1. Overview

The process of extracting business insight is quite important, and the effort carried out in statistical modelling might not be captured without it. Recall that part of this case study's objective is to ensure we solve the business problems that arise in the organisation. Solving this will require a series of steps, which include reshaping the predicted dataset, transforming it based on our business objectives and creating scenarios in line with these objectives. This analysis will cover this in three faces: Methodology, Customer risk segmentation, and Business Focus, i.e. applicable scenarios, to arrive at a solid business solution for Amex.

5.2. Methodology

5.2.1. Ventile-Based Profit Analysis

Ventile-based profit analysis divides data into 20 equal parts, allowing for finer granularity. The prediction output was transformed by sorting the non-defaults in descending order and assigning a ventile ranking from 1 to 20. New columns were then created using the following metrics:

- Count of Ventiles
- Count of Defaults (i.e., bad loans) in each ventile
- Mean Prob_0
- Prob_threshold
- Cumulative Defaults (bad)
- Cumulative Non-Defaults (good)
- Cumulative Percentage of Good and Bad
- Cumulative Percentage of Default Occurrence in each ventile

Furthermore, business scenarios were created with two assumptions:

- Loss from bad loans = $50
- Profit from good loans = $5

The business profit for each ventile was calculated using the formula:

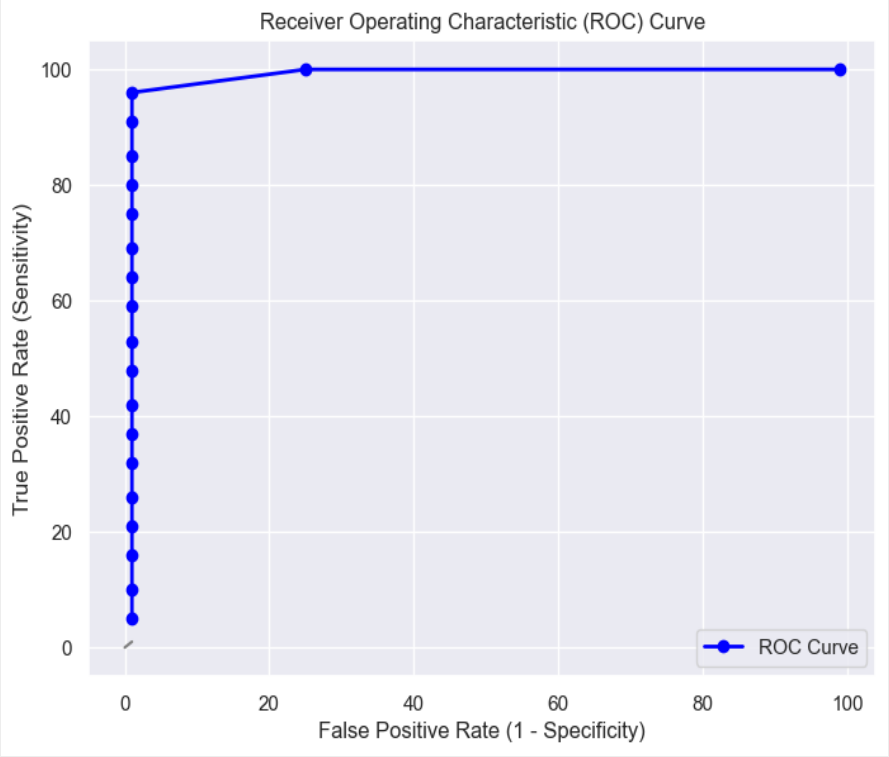
- Business Profit = $(\text{Cumulative Good} \times \text{Profit from Good Loans}) - (\text{Cumulative Bad} \times \text{Loss from Bad Loans})$

The new data, reflecting these calculations and scenarios, is presented below.

	Ventile	Count_of_Ventile	Bad	Prob_threshold	Prob_threshold%	Good	Cumulative_Bad	Cumulative_Good	Cumulative_Bad%	Cumulative_Good%	Cumm_%_of_Bad_to_occur	Business_Profit
0	1	569	0.0	1.000000	100	569.0	0.0	569.0	0	5	1	2845.0
1	2	569	0.0	1.000000	100	569.0	0.0	1138.0	0	10	1	5690.0
2	3	569	0.0	1.000000	100	569.0	0.0	1707.0	0	16	1	8535.0
3	4	569	0.0	1.000000	100	569.0	0.0	2276.0	0	21	1	11380.0
4	5	569	0.0	1.000000	100	569.0	0.0	2845.0	0	26	1	14225.0
5	6	569	0.0	1.000000	100	569.0	0.0	3414.0	0	32	1	17070.0
6	7	569	0.0	1.000000	100	569.0	0.0	3983.0	0	37	1	19915.0
7	8	569	0.0	1.000000	100	569.0	0.0	4552.0	0	42	1	22760.0
8	9	569	0.0	0.999928	99	569.0	0.0	5121.0	0	48	1	25605.0
9	10	569	0.0	0.999878	99	569.0	0.0	5690.0	0	53	1	28450.0
10	11	569	0.0	0.999744	99	569.0	0.0	6259.0	0	59	1	31295.0
11	12	569	0.0	0.999480	99	569.0	0.0	6828.0	0	64	1	34140.0
12	13	569	0.0	0.998915	99	569.0	0.0	7397.0	0	69	1	36985.0
13	14	569	0.0	0.997758	99	569.0	0.0	7966.0	0	75	1	39830.0
14	15	569	0.0	0.995282	99	569.0	0.0	8535.0	0	80	1	42675.0
15	16	569	0.0	0.989239	98	569.0	0.0	9104.0	0	85	1	45520.0
16	17	569	0.0	0.969369	96	569.0	0.0	9673.0	0	91	1	48365.0
17	18	569	0.0	0.888663	88	569.0	0.0	10242.0	0	96	1	51210.0
18	19	569	209.0	0.518980	51	360.0	209.0	10602.0	26	100	25	42560.0
19	20	572	572.0	0.000000	0	0.0	781.0	10602.0	100	100	99	13960.0

5.2.2. Validation

To validate the results of our ventile-based analysis, we compare the ROC curve from our logistic regression model to the ROC curve for the ventile-based profit analysis. The high AUC values in both curves confirm the model's excellent performance and the reliability of the ventile-based profit analysis. Therefore, these results are suitable for decision-making.



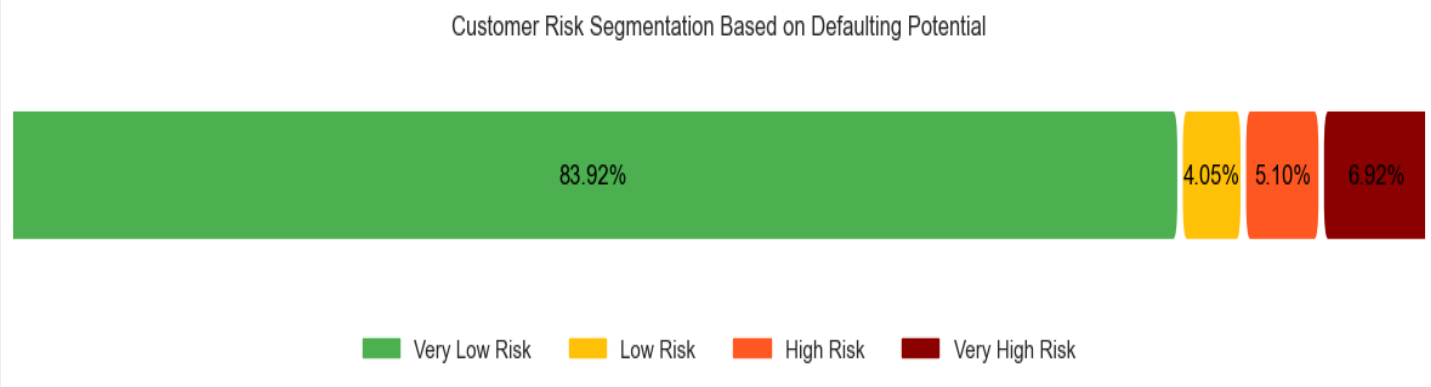
5.3. Business Implication

5.3.1. Customer Risk Segmentation

A probability threshold metric was adopted for segmentation to identify each customer's risk level while considering profitability and customer retention. The highest threshold to consider was 96%, and the lowest was 51%. Customers were segmented into the following risk categories:

- Very Low Risk (>= 96%)
- Low Risk (<96% - >= 88%)
- High Risk (<88% - >= 51%)
- Very High Risk (<51%)

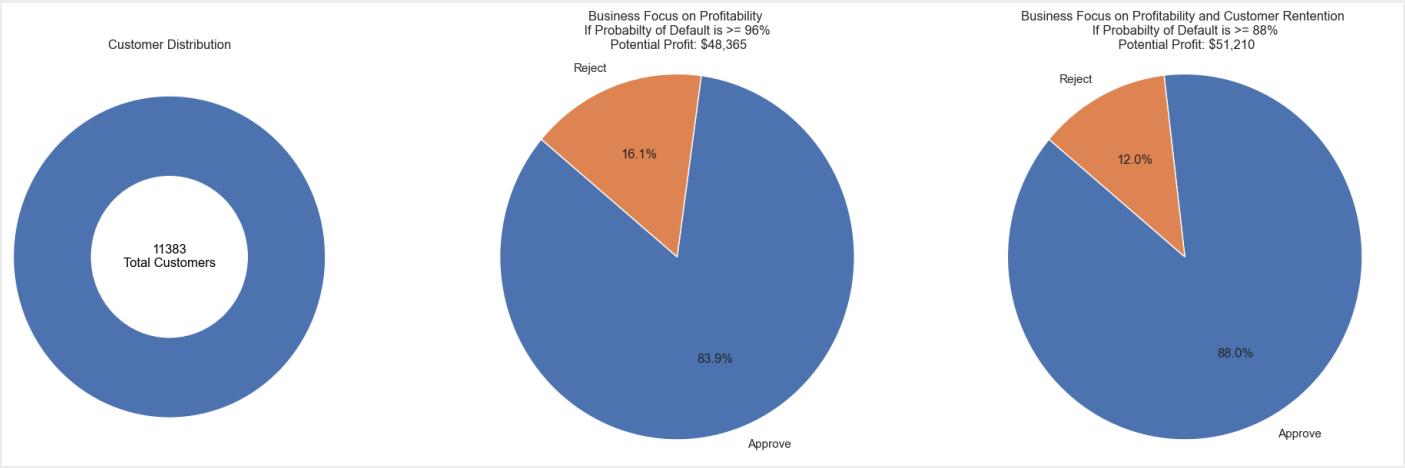
Further analysis shows that 83.92% of the customers fall into the very low-risk category, 4.05% fall into the low-risk category, 5.10% fall into the high-risk category, and 6.92% fall into the very high-risk category. This segmentation is visually represented in the image below:



5.3.2. Business Recommendations

Based on the analysis, the following actions are recommended:

- Focus on Profitability: If the business focuses solely on profitability, it is recommended that loans for customers with a probability threshold of 96% or above be approved. This will result in credit request approval for 83.9% of the customers and a rejection rate of 16.1%, leading to a potential profit of $48,365.
- Focus on both Profitability and Customer Retention: If the business aims to balance profitability and customer retention, it is recommended that customers have a probability threshold of 88% or above before approval. This strategy will lead to a 5% customer increase, resulting in a potential profit of $51,210.



5.3.3. Output

Depending on the choice adopted by management, the output for both Profitability and Customer retention can be downloaded here in XLSX format.

6. Conclusion

This analysis aimed to improve American Express's credit risk assessment process by leveraging historical customer data. By developing predictive models, specifically a highly accurate logistic regression model, the study successfully identified customers likely to default on their credit. The model's high accuracy (0.97) and robust performance metrics (precision: 0.76, recall: 0.95, F1 score: 0.85) confirmed its efficacy in distinguishing between defaulting and non-defaulting customers.

The business evaluation highlighted two primary strategies for American Express:

- Focusing on Profitability: Approving 83.9% of credit requests could yield a potential profit of $48,365 with a rejection rate of 16.1%.
- Balancing Profitability and Customer Retention: Approving 88% of credit requests could result in a higher potential profit of $51,210 with a lower rejection rate of 12%.

These findings provide actionable insights for American Express, enabling optimisation of their credit approval process while balancing profitability and customer retention. This approach mitigates potential losses and enhances overall business efficiency and customer satisfaction.

ABOUT THE AUTHOR

Temidayo Olowoyeye is a distinguished professional with a dual master's in Environmental Engineering and Agricultural Economics. With over 5+ years of experience, he is a proficient Data/Business Analyst dedicated to assisting individuals and

His expertise includes building compelling KPIs and target dashboards with PowerBI or Tableau, modifying and exploring relational databases using SQL, analyzing big data and building models using Python or R, and preparing reports with key findings and recommendations that can increase business efficiency.

His unwavering commitment to excellence and passion for data-driven insights make him an asset to any organization seeking a seasoned Analyst with a diverse skill set.

Read more