AMEX Customers Credit Card Default Prediction Modelling

Introduction

Credit cards are a common financial offering from traditional banks, but these institutions often hesitate to issue them to individuals likely to default on their payments. To mitigate this risk, banks collect and evaluate customer information to determine the potential for default.

It is common in the financial industry to build predictive models for credit card default based on various methods including machine learning techniques. Researchers have used predictors like income, job, residence, and credit situation using machine learning techniques like gradient boosting, support vector machines and random forest to predict the customers with a higher probability of default (Sayjadah et al., 2018). Another study utilized the logistic regression to model the impact immigration status has on the likelihood of the default where the results showed that the probability of default at a higher income level, immigrants were less likely to default while at the lower income level the probability of default was much higher (Chengan, 2018).

This report aims to build upon the current literature of employing the powerful generalized linear model framework and explore that what are the key predictors of credit card default among American Express customers, and how does the interaction between demographic factors impact the likelihood of default?

Methods

The study was conducted with a systematic approach to develop the final generalized linear regression model. The initial step involved an exploratory data analysis (EDA) on the dataset to gain a comprehensive understanding of its characteristics. Histograms and bar graphs provided an overall picture, while facilitating the identification skewness among the predictors. Additionally, a correlation matrix was generated to identify potential confounders and assess interaction term necessity within the model[1].

We utilized a logistic regression model using the Generalized Linear Model (GLM) framework in R, targeting a binary outcome through binomial distribution. Our objective was to establish a statistical model that accurately captures the relationship between predictors and the binary response. For optimal model selection, we used three methods: forward and backward selection to minimize the Akaike Information Criterion (AIC), ensuring a balance between model simplicity and fit; the Bayesian Information Criterion (BIC) for a more rigorous selection, preventing overfitting by penalizing complexity; and Lasso regularization, which effectively shrinks less important coefficients to zero, thereby emphasizing crucial predictors, minimizing overfitting, and enhancing the model's predictive performance. These methods collectively contributed to refining our model, ensuring it robustly predicts the binary outcome with a focus on key variables.

[1] Refer to Appendix.

The final model was chosen based on its parsimony, ensuring it was not overly complex while still retaining the optimal number of predictors. This selection process was critical to avoid the inclusion of unnecessary variables that could obscure the model's interpretability. Furthermore, a potential confounder, initially removed during variable selection, was reintroduced due to its contextual significance.

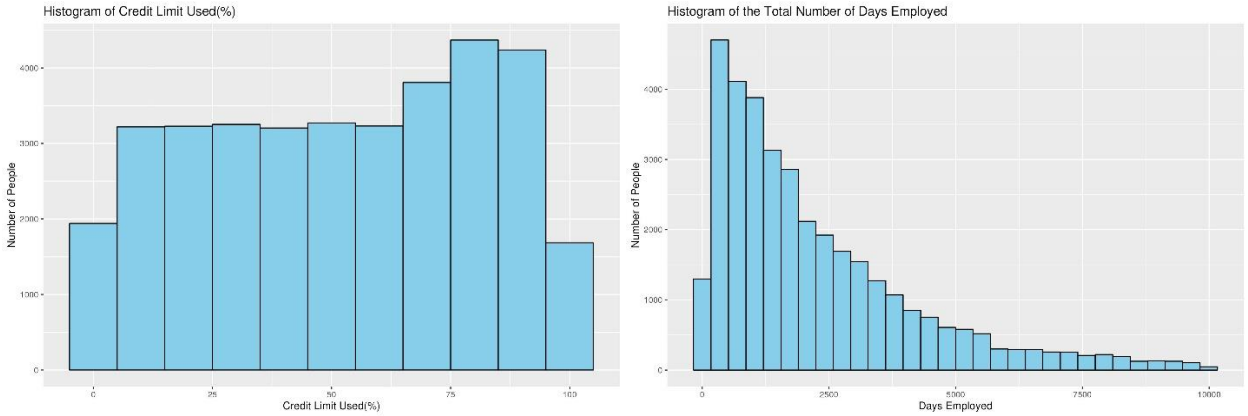
To evaluate the model's performance, calibration plots were constructed for each candidate model to assess the relationship between observed outcomes and predictions, ensuring the model's reliability across different probability levels. The Receiver Operating Characteristic (ROC) curve analysis was also conducted to measure the model's discriminative ability, specifically its capability to distinguish between the outcome classes correctly. Lastly, dffits plot of the final model was generated to identify influential data points that could unduly affect the model's parameters, to assess the limitation of the regression analysis. The final regression model was rigorously developed and validated, embodying a comprehensive and systematic approach to statistical modeling.

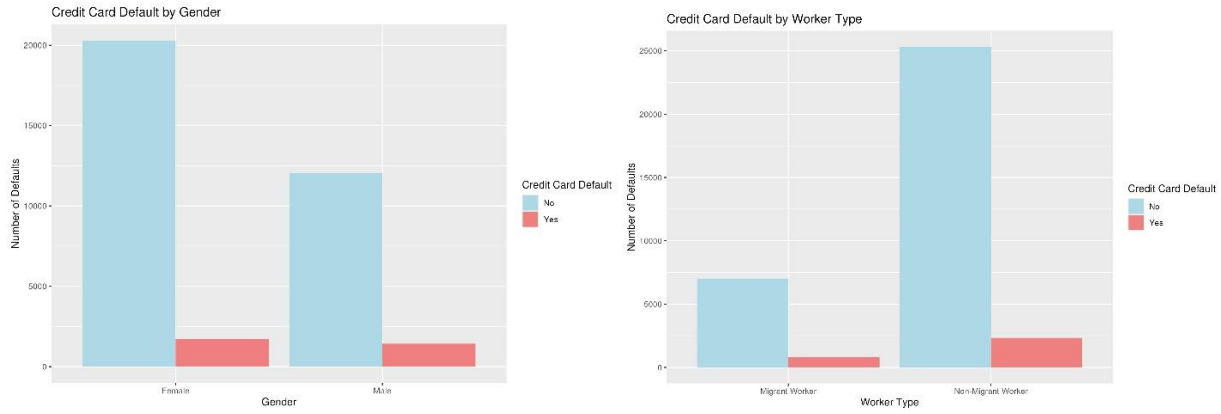
Finally, the dataset utilized to investigate the research question was retrieved from Kaggle, a repository for data which encompasses customer demographics and financial attributes of American Express customers. It includes variables like age, gender, car and house ownership, immigration status, credit score, default status, and credit limit utilization rate etc.

Results

The dataset contained 45,528 unique observations and following extensive data wrangling, visualizations were plotted for key predictors. The data exhibited a near-normal distribution with distinct separation and variable skewness which coincides with GLM assumption of normality. Bar plots revealed a marginally higher likelihood of credit default among females and non-migrant workers, suggesting potential lower risk aversion among non-migrant workers.

Figure 1: Exploratory Data Analysis





Examining the correlation matrix, a negative correlation was noted between the number of days employed and the immigration status of the customers. To address potential multicollinearity and control for confounding variables, an interaction term was integrated into the model.

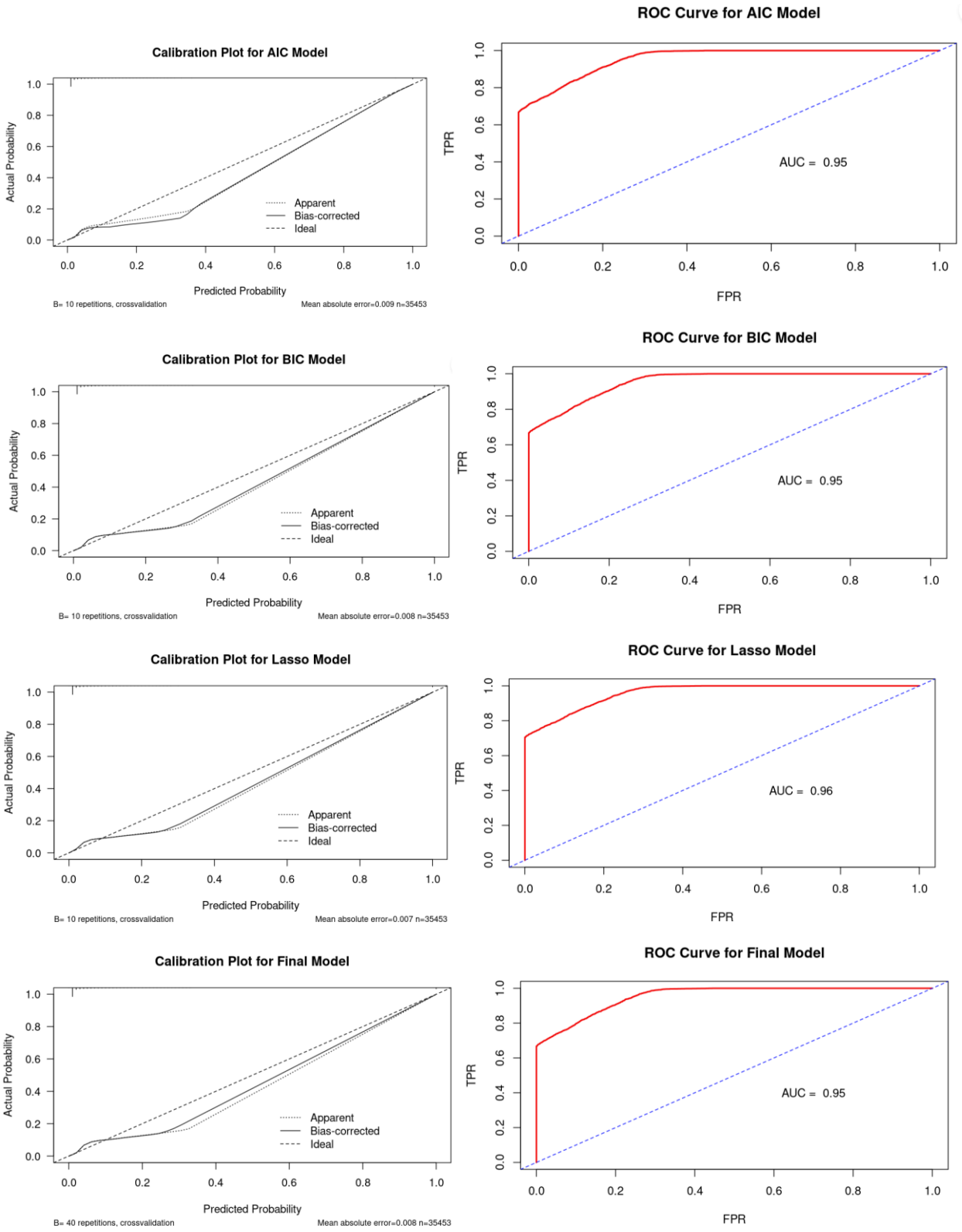
For the selection of these predictors, we employed various variable selection methods, based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Lasso. Each method yielded a distinct set of predictors, detailed in Table 1. Notably, AIC identified the most predictors, while Lasso identified the fewest.

Table 1: Variables selected based on the three methods

Selection Method	Variables
Akaike Information Criterion	gender, owns_car, owns_house, no_of_children, net_yearly_income, no_of_days_employed, total_family_members, migrant_worker, credit_limit, credit_limit_used(%), credit_score, prev_defaults, net_yearly_income:credit_limit, gender:owns_house, no_of_days_employed:migrant_worker
Bayesian Information Criterion	gender, owns_car, no_of_days_employed, migrant_worker, credit_limit_used(%), credit_score, prev_defaults, no_of_days,_employed:migrant_worker
Lasso Variable Selection	credit_limit_used(%), credit_score, prev_defaults, default_in_last_6months
Final Model	age, gender, owns_car, no_of_days_employed, migrant_worker, credit_limit_used(%), credit_score, prev_defaults, no_of_days,_employed:migrant_worker

We performed each models' validity by observing the calibration plots and ROC curves (see figure 2). The calibration plots were largely similar among all the models' showing points below the diagonal line, and that the models overestimated the probabilities where the actual occurrences are less frequent than predicted. Furthermore, the ROC curve for the AIC and BIC models yielded an AUC of 0.95 while Lasso had an AUC value of 0.96 indicating strong discriminative ability among all the models to predict the likelihood of credit default.

Figure 2: Calibration Plot and ROC Curves



Ultimately, the model chosen through BIC was considered the most optimal because the BIC model was the most parsimonious, offering a balance between simplicity and explanatory power, while facilitating easier interpretability.

Furthermore, age was incorporated as a predictor due to its logical relevance in the final model. The decision was based on the observation that younger individuals are more likely to default on credit payments. This tendency can be attributed to the less stable employment and inexperienced credit management among younger individuals while older individuals often benefit from more stable income sources including retirement funds.

Discussion

In our analysis, summarized in Table 2 that uses logistic regression, we found a significant association between gender and credit card default, with males 1.504 times more likely to default than females, indicating a 50.4% higher likelihood (coefficient = 0.408, odds ratio ≈ 1.504). This result aligns with the observed tendency of men to engage in riskier financial behaviors, potentially leading to increased default rates. Similarly, we found a significant association between credit score and credit card default, with each one-point increase in credit score reducing the likelihood of default by approximately 4.5% (coefficient = -0.046, odds ratio ≈ 0.955). This result suggests that higher credit scores are associated with a reduced risk of defaulting on credit cards, reflecting the general tendency for individuals with higher credit scores to have more stable financial behaviors and better debt management, leading to lower default rates. In general car ownership, number of employed days, credit limit utilization rate alongside gender and credit score provides us with significant predictors to predict and explain credit default likelihood which was the primary aim of this study.

Table 2: Regression Results based on the Bayesian Information Criterion (BIC)

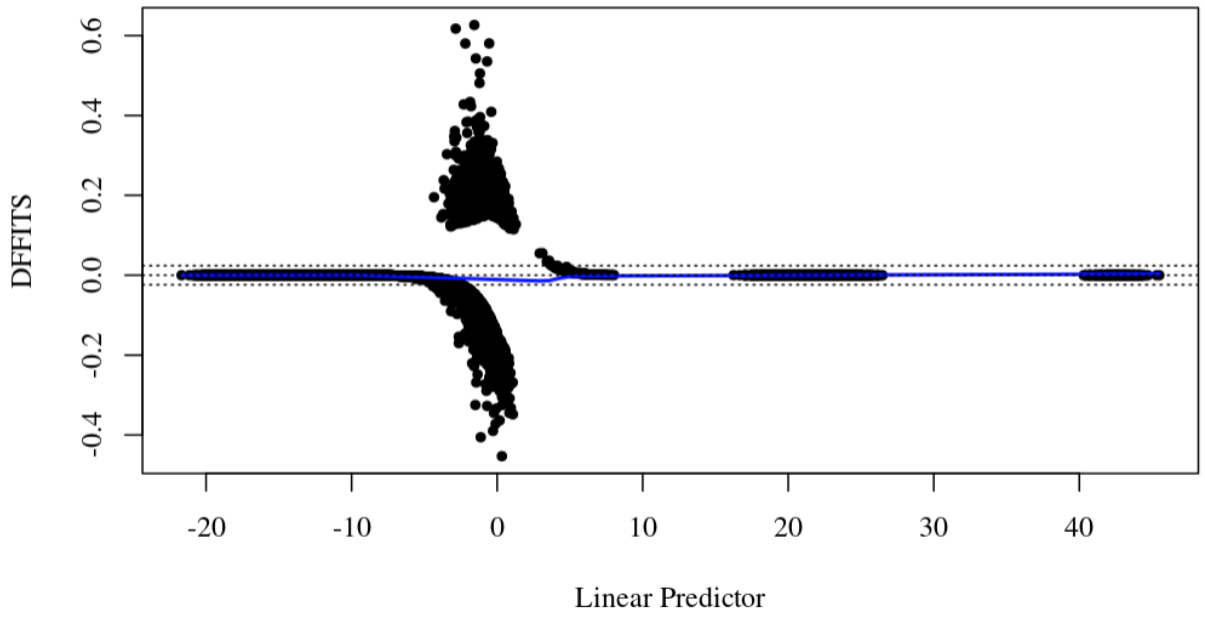
Coefficients:	**Estimate**	**Std. Error**	**Z Value**	**Pr(>|z|)**
(Intercept)	2.341e+01	1.190e+00	19.677	< 2e-16 ***
Age	5.092e-03	4.526e-03	1.125	0.260596
Gender	4.077e-01	9.344e-02	4.363	1.28e-05 ***
Owns_car	-3.606e-01	9.741e-02	-3.702	0.000214 ***
No_of_days_employed	-1.415e-06	2.812e-07	-5.033	4.83e-07 ***
Migrant_worker	3.008e-01	1.549e-01	1.942	0.052182
Credit_limit_used(%)	7.557e-02	3.081e-03	24.531	< 2e-16 ***
Credit_score	-4.601e-02	1.749e-03	-26.299	< 2e-16 ***
Prev_defaults	1.845e+01	2.812e+02	0.066	0.947674
No_of_days_employed: Migrant_worker	-1.369e-04	7.278e-05	-1.881	0.59955

*Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Metric	**Value**
Null deviance	24804.5
Residual deviance	4272.6
Degrees of freedom	44021
AIC	4290.6
Number of Fisher Scoring Iterations	19

To assess our final model's limitations, DFFITS plot served as a diagnostic to evaluate individual data points' influence on the logistic regression model, critical for identifying outliers and leverage points that could bias parameter estimates. In figure 3, a concentration near zero suggests limited influence for most observations. However, notable deviations indicate potential outliers with disproportionate impact. Given the identified class imbalance, these outliers may skew predictions toward the majority class, affecting model validity and our estimates. While recognizing the significance of class imbalance on parameter bias, its comprehensive analysis including sampling techniques was beyond our paper's scope, warranting further investigation in future research.

Figure 3: Dffits of the Final Model



Word Count: 1220 words

References

Zhou, X., Zhang, W., & Jiang, Y. (2020). Personal credit default prediction model based on convolution neural network. Mathematical Problems in Engineering, 2020, 1–10. https://doi.org/10.1155/2020/5608392

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-98361-6

Chengan, R. (2018). Migration Status and Loan Default. 39.

Sayjadah, Y., Hashem, M., Alotaibi, F., & Kasmiran, K. A. (2018). Credit Card Default Prediction using Machine Learning Techniques. IEEE Xplore. https://doi.org/10.1109/icaccaf.2018.8776802

Appendix

Correlation Matrix Heatmap