Math 42 Final Project

Credit Cards: Are you Getting Approved?

By: America Barrera, Sarina Doss, Aashman Rastogi, and Chinmay Varshneya

University of California, Los Angeles

405 Hilgard Ave, Los Angeles, California 90025

Contents

Abstract

Credit cards play a vital role as indispensable tools that enable individuals to make significant purchases that would otherwise be impossible. For our project, we focused on studying credit card late fees by using personal characteristics of individuals. Our aim was to develop a predictive model that credit card companies can leverage to forecast whether a person will incur late fees based on their provided information. To achieve this, we thoroughly evaluated our dataset using a range of models to identify the most effective approach for predicting the likelihood of late fees. Through our analysis, we determined that the optimal model is a KNN model utilizing the Manhattan distance metric that considers 9 neighbors. This carefully crafted model exhibited an impressive accuracy of approximately 0.741, showcasing its potential for accurate predictions to help companies determine if a person should or should not be approved for a credit card.

Problem Description

According to Pokora (2023), in 2021, 84% of U.S. adults held a credit card, and by the age of 25, approximately 73% of Americans had obtained one, making credit cards the most common initial credit experience for young adults. Credit cards are vital financial tools that offer numerous benefits to individuals. The approval of credit card applications varies based on factors such as credit history, income, and the specific requirements and policies of the credit card issuer. Credit card approval holds great significance as it enables individuals to establish a credit score. Building good credit proves highly advantageous as it qualifies people for loans to purchase assets like houses and cars that would otherwise be unattainable without borrowing.

In this project, we aim to construct a model that determines whether our hypothetical bank should approve an applicant for a credit card based on their individual attributes, specifically focusing on their history of late payments. This project is interesting and relevant, particularly for college students who currently possess credit cards or have plans to obtain one in the future. We will utilize various models to accurately predict whether a person is likely to incur late fees and compare these predictions with the actual outcomes.

Simplifications and Assumptions

The simplifications made in the code dataset for employment, education, marriage, gender, car ownership, and real estate ownership can be summarized as follows. The distinction between money earners and non-money earners was simplified to a binary classification, where anyone with a job or pension is considered a money earner (assigned a value of 1), while students and the unemployed are considered non-money earners (assigned a value of 0). Similarly, the level of education was simplified into four categories, with higher education/academic degree assigned the highest value (3), incomplete higher education assigned a value of 2, secondary/special secondary education assigned a value of 1, and incomplete secondary education assigned a value of 0. The marital status was reduced to a binary classification, where being married indicates dual income (assigned a value of 1), while being single, widowed, or divorced indicates no dual income (assigned a value of 0). Gender was also simplified to a binary classification, with males assigned a value of 1 and females assigned a value of 0. Ownership of a car and real estate were also reduced to binary classifications, where owning a car is represented by 1 and not owning a car is represented by 0, similarly for owning real estate. We also decided to drop rows including missing data and dropped the housing column as it was too broad to work with. These simplifications were made for the purpose of the code dataset, potentially for ease of analysis or computational reasons, but it's important to note that they may not capture the full complexity of the respective variables in reality.

Mathematical Model

We conducted an analysis using five different models to determine the most accurate one in predicting the likelihood of incurring late fees, and we compared these predictions with the actual outcomes. Our first model choice was the Logistic Regression model. Logistic regression is well-suited for credit card approval prediction because it effectively handles binary classification problems and large datasets. It provides a probabilistic framework that allows us to interpret feature coefficients and understand the impact of different factors on credit card approval. However, it has limitations as it doesn't consider non-linear relationships between features and is sensitive to outliers.

The second model we employed was the Random Forest model. Random Forest is a strong contender for credit card approval prediction due to its ability to handle complex relationships and generate robust predictions. It performs well with imbalanced data by balancing class weights, reducing the risk of overfitting. Nonetheless, Random Forest models can be computationally expensive and have longer training times compared to simpler models. Accurate results require tuning hyperparameters, and they generally excel when dealing with lower-dimensional data.

Our third model choice was the Gaussian Naive Bayes model. Gaussian Naive Bayes is a suitable option for credit card approval prediction due to its simplicity, efficiency, and capability to handle continuous and normally distributed features. However, it assumes independence between features, which may not hold true in real-world scenarios. It might not capture complex relationships between features or handle interactions effectively.

The fourth model we utilized was the XGBoost model, an algorithm with powerful capabilities for credit card approval prediction. XGBoost excels in handling complex relationships, imbalanced data, and provides high predictive accuracy. It employs an ensemble of decision trees and incorporates gradient boosting techniques, allowing it to capture nonlinear interactions, feature importance, and handle large feature spaces. However, XGBoost can be computationally intensive and requires careful tuning of hyperparameters for optimal performance.

Lastly, we employed the K-nearest neighbors (KNN) model, which is a good choice for credit card approval prediction due to its simplicity, non-parametric nature, and ability to capture local patterns in the data. KNN makes predictions based on the similarity of a new data point to its nearest neighbors in the training set. However, KNN can be sensitive to the choice of distance metric and the number of neighbors (K). It may not perform well with high-dimensional or sparse datasets. Additionally, the prediction process can be computationally expensive, especially with large training sets.

To find the best hyperparameters for the KNN classifier, we performed Hyperparameter Optimization using a grid search and cross-validation. We then trained a new KNN classifier with the best hyperparameters, made predictions on the test set, and calculated the accuracy of the predictions. In our case, we used the Manhattan Distance as a metric for computing distances in KNN. The Manhattan Distance calculates the sum of the absolute differences between corresponding feature values of two points. It measures the total length of the sides of a right-angled triangle when moving from one point to another in a grid-like fashion.

The best model we arrived on for classification was a KNN model using Manhattan distance and 9 neighbors. It resulted in an accuracy of approximately 0.741.

Solution of the Mathematical Model

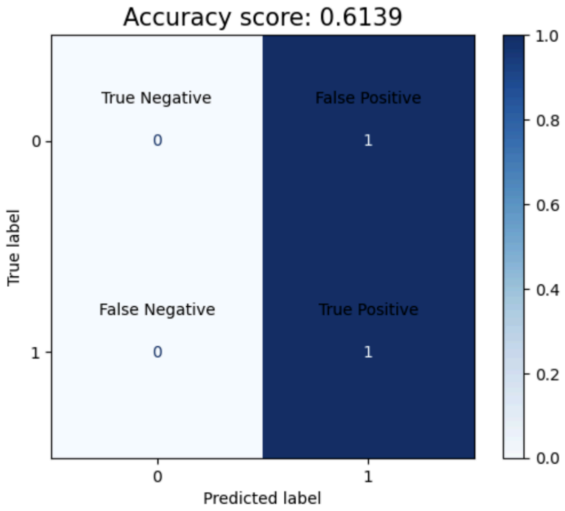
We used One Hot and Integer encoding on our data regarding the features of users so that we could apply common machine learning algorithms to them. We did not have to use any new analytical method aside from our judgment when deciding how exactly to encode various categorical variables in terms of which integers to assign and what they indicate about the relationship of one result to another. When it comes to the actual solution techniques, all the machine learning algorithms used were either in the scikit-learn Python library or the XGBoost Python library. We did optimize the best performing model (KNN) using a grid search technique. The techniques utilized in this project were not conversed in class but the overall structure of running a machine learning prediction model was explained by the TA Raymond Chu in discussion. We followed the standard steps of selecting and manipulating the data, then splitting it and then training the models and comparing the trained model with the test data.

In addressing the mathematical problem at hand, we employed various techniques to analyze and solve it. The problem specifically focused on predicting whether a person is likely to incur late fees in the context of credit card approval. To make accurate predictions, we explored the use of five different models stated above. We used machine learning algorithms to solve this problem.

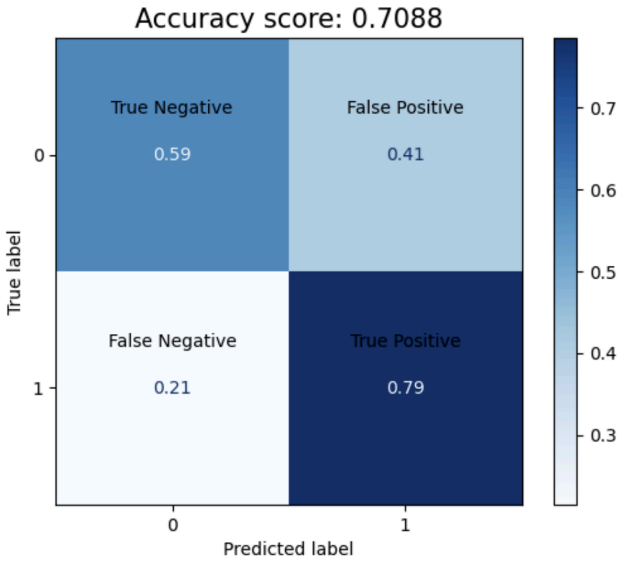
Results

The models and their corresponding graphs with accuracy scores are shown below.

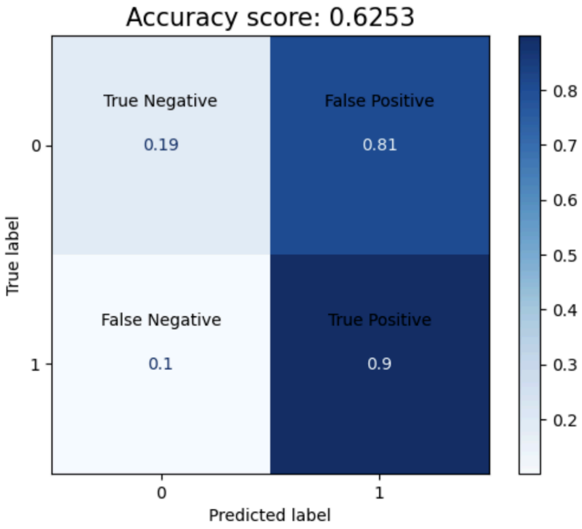
Logistic regression Model:



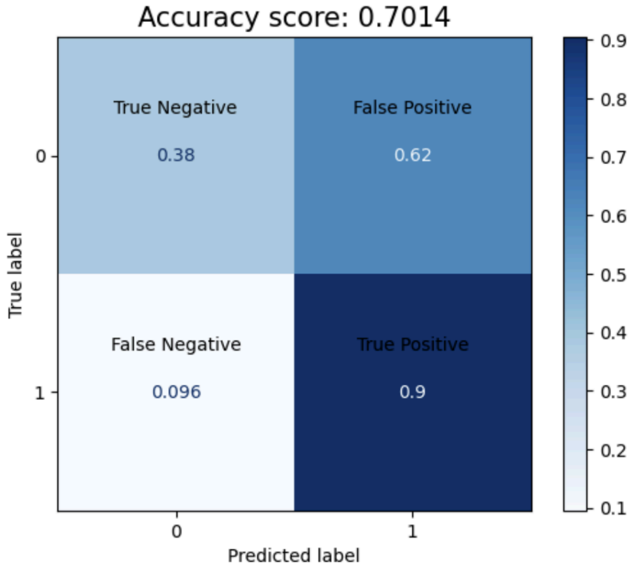
Random Forest:



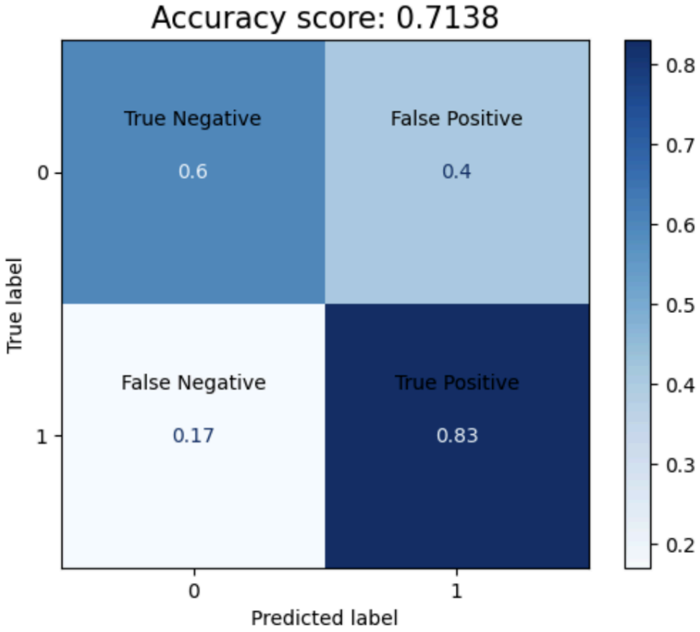
Gaussian Naive Bayes:



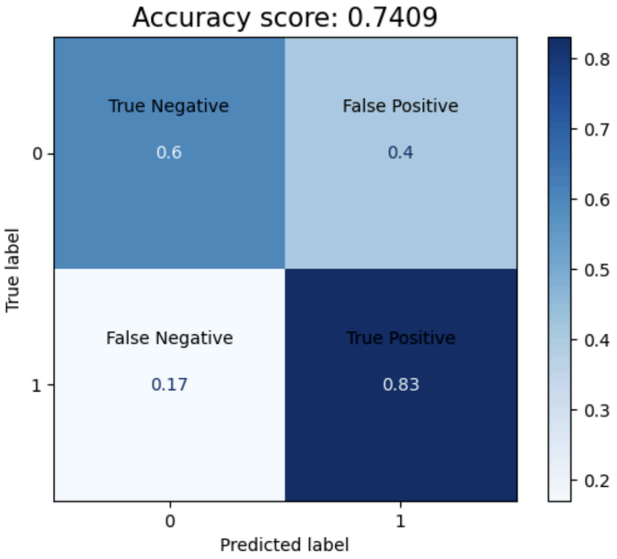
XG Boost:



K Nearest Neighbors:



KNN Hyperparameter Optimization:



As we can see, the best model we arrived at for classification was a KNN model using Manhattan distance and 9 neighbors. It resulted in an accuracy of approximately 0.741. Therefore we can conclude that KNN is the best model to accurately predict whether a person is likely to incur late fees meaning that it will be the most successful for the credit card company to use.

Improvements

Due to the low accuracy score of our model, there are several areas and aspects that require improvement. One approach is to create a Neural Network for our model. This involves adding more layers to enhance its complexity and constructing a surrogate model to avoid repetitive evaluations for each combination, thereby saving computation time. By reducing computation time, we can conduct more efficient explorations, exploiting patterns to identify trends and dependencies more effectively.

Furthermore, we can enhance our model by assigning greater significance to specific features during scaling. For example, considering our understanding of credit card data, we can prioritize scaling a person's income level over their marital status. Additionally, to improve the model further, we can test additional features. This can be achieved by acquiring a larger dataset with more features, allowing us to examine 20 features instead of the current 10.

Finally, one of the primary means to enhance our project is through improved engineering. In our initial approach, we oversimplified many of the columns to facilitate comprehension and workability. However, conducting a more in-depth analysis of the diverse columns would have been beneficial, considering the substantial variation among them. For instance, some columns exhibited a linear relationship while others did not. By running different models on specific columns, we could have improved the accuracy of our predictions.

Conclusion

In conclusion, credit cards serve as crucial tools that empower individuals to make substantial purchases they otherwise couldn't afford. Our project focused on studying credit card late fees by analyzing personal characteristics of individuals. The objective was to develop a predictive model that credit card companies could utilize to forecast the likelihood of late fees based on provided information. Through a comprehensive evaluation of the dataset using various models, we identified the KNN model with Manhattan distance and 9 neighbors as the optimal approach. This well-designed model achieved an impressive accuracy of approximately 0.741, demonstrating its potential to accurately predict whether an individual is likely to incur late fees.

This valuable insight can assist credit card companies in making informed decisions regarding credit card approvals and managing risk effectively.

Resources

“Choosing the Best Machine Learning Classification Model and Avoiding Overfitting.” Choosing the Best Machine Learning Classification Model and Avoiding Overfitting - MATLAB & Simulink, www.mathworks.com/campaigns/offers/next/choosing-the-best-machine-learning-classification-model-and-avoiding-overfitting.html. Accessed 11 June 2023.

II. Credit Cards – General Overview - FDIC, www.fdic.gov/regulations/examinations/credit_card/pdf_version/ch2.pdf. Accessed 12 June 2023.

Lake, Rebecca. “Applying for a Credit Card? Here’s How Your Approval Odds Stack Up.” Investopedia, 25 May 2023, www.investopedia.com/applying-for-a-credit-card-your-odds-of-being-approved-4684901.

Pokora, Becky. “Credit Card Statistics and Trends 2023.” Forbes, 9 Mar. 2023, www.forbes.com/advisor/credit-cards/credit-card-statistics/#sources_section.

Seanny. “Credit Card Approval Prediction.” Accessed 11 June 2023. Kaggle.

White, Alexandria. “What Is a Credit Card?” CNBC, 5 May 2023, www.cnbc.com/select/what-is-a-credit-card/.