MAS 648

FINAL Project - FRAUD

Machine Learning Models

Peiliang (Mark) Yu, Julian Pollak, Andrew Wolfe, Andreina Diaz

Abstract

This report explores the application of different Machine Learning models to identify contributing factors that detect credit card fraud. Using a the Fraud dataset which encompasses a binary fraud indicator (response variable) along with multiple features including dates, product types, state information, domestic (US) or international transactions, transaction amounts, card types, income, cardholder FICO scores, and balances, five machine learning models were trained, evaluated, and compared. The primary objective of this report included constructing two distinct models: an interpretable model, and a predictive model for efficient fraud detection.

This study used preprocessing techniques, along with feature engineering such as creating an "Age" variable measuring the account duration in months, and model selection methodologies to optimize model performance. Employing Logistic Regression (normal Logistic, Ridge, and Lasso), K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, and XGBoost (Extreme Gradient Boosting), these models were implemented to predict fraudulent activities.

Evaluation of the models was based on Accuracy, AUC, F-1 Score, and precision on the test datasets. Notably, the Logistic model was the most interpretable model and the XGBoost model emerged as the most proficient predictive model, showcasing superior accuracy, precision ,AUC, and overall performance for fraud

detection, solidifying its position as the preferred choice for robust and accurate predictions in this assessment.

Introduction

The shift towards digitalization in financial transactions has revolutionized convenience while also engendering a surge in credit card fraud occurrences. This poses a substantial challenge to financial security. Addressing this escalating threat, this report explores five different Machine Learning (ML) models that aim to comprehend the crucial factors in credit card fraud detection.

The dataset utilized includes transaction dates, product categories, locations, card details, and cardholder demographics. Using this data, the aim is to create and thoroughly evaluate two Machine Learning models. Model 1 is the best interpretable model, and model 2 is the best predictive Model. By comparing five different models for effectiveness, were were able to select the best models for both tasks. One model prioritizes interpretability, while the other focuses on accurate predictions. These models are vital for strengthening financial security and fraud detection.

This study aims to reveal hidden patterns in the Fraud dataset by using diverse Machine Learning methods, uncovering the complexities within fraudulent transactions. Through descriptive analytics, we were able to see the relationship between the variables. Barplots, scatterplots, correlation plots were used to identify relationships between variables. Five Models underwent rigorous evaluation based on Accuracy, F1 Score, precision, and AUC aiming to identify models adept at balancing accuracy and interpretability—essential for understanding fraud detection mechanisms.

Ultimately, this study aims to strengthen fraud detection capabilities by diving deeper into the complexities of fraudulent transactions.

Descriptive Analysis

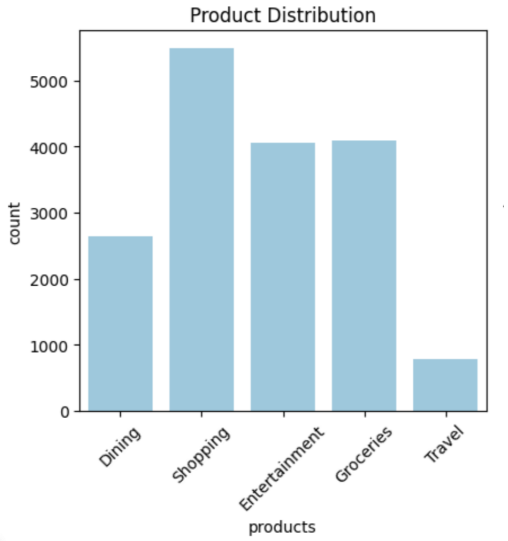

Figure 1. *Showcases the distribution of credit card product categories, revealing 'Shopping' as the predominant category, while 'Travel' ranks as the least frequent category.*

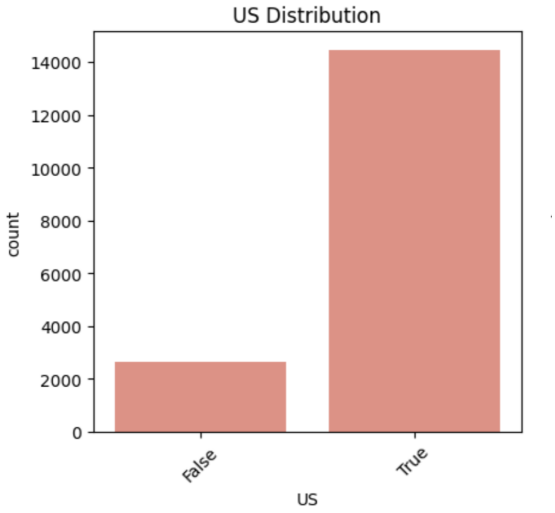

Figure 2. *Depicts the Dominance of US Credit Card Transactions in the Fraud Dataset.*

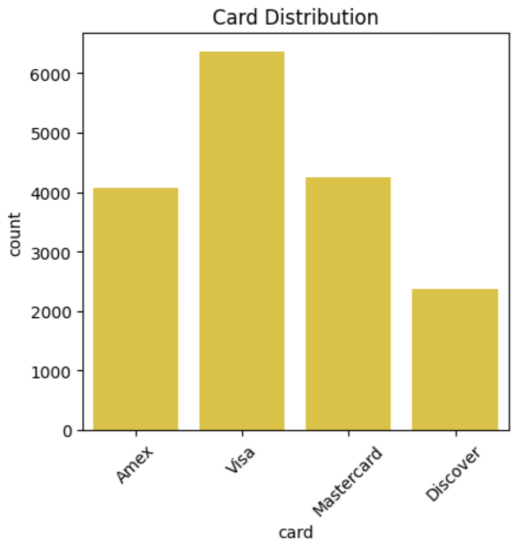

Figure 3. *Illustrated the Card Type Distribution among Major Credit Cards (Visa, Mastercard, Discover, and American Express), spotlighting Visa as the Dominant Choice, with Discover Representing a Smaller Proportion.*

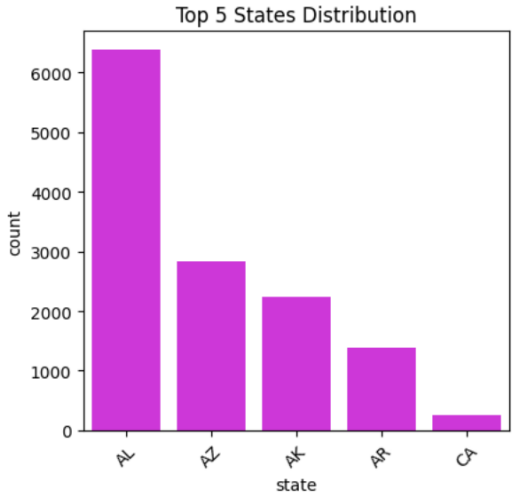

Figure 4. *Showcases the Distribution among the Top 5 States (Alabama, Arizona, Alaska, Arkansas, California), emphasizing Alabama as the Highest and California as the Lowest State in Representation within this dataset.*

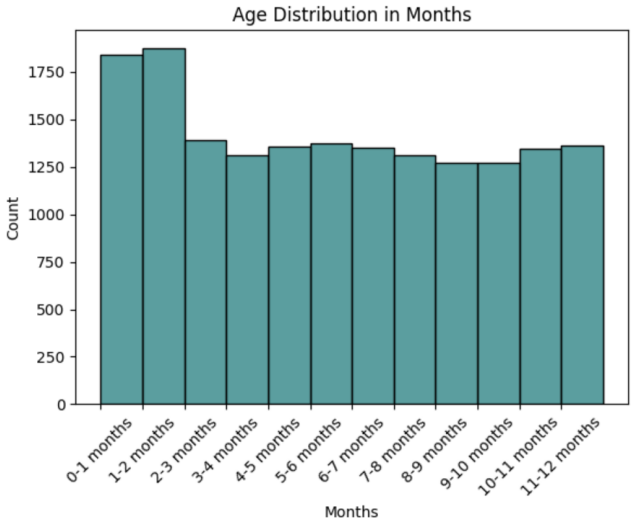

Figure 5. *Depicts the distribution of credit account ages in months reveals a concentration around newly opened accounts.*

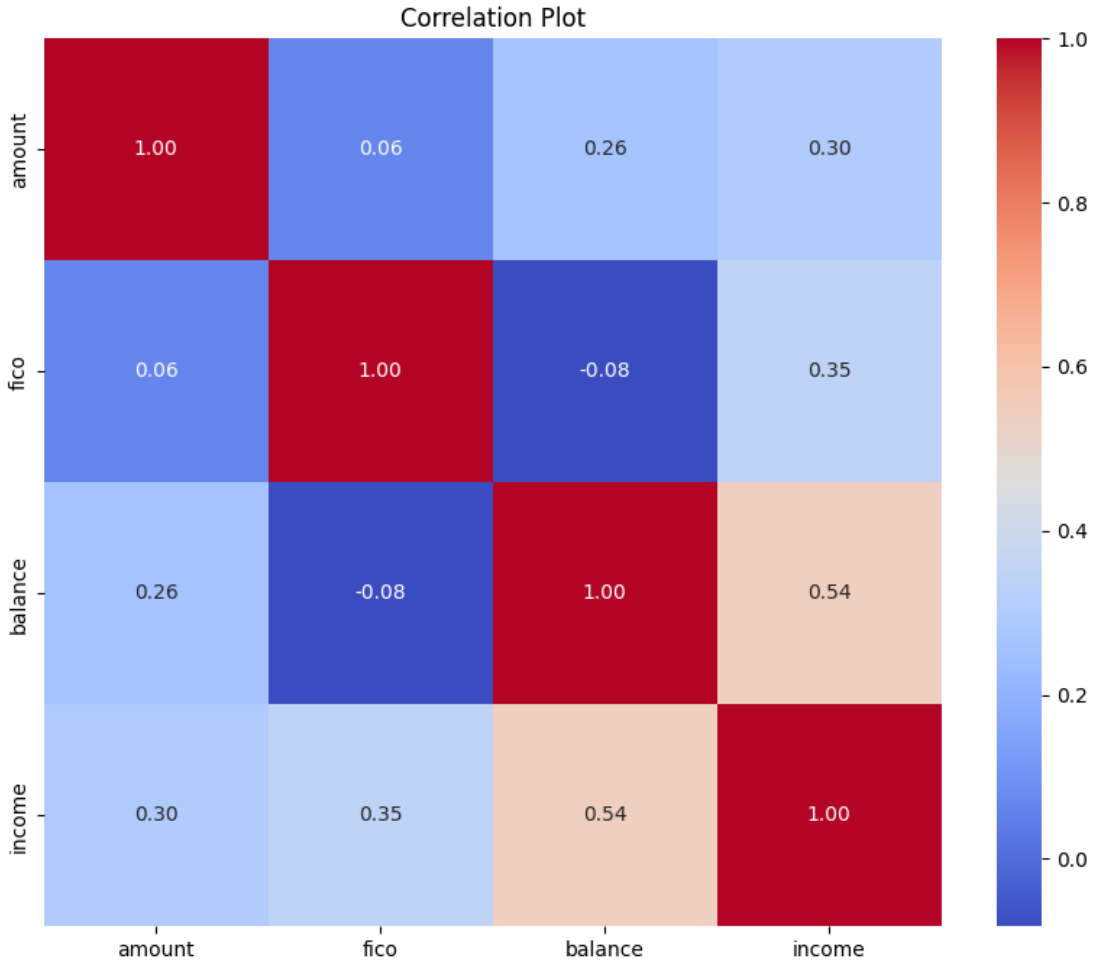

Figure 6. *Illustrates the Correlation Plot, highlighting a significant positive association between 'Income' and 'Balance' (correlation coefficient: 0.54).*

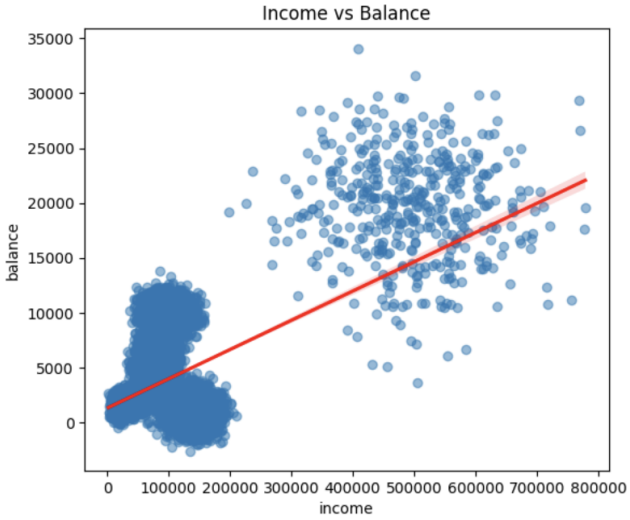

Figure 7. *Depicts a scatter plot depicting a moderately positive correlation (correlation coefficient = 0.54) between 'Income' and 'Balance'. The visualization suggests that higher income levels are associated with increased account balances.*

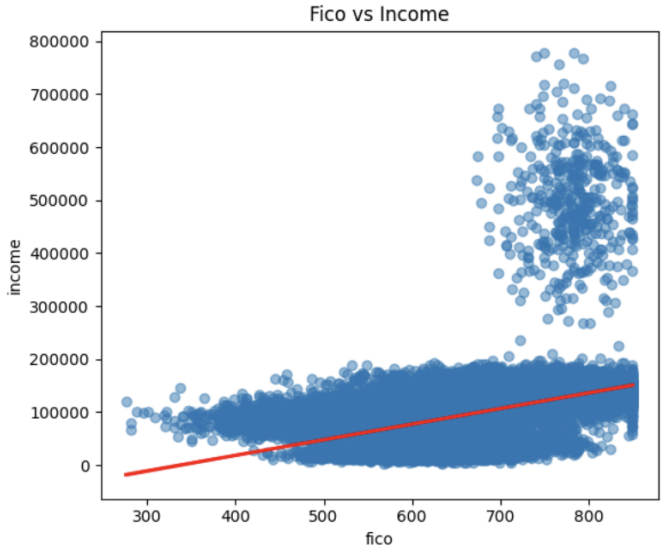

Figure 8. *Illustrates a scatter plot that reveals a moderate positive correlation (correlation coefficient = 0.35) between 'Fico' and 'Income'.*

Figure 9. *Depicts a scatter plot with a mild positive correlation (correlation coefficient = 0.30) between 'Amount' and 'Income'. The scatterplot suggests at a subtle tendency where higher income levels may correspond to slightly increased transaction amounts.*

Preprocessing

Data Cleaning: Categorical variables underwent transformation into dummy variables, while ensuring the absence of missing values. Additionally, all states except the top 5 states - Alabama, Arizona, Alaska, Arkansas, and California - were removed from the dataset.

Feature Engineering: In the process of feature engineering, an "Age" variable was created to capture the account age of credit card holders. This age data was segmented into bins to generate a visual barplot for analysis. Furthermore, feature scaling was applied to optimize the features for machine learning modeling.

Train-Test Split: All models were segregated data into training and test sets using an 80-20 split ratio and random state = 42.

Modeling: Selected and implemented Machine Learning classification algorithms such as Logistic, K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, and XGBoost (Extreme Gradient Boosting).

Evaluation Metrics: Employed critical evaluation metrics like Accuracy, F1 Score, AUC, and precision to comprehensively assess the performance of the models.

Modeling 1 - Best Interpretable Model

The analysis explored three logistic models—Logistic, Ridge, and Lasso—all of which yielded identical model summaries. This consistency across models suggests robust and stable predictive performance. Notably, at a significance level of 0.05, several variables demonstrated strong predictive significance for fraud, including “fico”, “balance”, “age”, “US_True”, “state_AK”, “state_AL”, “state_AR”, “state_AZ”, “state_CA”, “products_Entertainment”, and “products_Shopping”.

These models, encompassing both L1 (Lasso) and L2 (Ridge) regularization techniques, achieved an overall accuracy rate of 92.12%. Their primary aim is to forecast fraudulent outcomes based on a range of potential predictors.

An in-depth precision analysis unveils distinct precision rates for non-fraudulent (class 0) and fraudulent (class 1) transactions—92% and 69%, respectively. The model excels in accurately identifying non-fraudulent transactions, achieving a flawless recall rate of 100%. However, it faces a challenge in identifying instances of fraud, with a modest recall rate of 4%.

Modeling 2 - Best Predictive Model

In the pursuit of identifying the most effective model for fraud prediction, two contenders emerged: the XGBoost (XGB) model and the Random Forest Classifier. Both models showcased exceptional capabilities, yet their distinctive strengths became apparent upon closer examination of critical metrics.

Random Forest Classifier:

With an accuracy of 96.13% and an AUC score of 96.68%, the Random Forest Classifier demonstrated its capability achieving robust overall classification performance. Its precision for detecting fraud instances at 90% reflected a high level of accuracy when predicting fraud. However, a 59% recall for fraud indicated a

potential for missing some instances of actual fraudulent activities, leading to a slightly lower recall rate.

XGBoost (XGB) Model:

The XGB model consisted of an accuracy of 96.10% and a AUC score of 97.64%, displayed superior capabilities in accurately ranking instances, particularly in binary classification like this one. With a precision of 82% for detecting fraud, it demonstrated a high accuracy in identifying true fraudulent cases. Moreover, its recall for fraud instances stood significantly higher at 67%, signifying its capability to capture a larger proportion of actual fraud cases.

The disparity in recall rates between the XGB model and the Random Forest Classifier is a critical differentiator. XGB's superior recall indicates its proficiency in identifying a higher percentage of actual fraud cases while maintaining a reasonably high precision. This is crucial in fraud detection, where missing even a small fraction of fraudulent transactions can have significant consequences.

While both models exhibit strengths—Random Forest with its higher precision for fraud and XGB with its higher recall and AUC—it's evident that XGB stands out in detecting more instances of fraud while maintaining a reasonably high level of precision. Thus, for the sensitive task of fraud detection where the priority lies in capturing as many fraudulent transactions as possible, XGBoost emerges as the model of choice.

In conclusion, considering metrics of precision, recall, accuracy, and AUC, the XGBoost model is the optimal choice for fraud prediction, showcasing its proficiency in identifying fraudulent transactions with a higher recall rate while maintaining a noteworthy level of precision.

Model Comparison

Model	Accuracy	Precision (Class 0/1)	Recall (Class 0/1)	F1-Score (Class 0/1)
Logistic	92.12%	92%/69%	100%/4%	96%/8%
KNN (K=12)	92.26%	93%/61%	99%/11%	96%/19%
Decision Tree	95.25%	97%/70%	97%/71%	97%/71%
XGBoost	96.10%	97%/82%	99%/67%	98%/73%
Random Forest	96.13%	97%/90%	99%/59%	98%/71%

Figure 10. *Performance Metrics Comparison of Various Models for Fraud Prediction: Comparison of model performance metrics highlights XGBoost as the optimal choice for fraud prediction, excelling in accuracy, precision, and recall. Notably, Random Forest demonstrates high precision but comparatively lower recall, impacting its suitability for specific fraud detection requirements.*

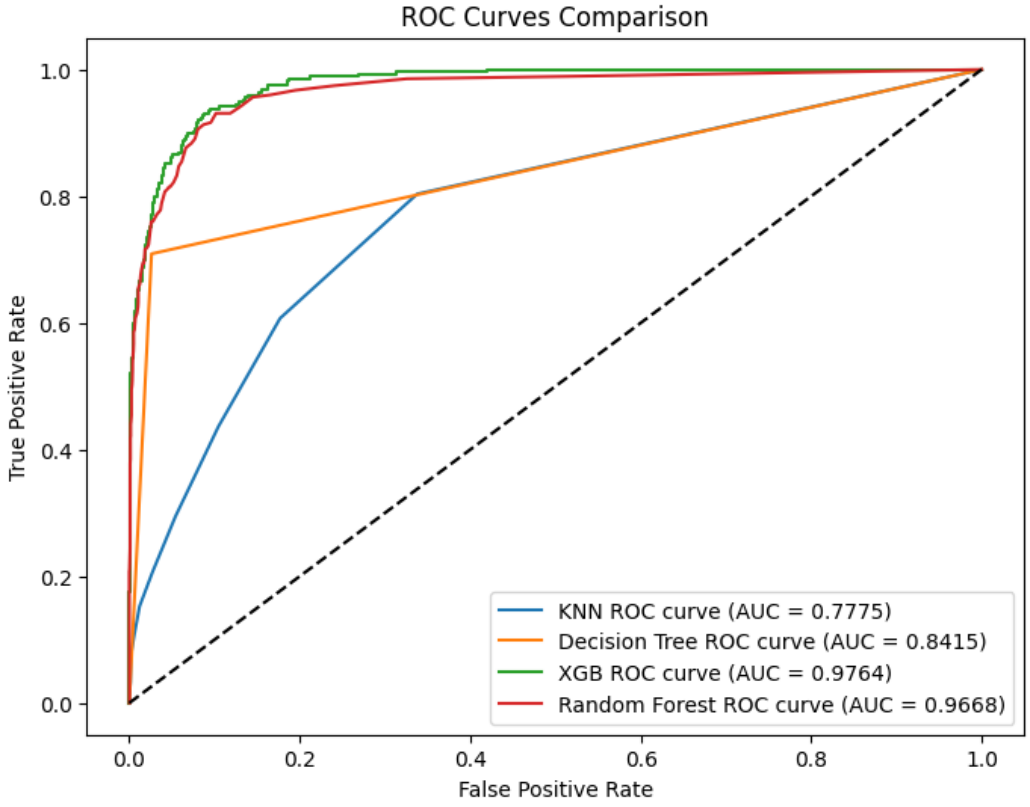

Figure 11. *Comparing AUC Scores of Fraud Prediction Models via ROC Curve. XGBoost leads with an AUC of 0.9764, showcasing superior classification ability, followed by Random Forest (0.9668) and Decision Tree (0.8415). KNN trails slightly, registering an AUC of 0.7775, indicating comparatively weaker discrimination between positive and negative classes.*

- **Logistic Model**

The logistic model demonstrated a pseudo R-squared of 0.1653, showcasing moderate explanatory power. Key predictors include 'fico', 'balance', 'age', 'US_False', various state indicators, and product categories like 'Entertainment' and 'Shopping'. However, the model demonstrates a significant imbalance in predicting fraud instances, achieving an accuracy of 92.12% but displaying notably lower recall (0.04) for fraudulent cases,

indicating its struggle in correctly identifying instances of fraud. The precision for fraud detection stands at 0.69, coupled with an AUC of 0.77, indicating moderate performance in distinguishing between positive and negative classes, with a tendency to misclassify actual fraud cases as non-fraudulent.

- **KNN (K=12) Model**

  The KNN model utilizing k=12 exhibited a 92.26% accuracy, showcasing a nuanced improvement in balancing precision and recall for fraud detection (Class 1) compared to the General KNN model. While displaying a commendable precision of 61% in identifying fraud, this model encountered limitations in accurately capturing instances of fraud, reflected by a recall of 11%.

- **Decision Tree**

  The Decision Tree Classifier achieved an overall accuracy of 95.25%, excelling in accurately predicting non-fraudulent cases (Class 0) with a precision of 97% and fraudulent cases (Class 1) with a precision of 71%. It was also very efficient in instances of fraud, reflected by a recall of 70% and demonstrates balance across all metrics. This model stood as the third best model in accuracy.

- **XGB**

  XGBoost stands out with the highest accuracy and consistently strong performance in precision, recall, and F1-scores for both classes. It demonstrates robustness and balance across metrics, making it a top-performing model. However, Random Forest Classifier follows closely and exhibits balanced performance across all metrics. Logistic Regression and K-Nearest Neighbors show strengths in certain aspects but have limitations in accurately predicting Class 1 instances.

- **Random Forest Classifier**

  Leveraging the Fraud dataset with a binary target ('Fraud'), this classifier assesses its performance using accuracy, precision, recall, F1-score, and support metrics. Achieving an impressive accuracy of 96.13%, it stands as the second-best model, slightly trailing behind XGB. It shares similarities with the XGB model, showcasing remarkable precision, recall, and F1-score for non-fraudulent cases but demonstrating comparatively lower performance in identifying instances of fraud. Despite being the second-best, its robustness in handling non-fraud cases and the overall effectiveness in binary classification underscore its strengths in comprehensive model performance assessment

Feature Importance

The top 10 important features in our XGBoost Classifier identified “product_Shopping” as the most important variable, followed by “amount”, “product_Entertainment”, “income”, and “card_Amex” in predicting fraud in our dataset. The variable “product_Shopping” had a feature importance of about 0.50, 0.40 units larger than any other variable, indicating that it is the most important feature in our model. These input variables are extremely important in predicting whether or not there is fraud for each circumstance. Feature importance helps us understand which variables have the most impact in predicting whether or not there is fraud. In this case, the variables below are all logical areas of concern when diving deeper into fraud. Feature importance not only leads to more comprehensibility of the model, but it also guides feature selection which is extremely important for creating highly successful machine learning models. Understanding the feature importance of each variable enables people to put their

focus on the most important variables, contributing to banks and fraud prevention companies knowing which elements to keep an eye on for detecting fraud in the future.

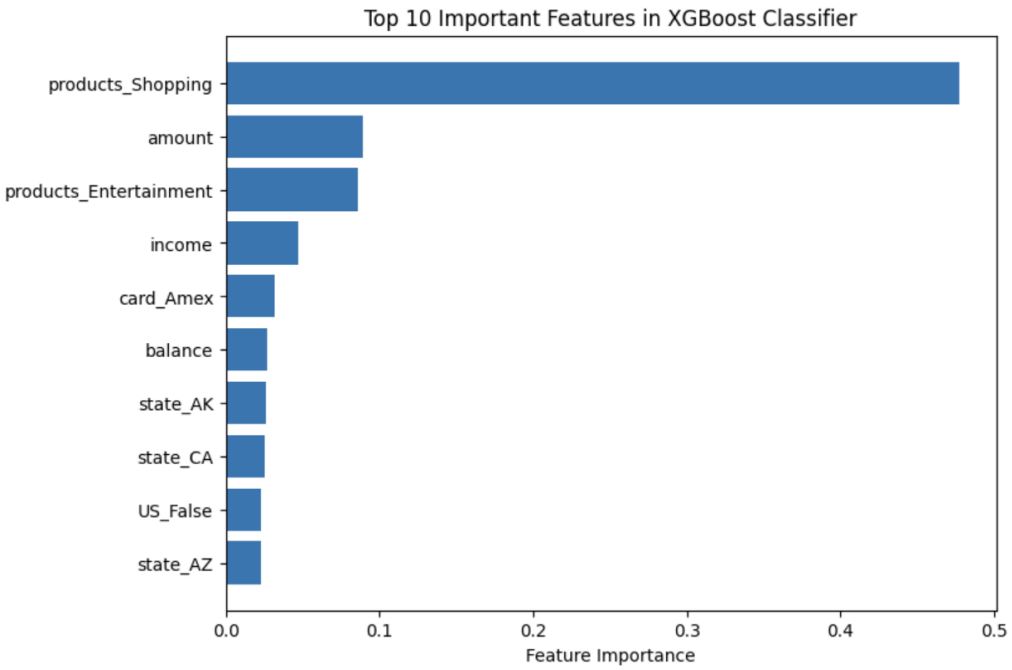

Figure 12. *Feature Importance Plot for XGBoost Classifier. This plot illustrates the relative importance of the top 10 features determined by the trained XGBoost classifier.*

Conclusion

This study explored five diverse Machine Learning models to detect credit card fraud, utilizing the Fraud dataset that consists of transaction details, cardholder information, and demographic indicators. Two primary models were constructed: one emphasizing interpretability, and another focusing on the best predictive capabilities.

Several preprocessing techniques and feature engineering strategies, including the creation of an "Age" variable to measure account duration, were utilized to optimize model performance. Only the top 5 states were used for all models. Logistic Regression Classification, K-Nearest Neighbors Classifier, Decision Tree Classifier, Random Forest

Classifier, and XGBoost were evaluated based on accuracy, AUC, F1-Score, and precision on test datasets.

Among these, the Logistic, Ridge, and Lasso models excelled in interpretability, achieving an accuracy of 92.12%. While proficient in identifying non-fraudulent transactions, they faced challenges in detecting instances of fraud, notably exhibiting modest recall rates.

The XGBoost and Random Forest Classifier emerged as formidable contenders. Both demonstrated exceptional capabilities, yet the XGBoost model exhibited superior recall for fraud instances (67%) compared to the Random Forest Classifier (59%). With its higher recall rate, coupled with a reasonably high precision (82%) and a AUC value of 0.97, the XGBoost model solidifies its standing as the optimal choice for fraud detection in this study.

Moreover, the XGBoost model highlighted "product_Shopping," "amount," and "income" as pivotal fraud predictors, shedding light on the influential variables impacting fraud prediction.

In summary, the XGBoost model outshines other contenders with its superior precision, recall, and overall accuracy in detecting fraud. While the Random Forest Classifier closely follows, the XGBoost's proficiency in capturing a larger percentage of actual fraud cases positions it as the premier choice for robust and accurate fraud predictions in this evaluation.