Banking Fraud Detection: Predictive Modeling

by Carlos Martinez

Overview:

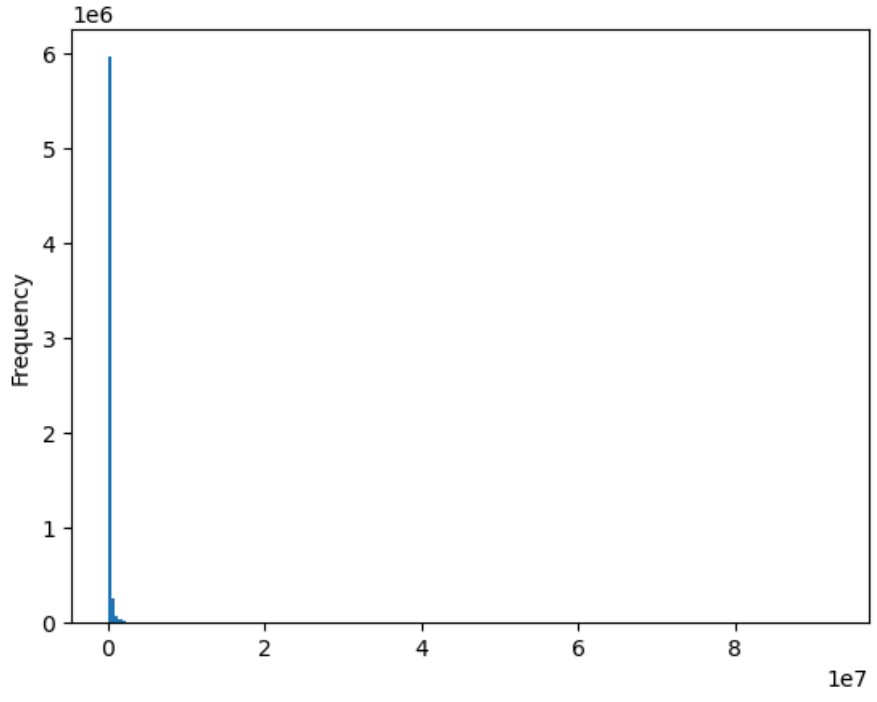
The project aimed to develop a fraud detection system for banking transactions using supervised machine learning techniques, specifically focusing on the Random Forest Classifier algorithm. Given the inherent imbalance in fraud detection datasets where fraudulent cases constitute a minority class, the challenge was to accurately identify fraudulent transactions while minimizing false positives. The dataset comprised of synthetic bank transaction records, containing both discrete and continuous variables.

Exploratory Data Analysis

To begin exploring the banking dataset, we start by conducting EDA to identify relationships among variables and to gain insight into feature distributions.

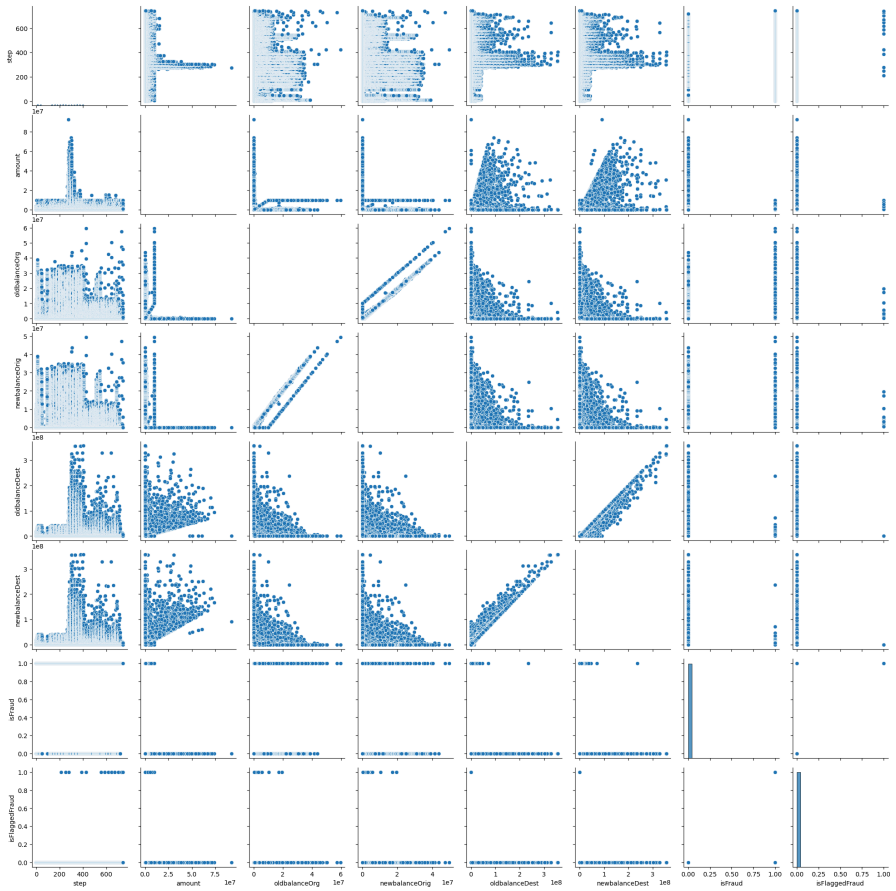
Histogram of Amount Observations:

- The amount variable is right skewed. This is an indicator that our dataset contains a majority of large amounts of money transferred.



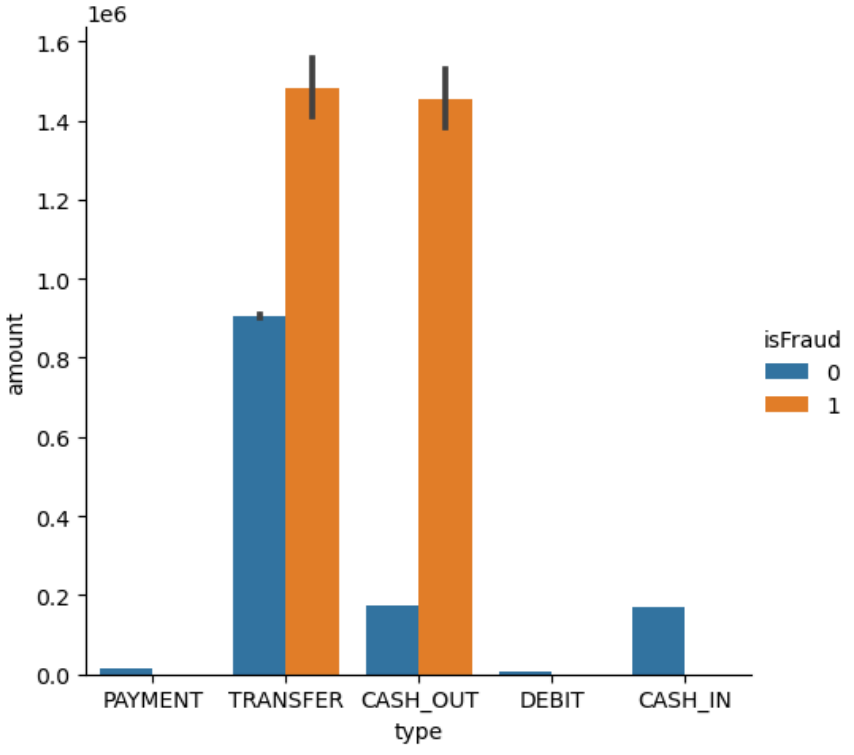
PairGrid Observations:

- There was not a lot of correlation between predictors, partially due to fraud being a minority occurrence in our dataset
- OldBalanceDest: The destination accounts for transactions that were fraud were to SMALLER amount accounts



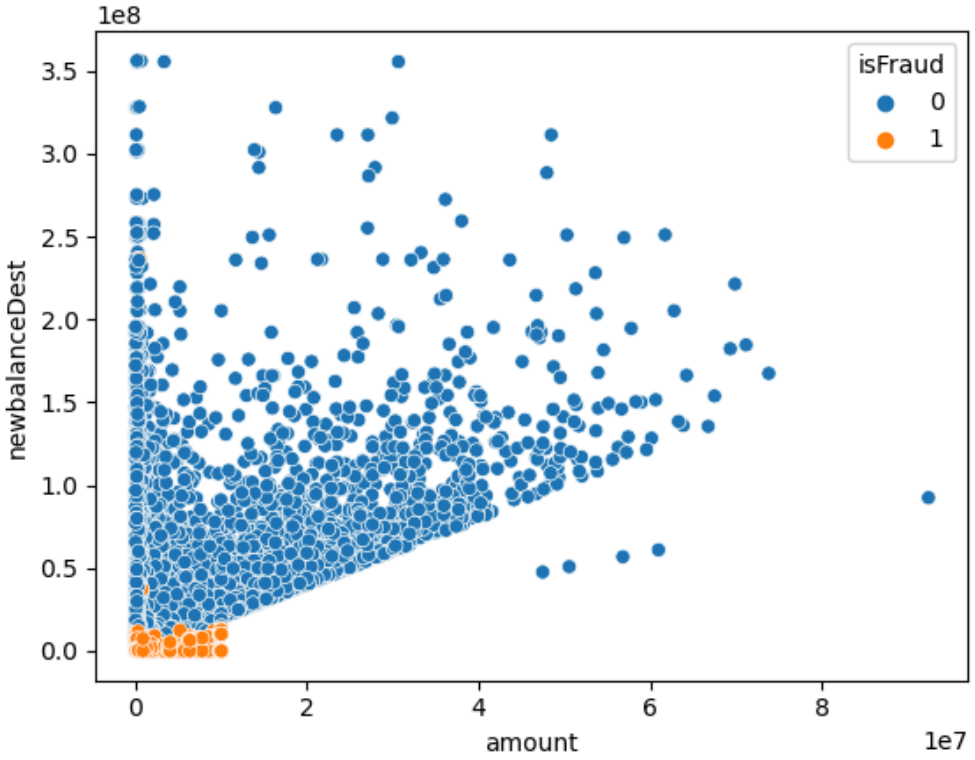
Bar Chart Type vs Amount Observations:

- This chart visualizes that most of the fraud transactions occur with the transfer of money.



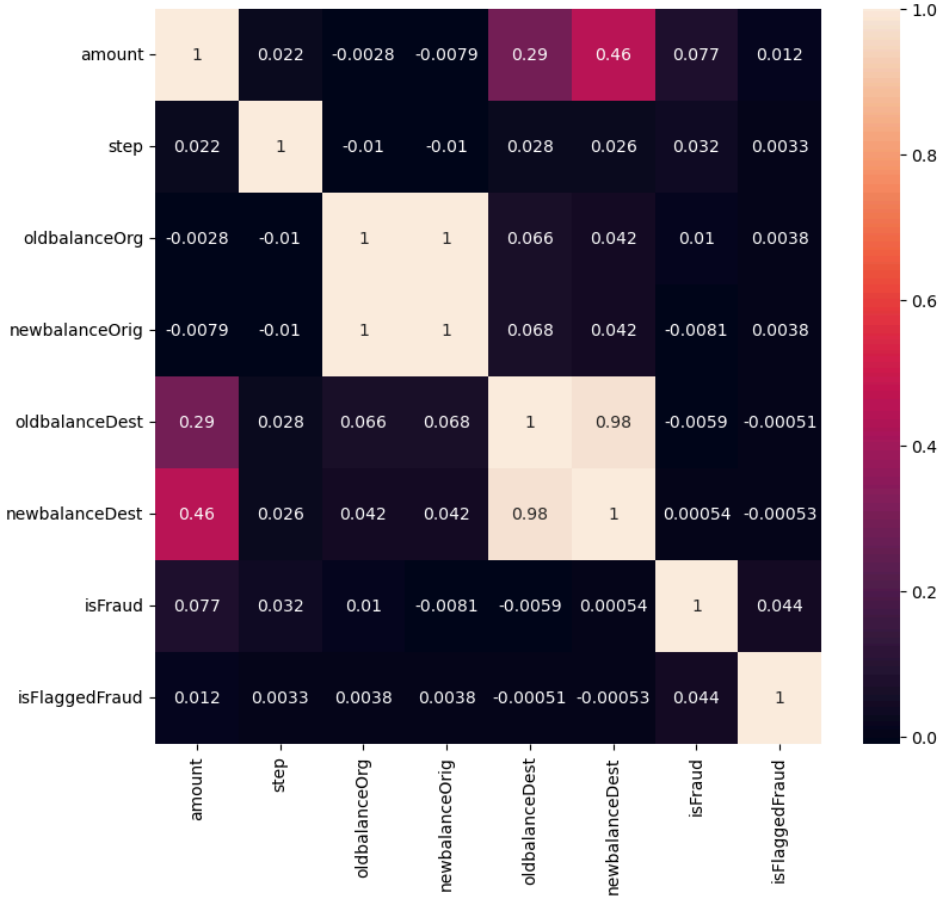
Scatterplot Amount vs Newbalancedest Observations:

- This scatterplot displays that most of the fraud cases occur with smaller amounts of money transferred to a destination account.



Heatmap Observations:

- This heatmap shows that some of the columns in our dataset have a high correlation (scores of 1) meaning that we can choose to drop them when cleaning our data to improve model processing.



Data Wrangling

The dataset then underwent preprocessing to encode categorical variables and drop columns in the dataset that would not contribute to our predictive modeling. During this phase, I decided to drop certain columns from the dataset based on what had been observed from exploring the data. These are the columns that were removed:

- **nameOrig and nameDest:** These columns contain account names, but we don't think they'll give us much insight into whether a transaction is fraudulent.
- **step:** This column represents the timestamp of transactions, but knowing when transactions occur doesn't seem to help us predict fraud.
- **newbalanceOrig and newbalanceDest:** These columns show account balances after transactions, but since we only know these values after the transaction takes place, they can't help us predict fraud beforehand.
- **isFlaggedFraud:** While this column might seem useful, this column reflects a naive model that flagged accounts for Fraud if the amount was greater than 200,000. This will not be useful for our predictive modeling.

Modeling

Random Forest Classifier was chosen because of its ability to handle imbalanced datasets effectively. The Random Forest Classifier was trained on the preprocessed dataset, utilizing techniques such as cross-validation to optimize hyperparameters and prevent overfitting. The model was trained to maximize sensitivity to capture fraudulent transactions while maintaining a low false positive rate.

In our initial model using a Random Forest Classifier, we obtained the following values:

- Model accuracy score of 0.999598435864471
- F1 score of 0.8248200205690779

This indicates that the model correctly classified approximately 99.96% of the observations and also that the model achieved a good balance between minimizing false positives and false negatives with a high F1 score. The discrepancies between the two values might be due to overfitting our model to the training data.

After proceeding with hyperparameter tuning using RandomSearchCV, we obtained the following values:

- Model accuracy score of 0.9994797740553419
- F1 score of 0.7555391432791729

After tuning our model, the F1 score decreased. This is an indicator that there was a trade-off with minimizing false positives.

Results and Discussion

This project aimed to develop a fraud detection system for banking transactions using the Random Forest Classifier algorithm. Initial exploratory data analysis (EDA) informed preprocessing steps, including dropping irrelevant columns. The initial model achieved high accuracy and F1 scores. However, after hyperparameter tuning, the F1 score decreased slightly, indicating a trade-off with minimizing false positives.

Some next steps could include:

- Fine-Tune Hyperparameters: Continue fine-tuning hyperparameters to find the optimal balance for model performance.
- Alternative models: Experiment with other supervised learning algorithms that support imbalanced datasets
- Feature Engineering: Explore additional features or transformations that could enhance model performance