Final Project – Credit Card Fraud

Due date: 12/4/2024

The objective of this final project study is to apply the concepts and tools learned during the semester in a real-case scenario. In this project, you will develop, compare, and choose the best option from supervised machine learning models.

Problem description

The final project for the Predictive Analytics course is designed to build, compare and interpret predictive models using a real[-ish] world dataset. For this project, you will use the provided credit card transactions data. Note that this dataset loosely resembles real transactional data from a major credit card company in the US.

Fraud is a problem for any bank. Fraud can take many forms, whether it is someone stealing a single credit card, to large batches of stolen credit card numbers being used on the web, or even a mass compromise of credit card numbers stolen from a merchant via tools like credit card skimming devices. Each of the transactions in the dataset has a field called *isFraud* (target variable).

For this project, you will need to upload on blackboard:

1. **Executive summary** (.pptx) should contain a maximum of 5 slides summarizing:
    a. Data descriptive analysis
    b. Modeling methodology and development
    c. Model performance
    d. Model interpretability
    e. Final recommendation

2. **Python notebook** (.ipynb) on Backboard as **LastName_Final.ipynb**. the notebook must address the questions below:

Set up

a. download and load your data (use your tool of choice: Google Collab or Jupyter notebook). I have provided the code for this step. Due to the size of the data, you will work with a sample of it (50%) – for the random state, use the last 3 digit (excluding any leading zeros) of you student ID. This data is in line-delimited JSON format

b. Please describe the structure of the data. Please provide summary statistics for each feature (Be sure to include a count of null, minimum, maximum, and unique values where appropriate.)

Data cleaning and preparation

a. Identify if there are missing/null values and display number of missing values per column. Choose to keep, replace or drop missing values. Explain why.

b. You will notice a number of what look like duplicated transactions in the data set. One type of duplicated transaction is a reversed transaction, where a purchase is followed by a reversal.

Another example is a multi-swipe, where a vendor accidentally charges a customer's card multiple times within a short time span.

I. Can you identify reversed and multi-swipe transactions?

II. What total number of transactions and total dollar amount do you estimate for the reversed transactions? For the multi-swipe transactions? (please consider the first transaction to be "normal" and exclude it from the number of transaction and dollar amount counts)

Descriptive analysis

Visualize the data in a way that you find most helpful to understand the relation between the features and the target (fraud). Elaborate your initial hypothesis for which variables appear to have a significant impact on the target variable. You need to provide your insights and summarize them in the notebook to get credit

Machine Learning modeling data prep

Note: The standardization of continuous variables is a common requirement for machine learning model development when the scale of variables varies significantly. However, for this case, the scale of the continuous variables does not appear to need standardization. Thus, it is not needed to be conducted for this case.

1. Drop any variable that is redundant, empty, or not relevant for the model development
2. Conduct the one-hot encoding of the categorical variables of the model
3. Conduct the standardization of continuous variables of the model, if you deem necessary
4. Create any additional feature or flag you think may be useful in the prediction of fraud
5. Split dataset into: `X` (features or independent variables) and `Y` (target or dependent variable)
6. Split the datasets into *X_train, X_test, y_train*, and *y_test*. Use a train-test random split of %70-%30.

Machine Learning Modeling: GBM and XGBoost

You will be developing and tuning two machine learning models: GBM and XGBoost. For each model, you should optimize and find the best hyperparameter combination to achieve the best model. You should provide:

a. Precision, recall and f1-score on the test sample
b. Confusion matrix
c. ROC curve and estimate the AUC
d. Plot feature/variable importance

Which model performed best and you recommend to the upper management of the company. Please compare the model performance and elaborate on your recommendation.

Machine Learning Interpretability: SHAP

For the final best model, create and interpret the following plots from the SHAP library:

a. Partial dependence (shap.partial_dependence_plot) for the top 5 variables
b. Mean shap bar plot
c. Beeswarm plot
d. Waterfall plot for the nth observations in the data (for n, use the last 3 digits in your student id)