Columbia University in the City of New York
MATHGR5430 Machine Learning for Finance, Spring 2022

Homework 2

Prof. Silva

Due by 02/20/2022

Please submit a single .pdf-file generated from your Jupyter-Notebook on CourseWorks.

1 Background

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days. The dataset has two classes, the positive class represents fraudulent transactions. It contains only numeric input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues. Features V1, V2,..., V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction amount. Feature *Class* is the **response variable** and it takes value 1 in case of fraud and 0 otherwise.

2 Exploratory Data Analysis

Load your data and perform simple exploratory data analysis (EDA).The data is located in the Files section of the canvas course page.
It is called hw2_credit_card_fraud.zip. It is a standard csv file. Answer the following:

1. What do you notice about the dataset?
2. What is the problem with what you observed and how can that affect the training of a ML model?
3. What is the difference and trade off of oversampling vs undersampling?

4. Plot the distributions for Amount, Transaction Time, and one of the other variables. What do you see and does it make sense? Plot two of the Vx variables together one on the x axis and the other on the y axis, what do you see?

5. What other thoughts do you have about the data, that if you had time you would investigate?

3 Classifying credit fraud

1. Split your data into train/test (Hint: from sklearn.model_selection import train_test_split).

2. What is the balance within the Class variable? (i.e. how many positives and negatives are there, is this relationship balanced?)

3. Perform a sampling method, based on your analysis above I.4, on only the training set.

4. Train a Logistic **Lasso** Regression model (Hint: from sklearn.linear_model import LogisticRegression).

5. What is the goal of running Logistic Lasso Regression?

4 Evaluation of results

1. Show the confusion matrix

2. Report performance metrics Accuracy, Precision, Recall

3. What metric would you maximize in search of the best performing model?

5 Bonus: Parameter Optimization

1. Use the sklearn documentation to perform a simple cross validation and grid search to optimize the parameters to your model. What is the lift you get from performing these additional steps? (Hint: from sklearn.model_selection import GridSearchCV, cross_val_score)