Machine Learning in Finance Assignment 2

Due date: 2/25/2023

February 15, 2023

1 Setup and Data Fetching (5)

- Goal: Learn and use different models to solve a multi-class classification problem. Please submit your code and a single .pdf-file generated from your Jupyter-Notebook on CourseWorks.
- Background: This data collects personal information of clients. The management wants to build an intelligent system to segregate the people into credit score brackets to reduce the manual efforts.
- Obtain and import the dataset (*HW2.csv*) from Courseworks as a Pandas dataframe.

2 Exploratory Data Analysis (20)

- Data cleaning: Some features have missing or invalid values. Choose the way you see as appropriate to clean the dataframe so that there are no missing or invalid values.
- Plot the distributions of two numerical features of your choice. What do you discover about those variables?
- Plot the correlation matrix of all numeric features. What do you discover?
- Appropriately encode all categorical features in the dataframe

3 Logistic regression (20)

- What is the use of validation dataset in machine learning?
- Make an appropriate split of the data $train, validation, test$ and fit a multi-class logistic regression model using the Scikit-learn library.
- Print the classification report of the model. What do you find in the report?

4 Decision Tree and Bagging (30)

- Describe the algorithm CART. What are the advantages and disadvantages of CART?
- Implement DecisionTreeClassifier from the sklearn library to train one decision tree. You can evaluate the accuracy of the validation set to tune model parameters. You should only evaluate your final accuracy on the test dataset.
- Use sklearn's *sklearn.tree.plot_tree* method and *matplotlib* to visualize your classification tree.
- Use 30 different random seeds to train 30 identical decision trees and record the test accuracies. Calculate and report the average accuracy and standard deviation across the 30 runs. What do you find using this bagging method?

5 Random Forest (25)

- What is the difference between bagging and random forest?
- Why is it important for individual estimators in the random forest to have access to only a subset of all features?
- Implement a random forest classifier to solve the classification problem again.
- Compare and contrast models from Section 3, 4 and 5.