Machine Learning in Finance Assignment 1

Due date: 2/12/2023

January 30, 2023

1 Setup and Data Fetching (5)

- Goal: Use machine learning to solve a binary classification problem using provided dataset. Please submit your code and a single .pdf-file generated from your Jupyter-Notebook on CourseWorks.
- Background: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).
- Obtain and import the dataset (*bank-additional-full.csv*) from Courseworks as a Pandas dataframe.
- Use *bank-additional-names.txt* doc associated with the dataset as complement information to get familiar with attributes.

2 Data Preprocessing (25)

- What is the balance within the y variable? (i.e. how many positives and negatives are there, is this relationship balanced?) How would you consider improving the dataset given your observations?
- Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques. (To-do) Choose the way you see as appropriate to clean the dataframe so that there are no missing values.
- Plot the distributions of two numeric independent variables of your choice. What do you discover about those variables?
- Plot the correlation matrix of all numeric features. What do you discover? How should you handle it in that situation?
- Appropriately encode all categorical features in the dataframe

3 Model Configuration and Testing (50)

- Split the data into train and test using 20% test size.
- Fit a logistic regression model using the Scikit-learn library (Hint: from sklearn.linear model import LogisticRegression).
- How does the model perform? Plot the confusion matrix and the ROC curve of both train and test dataset.
- Print the classification report of the model.
- What do you find in the report? And what is the reason behind that? What metric would you maximize in search of the best-performing model?

4 Logistic regression (20)

- Explore attributes of your model. Print out the coefficients, intercept of the model, as well as probability estimates of test dataset.
- Use linear combination and sigmoid function to calculate probability estimates of test dataset manually and check if the two match.

  - intercept: a_0
  - coefficients: $a_1, a_2, \cdots, a_m$
  - input: $x_1, x_2, \cdots, x_n, \quad$ where $x_i = (x_{i1}, \cdots, x_{im}).i = 1, \cdots, n$
  - probability estimates: $y_i = \dfrac{1}{1 + e^{-(a_0 + \sum_{k=1}^{m} a_k x_{ik})}}, \quad i = 1, \cdots, n$

  What is the relationship between logistic regression and linear regression?

5 Bonus (20)

- What is overfitting? What is underfitting? Does the model you trained exhibit any of these problem?
- 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)