CF969-7-SU-CO
Big-Data for Computational Finance
Academic Year: 2023/24
Assignment 2

Before you begin

Please refer to the Student's handbook on the School's Policy on Plagiarism and Late Submission.

The two deliverables below must be uploaded on FASER by the deadline as independent items, i.e., not bundled together in a zip file.

Assignment – Purpose

In this assignment, you will use data from a random sample of loans obtained from a peer-to-peer lending platform with the goal to predict who will default at its loan using features provided in the data.

On the CF969 moodle page ("Assessment Information" tab) there are three .csv data files:

- trainData: This is the dataset on which you will train all your models.
- testData: This is the dataset on which you will evaluate your model's fit.
- varDescription: This contains a description of the features

Task 1 (40%)

This task requires estimating several models to identify the best model to predict default based on the available data.

Create a new variable from trainData called "y" which takes the value = 1 if the column "loan status" has the value "Charged Off" and 0 otherwise. All other variables provided to you other than the loan status are features or "predictors". Consider whether you would like to transform your variables; for example, consider converting some of the categorical variables into a continuous variable.

1.1 *Linear Regression Model*

Fit a linear regression model to the trainData, with y as the outcome variable, with all the predictors.

(a) What is the Mean Squared Error for the training data?

(b) What is the Mean Squared Error for the testing data?

1.2 *Ridge Regression Model*

Fit a ridge regression model to the trainData, with y as the outcome variable, with the predictors. Explore all values of hyperparameter λ (lambda) ranging from 0.01 to 3 with an increment of 0.01.

(a) What is the Mean Squared Error for the "best" model of this class for the training data?

(b) What is the Mean Squared Error for the "best" model of this class for the test data?

1.3 *Lasso Regresion Model*

Fit a Lasso regression model to the trainData, with y as the outcome variable, with the predictors. Explore the same values of hyperparameter λ (lambda) as for the ridge regression.

(a) What is the Mean Squared Error for the "best" model of this class for the training data?

(b) What is the Mean Squared Error for the "best" model of this class for the test data?

1.4 *Random Forest*

Fit a randomForest to the trainData, with y as the outcome variable, with the predictors. Explore and fit the best model of this class. Please explain any estimation assumptions you take.

(a) What is the Mean Squared Error for the "best" model of this class for the training data?

(b) What is the Mean Squared Error for the "best" model of this class for the test data?

(c) How important are the variables in predicting default?

1.5 Neural Network

Select a Neural Network model and follow the same process as in the previous models.

(a) What is the Accuracy for the model for the training data?
(b) What is the Accuracy for the model for the testing data?
(c) Explain why you chose the particular Neural Network model

1.6 Evaluation

Compare and contrast the predictive power of all approaches and identify the best model to predict default from the given data.

Task 2 (60%)

Task 2 is where you present your answers to the questions stated in Task 1. You are required to combine all the work done for Task 1 and submit a report for predicting default for borrowers from the data platform. Discuss also how the variables are correlated with the "loan status". Identify the 10 most correlated and the 10 least correlated variables. Exploit all the information you generated for Task 1 to write **a report no longer than 5 pages** (excluding any references or optional appendices) and present your best model. Pay attention to explaining why it is the best model and present the best model's performance compared to the other models on hand.

What am I grading you on?

On your ability to use machine learning models learned during the second part of the module and on your ability to formulate your insights in a concise way into a report. Your ability to implement the models is less than half of the challenge. Your interpretation of the findings matters substantially.

Deliverables for Assignment 2:

- A document of no more than 5 pages discussing your implementation, your choices, and the results. This document **must** be a .pdf file
- The source code, either a .py file or a .ipynb file.