# CF969-7-SU-CO

# **Big-Data for Computational Finance**

# Academic Year: 2023/24

# **Assignment 02**

|  |  |
| --- | --- |
| **Name:** |  |
| **ID:** |  |
| **Date:** |  |

# **Contents**

1. Overview
2. Data Analysis
   1. Description of Training Data
   2. Description of Testing Data
3. Methodological Approach
   1. Data Cleaning and Preparation
   2. Model Building and Evaluation
4. Predictive Models
   1. Linear Regression
   2. Random Forest
   3. Ridge Regression
   4. Lasso Regression
   5. Neural Network
5. Comparative Analysis of Models
6. Correlation Study
   1. Variables with High Correlation
   2. Variables with Low Correlation
7. Summary and Insights

# **Overview:**

The rise of peer-to-peer (P2P) loaning stages has changed loaning by straightforwardly interfacing borrowers with loan specialists, expanding get to to credit. Be that as it may, this too presents the hazard of advance defaults, affecting lenders' returns and stage steadiness. This task points to foresee advance defaults utilizing information from a P2P loaning stage by building and assessing different machine learning models. The objective is to supply experiences to banks and the stage to move forward decision-making and hazard administration.

# **Data Analysis**

**2.1. Description of Training Data**

The training dataset is used to develop and train models for predicting loan defaults. It includes various features such as loan amount, interest rate, monthly installment, loan grade, employment length, home ownership status, annual income, credit status, debt-to-income ratio, number of delinquencies, revolving balance, total payments, interest received, and more. This dataset may contain missing or inconsistent values, necessitating thorough preprocessing.

**2.2. Description of Testing Data:**

The testing dataset is used to evaluate the performance of the models trained on the training data. It has the same features as the training data, such as loan amount, interest rate, monthly installment, loan grade, employment length, home ownership status, annual income, loan status, debt-to-income ratio, and number of delinquencies. This dataset helps assess how well the models generalize to new, unseen data. Like the training data, it may also contain unclean data that requires preprocessing.

# **Methodological Approach**

* 1. **Data Cleaning and Preparation:**

Preparing the dataset for model training involves several steps:

* + 1. **Column Verification:** Ensure the column names match the expected schema to identify any inconsistencies or unexpected columns.
    2. **Handling Missing Values:** Identify and address missing values using strategies like imputation with mean, median, or mode for numerical columns, and using mode or a placeholder for categorical columns.
    3. **Conversion of Categorical Variables:** Convert categorical variables into numerical representations using techniques such as one-hot encoding or label encoding.
    4. **Value Distribution Analysis:** Examine value counts and unique values for each column to understand the data distribution and identify any anomalies.
    5. **Outlier Detection and Handling:** Detect and address outliers in numerical columns to prevent skewing the analysis or model training.
    6. **Normalization and Scaling:** Normalize or scale numerical columns to ensure they have a similar range, improving model performance. Common techniques include min-max scaling and standardization.

# **Predictive Models**

* 1. **Linear Regression**

Linear regression is an essential AI model that predicts the objective variable in light of direct connections between the information highlights and the objective.

|  |  |  |
| --- | --- | --- |
| a) | **Mean Squared Error for training data:** | 0.06786949239601085 |
| b) | **Mean Squared Error for testing data:** | 0.06865236683288277 |

* 1. **Ridge Regression**

Ridge regression is a regularized version of linear regression that incorporates a punishment term to forestall overfitting by contracting the coefficients.

|  |  |  |
| --- | --- | --- |
| a) | **Best lambda:** | 0.01 |
| b) | **MSE for training data with best lambda:** | 0.06786987720806789 |
| c) | **MSE for test data with best lambda:** | 0.06865289293016083 |

* 1. **Lasso Regression**

Lasso regression is another regularized linear model that incorporates a penalty term, which can likewise result in highlight determination by contracting a few coefficients to nothing.

|  |  |  |
| --- | --- | --- |
| a) | **MSE for training data with best lambda:** | 0.06973031757438113 |
| b) | **MSE for test data with best lambda:** | 0.07052252421885377 |

* 1. **Random Forest**

Random forest is a troupe learning strategy that builds multiple decision trees and merges their results to improve predictive accuracy and control overfitting.

|  |  |  |
| --- | --- | --- |
| a) | **Mean Squared Error for training data:** | 0.0 |
| b) | **Mean Squared Error for testing data:** | 0.0 |

* 1. **Neural Network**

Neural networks are complex models inspired by the human brain, capable of capturing non-linear relationships in the data through multiple layers of neurons.

|  |  |  |
| --- | --- | --- |
| a) | **Neural Network Training MSE:** | 60203852.74096 |
| b) | **Neural Network Test MSE:** | 41017038.308935866 |

# **Comparative Analysis of Models**

Evaluate the performance of the different models based on metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve. Compare and contrast their strengths and weaknesses to identify the best-performing model.

Model best\_params mse\_train mse\_test mae\_test r2\_test

0 LinearRegression {} 1.543633e-31 0.078018 0.151777 0.133137

1 RidgeRegression {‘model\_\_alpha’: 0.01} 2.676650e-08 0.078075 0.151792 0.132498

2 LassoRegression {‘model\_\_alpha’: 0.01} 1.000000e-04 0.099344 0.102667 -0.103827

3 RandomForest {‘model\_\_max\_depth’: 10, ‘model\_\_min\_samples\_l… 1.604000e-02 0.074810 0.121000 0.168778

4 NeuralNetwork {‘model\_\_activation’: ‘tanh’, ‘model\_\_alpha’: … 1.414540e-03 0.353705 0.468059 -2.930056

# **Correlation Study**

* 1. **Variables with High Correlation**

Identify variables that have strong correlations with loan defaults, such as interest rate, debt-to-income ratio, revolving line utilization, monthly instalment, total payment, total interest received, late fees, recoveries, and collection recovery fees.

* 1. **Variables with Low Correlation**

Identify variables with weak correlations to loan defaults, such as loan ID, member ID, application type, home ownership status, total principal received, collections excluding medical expenses, current delinquent accounts, total collection amount, total current balance, and total revolving high credit limit.

# **Summary and Insights**

The test gives important information for lenders and the P2P lending platform to make better decisions and manage risks. Important findings show that the credit history of the borrower, their ability to make consistent payments, and their debt compared to their income are all important in predicting loan defaults. These experiences support using better screening forms and more accurate estimating models to reduce risks effectively. We should use real-time analytics and helpful resources to get lenders interested. In the future, changes will focus on improving prediction models, conducting long-term studies, and using advanced machine learning methods to better assess risks and improve strategies.