# CF969-7-SU-CO

# **Big-Data for Computational Finance**

# Academic Year: 2023/24

# **Assignment 2**

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

## **Table of Contents**

1. Introduction
2. Data Description
   * Train Data
   * Test Data
3. Methodology
   * Data Preprocessing
4. Models and Analysis
   * Linear Regression
   * Ridge Regression
   * Lasso Regression
   * Random Forest
   * Neural Network
5. Model Evaluation and Comparison
6. Correlation Analysis
   * Most Correlated Variables
   * Least Correlated Variables
7. Conclusion

# **Introduction:**

In today's monetary scene, peer-to-peer (P2P) loaning stages offer an elective to conventional managing an account by specifically interfacing borrowers with banks, growing get to credit for numerous people and businesses. Be that as it may, this moreover presents the hazard of advance defaults, affecting lenders' returns and stage solidness. This venture points to foresee credit defaults utilizing information from a P2P loaning stage by building and assessing different machine learning models to decide the foremost compelling one. The objective is to supply important experiences for moneylenders and the stage to move forward decision-making and hazard administration forms.

# **Data Description:**

1. **Train Data:**

The train Data dataset trains models for predicting loan defaults, including features like loan amount, interest rate, and employment length. It also contains loan status, debt-to-income ratio, and payment information. The dataset is unclean, requiring pre-processing to handle missing or inconsistent values.

1. **Test Data:**

The test Data dataset evaluates the models trained on train Data, containing the same features like loan amount, interest rate, monthly instalment, and loan status. It helps assess model performance on new data and may also require preprocessing due to unclean data.

1. **Methodology:**

**Data Preprocessing**

The information preprocessing steps include planning the dataset for show preparing by taking care of lost values, changing categorical factors, and guaranteeing the information is clean and reliable. Here may be a nitty gritty portrayal of the steps taken:

* ***Column Confirmation:***
* Begin by confirming the column names to guarantee they coordinate the anticipated construction. This helps identify any inconsistencies or startling columns within the dataset.
* ***Dealing with Lost Values:***
* Check each column for invalid or lost values. Utilize procedures such as tallying a number of lost values in each column to induce a diagram of information quality.
* **Handle lost values based on the sort of information:**
* For numerical columns, consider filling lost values utilizing the cruel, middle, or mode of the individual column.
* For categorical columns, fill lost values with the mode (most visit esteem) or a placeholder esteem demonstrating lost information.
* ***Categorical to Continuous Conversion:***
* Recognize categorical factors such as review and homeownerships.
* Change over these categorical factors into numerical representations using strategies like one-hot encoding or name encoding.
* ***Value Counts and Unique Values:***
* For each column, look at the esteem tallies and special values to get it the conveyance of the information. This makes a difference in recognizing irregularities and understanding the spread of diverse categories or numerical ranges.
* ***Outlier Detection and Handling:***
* Recognize any exceptions in numerical columns that might skew the examination or show preparing. Choose whether to evacuate, cap, or change these exceptions.
* ***Normalization and Scaling:***
* Normalize or scale numerical columns to guarantee they have a comparative extend, which makes a difference progress demonstrate execution. Common strategies incorporate min-max scaling and standardization.

# **Models and Analysis:**

* **Linear Regression**

Linear regression is a fundamental machine learning model that predicts the target variable based on linear relationships between the input features and the target.

Linear Regression Training MSE: 5.120248943261626e-30

Linear Regression Test MSE: 0.049999386477793305

* **Ridge Regression**

Ridge regression is a regularized version of linear regression that includes a penalty term to prevent overfitting by shrinking the coefficients.

Best Ridge Regression Training MSE: 4.5211051419650384e-11

Best Ridge Regression Test MSE: 0.04999975542616624

* **Lasso Regression**

Lasso regression is another regularized linear model that includes a penalty term, which can also result in feature selection by shrinking some coefficients to zero.

Best Lasso Regression Lambda: 3.0

Best Lasso Regression Training MSE: 6.584115154947355e-05

Best Lasso Regression Test MSE: 0.04988445651187996

* **Random Forest**

Random forest is an ensemble learning method that forms numerous decision trees and consolidates their outcomes to improve predictive accuracy and control overfitting.

Random Forest Training MSE: 0.006244999999999999

Random Forest Test MSE: 0.040955

Random Forest Feature Importance:

[0. 0. 0. 0. 0. 0.

0. 0. 0. 0.01492537 0. 0.13432836

0. 0.10447761 0. 0. 0. 0.02985075

0. 0. 0.34328358 0.37313433 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0.]

* **Neural Network**

Neural networks are mind boggling models propelled by the human cerebrum, fit for catching non-linear connections in the information through different layers of neurons.

Neural Network Training MSE: 60203852.74096

Neural Network Test MSE: 41017038.308935866

# **Model Evaluation and Comparison:**

Assess the presentation of the various models in light of measurements, for example, exactness, accuracy, review, F1-score, and AUC-ROC bend. Look into their strengths and shortcomings to distinguish the best-performing model.

Best Model Based on Test MSE:

Model: Random Forest

Test MSE: 0.040955

# **Correlation Analysis:**

* **Most Correlated Variables:**

1. **int\_rate:** Higher interest rates are as often as possible connected with higher risks, consequently relating with loan defaults.
2. **dti:** A higher debt-to-income ratio demonstrates higher monetary pressure, making default more probable.
3. **revol\_util:** High revolving line utilization suggests over-influence, expanding the gamble of default.
4. **installment:** Greater consistently planned instalments may be all the more constantly for borrowers to make due, provoking higher default rates.
5. **total\_pymnt:** Total payment made can demonstrate the borrower's capacity to reimburse advances; lower instalments might relate with defaults.
6. **total\_rec\_int:** Total interest received might indicate the length and repayment illustration of advances.
7. **total\_rec\_late\_fee:** High late fees suggest past portion issues, exhibiting higher default risk.
8. **recoveries:** Sums recovered from previous defaults correlate with loan performance.
9. **collection\_recovery\_fee:** High recovery charges might relate with incessant defaults and assortment endeavours.
10. **loan\_amnt:** Larger loan amounts may be all the more difficult for borrowers to repay, further developing default likelihood.

* **Least Correlated Variables:**

1. **id:** “Unique loan identifier”, has no significance to loan performance.
2. **member\_id:** “Unique member identifier”, inconsequential to loan status.
3. **application\_type:** Whether the application is individual or joint doesn't emphatically influence default rates.
4. **home\_ownership:** The ownership status of the home may not essentially influence the probability of default.
5. **total\_rec\_prncp:** Total principal received does not firmly show future defaults.
6. **collections\_12\_mths\_ex\_med:** Collections in the past 12 months excluding clinical expenses may not fundamentally influence default prediction.
7. **acc\_now\_delinq:** Current delinquent accounts might not have major areas of strength with new loan defaults.
8. **tot\_coll\_amt:** Total collection amount may not be unequivocally demonstrative of loan performance.
9. **tot\_cur\_bal:** Total current balance could insignificantly affect credit status.

10.**total\_rev\_hi\_lim:** Total revolving high credit limit has little impact on predicting loan defaults.

# **Conclusion:**

The examination uncovers vital bits of knowledge for moneylenders and the P2P loaning stage to optimize decision-making and hazard administration. Key discoveries emphasize the importance of borrower credit history, pay solidness, and debt-to-income proportions in anticipating advance defaults. These experiences advocate for improved screening forms and energetic estimating models to relieve dangers viably. Suggestions incorporate executing real-time analytics systems and instructive assets to engage moneylenders. Future changes seem centre on refining prescient models, conducting longitudinal considers, and joining progressed machine learning strategies for more nuanced chance appraisal and methodology refinement.