# 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
8. **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 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.

1. **Data Description:**

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 preprocessing to handle missing or inconsistent values.

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 the number of lost values in each column to induce an diagram of the 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 utilizing methods 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.

1. **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.

* Ridge Regression

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

* 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.

* Random Forest

Random forest is an ensemble learning method that builds multiple decision trees and merges their results to improve predictive accuracy and control overfitting.

* 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.

1. **Model Evaluation and Comparison:**

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.

1. **Correlation Analysis:**

* **Most Correlated Variables:**

1. int\_rate: Higher interest rates are often associated with higher risk, thus correlating with loan defaults.
2. dti: A higher debt-to-income ratio indicates higher financial stress, making default more likely.
3. revol\_util: High revolving line utilization suggests over-leverage, increasing the risk of default.
4. installment: Larger monthly instalments may be harder for borrowers to manage, leading to higher default rates.
5. total\_pymnt: Total payment made can indicate the borrower’s ability to repay loans; lower payments may correlate with defaults.
6. total\_rec\_int: Total interest received might indicate the duration and repayment pattern of loans.
7. total\_rec\_late\_fee: High late fees suggest previous payment issues, indicating higher default risk.
8. recoveries: Amounts recovered from past defaults correlate with loan performance.
9. collection\_recovery\_fee: High recovery fees may correlate with frequent defaults and collection efforts.

10.loan\_amnt: Larger loan amounts may be more difficult for borrowers to repay, increasing default likelihood.

* **Least Correlated Variables:**

1. id: Unique loan identifier, has no relevance to loan performance.
2. member\_id: Unique member identifier, unrelated to loan status.
3. application\_type: Whether the application is individual or joint doesn’t strongly affect default rates.
4. home\_ownership: The ownership status of the home may not significantly impact the likelihood of default.
5. total\_rec\_prncp: Total principal received does not strongly indicate future defaults.
6. collections\_12\_mths\_ex\_med: Collections in the past 12 months excluding medical expenses may not significantly impact default prediction.
7. acc\_now\_delinq: Current delinquent accounts may not have a strong correlation with new loan defaults.
8. tot\_coll\_amt: Total collection amount may not be strongly indicative of loan performance.
9. tot\_cur\_bal: Total current balance might have a minimal effect on loan status.

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

1. **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 center on refining prescient models, conducting longitudinal considers, and joining progressed machine learning strategies for more nuanced chance appraisal and methodology refinement.