REPORT

Big Data Analysis for Computational Finance (CF969-7-SU-CO)

INTRODUCTION

The purpose of the assignment is to develop and compare forecasting models to forecast credit defaults based on multiple parameters, including annual income, home ownership, loan amount, interest rate, and service life. The provided dataset contains traindata for model training, testdata for model evaluation. The main goal is to predict the default situation of the loan, especially if the loan will be repaid using different machine learning techniques. Accurate forecasting models can significantly reduce financial losses caused by negligence. In this context, "discharged" refers to loans that the credit system deems unlikely because the borrower is unable to meet his repayment obligations. By identifying these high-risk loans in advance, the company can make informed decisions about loan approval and interest rates. These methods include linear regression, Ridge regression, lasso regression, random forest, and neural networks. Each model is trained and evaluated on the respective data.

PRE PROCESSING

The preprocessing steps for handling missing values and encoding categorical variables were crucial for the train_data and test_data datasets. In train_data, missing values in numerical features were imputed with the mean, while missing values in categorical columns were filled with the mode. On the other hand, the test_data dataset initially filled missing values using forward fill, followed by mean imputation for numerical columns and mode imputation for categorical columns. Both datasets underwent label encoding for categorical variables to unify their representation in numerical form. These steps were essential in ensuring the datasets were properly prepared for machine learning model training and evaluation. The collected data was thus cleaned, complete, and in a structured format suitable for predictive model development.

LABEL ENCODING

The code encodes the categorical columns of both the training and test datasets to convert them to numbers. Categorical columns in the training dataset are initially identified using the `select_dtypes (include=['object'])' method. The code then iterates over these identified columns, applying the "fit_transform" parameter to the training data and the "transform" parameter to the test data. This approach ensures a consistent numerical

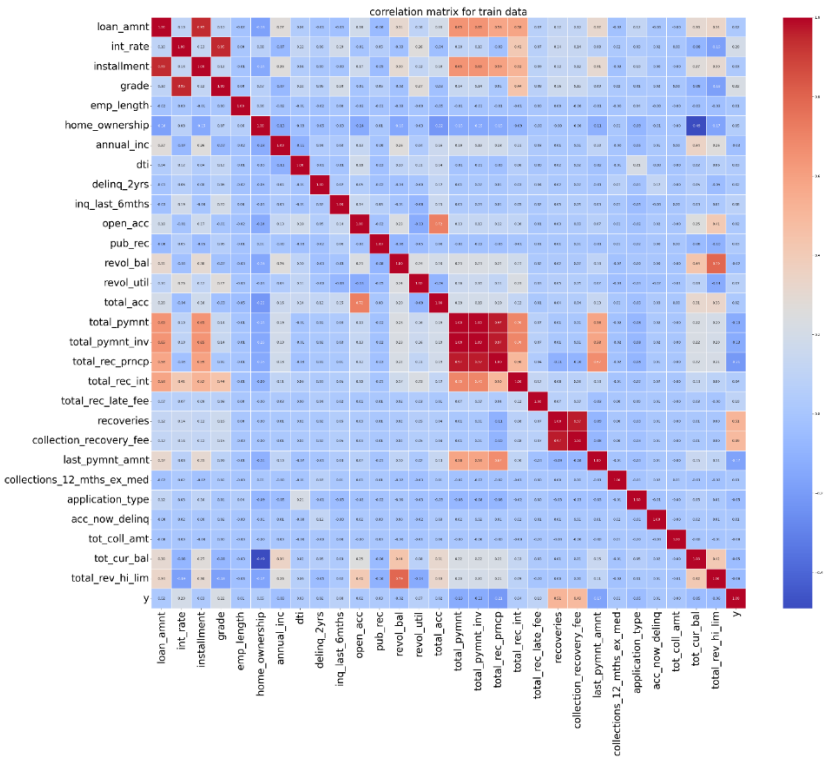
mapping of categorical variables in both datasets. Next, the "loan_status" column is removed from both the training and test datasets because it is assumed to be a target variable and is not needed for further processing in this context. This procedure ensures that all categorical variables are coded appropriately, facilitating their use in subsequent machine learning model training and evaluation.

VISUALIZATION

Visualizations provide a comprehensive overview of both numerical and categorical characteristics of the dataset. The first set of graphs illustrates the distribution of various numerical properties, highlighting their distribution, possible skewness and outliers. This is followed by a breakdown of categorical functions such as emp_length, application_type, home_ownership, and grade. These bar graphs show the frequency of different categories of each characteristic, showing the prevalence of certain categories. Finally, a plot looking at the relationship between annual income and interest rate shows a concentration of data points at lower income levels and different interest rates. This spread suggests a complex relationship between income and interest rates, as most applicants have lower incomes but a wide range of interest rates.

CORRELATION MATRIX

The correlation matrix provides a comprehensive visualization of the relationships between the various features of the combined train and test dataset. This highlights the strong positive correlation between features such as total_pymnt, total_rec_prncp and total_pymnt_inv, indicating that these financial metrics are closely related. In contrast, features such as acc_now_delin, delinq_2yrs, and pub_rec have weak negative correlations with other metrics. The analysis identified the top 10 traits that correlated most with the target variable y and the 10 traits that correlated the least with it, indicating the most and least influential predictors. This correlation analysis is crucial for feature selection and design and helps in the preprocessing and modeling stages by highlighting the most important features to predict the target variable. An annotated heat map with color gradients allows easy interpretation of the strengths and directions of these correlations.



The main task here is to find the 10 most correlated values and the 10 least correlated values, so we get the 10 best and 10 bottom correlated values. This is as follows:

Top 10 Correlated Values	Correlation Coefficient	Bottom 10 Correlated Values	Correlation Coefficient
Recoveries	0.514396	total_rec_prncp	-0.214704
collection_recovery_fee	0.490135	last_pymnt_amnt	-0.174928
grade	0.220607	total_pymnt	-0.130961
int_rate	0.196532	total_pymnt_inv	-0.130823
total_rec_late_fee	0.099707	total_rev_hi_lim	-0.057475
inq_last_6mths	0.084136	tot_cur_bal	-0.052145
revol_util	0.068672	application_type	-0.051029
home_ownership	0.049338	annual_inc	-0.034606
total_rec_int	0.040108	revol_bal	-0.019608
dti	0.033825	tot_coll_amt	-0.000018

DATA SPLITTING

Here we have to divide the data into two parts. The first is X_train, which contains only t he features while omitting the 'y' variable, and the second is Y_train, which contains the fixed variable, in this case 'y'. We did the same with our test data.

LINEAR REGRESSION

The Linear Regression section describes the use of this basic but powerful prediction m odel that analyzes the linear relationship between the target and independent variables t o reduce prediction error. The performance of the model is evaluated using the mean sq uared error (MSE) and the Rsquared index (R2), as a benchmark for more advanced m odels. The MSE value is 0.0686 for the training data and 0.0679 for the test data, indicat ing that the linear regression model captures the relationships between the variables we ll.

RIDGE REGRESSION

Ridge regression, a type of linear regression, regularizes the coefficients by adding a penalty term to the ordinary least squares (OLS) objective function. The accuracy achieved was 0.0679 for the test data and 0.0686 for the training data, which is comparable to linear regression. This accuracy represents the average squared distance between the actual target values and the predicted values.

LASSO REGRESSION

Lasso regression, also known as L1-regularized linear regression, is a statistical method used for regression analysis. It addresses the common problem of overfitting in linear regression. The mean squared error (MSE) values of 0.0696 for the training data and 0.0691 for the test data suggest that the model fits the data reasonably well, indicating a good fit between the variables.

RANDOM FOREST REGRESION

Random Forest is a powerful machine learning algorithm and a prominent ensemble learning technique used extensively in both regression and classification tasks. I have provided the tuning parameters: 100 estimators, a random state of 42, and a depth of 10. The resulting mean squared error (MSE) values—0.0273 for the training data and 0.1762 for the test data—are exceptionally low, indicating a highly promising performance. Therefore, compared to other models, the Random Forest model demonstrates the best fit for the data.

NEURAL NETWORK

The Neural Networks section describes the architecture and ability of neural networks to handle complex and nonlinear relationships in data. We cover layers, neurons, activatio n functions, and the learning process using regression to minimize prediction error. The ability of neural networks to learn complex patterns is a powerful tool for predictive tasks .

The Reason for Model Selection:

The following factors led to the selection of the Keras Sequential model for binary classification: its suitability for binary classification due to its specific design; its use of binary cross-entropy loss, which is perfect for binary classification tasks; its easily interpretable direct accuracy output as a metric; its balanced architecture (64 -> 32 -> 1), which strikes a balance between computational efficiency and model complexity; its use of a sigmoid activation function in the final layer, which outputs probabilities between 0 and 1; the flexibility Keras offers in model design, making it appropriate for both small and large datasets; and its ease of modification and experimentation with various architectures and hyperparameters.

CONCLUSION:

Final Comparison With models lowest test Mean Squared Error (MSE) of 0.1762, the Random Forest model outperformed the Linear Regression, Ridge Regression, Lasso Regression, and Neural Network models in forecasting loan defaults. Its ensemble learning method improves generalisability and minimises overfitting by averaging several decision trees. The model is the best option since it can accurately depict intricate non-linear linkages and interactions in financial data and offers insights into the significance of individual features. Future research should concentrate on enhancing the Random Forest model's tuning and investigating new ensemble methods in order to maintain and raise prediction accuracy.