CF969-7-SP-CO

Big-Data for Computational Finance

Academic Year: 2023/24

Assignment 2

Melek KURU – 2315873

1. Introduction

This project explores data from a website where individuals lend and borrow money directly, without involving banks. The main challenge is to accurately predict which borrowers are likely to not repay their loans. We use three different data sets: trainData.csv to train our models, testData.csv to test them, and varDescription.csv to understand the variables involved.

Our goal is to identify patterns that help us predict loan defaults. We have tested various methods of analysis, from examining detailed data such as income and loan size to trying out different prediction techniques, including machine learning methods.

This report aims to identify the best method for predicting who might default on their loans by comparing different strategies and explaining why some are more effective. We focus on the most important findings and what we have learned in a clear and concise manner.

2. Data Preprocessing

The first step in our analysis was to clean and prepare the data for modeling. This process, known as data preprocessing, involved several key tasks:

1. Reading the Data: We started by loading the `trainData.csv` and `testData.csv` files into Pandas DataFrames. This gave us a structured format to work with, making it easier to manipulate and analyze the data.

2. Handling Missing Values: Our dataset had missing values in several columns. We chose specific strategies for each:

- Columns with a high percentage of missing values (`id`, `member_id`, and `mths_since_last_delinq`) were removed, as their absence in most records made them unreliable for analysis.
- For columns with a moderate amount of missing data, such as `emp_length` and various financial metrics, we imputed missing values. Numeric columns were filled with the median value, while categorical data was handled by assigning a special category or the most frequent category, depending on the context.

3. Feature Engineering: We transformed certain features to better suit our analysis:
- The `loan_status` column was encoded into a binary format where 'Charged Off' was marked as 1 (default) and all other statuses as 0 (non-default).
- Categorical variables like `grade` and `home_ownership` were converted into numeric formats using techniques such as label encoding for ordered categories and one-hot encoding for nominal categories.

4. Normalization: We scaled the features to ensure that all variables contributed equally to the analysis. Using MinMaxScaler, we transformed the feature values to a common scale without distorting differences in the ranges of values.

3. **Linear Regression Model**

In our analysis, the Linear Regression model was initially applied to predict loan defaults. Training the model on our dataset, we observed **a Mean Squared Error (MSE)** of **0.06685**, which spiked dramatically to **3820.7** on the test dataset. This significant increase in MSE from training to testing suggests that our model, while fitting the training data closely, fails to generalize to new, unseen data. This condition, known as overfitting, indicates that the model may be capturing noise as if it were a true signal, making it unreliable for predicting future defaults.

The stark contrast between the training and testing MSE underscores the need for adjustments, specifically through regularization techniques. Regularization can help mitigate overfitting by penalizing larger coefficients, thereby encouraging a more balanced model that focuses on the most relevant features.

Examining the model's coefficients, it's clear that some variables have a disproportionate impact on predictions. Notably, **total_pymnt** stands out with a coefficient of **19121.38291030886**, significantly higher than others like **total_pymnt_inv**, which has a coefficient of **1.1350196886983877**. This discrepancy in coefficients suggests some variables heavily influence the model's predictions, potentially skewing its performance.

In conclusion, while the Linear Regression model provides a useful starting point for our analysis, its tendency to overfit and the uneven influence of certain variables highlight the need for a more sophisticated approach. The insights gained from this initial model set the stage for further exploration with models that incorporate regularization and potentially offer a more reliable prediction of loan defaults.

4. **Ridge Regression Model**

Applying Ridge regression with a variety of lambda values highlighted that the optimal performance was achieved with the minimum lambda of **0.01**, striking a balance that minimized Mean Squared Error (MSE) to **0. 0.06684480261342572** on the training set. This optimization suggests a model finely tuned to avoid overfitting while maintaining predictive accuracy. When tested against a separate dataset, the MSE was similarly low at **0.06781619989983911**, demonstrating the model’s ability to generalize effectively to new data. This is a marked improvement over the Linear Regression model, indicating successful regularization.

A deeper look into the coefficients revealed a significant reduction in their magnitude, particularly for features such as **total_pymnt**, where the coefficient decreased from **19121.382906361545** to **1.6439962408010707**. This reduction showcases the Ridge model's ability to penalize large weights, ensuring a more balanced and realistic influence of each feature on the prediction outcome.

This analysis proves Ridge regression's efficacy in addressing overfitting, evidenced by its robust performance on both training and testing datasets. The optimal lambda value of **0.01** enabled the model to capture the essence of the data without being misled by its complexities or peculiarities, making it a suitable approach for predicting loan defaults with greater accuracy and reliability.

5. **Lasso Regression Model**

In the exploration of Lasso regression for predicting loan defaults, we found a model that remarkably simplifies complexity while retaining predictive accuracy. Notably, at a lambda value of **0.01**, the model achieved an MSE **of 0.09623** for the training data and **0.096662** for the test data. These closely aligned errors across both datasets underscore the model's ability to generalize effectively to new observations, a cornerstone of successful machine learning applications.

A standout feature of the Lasso regression model is its sparsity, with most coefficients reduced to zero except for those associated with "**grade**" and "**total_rec_prncp**." This outcome highlights the significant influence of these variables over others in predicting loan defaults. The "**grade**" variable, reflecting the loan's risk classification, and **"total_rec_prncp**," indicating the principal amount received to date, intuitively play crucial roles in determining the likelihood of default. Their prominence in the Lasso model underscores their practical relevance in assessing default risk.

The model's simplicity, achieved through the selective shrinkage of coefficients to zero, aligns with the principle of Occam’s Razor: the simplest solution is often preferable. This adherence to simplicity, coupled with the minimal deviation in error between the training and testing phases, positions the Lasso regression model as an efficient and interpretable tool for default prediction. It strikes a balance between reducing complexity and maintaining a low error rate, making it a compelling choice for this task, especially when considering the value of straightforward model interpretation in practical applications.

6. **Random Forest Model**

In our analysis, we used the Random Forest model to predict who might not pay back their loans. We experimented with different settings of the model to find the best performance. The Random Forest is good at working with complex data and can prevent overfitting thanks to its method of using many trees to make decisions. We tested various numbers of trees and different maximum depths for these trees.
Our exploration found optimal settings across different configurations:

With **50 trees** and unrestricted tree depth, we achieved an **MSE of 0.0029506925765310794** on the training data and **0.022697161667885886** on the test data.
Increasing to **100 trees**, the training data **MSE** slightly decreased to **0.0027779327002250644**, and the test data **MSE** improved to **0.022310210314557426**.
Further increasing to **150 trees**, we maintained the trend of decreasing **MSE** on training data, underscoring the model's robustness and ability to learn from the data without overfitting significantly.

A crucial aspect of our Random Forest analysis was the evaluation of feature importances, which provides insight into which variables most significantly influence the prediction of loan defaults. Notably, features such as recoveries, int_rate, and installment were identified as having the highest importance, indicating their critical role in predicting loan defaults. In contrast, variables like **collection_recovery_fee** and **home_ownership_NONE** had minimal impact.

The Random Forest model demonstrated exceptional predictive power, with the ability to generalize well to unseen data, as evidenced by the consistency of MSE across training and test datasets. The optimal model, with **100 trees** and unrestricted depth, offered a balance between complexity and predictive accuracy, making it a strong candidate for predicting loan defaults. The analysis of feature importances further enriched our understanding, pointing to specific factors that are most predictive of loan defaults. This insight can guide the development of more targeted strategies for loan approval and risk management.

7. **Neural Network Model**
We employed a Multilayer Perceptron (MLP) neural network with a single hidden layer of 100 neurons to predict loan defaults, training it on scaled features for better uniformity and convergence. The MLP model showed excellent performance, achieving a **96.64% accuracy** on the training dataset and **95.93%** on the test dataset, indicating its high capability in classifying loan statuses and generalizing well to new data. This performance demonstrates the model's robustness and effectiveness, making it a valuable tool for practical applications in loan approval and risk assessment. The MLP's straightforward architecture also lends itself well to real-world deployment, combining strong predictive power with ease of interpretation.

8. **Evaluation**
In our comprehensive analysis of various predictive models, we aimed to identify the most effective approach for forecasting loan defaults within a peer-to-peer lending context. Our exploration encompassed Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and Neural Network (MLP) models, each subjected to rigorous training and testing to assess its predictive performance.

Linear and Ridge Regression models provided foundational insights but struggled with overfitting, with Ridge offering some improvement through regularization. Lasso Regression took a step further by simplifying the model, reducing many coefficients to zero, and highlighting key predictive features.

Random Forest emerged as a powerful contender, significantly enhancing predictive accuracy while offering insights into feature importance. This model's strength lies in its ability to handle complex data without overfitting, making it highly suitable for real-world applications.

The Neural Network (MLP) model demonstrated the highest accuracy, showcasing its capability to model complex, non-linear relationships in the data. However, despite its predictive prowess, the interpretability and practical application considerations led us to favor the Random Forest model.

Best Model: Random Forest
The Random Forest model was identified as the best overall choice due to its balanced offering of high accuracy, robustness, and interpretability. It stands out for its ability to provide valuable insights into which factors most influence loan default predictions, a critical attribute for decision-makers. This model combines the advantages of ensemble learning with practical applicability, making it ideally suited for predicting loan defaults on a peer-to-peer lending platform.

In summary, while each model contributed valuable perspectives, the Random Forest model's combination of accuracy, generalization capability, and interpretability makes it the most appropriate tool for this task. Its effectiveness underscores the utility of ensemble methods in addressing complex prediction challenges in the financial domain.