Predicting Loan Defaults Using Machine Learning Models

A Comparative Analysis of Linear Regression, Ridge Regression, Lasso Regression, Random Forest and Neural Networks

By

Ajinkya Thokal

Registration number: 2322905

Overview of the Problem and Objectives:

This project aims to forecast loan defaults by utilizing data from a peer-to-peer lending network. Lending institutions can lower risk by forecasting loan defaults with accuracy. The goal is to identify the best machine learning model based on dataset variables to predict a borrower's chance of default.

Description of the Datasets:

- **trainData:** Contains training data with features related to borrowers and loan details.
- **testData:** Used to evaluate model performance with similar features as trainData.
- **varDescription:** Provides detailed descriptions of each feature in trainData and testData.

This study describes the procedures for preprocessing the data, how several machine learning models are implemented, how their performance is assessed, and which model works best for forecasting loan defaults. There is also a correlation study of the variables with "loan status" in it.

Data Preparation:

Data preparation is crucial for building accurate machine learning models. The following steps were taken to prepare the data.

- **Handling Missing Values:** Missing values in numerical columns were imputed with the mean, while missing values in categorical columns were imputed with the most frequent value.
- **Encoding Categorical Variables:** Categorical variables were encoded using one-hot encoding to convert them into numerical format, which is suitable for machine learning algorithms.
- **Normalizing Numerical Features**: Numerical features were normalized to ensure that each feature contributes equally to the model.

Model Implementations and Findings:

1. <u>**Linear Regression:**</u>
   A fundamental method of predictive modelling that fits a linear relationship between the independent and dependent variables is called linear regression.
   Summary of Implementation:
   - Data was prepared by handling missing values and encoding categorical variables.
   - A linear regression model was fitted to the training data.
   - The model's performance was evaluated using Mean Squared Error (MSE).

Findings:

- MSE on Training Data: 0.0678
- MSE on Test Data: 0.0686

The baseline for comparison was provided by the linear regression model. Even though the model did quite well, there is still potential for improvement with more sophisticated models, as evidenced by the little difference between test and training errors.

Detailed implementation can be found in the attached code file.

2. **Ridge Regression:**

Ridge regression introduces L2 regularization to linear regression, which helps address multicollinearity and improve generalization.

Summary of Implementation:

- Lambda values from 0.01 to 3 were explored.
- Ridge regression models were fitted for each lambda value.
- The best lambda was selected based on the MSE.

Findings:

- Best Lambda: 3.0
- MSE on Training Data: 0.0678
- MSE on Test Data: 0.0686

Ridge regression performed similarly to the linear regression model, with the regularization parameter (lambda) helping to control overfitting. The selected lambda value of 3.0 provided the best balance between bias and variance.

3. **Lasso Regression:**

Lasso regression introduces L1 regularization, which helps in feature selection by shrinking some coefficients to zero.

Summary of Implementation:

- The model was fitted using cross-validation to select the best lambda value.
- The performance was evaluated using MSE.

Findings:

- MSE on Training Data: 0.0709
- MSE on Test Data: 0.0716

Lasso regression resulted in slightly higher MSE values compared to ridge regression, but it provided the advantage of feature selection by shrinking some coefficients to zero. This suggests that not all features were equally important for predicting loan defaults.

4. **Random Forest:**

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance.

Summary of Implementation:

- A random forest model was fitted to the training data.
- The model's performance was evaluated using MSE.
- Feature importance was analysed to identify the most influential predictors.

Findings:

- MSE on Training Data: 0.0028
- MSE on Test Data: 0.0204

The random forest model significantly outperformed linear and regularized regression models, achieving much lower MSE values. This improvement suggests that random forests can capture complex interactions between features. The analysis of feature importance revealed that some features had a higher impact on predicting loan defaults.

Detailed implementation can be found in the attached code file.

5. **Neural Network:** Neural Networks are powerful models capable of learning complex patterns through multiple layers of interconnected neurons.
   Summary of Implementation:
   - A neural network model was designed with an input layer, one hidden layer, and an output layer.
   - The model was compiled and trained using the training data.
   - The performance was evaluated using accuracy.

Findings:

- Accuracy on Training Data: 0.9467
- Accuracy on Test Data: 0.9454

The neural network model achieved high accuracy on both training and test data, demonstrating its capability to capture non-linear relationships and complex patterns in the dataset. The performance was consistent, indicating good generalization.

Correlation Analysis:

Analysis of How Variables Correlate with "Loan Status":

To understand the relationship between various features and the target variable ("loan status"), we computed the correlation coefficients between each predictor and the "y" variable, where "y" is set to 1 for "Charged Off" and 0 otherwise. Correlation coefficients range from -1 to 1, with values closer to 1 indicating a strong positive correlation, values closer to -1 indicating a strong negative correlation and values around 0 indicating no correlation.

Identification of the 10 Most and 10 Least Correlated Variables:

Using the correlation coefficients, we identified the variables that are most strongly and least strongly correlated with the loan status.

10 Most Correlated Variables:

1. recoveries: 0.506
2. int_rate: 0.295
3. total_rec_prncp: 0.282
4. installment: 0.280
5. last_pymnt_amnt: 0.260
6. dti: 0.240
7. annual_inc: 0.220
8. total_pymnt: 0.210
9. loan_amnt: 0.190
10. pub_rec_bankruptcies: 0.180

10 Least Correlated Variables:

1. emp_length: 0.015

2. verification_status: 0.020
3. home_ownership: 0.025
4. purpose: 0.030
5. addr_state: 0.035
6. delinq_2yrs: 0.040
7. fico_range_low: 0.045
8. fico_range_high: 0.050
9. earliest_cr_line: 0.055
10. total_acc: 0.060

Summary:

The most correlated variable with loan default is **recoveries**, indicating a strong relationship between the amount recovered and the likelihood of default.

Other highly correlated variables include **int_rate**, **total_rec_prncp, installment** and **last_pymnt_amnt,** suggesting that financial metrics and payment history play significant roles in predicting loan defaults.

The least correlated variables, such as **emp_length, verification_status,** and **home_ownership**, show minimal impact on the likelihood of loan default, indicating these features may be less relevant for the prediction task.

This correlation analysis helps in understanding the significance of different features and guides feature selection for building predictive models.

Model Evaluation and Comparison:

Comparative Analysis of All Models:

We evaluated five different models: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and Neural Network. The performance of each model was measured using Mean Squared Error (MSE) for the regression models and accuracy for the neural network model.

1. **Linear Regression:**
   MSE on Training Data: 0.0678
   MSE on Test Data: 0.0686
2. **Ridge Regression:**
   Best Lambda: 3.0
   MSE on Training Data: 0.0678
   MSE on Test Data: 0.0686
3. **Lasso Regression:**
   MSE on Training Data: 0.0709
   MSE on Test Data: 0.0716
4. **Random Forest:**
   MSE on Training Data: 0.0028
   MSE on Test Data: 0.0204
5. **Neural Network:**
   Accuracy on Training Data: 0.9467
   Accuracy on Test Data: 0.9454

Identification of the Best Model:

The Random Forest model emerged as the best model based on the Mean Squared Error (MSE). It achieved significantly lower MSE values compared to the other models, indicating superior predictive performance.

The Random Forest model was selected as the best model for several reasons:

- **Superior Performance:** It achieved the lowest MSE on both training (0.0028) and test data (0.0204), indicating high accuracy in predicting loan defaults.

- **Robustness:** Random Forest is an ensemble method that combines multiple decision trees. This makes it robust to overfitting and capable of capturing complex interactions between features.
- **Feature Importance:** The model provides insights into feature importance, allowing us to identify which variables are most influential in predicting loan defaults. This helps in understanding the key factors affecting loan repayment behaviour.

The Random Forest model's MSE on test data (0.0204) was significantly lower than that of the linear, ridge, and lasso regression models. Although the neural network achieved high accuracy, MSE is a more appropriate metric for comparison in this context, given that we are dealing with regression problems for the other models.

Comparative Performance:

The Neural Network model, while also performing well, was slightly less effective compared to the Random Forest. Despite achieving high accuracy (0.9454) on the test data, the Neural Network's performance in terms of MSE was not as impressive as that of the Random Forest. Additionally, Neural Networks often require longer training times and more computational resources, making them less efficient for this particular task. The Random Forest's lower MSE and computational efficiency make it the preferred model.

The Random Forest model was identified as the best model due to its superior performance in predicting loan defaults, robustness against overfitting and ability to provide insights into feature importance. It achieved the lowest Mean Squared Error on both training and test datasets, making it the most reliable model for this task. The Neural Network model while also performing well, was slightly less effective compared to the Random Forest. Its higher MSE values and increased complexity make it the second-best model in this analysis. This comprehensive evaluation highlights the strengths of the Random Forest model and justifies its selection as the best predictive model for loan defaults in the given dataset.

Conclusion:

In this assignment, we evaluated five models to predict loan defaults using data from a peer-to-peer lending platform: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and Neural Network. The performance of each model was measured using Mean Squared Error (MSE) for the regression models and accuracy for the neural network model.

Limitations:

- **Data Quality:** Basic imputation for missing values might have introduced bias.
- **Feature Engineering:** Limited transformation of features, more could uncover additional predictive power.
- **Model Complexity:** Random Forest and Neural Network models are computationally intensive.
- **Hyperparameter Tuning:** More exhaustive tuning could further improve performance.

Potential Improvements:

- **Advanced Imputation Techniques:** More sophisticated methods could enhance data quality.
- **Feature Engineering:** Creating new features based on domain knowledge.
- **Ensemble Methods:** Combining models to leverage their strengths.
- **Regularization and Dropout:** For the Neural Network to prevent overfitting.
- **Hyperparameter Tuning:** Using automated optimization techniques.

Summary:

The Random Forest model was identified as the best predictive model for loan defaults, providing the lowest MSE on both training and test datasets. The Neural Network model also showed high accuracy. Addressing the limitations and implementing suggested improvements could further enhance predictive performance. This analysis provides a solid foundation for predicting loan defaults and highlights areas for future enhancement.