A Comparison of Logistic Regression and Random Forest For Loan Approval Prediction

Coursework: Machine Learning Name: Mohd Arifullah

Description and Motivation

- Predicting the probability of loan approval with Machine Learning algorithms (logistic regression and random forest). Using historical data, models, and decision criteria to assess and measure consent outcomes.
- By accurately predicting loan approval, the project serves to streamline the decision-making process for financial institutions, enhance the customer experience and potentially reduce the risk of default by identifying high-risk applcations.
- Considering the results of Logistic Regression and Random Forest using different performance metrices.

Initial Analysis of Dataset & Basic Statistics

- The dataset is Loan Approval Prediction from Kaggle. The original dataset consists of 4,268 rows and 13 columns with 3 Categorical columns and 9 Numeric columns and 1 target variable with the values **Approved** or or **Rejected**.
- The Tab 1 describes the statistical information about the mean, standard deviation, max and max of the numerical features. There was no missing values but some negative values which we will be handled later.
- The Correlation heatmap (Fig 1), describes the correlation with inner variables.
- It can be observed from the heatmap that the number of **dependents** and **loan_amount** seems to have a large influence on predicting the approval of loan followed by income and cibil score.
- The histogram (Fig 2) provides the information of the number of loans approved or rejected which is **Accepted** rate of 62.3% and **Rejected** rate of 37.7%.
- The boxplots (Fig 3) can help to analyse distribution of the numeric features.
- The count plots (Fig 4) is to analyze the distribution of the categorical features.
- There is a high chance of loan rejectetion if the cibil score is below 600.
- It can be observed that if the no of dependents increases the chance of loan rejection also increases parallely.
- It shows that banks like approving loans that are easier to pay back quickly which means loan with less no of term are likely to be get approved. If a loan is small or takes a long time to repay, it's more likely to be rejected.
- It can observed that more assets, both movable (bank assets, luxury assets) and immovable (residential assets, commercial assets), make banks more likely to approve a loan. Having more stuff makes getting a loan easier.

Feature	Min	Max	Mean	Std
dependents	0	5	2	2
income_annual	20000 0.0	990000 0.0	505044 3.78	272875 4.48
loan_amount	30000 0	395000 00	151334 50.45	904336 2.98
loan_term (years)	2	20	11	6
cibil_score	300	900	600	172
residential_assets_value	10000 0.0	291000 00.0	760328 8.84	642771 6.72
commercial_assets_value	0	194000 00	497315 5.30	438896 6.08
luxury_assets_value	30000 0	392000 00	151263 05.92	910375 3.66
bank_asset_value	0	147000 00.0	496086 5.17	313067 8.50

Table 1. All Features

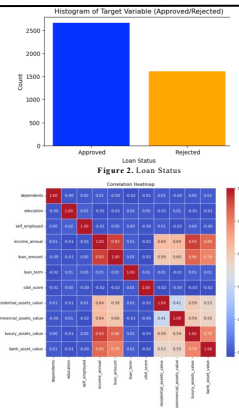

Figure 2. Loan Status

Figure 1. All Correlation Heatmap

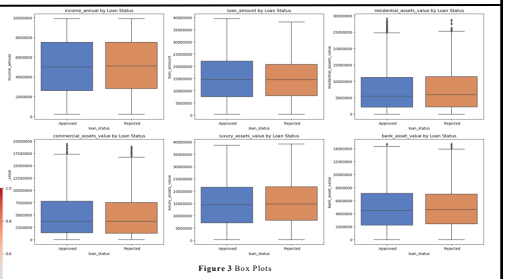

Figure 3 Box Plots

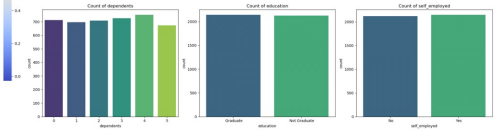

Figure 4. Count Plots

Hypothesis Statement

- LR is easier to train and doesn't require as many computational resources. RF, on the other hand, consumes more computer power.
- Random Forest is expected to deliver a more complex, possibly multidimensional decision boundary than Logistic Regression.

Training Choice & Evaluation

- Dataset was splitted into 80:20 for training and testin with training set consists of 3,415 rows while the testing set consisting of 853 rows. The test data remains unseen to models until end.
- Optimize model by Feature Engineering for estimating the accuracy, precision and other performance metrices on the models.
- Calculating optimal hyperparameters for the models based on validation accuracy, confusion matrices and other performance metrices.
- After determining the hyperparameters, the training set is assessed using training, validation, and confusion matrices.
- Then the models will be tested with their best optimal parameters to find the performance.

Parameters Choices & Results

Logistic Regression

- Model is trained by doing standard scaling to standardize the features, ensuring that they have zero mean and unit variance.
- **Rigid** regularization technique (penalty L1) was used with Grid Search to find the best parameters to optimize the model's performance.

Choice of parameters

- Choosing the regularization technique L1 (Rigid) from (Rigid, Lasso) optimization and C (Inverse of Regularization) with 10 in range of (1,10).

Random Forest

- Fitting the model algorithm with Gini's diversity index for splitting criterion.
- Grid search is to tune the hyperparameters by exploring the number of trees in the forest across four log-scaled values in 10, 50, 100, and 200.

The choice of parameters

- Bagging Technique is used to train the multiple models
- The number of trees in the optimized RF is 200.

	LR	RF
Accuracy	0.92	0.60
AUC	0.96	0.55
Train Time(s)	1.04	0.08
Predict Time	0.01	0.18
Test Error	0.01	0.40

Table 2. Prediction results

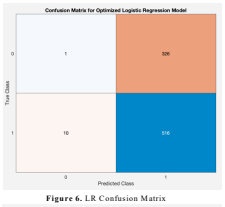

Figure 6. LR Confusion Matrix

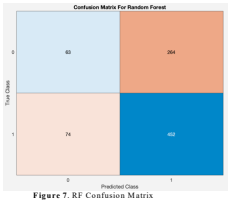

Figure 7. RF Confusion Matrix

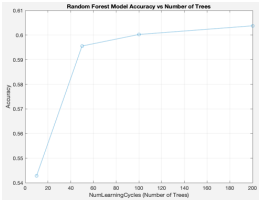

Figure 5. Accuracy vs No Trees

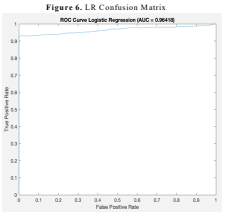

Figure 8. RUC (LR)

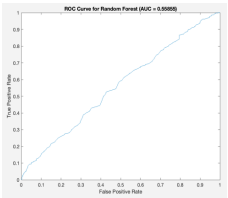

Figure 9. RUC Graph (RF)

Lessons Learned & Future Work

Lessons learned

- It is important to consider other results in addition to accuracy, such as the confusion matrix, the training error, the testing time, and the AUC curve.
- While hyperparameters tuning, in case for RF, be aware that there are many different parameters and methods to implement, so it may take a lot of time.

Future work

- Gathering more data to handle the cost of misclassification.
- Using variety range of hyperparameters tuning to enhance the model performance especially in Random Forest.

References

[1] Puneeth.B.R, A.K,A.Kumar, B.Rao, P.S.K and S.A.P, An Approach to Predict Loan Eligibility using Machine Learning, 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE), Karkala, India, 2022, pp. 23-28, doi: 10.1109/AIDE57180.2022.10059881.

[2] Kumar Arun, Garg Ishan, Kaur Sanmeet "Loan Approval Prediction based on Machine Learning Approach"IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 79-81 www.iosrjournals.org

[3] Hamayel, Mohammad & Moreb, Mohammed & Abumohsen, Mobarak. (2021). Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine. 10.1109/ICIT52682.2021.9491636.

[6]Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News(http://CRAN.R-project.org/doc/Rnews/), 2(3):9–22, 2002.

Implementing models with pros and cons

Logistic Regression(LR)

- Logistic regression is a method for analysing data where the outcome is influenced by one or more variables.
- This model is easy to understand because its coefficients reveal how the variables are related to the outcome.
- Logistic regression doesn't consider a linear relation between the dependent and independent features, unlike other regressions.
- This classification strategy uses a logistic function to predict binary Result given an independent variable[1]

Pros

- Logistic Regression is more effective when data is linearly separated.
- Its high interpretability means that it doesn't require a lot of processing power

Cons

- It assumes that the log odds are linearly related to the independent variables. In cases where this linearity assumption does not hold true, logistic regression can give poor performance
- Logistic regression models are highly explainable but they are not as strong as non-linear models such as deep neural networks, or gradient boosting[4] .
- Results may become skewed as a result of outliers and substantial impact on the result when working with complex and non-linear relationships found in high-dimensional data, it may underfit.

Random Forest (RF)

- A supervised learning technique mostly applied to regression and classification issues.
- Random forests are a group learning system for characterization (and relapse) that work by building a large number of Decision trees at preparing time and yielding the class that is the mode of the classes yield by individual trees[2].
- The random forest algorithm estimates the importance of a variable by looking at how much prediction error increases when (OOB) data for that variable is permuted while all others are left unchanged[6].

Pros

- By building multiple trees it provides high accuracies.
- For large data sets, it generates high accurate predictions.

Cons

- High Risk of overfitting if not properly tuned.
- Random Forest not suitable with sparse data.

Analysis & Results Evaluation

- Random Forest accuracy was observed to be 98%, due to this much accuracy it is likely to said that the model was overfitted. and in order to counter this problem I have find out the top two features which were highly correlated with my target variable and then removed them.
- After Feature Engineering, the RF model's accuracy was further reduced to 59%, which is less as compared to the prior study done by Puneeth.B.R, and Ashwitha K (2022)[1].
- In our Random Forest the risk of overfitting may be high due to the high number of trees as each tree may end up memorizing the noise from the training set rather than capturing the underlying patterns. One more reason for such high accuracy could be , like Random Forest assigns score to features based con how much they contribute to the overall performance. So if there are some noise or irrelevant features, RF model can assign them high score leading to overfitting.
- Further refinement of the RF model was conducted using GridSearch. GridSearch is basically a technique to which uses a set of hyperparameters space and finds the optimal one out of all possible combinations. The hyperparameter space used includes the no of tree in the range of 50-200. As we can see from the Fig 5, in which the best accuracy was found with 200 trees which is 61%.
- One of the major drawback of GridSearch was the time consumption. As GridSearch explores all the possible combination of hyperparameters and in our RF, we have used a hyperparameters of max no of trees, fitting these parameter can be computationally expensive and time-consuming. The RF model with GridSearch optimization took nearly 7.3 seconds.
- The accuracy of the logistic regression model was 92.2% which surpassed benchmarks set by Puneeth.B.R & Ashwitha K (2022) .This success could be due the detailed customer personal data, encompassing assets such as residential, commercial, and luxury properties, along with a randomized approach to test-train splitting.
- With such higher accuracy Logistic Regression demonstrated excellent performance metrics. Its effectiveness as a baseline model is noteworthy, it implies that the decision boundaries between classes are probably linear, which may eliminate the need for more sophisticated models in some circumstances.
- Optimization of the logistic regression model was achieved through GridSearch, which facilitated the fine-tuning of hyperparameters. The best parameter set identified was ridge as the penalty and c which is the inverse of regularization was 10 in a range of 1-10. Also the time taken was comparatively less as compared to the optimized RF model. The improved accuracy of the LR model was observed to be 92.7%.
- The performance of the optimized models can be seen in Table 2.
- In the final comparative analysis, the RUC graph (Fig 8) for the LR model exceeded that of the RF (Fig 9) model ,this could be due to the feature engineering which was applied on Random Forest which reduces its accuracy further. The area under the curve's slope indicates how well a model can forecast the future[4].
- High accuracies of the models suggested a potential for overfitting, necessitating the use of additional evaluative methods such as Receiver Operating Characteristic (ROC) curves and confusion matrices.
- Majority of correct classifications in both models observed in the 'Approved' category, likely due to a larger number of learning labels in this class, enhancing classification confidence as it can be seen in the Fig 6 and Fig 7 (Confusion matrix).
- Further investigation and refinement recommended to improve the robustness of the models, considering these findings and the expanded data dimensions.

References (continued)

[4] Z. Ereiz, "Predicting Default Loans Using Machine Learning (OptiML)," 2019 27th Telecommunications Forum (TELFOR), 2019, pp. 1-4, doi: 10.1109/TELFOR48224.2019.8971110.

[5] S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas and B. Koroglu, "A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 326-331, doi: 10.1109/UBMK52708.2021.9558894