Final Report: Predictive Modeling for Loan Default Risk Assessment

Introduction

Background Information

Banks play a critical role in the financial industry, with a major part of their business coming from lending money. When borrowers cannot pay back the loans, banks face financial losses and operational risks. This project will use machine learning techniques to build a model that can predict if a borrower will default on their loan, helping banks make better lending decisions and manage risks effectively.

Objective

The objective is to construct and refine a model that can precisely predict the chances of loan defaults. By doing so, banks can substantially diminish their vulnerability to bad debts, ultimately improving their financial health. The research question for this objective: "Using advanced techniques, how accurately can we predict the chances of loan defaults by using various information of borrowers and loans?"

Dataset Source (Dataset title: Loan Default Dataset**)**

Data Statistics

Description of dataset

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ID	1	148670	99224.50	42917.48	99224.50	99224.50	55104.54	24890	173559.00	148669.00	0.00	-1.20	111.31
year	2	148670	2019.00	0.00	2019.00	2019.00	0.00	2019.00	2019.00	0.00	NaN	NaN	0.00
loan_amount	3	148670	331117.74	183909.31	296500	313493	177912	16500	3576500.00	3560000.00	1.67	9.13	476.97
rate_of_interest	4	112231	4.05	0.56	3.99	4.03	0.54	0.00	8.00	8.00	0.39	0.34	0.00
Interest_rate_spread	5	112031	0.44	0.51	0.39	0.42	0.51	-3.64	3.36	7.00	0.28	-0.18	0.00
Upfront_charges	6	109028	3225.00	3251.12	2596.45	2745.90	3126.30	0.00	60000.00	60000.00	1.75	6.37	9.85
term	7	148629	335.14	58.41	360.00	351.78	0.00	96.00	360.00	264.00	-2.17	3.17	0.15
property_value	8	133572	497893.47	359935.32	418000	447440.75	252042	8000.00	16508000	16500000	4.59	73.22	984.84
income	9	139520	6957.34	6496.59	5760.00	6128.95	3380.33	0.00	578580.00	578580.00	17.31	885.25	17.39
Credit_Score	10	148670	699.79	115.88	699.00	699.73	148.26	500.00	900.00	400.00	0.00	-1.20	0.30
LTV	11	133572	72.75	39.97	75.14	74.00	18.55	0.97	7831.25	7830.28	120.61	19978	0.11
Status	12	148670	0.25	0.43	0.00	0.18	0.00	0.00	1.00	1.00	1.18	-0.62	0.00
dtir1	13	124549	37.73	10.55	39.00	38.33	10.38	5.00	61.00	56.00	-0.55	0.38	0.03

Table 1. Descriptive and summary statistics of Loan Default Risk Dataset.

Table 3 shows a screenshot of descriptive statistics of the dataset. We have removed all non-numeric features because their statistics are not relevant. As a result, the table above shows statistics for 13 numerical features, rounded to 2 decimal points. From the statistics above, we can see zeros in some of the fields, but it is normal if you understand the dataset. For example, median, mad, min value of feature “Status” has value 0 which may lead us to assume there are missing values. However, it is an appropriate value because feature “Status” indicates whether loan is approved or not and is represented as “0” and “1”.

The statistics above also helped distinguish features that are not meaningful in the analysis. Features **'ID'** and '**Year'** are not incorporated into future data modelling with 'ID' being nominal data, as well as the 'Year' entries all showing 2019 submissions without the month or day.

Features with Missing Values (N/A)

Feature	N/A count	N/A Percentage (%)
rate_of_interest	36439	24.51
Interest_rate_spread	36639	24.64
Upfront_charges	39642	26.66
property_value	15098	10.16
LTV	15098	10.16
dtir1	24121	16.22

Table 2. Data Features with 'N/A' and percentage of 'N/A' for each.

In our analysis of 148,670 applications, three critical features stood out due to a high rate of missing data. 'Rate of Interest', 'Interest Rate Spread', and 'Upfront Charges' showed similar pattern with about 25% of its data points missing. These features were removed from the analysis.

Features with Empty Data

Feature	Empty count	Empty Percentage (%)
loan_limit	3344	2.2
approv_in_adv	908	0.61
loan_purpose	134	0.09
Neg_ammortization	121	0.08
age	200	0.13

Table 3. Dataset features with empty values and their percentages.

Five features contained empty values in the dataset, but the percentage is insignificant compared to N/A values. The sum of empty data in the whole dataset is about 3%. The rows with empty data were removed from the dataset for modeling.

Exploratory Data Analysis

Finding #1: Significant Imbalance in Loan Status Distribution

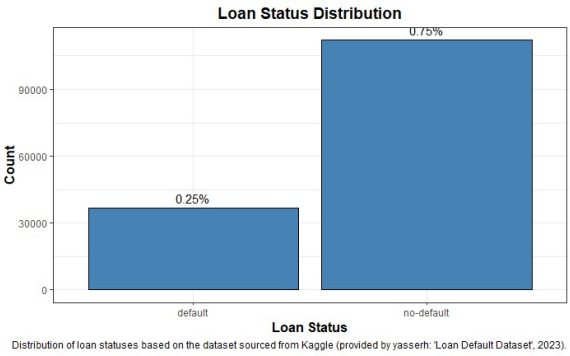


Figure 1A. Loan Status Distribution

We began a thorough exploration of a dataset revolving around loan default. The initial phase involved understanding the distribution of the target variable, which revealed a significant imbalance: approximately 75% of loans were repaid without default, while the remaining 25% defaulted. Such an imbalance raised awareness that our models might be biased towards predicting the majority class (no-default). Accurate prediction of the minority class (default) is crucial given that the financial consequences of incorrect predictions are detrimental.

Finding #2: Uniformity in Loan Terms and Credit Scores

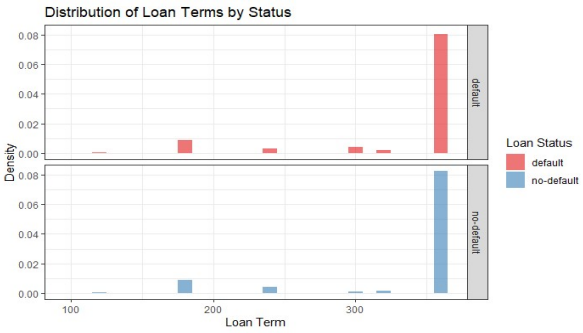


Figure 2A. Distribution of Loan Term by Status Status

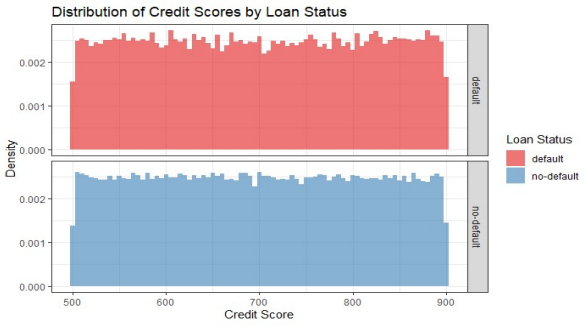
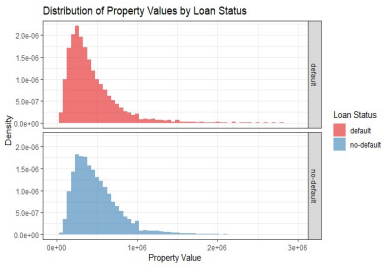
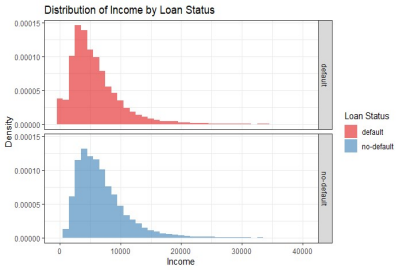


Figure 2B. Distribution of Credit Scores by Loan

In our exploratory data analysis (EDA) of numerical variables, we observed that the variables "term" and "Credit_Score" exhibit surprisingly similar distribution patterns for both default and non-default groups. The uniformity in the distribution of loan terms suggests that the duration of a loan, by itself, might not be a critical factor in predicting the likelihood of a default. Similarly, upon examining credit scores, we noticed that both the default and non-default groups are similar across the entire range of scores. This observation was unexpected, as it challenges the conventional assumption that lower credit scores would be more prevalent in the default group. The similarity in credit score distributions across both categories implies that credit scores may also not be a decisive predictor in determining the risk of a loan default.

Finding #3: Distinct Patterns in Other Financial Variables





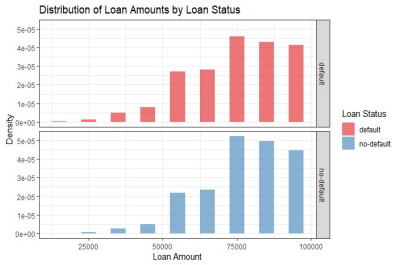
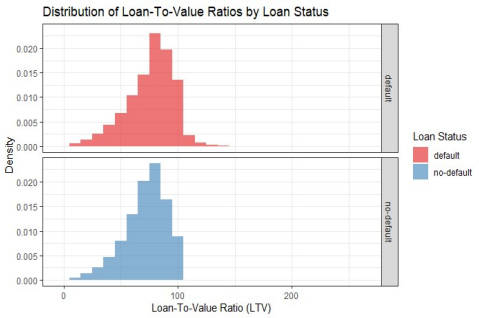


Figure 3A.Distribution of Property Values ***Figure 3B.Distribution of income*** ***Figure 3C.Distribution of Loan Amount***

Our analysis extended to other variables, such as property values, incomes, loan amounts, Loan-to-Value (LTV) ratios, and Debt-to-Income Ratios (DTIR). We observed noticeable differences in these variables between the default and no-default groups. Particularly, higher property values, incomes, and loan amounts seemed to lean towards the non-default group, suggesting they are likely significant factors in determining loan default. It makes sense, as those with higher incomes tend to purchase more expensive properties and secure larger loan amounts.



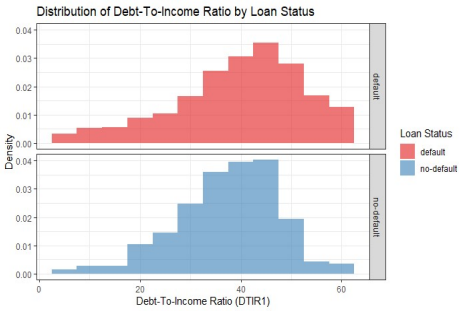


Figure 3D. Distribution of Loan-To-Value Ratios ***Figure 3E. Distribution for Debt-To-Income Ratio***

Regarding LTV ratios, we observed a distinct skew towards higher ratios in the default group, indicating that a larger portion of their properties is financed through loans. This pattern suggests that higher LTV ratios might be an indicator of increased default risk. Similarly, in the DTIR distributions, the default group demonstrated a skew towards higher ratios, implying that a substantial portion of their income is committed to debt payments. This finding indicates that higher DTIRs could be a warning sign of potential default.

Finding #4: Variability in Categorical and Control Variables

In our EDA of categorical and control variables, we observed varying levels of variance across variables. Variables such as 'construction_type', 'Secured_by', and 'Security_Type' predominantly leaned towards one category, suggesting their possible limited role in the modeling. In contrast, variables like 'loan_type', 'loan_purpose', and 'submission_of_application' showed substantial variances, indicating their potential significance. The control variables 'age', 'Gender', and 'Region', while showing minimal variance, are retained because we thought they are needed for adjusting any potential confounding effects. Overall, the dataset is comprised of a diverse mix of variables, with some flagged for potential exclusion and others for their potential significance.

Finding #5: Missing Data and Potential Outliers

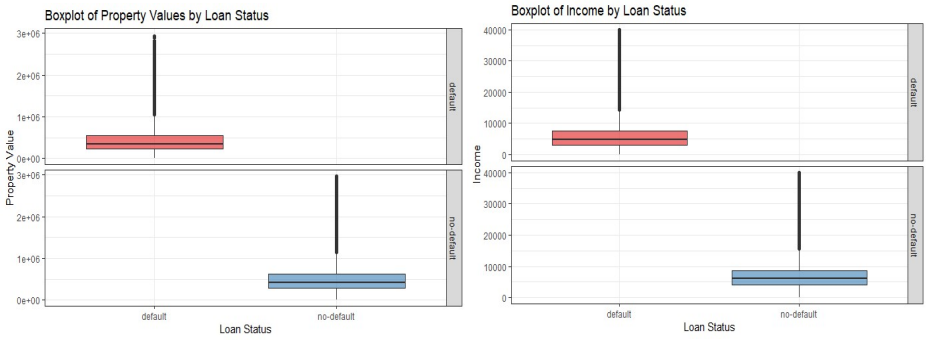


Figure 4A. Boxplot of Property Values by Loan Status ***Figure 4B. Boxplot of Income by Loan Status***

Our EDA also spotlighted areas with missing data and potential outliers, which could have a significant impact on the modeling process. For certain variables like property values and income, the presence of seemingly high values raised concerns. Upon more in-depth investigation, they were not unrealistically high; therefore, they were retained.

Data Cleaning and Preprocessing

After EDA, we removed several categorical variables that were identified to be univariant, did not contribute much to model variation. Specifically, columns such as 'ID', 'Year', 'construction_type', 'rate_of_interest', 'Interest_rate_spread', 'Upfront_charges', 'term', 'Secured_by', and 'Security_Type' were excluded. Following this, we converted the Status column into a binary categorical variable for subsequent logistic regression modeling processes.

To manage missing values, imputation was used. Instead of imputing them with mean or median values, we implemented a more robust technique using bagged models. This technique leverages multiple instances of a model to predict and impute missing data in columns like 'income', 'dtir1', 'LTV', 'property_value', 'loan_limit', 'approv_in_adv', 'loan_purpose', 'Neg_ammortization', 'age', and 'submission_of_application'. With the imputed dataset ready, we proceeded to conduct logistic regression tests on various data samples for evaluation of variable significance.

Logistic Regression for Variable Significance

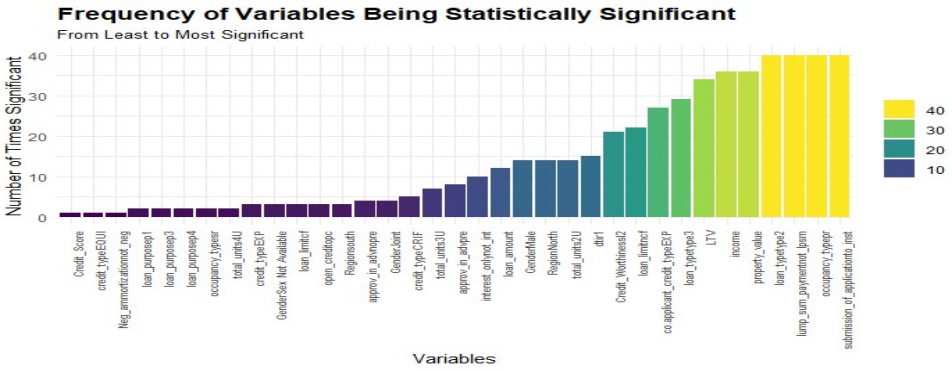


Figure 5. Frequency of Variables Being Statistically Significant

To ensure robustness and account for potential variability, we iteratively ran logistic regression models 40 times on different random samples of the dataset. Each time, we recorded the variables that were statistically significant in predicting the loan status.

The chart shown above effectively communicates which variables might be most important in predicting the loan defaults and which are less so, guiding decisions for feature selection in subsequent modeling efforts. At the lower end of the frequency spectrum, variables like 'Credit_Score', 'Neg_ammortizationnot_neg', 'credit_typeEQUI', 'loan_purposep1', 'loan_purposep3', and 'loan_purposep4' were found to be significant a minimal number of times (less than 5 times). This suggests they may not be strong predictors in the context of the model used. At the highest end of the spectrum, 'loan_typetype2', 'lump_sum_paymentnot_lpsm', 'occupancy_typepr', and 'submission_of_applicationto_inst' stand out, each with a frequency of 40.

Based on these findings, we hypothesize that variables that achieved the highest frequency of significance are likely to also be pivotal in forecasting loan defaults in ensemble models.

Data Modeling

To develop our Random Forest, LightGBM, and logistic regression models, we started with a dataset that had previously undergone cleaning and processing during logistic regression analysis, where we assessed the significance of variables. To recall, this dataset was carefully refined by removing non-informative categorical variables and enhancing it through advanced bagged model imputation techniques to handle missing values. Also, we converted categorical variables into dummy variables to transform qualitative data into a numerical format.

Our approach began with a strategic decision to utilize a smaller subset of our dataset, aiming to explore the hyperparameter space efficiently. This preliminary step was pivotal, as it allowed us to discern which hyperparameters were most influential without the substantial time investment typically associated with hyperparameter tuning.

1. Random Forest Model

In our approach to optimizing the Random Forest model, we first established a hyperparameter tuning grid using the grid_latin_hypercube method. This allowed us to explore a diverse range of hyperparameters, including the number of trees, the mtry parameter (number of variables tried at each split), and the minimum node size.

To assess the model's performance across different hyperparameters, we used ten-fold cross-validation, ensuring a balanced representation of our target variable, 'Status'. The cross-validation process was accelerated using parallel processing. The tune_grid function was then applied to systematically evaluate the model across the parameter grid, capturing a wide spectrum of possible configurations. This careful setup was critical for the subsequent analysis of how each hyperparameter influenced the model's accuracy and ROC AUC.

Hyperparameter Space Exploration

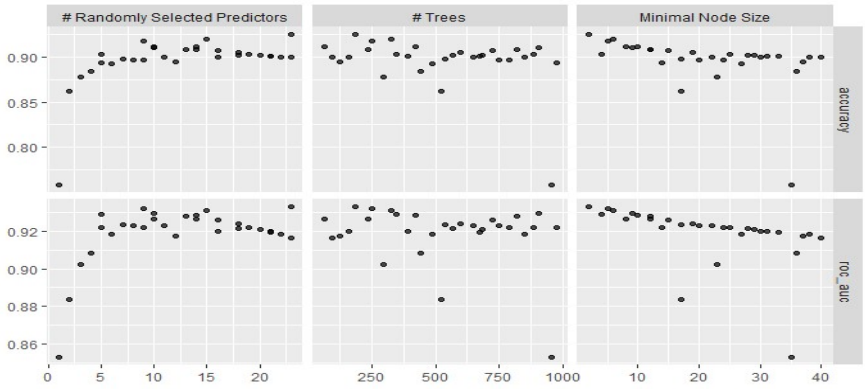


Figure 6. Plots of accuracy and roc_auc of hyperparameters

From the graph above, it is observed that the number of randomly selected predictors (mtry) shows a positive relationship with both accuracy and the area under the ROC curve (ROC AUC), indicating an increase in model performance with higher mtry values. On the contrary, the minimal node size demonstrates a subtle yet inverse relationship with model performance, with larger node sizes slightly detracting from model accuracy and ROC AUC.

Considering these observations, the range for mtry was chosen to be between 10 and 20, optimizing for the observed plateau in performance improvement. Similarly, the minimal node size was confined to the range of 0 to 20 to avoid the diminutive decline in performance associated with larger node sizes. Also, we configured the model with a fixed number of trees (500) to ensure sufficient model complexity.

2. Light Gradient Boosting Machine Model

We implemented the Light Gradient Boosting Machine (LightGBM) model due to its performance edge over XGBoost and computational efficiency. We configured the model with 500 trees and focused on tuning key hyperparameters including tree depth, minimum node size, loss reduction, sample size, mtry, and learning rate.

For efficient hyperparameter optimization, we employed the grid_latin_hypercube method to define a tuning grid. In line with our previous approach for random forest, we maintained a ten-fold cross-validation setup and leveraged parallel processing for faster execution. Subsequently, we systematically assessed LightGBM's performance across the parameter grid using the tune_grid function.

Hyperparameter Space Exploration

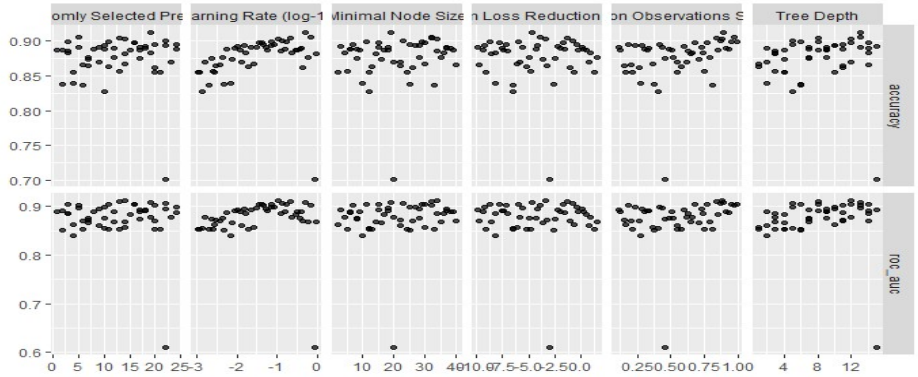


Figure 7. Plots of accuracy and roc_auc of hyperparameters

While most hyperparameters did not significantly impact model performance (accuracy and ROC AUC), the number of randomly selected predictors (mtry) and the learning rate displayed significant patterns. Notably, the learning rate exhibited a quadratic relationship with model performance, suggesting that both excessively low and high learning rates impede the model's effectiveness. Conversely, a middle range of learning rates clearly results in optimal performance. Based on these observations, we have narrowed the learning rate to a range between -2 and -0.5 for the subsequent phase of model tuning.

3. Logistic Regression Model

The logistic regression model, log_spec, was modeled to serve as a baseline for comparison with ensemble models. It was set up using the generalized linear model (GLM) approach with a binomial family to predict binary outcomes. The model was constructed without hyperparameter tuning as the logistic regression does not involve complex iterative processes like Random Forest or LightGBM.

Addressing Unbalanced Data: SMOTE's Impact on Logistic Regression

Model	**Metric**	**Without SMOTE**	**With SMOTE**
Logistic Regression	Accuracy	0.8635	0.8258
Logistic Regression	ROC AUC	0.8313	0.8369

Table 4. SMOTE's Impact on Logistic Regression

In addressing the imbalance in our dataset, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to evaluate its impact on the model's predictive accuracy for default cases. The application of SMOTE led to a marginal increase in the ROC AUC, from 0.8313 to 0.8369, indicating a slight improvement in the model's ability to distinguish between default and non-default cases. However, this benefit was offset by a decrease in mean fold accuracy, which dropped from 0.8635 to 0.8258. In the context of our project, accuracy is of paramount importance, as it reflects the model's overall ability to correctly identify both default and non-default cases. This is critical in financial risk modeling, where false positives (predicting a default incorrectly) and false negatives (failing to predict an actual default) can have significant

implications. Consequently, given the higher importance assigned to accuracy and its notable decrease with SMOTE, we decided against incorporating this technique into our final model. This decision was made to ensure that the model remains highly reliable in identifying actual loan defaults, which is essential for practical applications in financial risk management.

Assessing Model Superiority: Random Forest vs. Light GBM vs. Logistic Regression

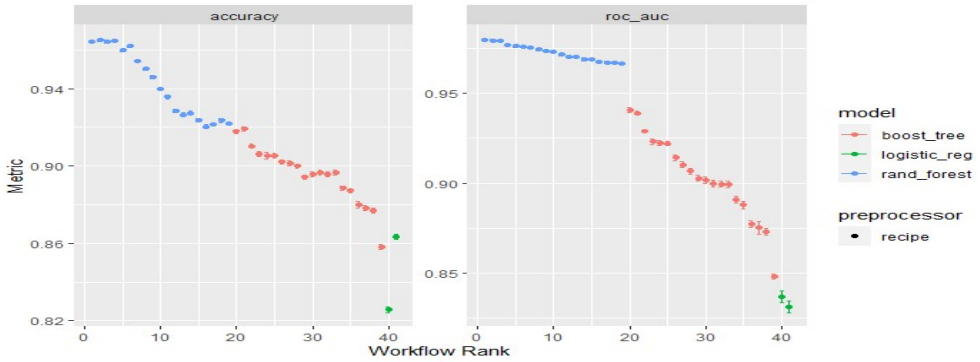


Figure 8. Plots for assessing model superiority

The graph displays two scatter plots, each ranking the performance of various models—Random Forest (rand_forest), Light GBM (boost_tree), and Logistic Regression (logistic_reg)—based on two different metrics: accuracy and roc_auc (Area Under the Receiver Operating Characteristic Curve). In both plots, each point represents a model's performance with a specific set of hyperparameters. Irrespective of the hyperparameter set used, all Random Forest models outperform others, achieving higher rankings for accuracy and ROC AUC, which indicates they are the superior models in this assessment.

Confusion matrices on the predictions Header

Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27680  1898
         1   258  7332

               Accuracy : 0.942
```

Light GBM

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27475  2793
         1   463  6437

               Accuracy : 0.9124
```

Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27632  4746
         1   306  4484

               Accuracy : 0.8641
```

Figure 9. Confusion matrices for Random Forest, LightGBM, and Logistic Regression

The demonstrated supremacy of the Random Forest model in the scatter plot rankings is further confirmed by the confusion matrix outcomes, where its predictive precision for loan defaults continues to outperform the other models. Our Random Forest model is the most reliable, demonstrating a strong balance between accurately identifying defaults and not mislabeling non-defaults. The Light GBM model, despite being reasonably accurate, falls short in detecting actual defaults as effectively as the Random Forest model. The Logistic Regression model has the most difficulty in recognizing defaults, which may present a challenge when trying to mitigate financial risk.

Final Model Selection

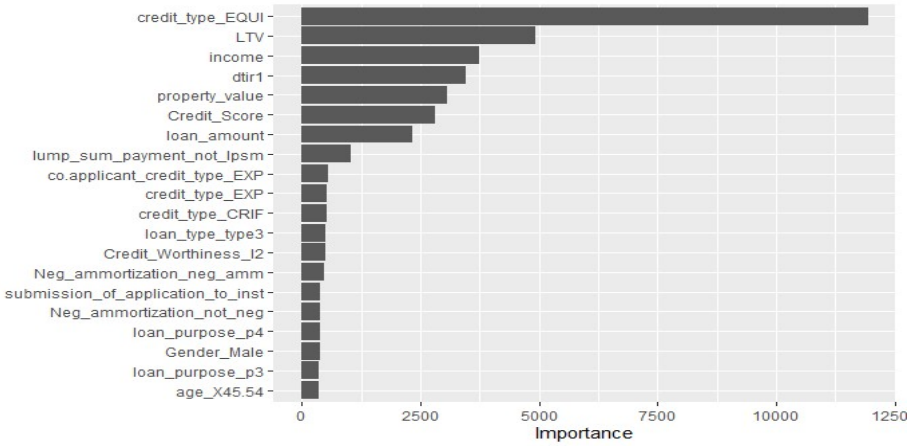


Figure 10. Variable importance plot for the Random Forest model

The variable importance plot for the Random Forest model illustrates that the key predictors for loan default are:

1. **credit_type**: Indicates the kind of credit issued, with a major influence on default prediction.
2. **LTV** (Loan to Value ratio): Suggests the loan amount in relation to the property's value, a critical risk assessment metric.
3. **income**: Reflects the borrower's earnings, directly impacting their loan repayment capability.
4. **dtir1**: Likely the debt-to-income ratio, showing the proportion of debt to borrower's income.
5. **property_value**: Represents the value of collateral, important for loan security.
6. **Credit Score**: A measure of creditworthiness based on credit history.
7. **loan_amount**: The size of the loan, which correlates with default risk.

These top variables significantly shape the model's predictions, indicating financial health and risk factors as primary determinants in assessing the likelihood of loan default.

Conclusion

Our in-depth study leverages advanced predictive modeling techniques to tackle the critical challenge of loan default risk within the banking sector. Our efforts have resulted in the creation of a robust Random Forest model that consistently outperforms other models in terms of predictive accuracy, as demonstrated by various performance metrics and confusion matrix analysis. Through the data preparation and preprocessing, which included the strategic imputation method for missing data, removal of non-informative features and the conversion of categorical variables, we have refined a dataset that enables the development of a highly reliable and precise model. The variable importance plot has played a crucial role in identifying the most influential predictors, including credit type, loan-to-value ratio (LTV), income, and credit score, highlighting the pivotal role of a borrower's financial profile in assessing default risk.