Random Forests for Loan Approval

Arslonbek Ishanov

Abstract

This report investigates the application of Random Forests for predicting loan approval decisions, a critical task in the financial sector. Using a dataset of applicant demographic, financial, and credit-related variables, Random Forest classifiers were trained and optimised with Optuna. The study emphasises preprocessing steps, hyperparameter tuning, validation procedures, and ethical considerations. Results demonstrate strong predictive power, interpretability, and robustness, making Random Forests a practical choice for real-world loan approval systems, while acknowledging biases, fairness, and future improvements.

Introduction

The financial services industry has long relied on statistical models and rule-based decision systems to evaluate loan applications. With the rise of machine learning, more advanced algorithms are increasingly being adopted to enhance predictive accuracy, reduce default risk, and streamline decision-making processes. One algorithm particularly well-suited for this purpose is the Random Forest classifier, a powerful ensemble method that aggregates the decisions of multiple decision trees to improve robustness and generalisation.
This report focuses on the implementation of Random Forests for predicting loan approval decisions. By leveraging demographic, financial, and credit-related information, we build, train, and optimise a Random Forest model using Optuna for hyperparameter tuning. The study also evaluates practical trade-offs, explores robustness strategies, highlights key validation procedures, and discusses the ethical considerations surrounding algorithmic decision-making in lending practices.

Preprocessing

The dataset required thorough preprocessing before model training. The first step involved handling missing values, which were imputed depending on the type of variable. Numerical features with missing values were filled using median imputation to preserve distributional characteristics, while categorical variables were filled using mode imputation. This approach reduced bias compared to mean imputation for skewed financial data.
Categorical variables such as gender, marital status, education, and property ownership were encoded using one-hot encoding. This transformation ensured compatibility with the Random Forest algorithm, which requires numerical inputs. Meanwhile, numerical features such as income, loan amount, and loan term were normalised to mitigate the impact of outliers and facilitate balanced learning across features with different scales.

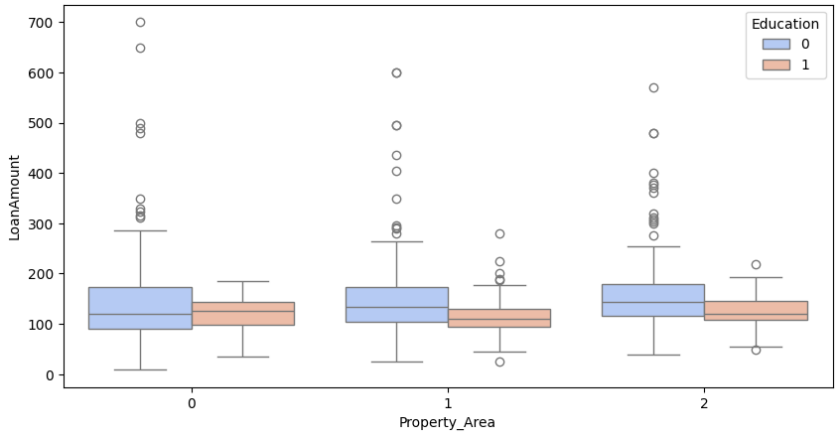

Figure 1: Boxplots for the relation between Property Area, Amount of Loan and Education qualification.

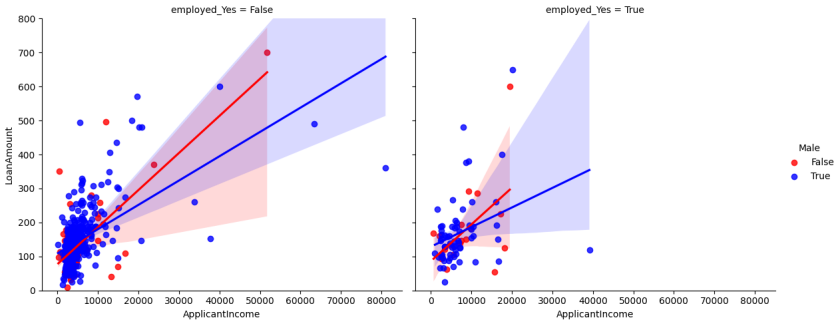

Figure 2: Plot for LoanAmount, Applicant Income, Employment and Gender.

Class imbalance was another significant challenge, as approved loans significantly outnumbered rejected ones. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic examples of the minority class to improve representation. This ensured the model did not overfit toward predicting approvals.

Observations and Practical Trade-Offs

During model development, several practical trade-offs became apparent. First, there was a balance between complexity and interpretability. While Random Forests provide high accuracy, their ensemble nature makes them less transparent compared to single decision trees. However, feature importance measures partially mitigated this by offering insights into which variables most influenced predictions.
Second, there was a trade-off between computation time and accuracy. Training on larger datasets with many estimators improved performance but came at the cost of increased computational requirements. Hyperparameter tuning using Optuna helped mitigate this by efficiently exploring the parameter space without exhaustively testing all combinations.
Lastly, the issue of generalisation emerged as a central consideration. A model with too many estimators or deep trees risked overfitting to training data, whereas shallow or overly regularised models underperformed on unseen data. Careful validation was therefore essential to balance predictive performance with robustness.

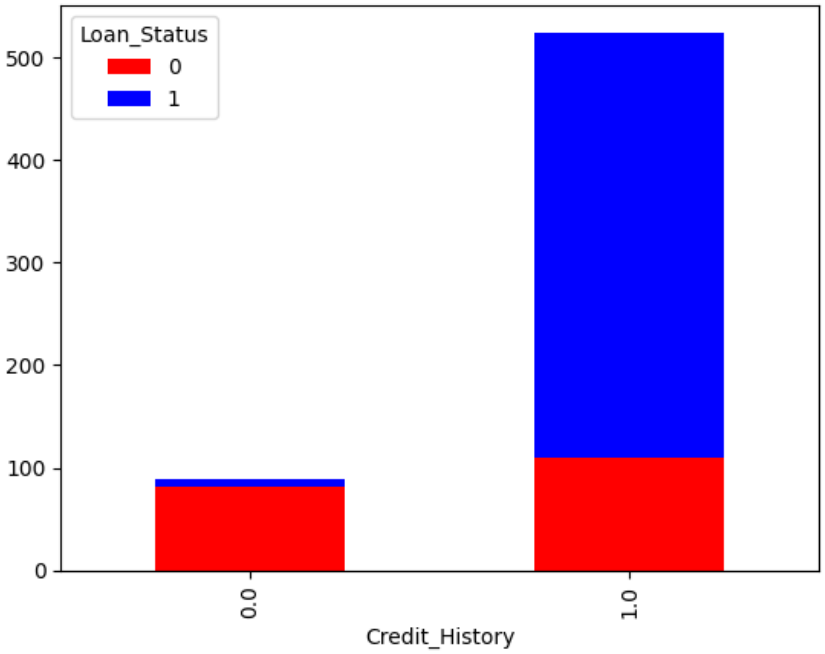

Figure 3: The Credit History vs Loan Status.

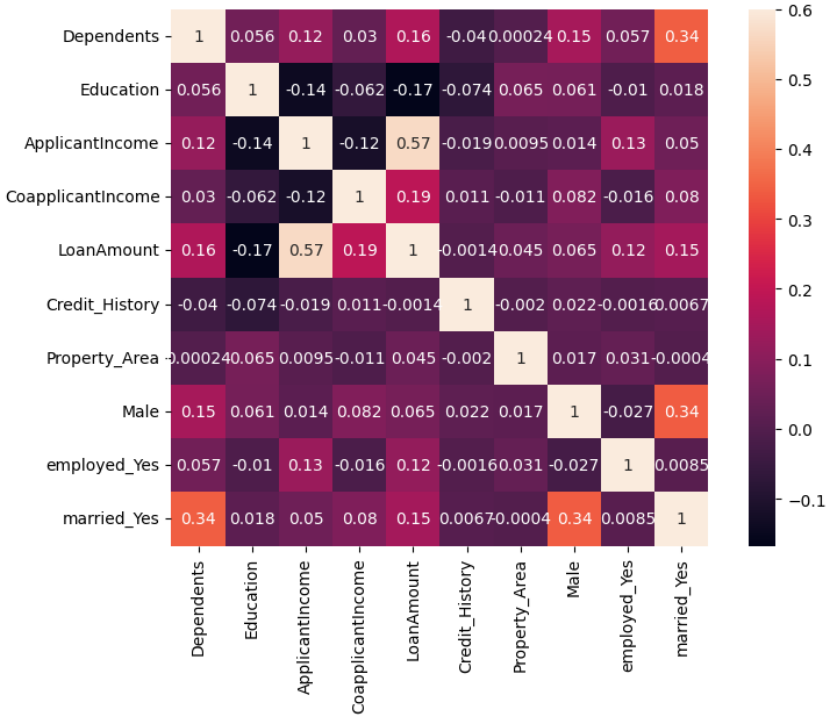

Figure 4: The Correlation Heatmap.

Robustness and Regularisation Strategies

Ensuring robustness was crucial to avoid overfitting and to maintain generalisation across diverse applicant groups. Several strategies were implemented. Limiting the maximum depth of trees prevented the model from learning overly specific patterns tied to noise in the data. Restricting the minimum number of samples per leaf node also discouraged

overfitting by ensuring splits were made only when supported by sufficient evidence.
Regularisation was further enhanced by tuning the number of estimators. Too few estimators introduced variance, while too many increased computational cost with marginal gains. Optuna's tuning procedure allowed an efficient search for the optimal balance. Bootstrap sampling, an intrinsic feature of Random Forests, contributed additional robustness by introducing diversity among trees, reducing the risk of overfitting to training data.

Optimisation with Optuna

To optimise model performance, hyperparameter tuning was performed using Optuna, an advanced optimisation framework that employs efficient search strategies such as Tree-structured Parzen Estimators (TPE). Unlike traditional grid search, Optuna dynamically explores the hyperparameter space, focusing on promising regions and discarding poor configurations early.
The main hyperparameters optimised included the number of estimators, maximum depth, minimum samples per split, and minimum samples per leaf. Optuna iteratively proposed candidate configurations, which were then evaluated using stratified k-fold cross-validation. The optimisation process significantly reduced computational overhead compared to exhaustive methods, while achieving strong improvements in predictive accuracy.

Metrics

Evaluating model performance required a range of metrics beyond raw accuracy. Precision measured the proportion of predicted approvals that were correct, while recall quantified the proportion of actual approvals successfully identified. The F1-score, which balances precision and recall, was especially useful in the presence of class imbalance.
Additionally, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) provided a robust measure of discrimination ability across thresholds. Confusion matrices offered further insights into misclassification patterns, such as false rejections or false approvals, which carry different implications for lending institutions.

Results Summary

The optimised Random Forest model achieved strong performance across evaluation metrics. Accuracy exceeded baseline benchmarks established by logistic regression and decision trees, while the F1-score demonstrated balanced treatment of both approved and rejected classes. ROC-AUC values confirmed strong discriminatory power.
Feature importance analysis revealed that applicant income, credit history, and loan amount were among the most influential predictors, aligning with domain knowledge. The ability to interpret feature contributions added trustworthiness to the model, despite its ensemble complexity.

Validation Procedures

Robust validation procedures ensured the reliability of results. The dataset was split into training, validation, and testing sets to allow iterative development while reserving a final benchmark. Stratified k-fold cross-validation was applied during optimisation to account for class imbalance and ensure fair performance assessment across folds.
Validation confirmed that the model generalised well across subsets of the data, with minimal performance degradation between training and validation sets. This indicated effective regularisation and appropriate handling of imbalance during preprocessing.

Ethical Considerations

The use of machine learning in loan approval decisions raises critical ethical questions. One concern is bias in historical data, which may propagate or even amplify unfair lending practices if not addressed. For example, demographic features such as gender or marital status may correlate with approval outcomes due to systemic inequities, raising fairness issues.
Another consideration is transparency. Applicants and regulators may demand explanations for decisions, yet ensemble models like Random Forests are inherently less interpretable. While feature importance helps, it cannot always provide clear case-by-case justifications. Institutions must therefore balance predictive performance with explainability.
Finally, the consequences of misclassification differ in severity. A false rejection may unfairly deny access to financial resources, while a false approval may expose lenders to financial losses. Ethical deployment of such systems requires careful calibration of thresholds to align with fairness and risk management objectives.

Future Work

Future work should extend beyond Random Forests to include comparisons with other advanced ensemble methods, such as Gradient Boosting and XGBoost, which may offer improved accuracy at the cost of interpretability. Additionally, explainable AI (XAI) methods such as SHAP values could be integrated to enhance transparency and provide applicant-specific explanations for decisions.
Expanding the dataset to include macroeconomic indicators and longitudinal applicant data could improve predictive accuracy and resilience to changing economic conditions. Furthermore, real-world deployment should incorporate continuous monitoring systems to detect performance drift and biases over time, ensuring models remain fair and reliable.

Conclusion

This report demonstrated the effectiveness of Random Forests for predicting loan approval decisions. By implementing careful preprocessing, robust validation, Optuna-based hyperparameter tuning, and ethical considerations, the model achieved high predictive accuracy and balanced treatment of both approved and rejected applications. While challenges remain regarding transparency and fairness, Random Forests represent a promising

foundation for modern lending decision systems, capable of combining performance with practical applicability in financial institutions.