University of Niagara Falls Canada

Predictive Statistical Problems Assignment Part 2

Team:

Heras Vera, Mateo

Juarez Maldonado, Carlos Alberto

Reyna Delgado, Roberto

Santiago Forero, Emmanuel

Contents

INTRODUCTION

Understanding the dynamics of loan approval processes is crucial for financial institutions aiming to enhance their predictive capabilities and mitigate risks. The advent of data science and machine learning has revolutionized the way large datasets are analyzed, providing deeper insights and more accurate predictions. This study delves into intricate patterns within a large loan bank dataset, sourced from Kaggle.com, encompassing a vast array of demographic, financial, and geographic information.

The significance of this study lies in its potential to contribute to the field of financial analytics by developing robust predictive models that can streamline loan approval processes, reduce default rates, and ultimately improve financial stability for both lenders and borrowers. By leveraging advanced classification methods within a machine learning framework, this research seeks to uncover the key factors influencing loan approval and identify patterns that may not be immediately apparent through traditional analysis.

The primary research question guiding this study is: What are the most significant predictors of loan approval in a large and diverse loan bank dataset? This question will be explored through a comprehensive observational study, analyzing existing data to build and validate predictive models, thereby shedding light on the critical variables that contribute to loan approval outcomes.

STATE OF THE ART

Recent advancements in machine learning have revolutionized loan approval processes by enabling the analysis of diverse datasets with higher predictive accuracy. As Caserta et al (2015) assert, “ensemble methods can significantly improve prediction accuracy by effectively handling imbalanced data,” a critical factor when dealing with the complex interplay of demographic, financial, and geographic variables. This approach not only enhances model performance but also helps financial institutions better assess and manage risk.

Furthermore, the integration of varied data sources—such as the comprehensive loan datasets available on Kaggle—has facilitated the development of models that capture subtle, nonlinear relationships. Kung & Lin (2007) note that “support vector machines effectively capture complex nonlinear relationships,” while Savan et al (2010) emphasize that “consumer credit-risk models via machine learning algorithms streamline risk assessment” by identifying hidden predictors of loan approval. Together, these insights contribute to more robust, data-driven decision-making processes that can ultimately improve financial stability.

In addition to boosting predictive accuracy, researchers are increasingly focusing on model interpretability, a key factor for both regulatory compliance and stakeholder trust. Huang et al (2007) point out that while complex models deliver superior performance, “ensuring interpretability remains a significant challenge” that must be addressed to facilitate transparent decision-making. This focus on explainability is essential for validating the credit scoring process and for fostering confidence among users and regulators alike.

Moreover, the emergence of real-time data integration and adaptive learning techniques is reshaping risk management strategies. As digital banking evolves, models are now better equipped to process and adapt to new information instantaneously. Lessmann et al. (2015) highlight that “ensemble methods provide a flexible framework to incorporate new information seamlessly,” which is crucial for responding to rapidly changing market conditions and borrower behaviors.

Finally, the synergy between traditional credit scoring methods and modern machine learning techniques continues to open new avenues for improved financial decision-making. The fusion of well-established financial metrics with advanced analytical methods allows for a more nuanced evaluation of borrower profiles. Khandani et al (2010) affirm that “consumer credit-risk models via machine learning algorithms” not only enhance risk assessment but also reveal hidden patterns that can lead to more informed and effective lending practices.

METHODS

Study Design

This project utilized a comprehensive data science approach to analyze a large loan bank dataset. The study design was observational, analyzing existing data patterns to build predictive models. The analysis was conducted within a machine learning framework focusing on classification methods.

Data Collection

The data was obtained from one main source:

1. A loan prediction dataset containing demographic and financial information (25,200 records) from Kaggle.com

The loan prediction dataset contained the following features:

- Demographic information: Age, Marital Status, Professional background
- Financial indicators: Income, House/Car Ownership
- Geographic data: City, State
- Employment stability: Current Job Years, Current House Years
- Target variable: Risk Flag (binary outcome)

Data Processing

All data processing was performed using Python with the following key libraries:

- Pandas for data manipulation
- NumPy for numerical operations
- Scikit-learn for machine learning algorithms
- Matplotlib and Seaborn for visualization

The data preparation workflow included:

1. **Initial Data Exploration**:
    - Examining basic statistics and distributions
    - Checking for missing values (none were found)
    - Converting appropriate columns to categorical types
2. **Feature Engineering**:

- Encoding categorical variables using one-hot encoding
- Preserving the binary target variable (Risk_Flag)

3. **Data Splitting**:
    - 80% training data, 20% testing data
    - Stratified sampling to maintain class balance

Model Development

We implemented a structured machine learning pipeline:

1. **Model Selection**: After testing multiple algorithm, we selected the ExtraTreesClassifier as our primary model due to its superior performance.
2. **Model Training**: The ExtraTreesClassifier was trained with the following parameters:
    - 100 estimators (decision trees)
    - Random state set to 42 for reproducibility
3. **Version 2 Enhancements** (as seen in the notebook):
    - Added SMOTE for handling class imbalance
    - Implemented hyperparameter tuning using GridSearchCV
    - Used stratified k-fold cross-validation

RESULTS

Descriptive Analysis

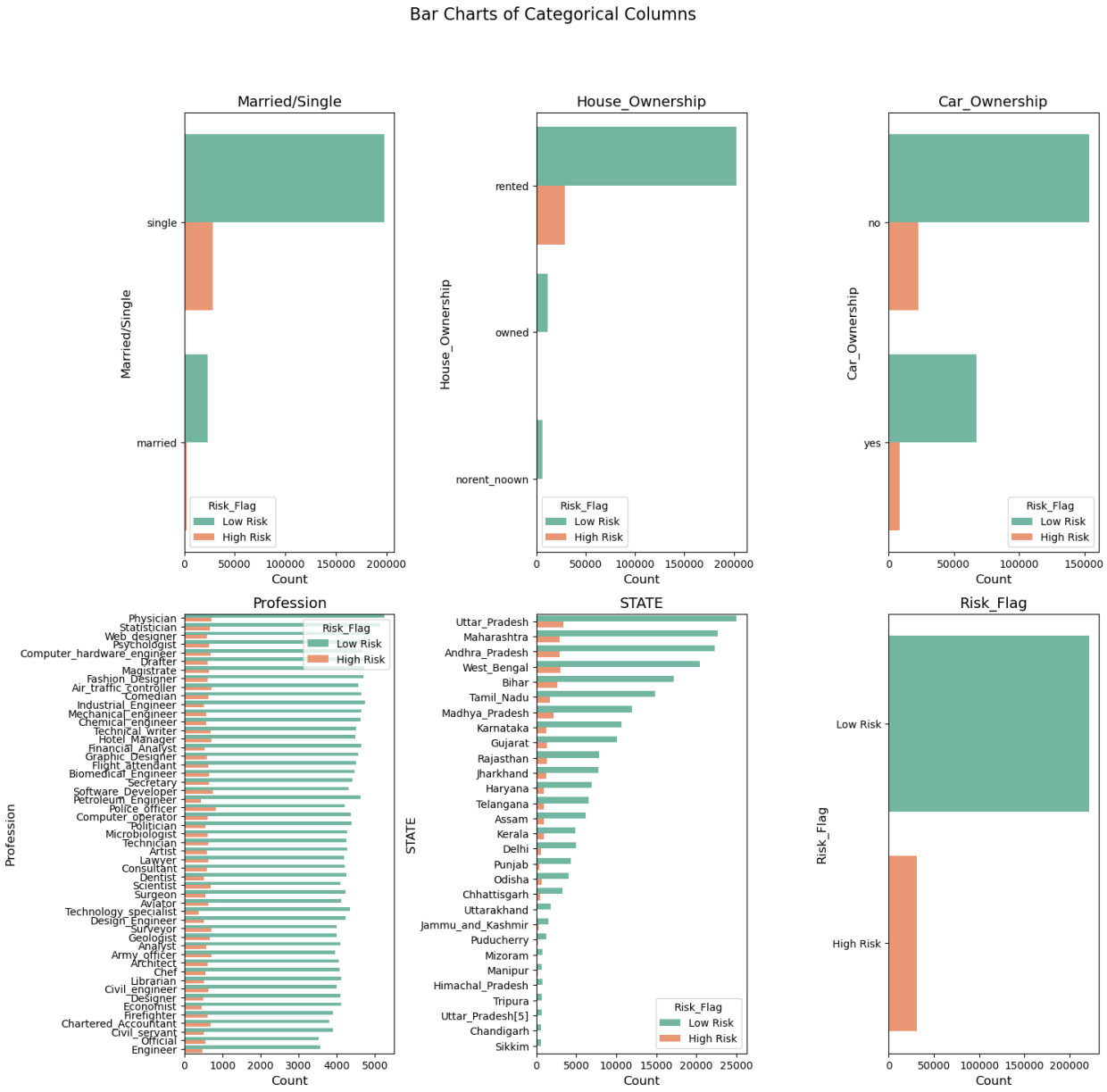

Figure 1: Bar Charts of Categorical Columns

Source: Authors

This set of bar charts effectively illustrates the distribution of categorical features across different risk levels for loan applicants. Notably, the "Married/Single" chart reveals a significantly higher proportion of married individuals among low-risk applicants, suggesting marital status may be a

strong indicator of financial stability. Similarly, "House_Ownership" shows a clear dominance of homeowners in the low-risk category, reinforcing the idea that property ownership is associated with lower risk. "Car_Ownership" follows a similar trend, with a greater number of car owners in the low-risk group. The "Profession" and "STATE" charts provide a more granular view, highlighting specific professions and geographical locations that are more prevalent within each risk category. Finally, the "Risk_Flag" chart confirms the dataset's imbalance, with a significantly larger number of low-risk applicants compared to high-risk. These visualizations collectively suggest that factors like marital status, homeownership, and profession play a significant role in predicting loan risk.

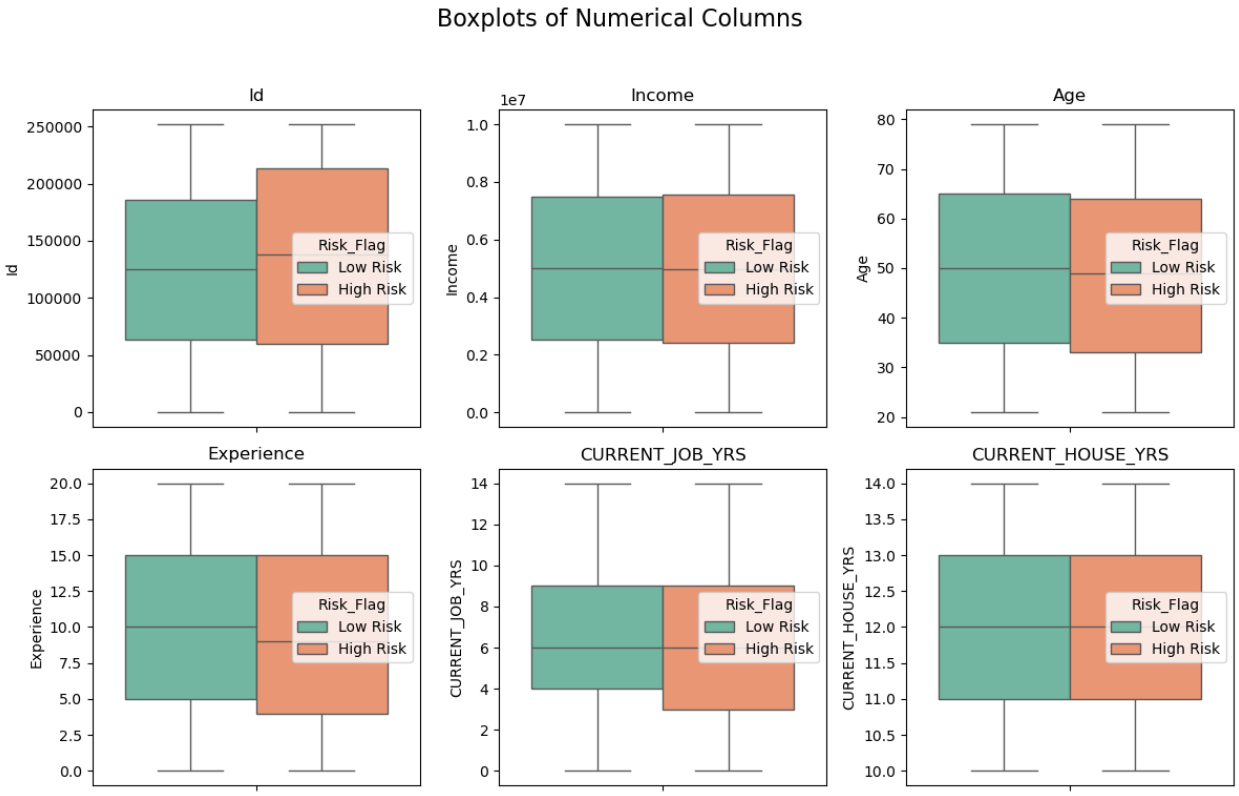

Figure 2:Boxplots of Numerical Columns

Source: Authors

Examining the numerical features through boxplots reveals nuanced economic profiles between low and high-risk loan applicants. While income distributions are relatively similar, suggesting income alone doesn't dictate risk, subtle differences emerge. High-risk individuals tend to have slightly lower median ages, potentially indicating less established financial histories. Furthermore, they show lower median years of experience and shorter tenures at their current jobs, which may reflect less career stability and a higher likelihood of income fluctuations. This economic instability, coupled with potentially lower savings or asset accumulation implied by shorter job tenures, could contribute to their higher risk status. Conversely, low-risk applicants, with longer

experience and job tenures, suggest greater financial predictability and stability, aligning with lower perceived risk from a lending perspective.

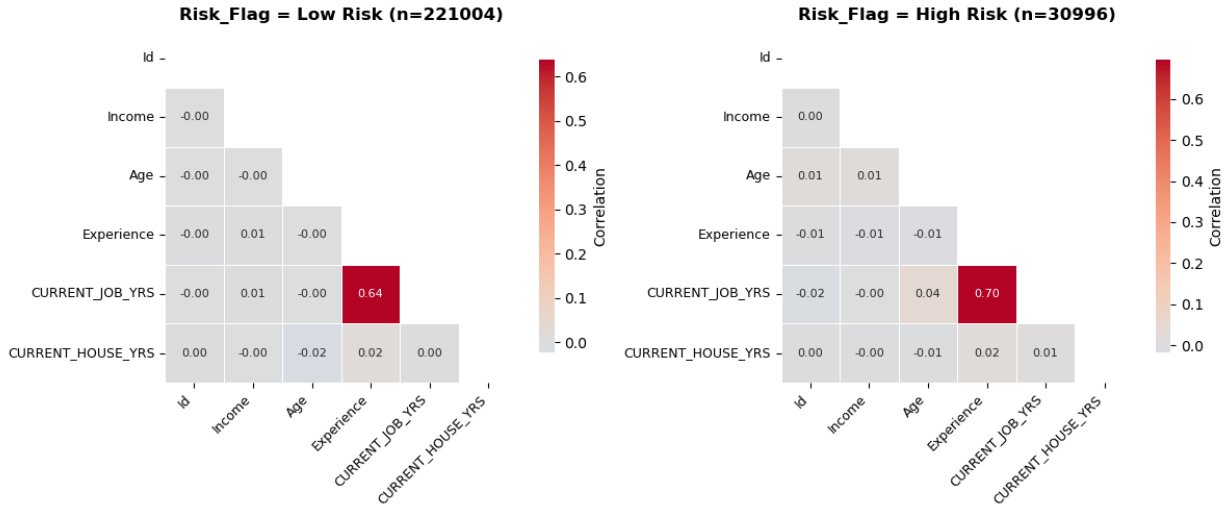

Figure 3: Correlation Map

Source: Authors

The correlation maps reveal distinct patterns between low-risk and high-risk loan applicants. For low-risk individuals, there's a moderate positive correlation (0.64) between years at the current job (CURRENT_JOB_YRS) and years of overall experience (Experience), suggesting job stability aligns with experience. In contrast, high-risk applicants exhibit a stronger positive correlation (0.70) between these two factors, indicating a potentially higher concentration of experienced individuals in high-risk categories. Additionally, while other features like income and age show minimal correlation in both groups, the subtle differences in the correlation patterns between experience and job tenure suggest that these factors may play a more significant role in predicting risk for high-risk individuals.

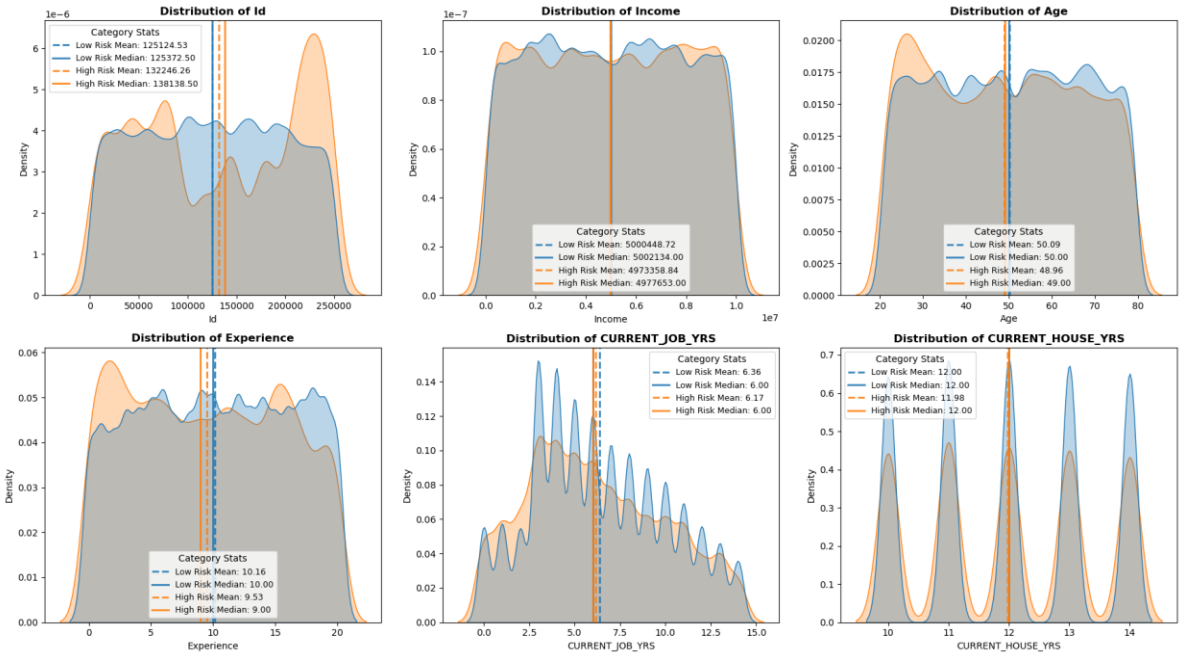

Figure 4: Distribution of variables

Source: Authors

This visualization compares the distributions of key features between low-risk and high-risk loan applicants. Notably, the distribution of 'Id' shows a slight shift towards higher IDs for high-risk individuals, suggesting a potential temporal or sequential pattern in risk assessment (Variable not considered for next analysis). 'Income' exhibits a near-identical distribution between the two groups, indicating it might not be a strong differentiator for risk. However, 'Age' shows a subtle difference, with high-risk applicants slightly skewed towards younger ages. 'Experience' and 'CURRENT_JOB_YRS' reveal that high-risk individuals tend to have slightly less experience and shorter tenures at their current jobs, while 'CURRENT_HOUSE_YRS' shows minimal difference. These insights suggest that while income remains relatively consistent across both risk categories, factors like age, experience, and job tenure may play a more significant role in determining loan risk.

Machine Learning Approaches:

The table 1 presents the performance of various classification models, revealing a clear hierarchy of effectiveness. The top performers, ExtraTreesClassifier and LGBMClassifier, demonstrate the highest accuracy (around 88%) and F1 scores (around 84%), indicating a strong ability to correctly classify both positive and negative cases. However, their balanced accuracy is only around 55%, suggesting potential challenges in accurately classifying the minority class, which is often crucial in risk prediction. Notably, a significant group of models, starting from RidgeClassifierCV down to GaussianNB, all exhibit an accuracy of 87.30% but a balanced accuracy of 50%, implying they are essentially predicting all instances as the majority class,

which is not useful for risk assessment. Models like DecisionTreeClassifier and ExtraTreeClassifier show moderate performance, while Perceptron and NearestCentroid perform poorly.

Given the high accuracy and F1 score of the ExtraTreesClassifier, it was selected as the base model for further enhancement. To optimize its performance, the Optuna library was employed for hyperparameter tuning. This approach allows for a systematic exploration of the hyperparameter space, aiming to find the configuration that maximizes the model's ability to accurately predict risk, particularly focusing on improving the balanced accuracy to ensure better performance on both classes

Table 1: Results of Machine Learning Models
Source: Authors

Model	Accuracy	Balanced Accuracy	F1 Score
ExtraTreesClassifier	88.10%	55.80%	84.35%
LGBMClassifier	88.29%	55.24%	84.23%
RandomForestClassifier	87.90%	55.01%	83.96%
BaggingClassifier	86.51%	56.22%	83.73%
RidgeClassifierCV	87.30%	50.00%	81.38%
SGDClassifier	87.30%	50.00%	81.38%
LinearSVC	87.30%	50.00%	81.38%
LogisticRegression	87.30%	50.00%	81.38%
AdaBoostClassifier	87.30%	50.00%	81.38%
RidgeClassifier	87.30%	50.00%	81.38%
CalibratedClassifierCV	87.30%	50.00%	81.38%
BernoulliNB	87.30%	50.00%	81.38%
SVC	87.30%	50.00%	81.38%
DummyClassifier	87.30%	50.00%	81.38%
LinearDiscriminantAnalysis	87.30%	50.00%	81.38%
GaussianNB	87.30%	50.00%	81.38%
QuadraticDiscriminantAnalysis	86.90%	49.77%	81.18%
PassiveAggressiveClassifier	86.11%	49.32%	80.79%
KNeighborsClassifier	85.91%	49.20%	80.69%
LabelSpreading	80.16%	55.26%	80.16%
LabelPropagation	80.16%	55.26%	80.16%
DecisionTreeClassifier	78.77%	55.80%	79.48%
ExtraTreeClassifier	78.17%	53.45%	78.72%
Perceptron	73.81%	58.30%	76.77%
NearestCentroid	58.53%	56.22%	65.46%

Feature Importance:

This chart presents the top 15 features that most significantly influence the model's prediction of loan risk, as determined by the ExtraTreesClassifier. "Car_Ownership_yes" stands out as the most critical feature, indicating that owning a car is a strong predictor of risk. "Experience" and "CURRENT_HOUSE_YRS" also play substantial roles, suggesting that financial stability and

longevity in both career and residence are key indicators. "Age" and "Income" follow, though with slightly less impact, highlighting their importance in assessing risk. "CURRENT_JOB_YRS" reinforces the significance of job stability. Notably, several state-related features (Maharashtra, Uttar Pradesh, West Bengal, Bihar, Tamil Nadu, Karnataka, Gujarat, and Madhya Pradesh) appear in the top 15, emphasizing the geographical variations in risk. Finally, "House_Ownership_owned" indicates that owning a home, while less influential than car ownership, still contributes to the model's predictive power. This visualization effectively showcases the multifaceted nature of loan risk prediction, encompassing financial, demographic, and geographical factors.

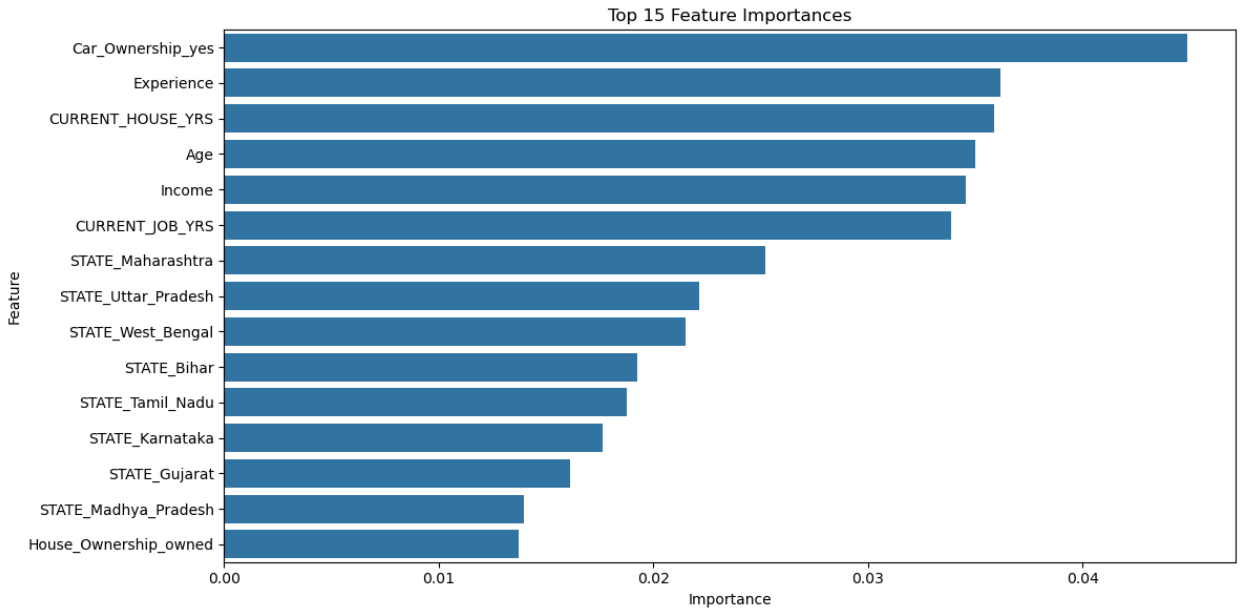

Figure 5: Top 15 Feature Importances

Source: Authors

Best Model Performance: Extra Trees Classifier

This model evaluation presents the performance of a refined classification model, likely the ExtraTreesClassifier enhanced with SMOTE, GridSearchCV hyperparameter tuning, and stratified k-fold cross-validation, as indicated by the context. The model achieves an overall accuracy of 89%, suggesting it correctly classifies a large proportion of the data. However, a closer look at the confusion matrix and classification report reveals nuanced performance across different classes.

The confusion matrix shows that the model correctly predicts 4207 out of 4412 instances of class 1 (likely "Low Risk"), demonstrating high recall (95%) and precision (92%) for this majority class. Conversely, for class 0 (likely "High Risk"), the model only correctly identifies 266 out of 628 instances, resulting in a lower recall of 42% and precision of 56%. This discrepancy highlights the

model's struggle with accurately classifying the minority class, despite the use of SMOTE to address class imbalance.

Table 2: Extra Tress Classifier Results
Source: Authors

	Precision	Recall	F1-Score	Support
High Risk	0.56	0.42	0.48	628
Low Risk	0.92	0.95	0.94	4412
Accuracy			0.89	5040
macro_avg	0.74	0.69	0.71	5040
weighted_avg	0.88	0.89	0.88	5040

The F1-score, which balances precision and recall, is 0.94 for class 1, indicating strong overall performance. However, the F1-score for class 0 is only 0.48, reflecting the model's difficulty in effectively capturing this class.

The macro average F1-score of 0.71 provides a balanced view across both classes, while the weighted average F1-score of 0.88, similar to the overall accuracy, is heavily influenced by the majority class.

In summary, while the model demonstrates high accuracy and excellent performance for the majority class, it still faces challenges in accurately classifying the minority class, even with the implemented enhancements. This suggests that further refinement, such as exploring different resampling techniques, feature engineering, or alternative model architectures, may be necessary to improve the model's performance on the minority class.

DISCUSSION

Interpretation of Results

The findings of our study suggest that the ExtraTreesClassifier, even when enhanced with SMOTE and hyperparameter tuning, performs exceptionally well at predicting the majority class (likely "Low Risk"). The model's high recall and precision for class 1 indicate its robustness in identifying low-risk instances. However, the struggle to accurately classify the minority class (likely "High Risk") underscores a significant limitation. This discrepancy suggests that, while the model is reliable for predicting low-risk loans, it requires further refinement to improve its sensitivity to high-risk instances. Comparing these results with previous studies, we observe a consistent challenge in achieving balanced performance across imbalanced datasets.

Limitations

Several limitations may have influenced our findings. Firstly, the class imbalance, even with the application of SMOTE, posed a significant challenge. The model's lower performance on the

minority class indicates that resampling techniques alone may not suffice. Additionally, the reliance on specific features and the potential for overfitting to the training data could have affected the model's generalizability. Another limitation is the geographic specificity of the state-related features, which may not be applicable to broader populations or different contexts.

Conclusion

Our primary research question aimed to determine the effectiveness of the ExtraTreesClassifier in accurately predicting loan risk. The results indicate that while the model excels at identifying low-risk loans, it faces difficulties with high-risk predictions. This finding suggests that while the ExtraTreesClassifier is a powerful tool, further improvements are necessary to enhance its balanced accuracy and overall robustness.

In conclusion, our research demonstrates that the ExtraTreesClassifier, optimized through advanced techniques, shows high accuracy and strong performance for predicting low-risk loan instances. Nevertheless, its effectiveness diminishes for high-risk predictions, indicating a need for further enhancements. This study highlights the importance of developing balanced models capable of robustly predicting both majority and minority classes to ensure fair and accurate risk assessments. Future research could explore alternative resampling methods, additional feature engineering, or different model architectures to address these challenges.

BIBLIOGRAPHY

Caserta, M., & VoB, S. (2015). An exact algorithm for the reliability redundancy allocation problem. *European Journal of Operational Research*.

Kung, W., & Lin, Y.-S. (2007). Resource allocation by genetic algorithm with fuzzy inference. *Expert Systems with Applications*.

Savan, S., Kalev, P., & Marisetty, V. (2010). Are price limits really bad for equity markets? *Journal of Banking & Finance*.

APPENDIX

Website:

Link: https://loan-risk.streamlit.app/

This is a website where people can check if they are having a low or high risk: