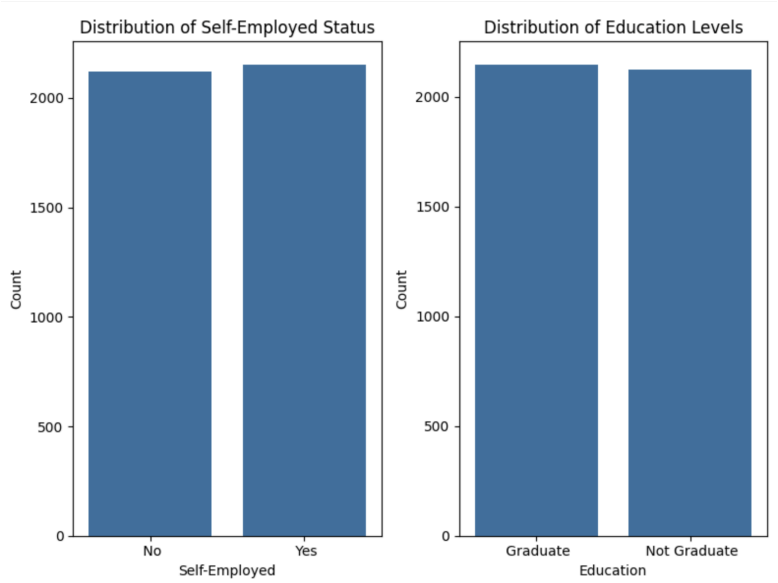
Malika Khakimova (mjk9984)
Data Bootcamp
Final Project
Prof. Jacob Koehler

Predicting the Outcome of Loan Application

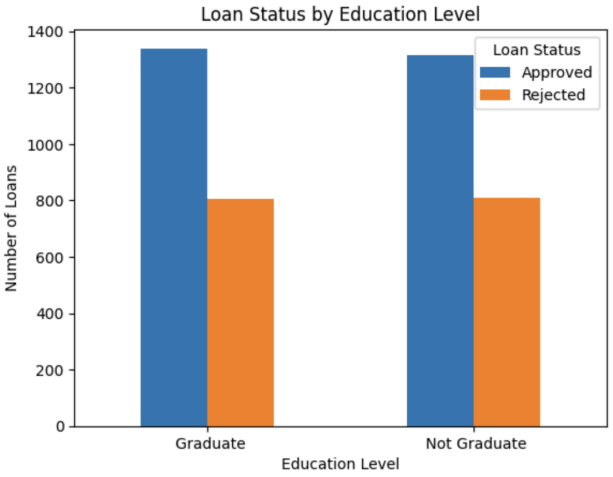
Introduction and Problem Statement: Loan approval is an important process in the financial sector, directly impacting both lending institutions and borrowers. For my final project, I plan to develop a predictive model that accurately estimates the likelihood of loan approval using a dataset of financial records and associated information. This dataset includes key features such as credit scores, income levels, employment status, loan terms, and asset values. The predictive model has practical applications for both lenders and borrowers. Lending institutions can use the model's insights to streamline decision-making, reduce default risks, and optimize resource allocation. Borrowers, on the other hand, can gain a clearer understanding of the factors that influence loan approval, helping them to make informed financial decisions. Moreover, the successful development of this model will enhance the predictive capabilities of the financial industry, supporting more efficient and equitable loan approval systems. By enabling data-driven decisions, the model can help stakeholders adapt their strategies to evolving market demands, fostering transparency and fairness in lending practices. In this study, several machine learning models were implemented and evaluated for their performance. Among these, the Random Forest model proved to be the most effective in predicting loan status based on the given features.

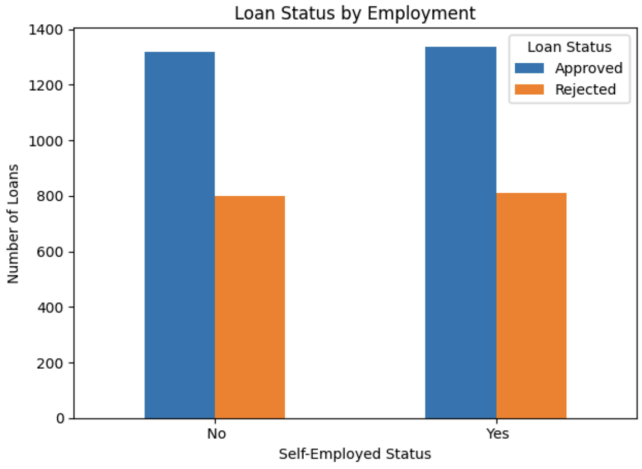
Data Description: The dataset, sourced from Kaggle in CSV format, contains 4,269 records with 13 attributes related to loan applications. The target variable for this task is **'loan status,'** a categorical variable indicating whether a loan is **'approved'** or **'rejected.'** While some preprocessing is required, such as removing whitespace from column names and encoding the target variable, most of the features used for modeling are numeric. The dataset includes attributes such as credit score, income, employment status, education level, loan term, loan amount, and value of different assets. On initial analysis, most loan applications from the dataset are approved, with rejections accounting for approximately 37.8% of cases.

Regarding the distribution of applicants' employment status and education level, the data is well-balanced across these categories.



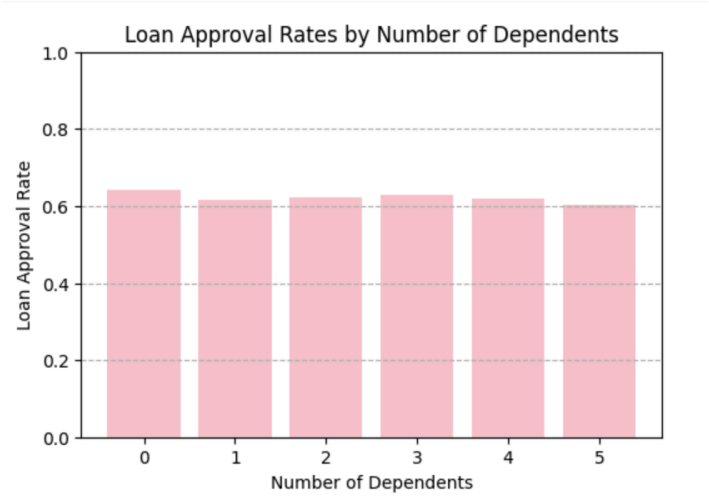
When analyzing loan status by employment and education levels, the approval rates remain consistently higher than rejection rates across all categories. This suggests that neither education level nor employment status significantly affects the likelihood of loan approval.



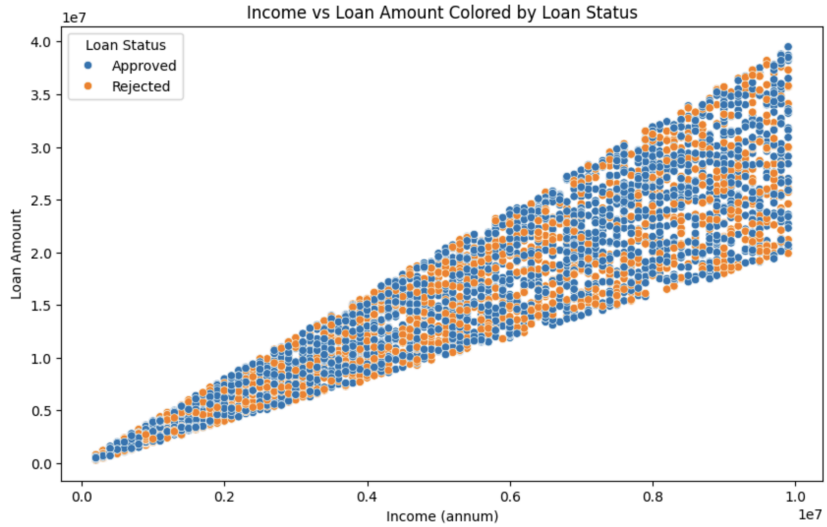


As for the number of dependents of the applicants in the dataset, the approval rate for loan applications remains relatively consistent across all categories of dependents, ranging from 60%

to 65%. There is little variation in approval rates, regardless of whether an applicant has no dependents or up to five dependents.

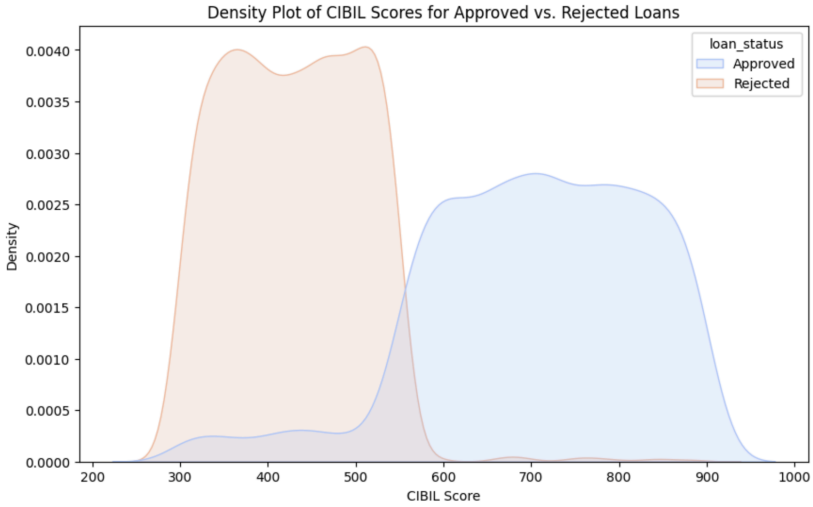


Expectedly, the relationship between annual income and loan amount shows a strong positive linear correlation, with a distinct triangular pattern where higher incomes are associated with a wider range of loan amounts. As income increases, the maximum loan amount tends to rise as well. Both approved and rejected loans are spread across all income levels and loan amounts, indicating that these two factors alone do not determine the loan approval status.

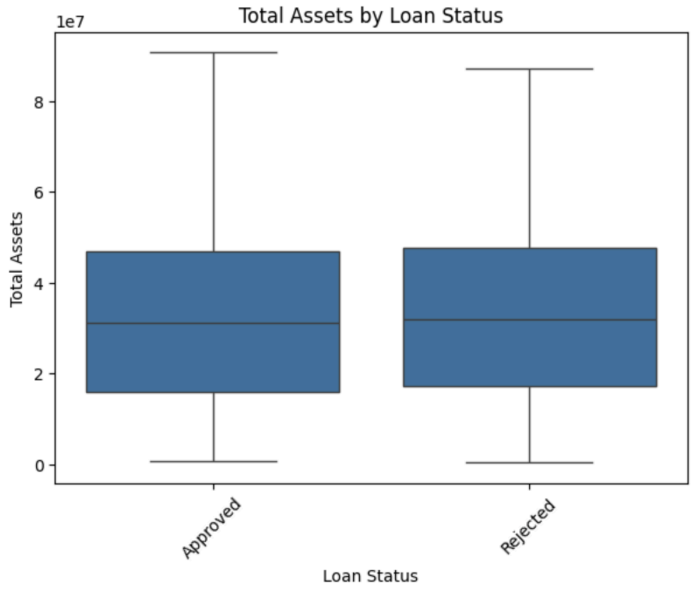


Next, I analyzed the impact of credit scores on loan approval by creating a density plot to visualize the distribution of CIBIL scores for approved and rejected loans. Approved loans are typically linked to higher credit scores, with a wider and more symmetrical distribution. In contrast, rejected loans are concentrated around lower scores, forming a sharper peak. A key

threshold appears around the 550–600 range, where the likelihood of approval increases significantly, though some overlap indicates flexibility in decision-making for borderline cases. Scores above 600 are strongly associated with approval, while those below 500 are mostly linked to rejections.



The final feature considered was assets, which include categories such as luxury, residential, commercial, and bank assets. For simplicity and consistency in predictive modeling and exploratory data analysis, these were combined into a single variable, ‘total assets’, representing the sum of all asset types for each applicant. An analysis using a box plot revealed that total assets show no significant differences between applicants whose loans were approved and those whose loans were rejected.



Just to check for any potential interesting differences, I calculated the mean asset values for approved and rejected loans across the four asset categories: Residential, Commercial, Luxury, and Bank assets. Overall, higher asset values do not consistently correlate with higher approval rates. There is no significant difference in the mean asset values between approved and rejected loans across these categories, with the variations being relatively small.

```
[ ] print(loan.groupby('loan_status')['residential_assets_value'].mean())
```

```
loan_status
Approved    7.399812e+06
Rejected    7.592498e+06
Name: residential_assets_value, dtype: float64
```

```
[ ] print(loan.groupby('loan_status')['commercial_assets_value'].mean())
```

```
loan_status
Approved    5.001355e+06
Rejected    4.926720e+06
Name: commercial_assets_value, dtype: float64
```

```
[ ] print(loan.groupby('loan_status')['luxury_assets_value'].mean())
```

```
loan_status
Approved    1.501660e+07
Rejected    1.530694e+07
Name: luxury_assets_value, dtype: float64
```

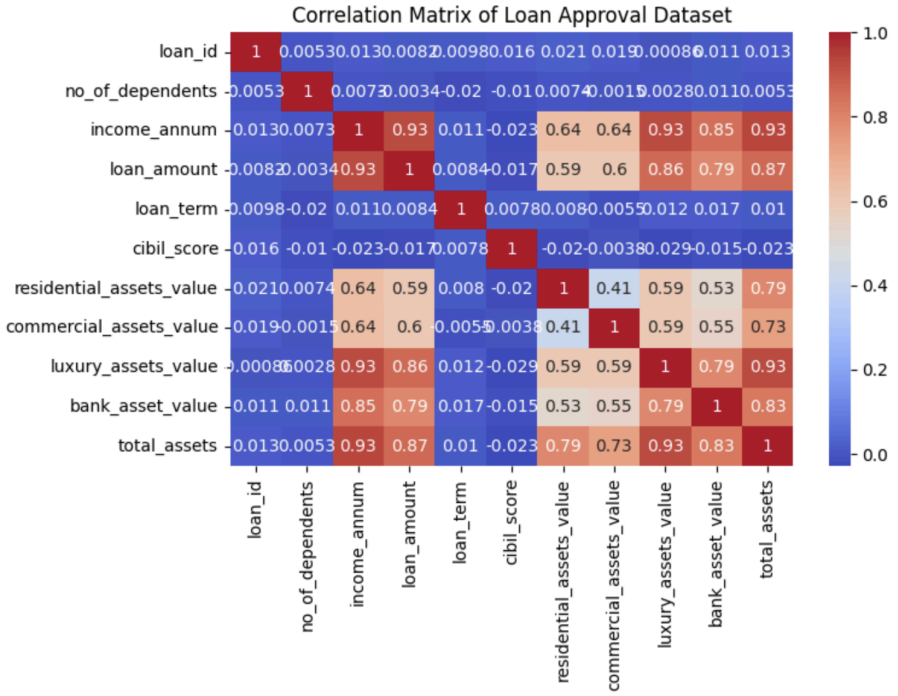
```
[ ] print(loan.groupby('loan_status')['bank_asset_value'].mean())
```

```
loan_status
Approved    4.959526e+06
Rejected    5.004960e+06
Name: bank_asset_value, dtype: float64
```

Finally, I examined the correlation matrix for the numeric attributes in the dataset. Strong positive correlations were observed between income and luxury asset value (0.93), total assets (0.93), and bank asset value (0.85). Loan amount showed a strong correlation with both income (0.93) and total assets (0.87), while luxury asset value was also strongly correlated with total assets (0.93). Residential and commercial assets had a moderate correlation of 0.41. In contrast, CIBIL score and demographic variables such as the number of dependents and loan term exhibited weak or minimal correlations with most other variables.

Overall, credit score seems to be the most influential factor in loan approval decisions. A potential limitation that arises from this analysis is the possibility of multicollinearity, as some attributes show strong correlations with each other, which could result in redundant information in predictive modeling. As I move forward with constructing predictive models, I am assuming that the dataset accurately reflects the financial records of loan applicants, even though it may

not fully capture other important factors, such as local economic conditions or more specific personal circumstances, which could also influence loan decisions.



Models and Methods: For this project, I selected Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) to predict loan approval based on various applicant characteristics. Loan approval prediction is a binary classification problem, where the goal is to determine whether an applicant will be approved or denied based on financial data and other features. Logistic Regression is great for this task as it handles binary outcomes well and offers clear insights into how input variables influence the probability of loan approval. Random Forest, on the other hand, excels at capturing complex interactions and non-linear relationships, while also providing feature importance scores that help identify the most influential predictors. KNN complements the other models by detecting local patterns and similarities, making it useful for identifying clusters of applicants with comparable profiles likely to share similar outcomes. Alongside these three models, I also created a baseline model that predicts the majority class for all test samples, which serves as a reference point for comparison.

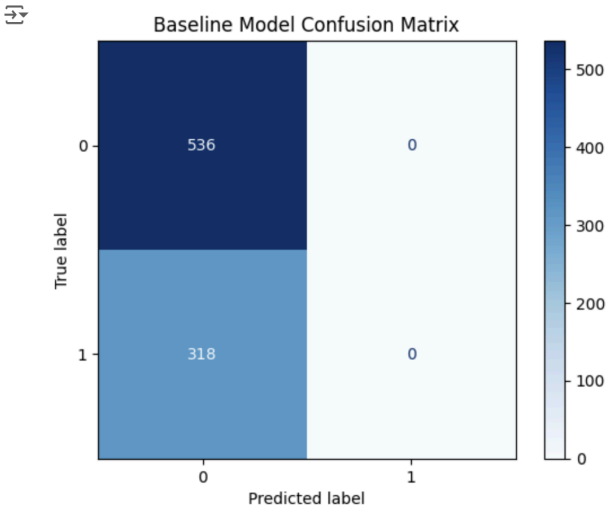
I started the construction of predictive models by encoding the target variable (loan_status) using LabelEncoder for binary classification. To ensure fair model performance, I standardized the

features using StandardScaler, which is essential for models sensitive to feature scaling, and split the data into training and testing sets, reserving 20% for testing and 80% for training. After training and testing the models, I evaluated them based on accuracy and precision, using confusion matrices to visualize predictive power. Accuracy provides an overall measure of model performance, while precision helps assess the reliability of positive loan approvals to minimize false positives. Finally, recognizing the Random Forest model as the most accurate, I applied 10-fold cross-validation to it to obtain robust performance metrics and reduce overfitting. I also extracted and analyzed the feature importances from the Random Forest model to check which features are most influential in predicting loan approval.

Results and Interpretation: After model construction, I get the following results for each model:

1) Baseline Model

Accuracy Score: 0.6276346604215457

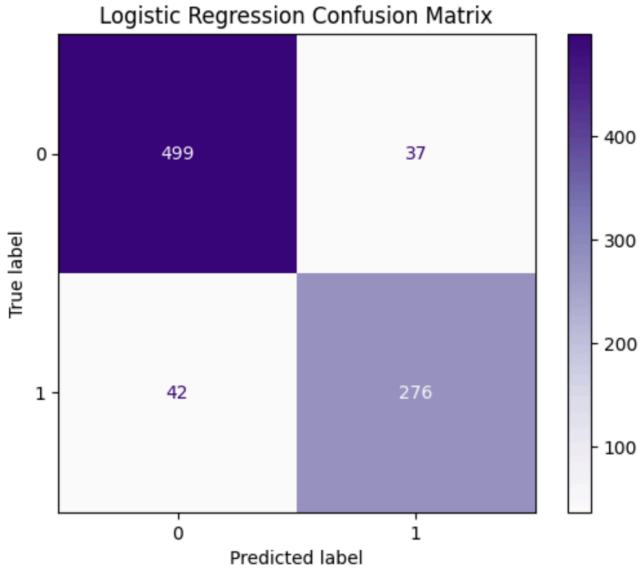


The ‘naive’ model predicts loan approval status based solely on the majority class, assuming all loans fall into the dominant category, either "approved" or "rejected." This approach achieves a moderate accuracy of **62.76%**, but it entirely fails to identify positive cases, resulting in zero precision. Although simplistic, the naive model serves as a baseline for evaluating the performance of other models.

2) *Logistics Regression Model:*

Logistic Regression Accuracy: 0.9074941451990632

Logistic Regression Precision: 0.8817891373801917

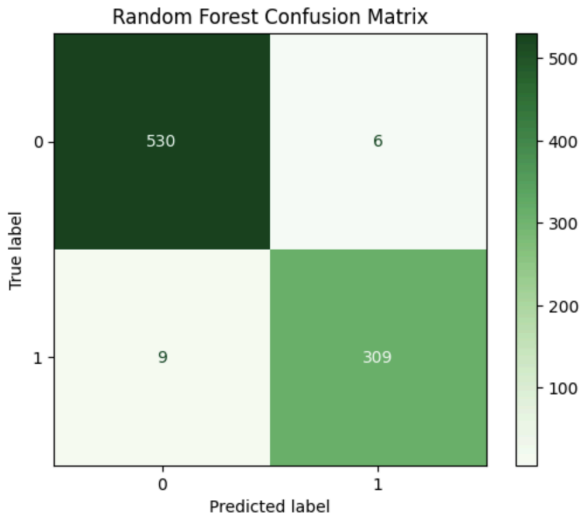


The logistic regression model significantly outperforms the baseline, demonstrating a higher accuracy and stronger precision. With 37 false positives and 42 false negatives, the model has a moderately low misclassification rate. The logistic regression model is both reliable and consistent, making it a solid choice for loan approval prediction.

3) *Random Forest*

Random Forest Accuracy: 0.9824355971896955

Random Forest Precision: 0.9809523809523809

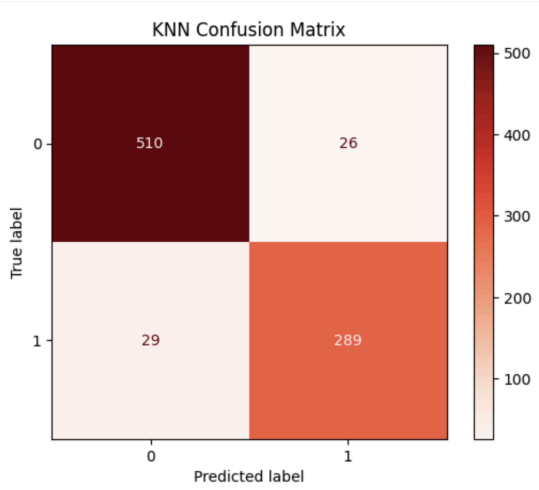


The random forest model delivers exceptional performance, achieving the highest accuracy and precision among the evaluated models. It misclassifies only 15 instances (6 false positives and 9 false negatives), showcasing high reliability and a low error rate. This makes it an excellent choice for predicting loan approval status with a high degree of confidence.

4) KNN Model

KNN Accuracy: 0.9355971896955504
KNN Precision: 0.9174603174603174

The KNN model performs well, ranking between the logistic regression and random forest models in terms of accuracy and precision. While it effectively identifies true negatives (class 0), it still misclassified 26 false positives and 29 false negatives. Although the model is robust, there is room for improvement.



As the Random Forest Model performed exceptionally well, I ensured its reliability across multiple subsets of the training data using K-fold cross-validation. Instead of evaluating the model on a single train-test split, cross-validation divides the data into 10 folds, iteratively training the model on 9 folds and testing it on the remaining one. Such a process reduces overfitting risks and validates model consistency.

Cross-validation accuracy scores: [0.98830409 0.98245614 0.97368421 0.99122807 0.97660819 0.98240469 0.98533724 0.99120235 0.99120235 0.97653959]

Average Cross-validation accuracy: 0.9838966918763183

The cross-validation accuracy scores, ranging from 97.37% to 99.12%, with an average of 98.39%, show that the Random Forest model is both highly accurate and consistently reliable across all data splits, making it a well-suited choice for predicting loan approvals.

To identify the features that the model relies on the most and to quantify their contributions to its predictive power, I applied the feature importance technique of the Random Forest model.

```
Top features based on importance:
             Feature  Importance
3       cibil_score    0.837400
4         loan_term    0.065763
2       loan_amount    0.037471
1      income_annum    0.025252
0      total_assets    0.024888
5  no_of_dependents    0.009225
```

The CIBIL score is a single most critical factor in loan approval, contributing 83.74% to the model's decisions. A high CIBIL score reflects strong creditworthiness and a reliable repayment history, making it the primary determinant of approval likelihood. Following this, the loan term (6.58%) plays a somewhat important role, as longer terms may suggest increased financial flexibility but could also pose higher risks for lenders. Lenders appear to favor terms that match an applicant’s repayment capacity. The loan amount (3.75%) also influences loan decisions moderately, with larger amounts potentially increasing financial risk, while reasonable amounts relative to the applicant’s financial profile improve approval chances. Annual income (2.53%) has a smaller impact compared to credit and loan-specific metrics. Lastly, total assets (2.49%) provide additional assurance of financial stability but are a relatively minor factor in determining approval outcomes.

These trends revealed by the model match with common lending practices, which prioritize risk minimization and ensuring borrowers have the capacity to repay loans. A high credit score, such as the CIBIL score, is a key indicator of creditworthiness, reflecting a history of timely repayments and responsible financial behavior. Lenders often view applicants with higher scores as lower risk. Financial stability, demonstrated through higher income and substantial assets,

reassures lenders that the borrower has the resources to meet repayment obligations without financial strain. At the same time, lower loan amounts and shorter repayment terms reduce the financial exposure and duration of risk for lenders, making these applications more likely to be approved.

Conclusion and Next Steps: This project aimed to develop a predictive model for loan approval using financial and demographic data. Through the analysis, several machine learning models — Logistic Regression, Random Forest, and KNN — were tested, with Random Forest being the most effective in predicting loan approval status. The model demonstrated an impressive accuracy of 98.24% and a precision of 98.09%. The primary predictor identified by the model was the applicant's CIBIL score, which accounted for 83.74% of the decision-making process, followed by loan term, loan amount, income, and assets.

These findings have important implications for both lenders and borrowers. By using the Random Forest model, lenders can make more accurate and efficient loan decisions, saving time and reducing the costs of manual approval processes. The model helps with minimizing risks and improving the overall performance of their loan portfolios, making the lending process better for both sides. The model also provides clear insights into what factors affect loan approval, helping borrowers understand how they can improve their financial situation before applying. This transparency can increase their chances of being approved.

One key limitation of this project is the dataset's lack of time and location data, which prevents the model from accounting for external factors such as regional economic conditions, employment rates, and inflation levels. These factors can significantly influence loan approval decisions, as they affect a borrower's ability to repay and lenders' risk tolerance. The absence of such context may reduce the model's accuracy in capturing regional or temporal variations in lending practices. For example, economic and regulatory conditions vary over time and between regions, so a model trained on data from one context may not generalize well to another. While the model may perform well in environments with similar lending practices, its predictions could be less accurate in regions with distinct economic or regulatory conditions.

For next steps, one recommendation is to test the model on a larger, more diverse dataset to validate its performance across various demographic groups and loan types, including time and

location-specific data. This will enable the model to account for regional and temporal factors, improving its generalizability across various economic and regulatory contexts. Moreover, integrating macroeconomic indicators such as inflation, unemployment rates, and GDP growth can provide a broader perspective on the economic conditions influencing loan approval decisions. Incorporating a debt-to-income ratio as a feature to assess how much of an applicant's income is allocated to existing debts would further enhance the model's predictive power, as it is a critical factor for assessing an applicant's repayment capability. These steps may help to create a more robust, context-aware predictive model that delivers reliable results across different environments.