Loan Approval Prediction: Analyzing Financial and Demographic Factors

Introduction

In the financial industry, effective risk evaluation is key for loan approvals. Loan approval datasets, which include applicant demographics, financial history, and credit scores, help assess creditworthiness and guide decision-making. By analyzing these datasets, institutions can identify risk factors, optimize lending strategies, and set appropriate credit limits and interest rates. Machine learning models are increasingly used to predict loan approval outcomes, improving operational efficiency and decision-making. This data-driven approach enhances accuracy, reduces losses, and supports regulatory compliance, ultimately contributing to the stability of the financial system.

Problem Statement

The objective of this project is to compare the performance of two statistical learning algorithms on a loan approval classification task using a financial dataset. The aim is to determine which model offers superior predictive accuracy and interpretability, which are critical in making risk-informed loan approval decisions.

Data Description

This synthetic dataset comprises 20,000 records of personal and financial data with 31 independent variables(fearures), designed to facilitate the development of predictive models for risk assessment. The dataset includes key variables relevant to loan applications, such as Application Date, Age, Annual Income, Credit Score, Employment Status, Education Level, Experience, Loan Amount, Loan Duration, Marital Status,Risk Factor, etc. The dataset used contains variables such as applicant demographics, income, credit history, loan amount, and other financial details to predict loan approval status. The key features include:

- *Independent Variables: Age, Annual Income, Credit Score, Experience, Loan Amount, Loan Duration, Number of Dependents, Monthly Debt Payments, Credit Card Utilization Rate, and Net Worth*. It also contains categorical features such as *Employment Status* (e.g., Self-Employed, Unemployed), *Education Level* (e.g., Doctorate, High School), *Marital Status* (e.g., Married, Single), *Home Ownership Status* (e.g., Own, Rent), and *Loan Purpose* (e.g., Education, Debt Consolidation).
- *Response Variable:* Loan approval status (approved or not approved) measured in a binary scale.

Methodology

Data Preprocessing-

1. *Data Reduction:*
    - Columns deemed irrelevant to loan analysis, such as those that do not significantly impact loan approval predictions, were removed. This helped streamline the dataset and focus on variables with the most predictive power, enhancing model efficiency.
2. *Checking for Missing Values:*
    - No missing values were detected in the dataset, allowing for a seamless transition to further preprocessing without the need for imputation or data cleaning for missing entries.
3. *Outlier Detection:*
    - No outliers were detected in the dataset, ensuring that all values fall within an acceptable range. This avoids potential biases or distortions in the predictive model, enabling a more accurate and consistent analysis.

Feature Engineering

1. *Encoding Categorical Variables:*
    - *One-Hot Encoding*- Applied to variables like **Employment Status**, **Education Level**, **Marital Status**, **Home Ownership Status**, and **Loan Purpose** to convert categories into binary features.
    - *Label Encoding*- Used for the target variable, **Loan Approval Status**, to represent approval as a numerical value.

2. *Scaling Numerical Features:*
    - *Standard Scaling*: Used to normalize continuous features for improved model performance, including *Annual Income, Experience, Loan Amount, Net Worth, Total Debt-to-Income Ratio,* and *Risk Score*.

Exploratory Data Analysis (EDA)

1. *Descriptive Statistics:*
    - Summary statistics were generated to provide an overview of key metrics, including measures of central tendency (mean, median) and spread (standard deviation) for continuous variables. This helped in understanding the general distribution and tendencies within the dataset.
2. *Data Visualizations:*
    - **Proportion of Loan Approvals and Denials**: A bar chart displayed the ratio of approved versus denied loans, offering a quick insight into the target variable's distribution.
    - **Loan Purpose Distribution**: A chart was generated to show the frequency of each loan purpose, helping identify common reasons for loan applications.
3. *Multicollinearity Check:*
    - To detect multicollinearity among independent variables, the **Variance Inflation Factor (VIF)** was calculated. Variables with VIF scores greater than 10 were identified as highly collinear, signalling presence of multicollinearity and potential instability in the model.
4. *Correlation Analysis for Collinear Variables:*
    - For variables with high multicollinearity, **correlation analysis** was performed . This identified pairs of highly correlated variables, guiding the removal of redundant predictors to ensure a more reliable model. After removing these variables, multicollinearity was rechecked, confirming all remaining VIF scores were below 10.
5. *Feature Selection Based on Target Correlation:*
    - Finally, variables with an absolute correlation less than 0.1 with the target variable (**Loan Approval Status**) were removed, as they contributed little predictive power.
    - This process yielded a refined subset of predictor variables, optimized for model fitting and predictive accuracy.

Model Fitting and Evaluation

1. *Data Splitting:*
    - The dataset was split into training and testing sets using a 60:40 ratio. This allowed the models to be trained on a majority of the data while reserving a portion for independent testing and performance evaluation.
2. *Logistic Regression:*
   The Logistic Regression model, using a balanced class weight, was trained and evaluated with metrics like Accuracy, Precision, Recall, F1 Score, and AUC-ROC for both training and testing datasets. Confusion matrix and ROC-AUC curve were used for further assessment.
3. *Support Vector Machine (SVM):*
   SVM with a linear kernel was trained and evaluated using the same metrics as Logistic Regression. A confusion matrix and ROC-AUC curve were used to analyze its performance.
4. *Random Forest Classifier:*
   Random Forest Classifier was used to predict loan approval, evaluated with the same metrics as the other models. Confusion matrices and ROC-AUC curves were also used for performance analysis.

Feature Importance Evaluation

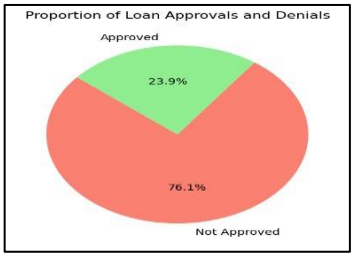
In *Logistic Regression*, feature importance is derived from the absolute values of the model's coefficients. A larger absolute coefficient signifies a stronger influence of that feature on the outcome. These values are unscaled, with higher magnitudes indicating more significant features. A separate bar chart was plotted to represent the importance of features based on the model's coefficients.

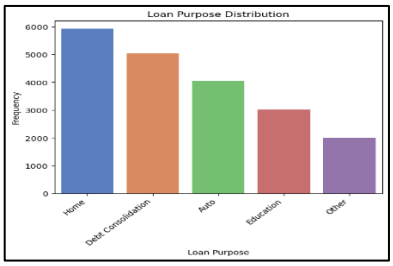
For the *Support Vector Machine (SVM)*, feature importance is also based on the absolute values of the coefficients in a linear SVM model. Similar to Logistic Regression, larger absolute values of coefficients indicate higher importance. A bar chart was used to show the feature importance for SVM.

For the *Random Forest Classifier*, feature importance is determined by the reduction in impurity (Gini index) during the tree-splitting process. Features that lead to larger reductions in impurity are deemed more important. The importance values range from 0 to 1, with higher values indicating greater influence on loan approval predictions. A bar chart was created to visualize the sorted importance of each feature.

Results and Interpretation

- **Descriptive Statistics:** The dataset includes 20,000 records with financial and personal data for loan applications. Key variables are age (18-80, mean 39.75), annual income (15,000-485,341, mean 59,161), credit score (343-712, mean 571.61), and loan amount (3,674-184,732, mean 24,882). Loan approval has a mean rate of 24%. Other financial details like debt payments, net worth, and risk score are also present, aiding in loan approval prediction analysis.
- **Visualizations:** The pie chart illustrates that 23.9% of loans were approved, while 76.1% were not approved. The bar chart shows that most loans are for Home and Debt Consolidation, indicating a focus on property and debt management. Auto and Education loans are also common, while "Other" covers miscellaneous needs. This suggests borrowers prioritize major investments and debt reduction.





- **Multicollinearity Analysis:** The initial multicollinearity check showed high VIF values for features like *Age (324.47)*, *CreditScore (97.36)*, *Experience (86.91)*, and *RiskScore (96.33)*, indicating strong multicollinearity. These features may need removal or adjustment. Moderate VIF values were observed for *LoanAmount (15.78)*, *LoanDuration (13.46)*, and *InterestRate (68.57)*, suggesting some correlation with other features. Rest independent variables showed very less VIF score.
- **Correlation Analysis:** A correlation matrix and heatmap identified strong pairwise correlations among these variables, suggesting redundancy. We removed *Age, Total Assets, Credit Score, Interest Rate,* and *Monthly Loan Payment* to reduce multicollinearity. Recalculated VIF scores for the remaining variables confirmed all values were below 10, ensuring stability.
- **Correlation Analysis with the Target Variable:** We calculated the correlation of each feature with the target variable, *LoanApproved*, and removed features with an absolute correlation below 0.1, as they offered minimal predictive value. This selection resulted in a refined set of features: *Annual Income, Experience, Loan Amount, Net Worth, Total Debt to Income Ratio, Risk Score*, and specific education levels (*Doctorate, High School, Master*), which were retained for model training.

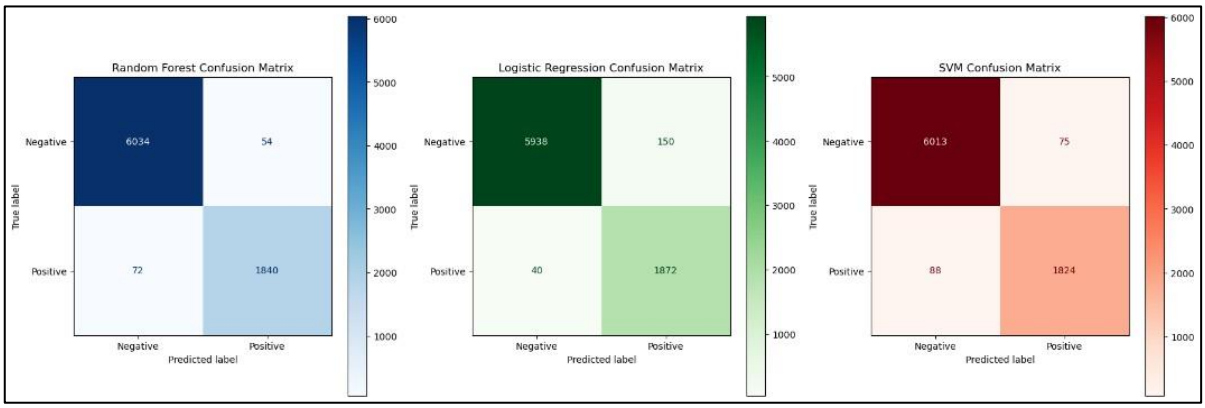
- **Model Performance and Evaluation Metrics:-**

Logistic Regression	Random Forest	SVM
Training Metrics: Accuracy: 0.98 Precision: 0.93 Recall: 0.98 F1 Score: 0.95 AUC-ROC: 1.00 Testing Metrics: Accuracy: 0.98 Precision: 0.93 Recall: 0.98 F1 Score: 0.95 AUC-ROC: 1.00	Training Metrics: Accuracy: 1.00 Precision: 1.00 Recall: 1.00 F1 Score: 1.00 AUC-ROC: 1.00 Testing Metrics: Accuracy: 0.98 Precision: 0.97 Recall: 0.96 F1 Score: 0.97 AUC-ROC: 1.00	Training Metrics: Accuracy: 0.98 Precision: 0.97 Recall: 0.96 F1 Score: 0.96 AUC-ROC: 1.00 Testing Metrics: Accuracy: 0.98 Precision: 0.96 Recall: 0.95 F1 Score: 0.96 AUC-ROC: 1.00

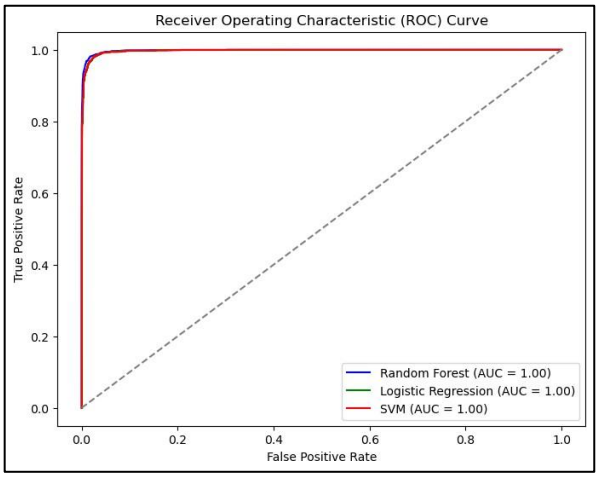
The logistic regression model achieved *98% accuracy* with high precision (93%) and recall (98%), reflecting its strong ability to correctly classify loan approvals. The *F1 score* of 95% and *AUC-ROC* of 1.00 show balanced performance and perfect classification ability.

The SVM model also showed strong results, with *98% accuracy* in both training and testing. *Precision* of 97% and *recall* of 96% indicate that the model is highly effective at predicting loan approvals, with a slightly higher precision than recall. The *F1 score* of 96% highlights the balanced performance. The *AUC-ROC score* of 1.00 confirms that the SVM model can perfectly separate the two classes (approved vs. denied loans).

The Random Forest model achieved *100% accuracy* on the training set, demonstrating an exceptional fit to the data. It achieved *perfect precision, recall, and F1 scores (100%)* on the training data, which suggests the model perfectly classified all training examples. While the testing metrics showed slightly lower scores, with *98% accuracy* and *97% precision*, the model still demonstrated strong performance. The *AUC-ROC score* of 1.00 indicates that it can distinguish between loan approvals and denials with perfect precision.

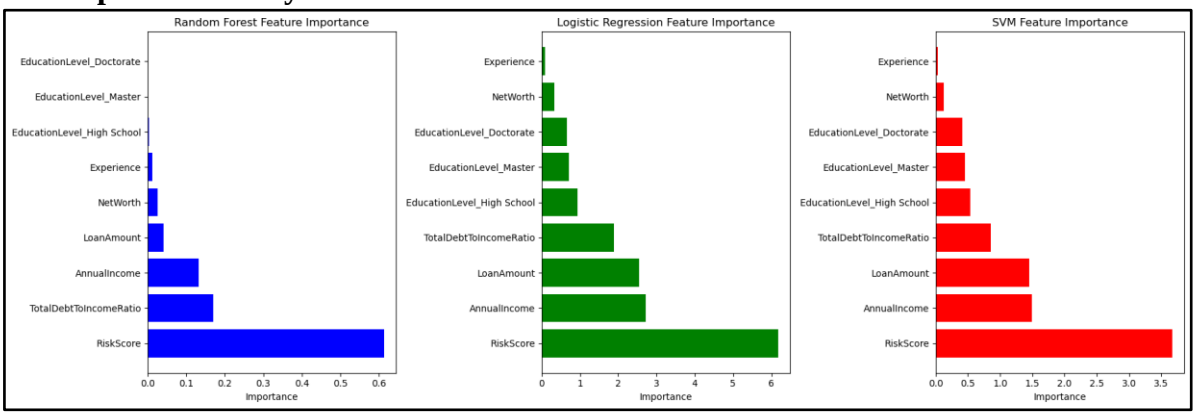


The confusion matrices compare how well three models Random Forest, Logistic Regression, and SVM—classify loan approvals ("Approved") and denials ("Not Approved"). **Random Forest**: Accurately classifies most cases with few errors. It has very few wrong approvals (54) and denials (72), showing high accuracy.**Logistic Regression**: Also performs well but makes slightly more wrong approvals (150) while missing fewer approvals (40). It's more sensitive in catching approvals but trades off with more false approvals. **SVM**: Similar to Random Forest, with slightly more missed approvals (88). It strikes a balance but is slightly less accurate in predicting approvals compared to Logistic Regression.



All three models perform well in classifying loan approvals, with Random Forest and SVM showing lower false positive and false negative rates. Logistic Regression has a slight edge in predicting approvals (lower false negatives) but has a higher false positive rate.

- **Feature importance analysis across three models:-**



Random Forest, Logistic Regression, and SVM highlights the *"RiskScore"* as the most critical factor in predicting loan approvals, consistently showing the highest importance across all models. Following *"RiskScore,"* factors like *"TotalDebtToIncomeRatio," "AnnualIncome,"* and *"LoanAmount"* also play significant roles, though their influence varies by model. Overall, this suggests that risk assessment and financial metrics, such as income level and debt ratios, are key drivers in determining loan approval, with *"RiskScore"* emerging as the primary predictor across different model types.

Conclusion

1. *Comparison of Models:*
   - **Accuracy and Performance**: All three models—Random Forest, Logistic Regression, and SVM—achieve high accuracy (98% on testing), with Random Forest showing perfect scores on the training set, which suggests possible overfitting. Both Logistic Regression and SVM deliver balanced results, with Logistic Regression achieving high recall (98%) and SVM showing a slightly higher precision (97%).
   - **Interpretability**: Logistic Regression is the most interpretable model, as it directly shows feature coefficients, allowing for clear insight into how each variable impacts predictions. Random Forest and SVM are more complex, with Random Forest providing some interpretability through feature importance but less so than Logistic Regression.
2. *Best Performing Model:*

   Random Forest is the top performer due to its high accuracy, perfect training fit, and strong feature importance insights, which highlight "RiskScore" as the most influential predictor. Its slightly higher testing precision (97%) compared to Logistic Regression and SVM further supports its reliability. Key factors contributing to its superior performance include:
   - **Ensemble Learning**: By combining multiple decision trees, Random Forest captures complex patterns and interactions in the data, which enhances its predictive power.
   - **Feature Importance**: Random Forest effectively identifies crucial features like "RiskScore," "TotalDebtToIncomeRatio," and "AnnualIncome," focusing on financial risk and income metrics that are key for loan approval decisions.
   - **Generalization Ability**: Despite its perfect training fit, the model maintains high accuracy (98%) on testing data, showing strong generalization without significant overfitting.

3. *Factors for Superior Performance:* Random Forest's ensemble nature allows it to capture complex patterns and interactions between features, especially important in financial data. Its focus on key variables like *"RiskScore"* and *"TotalDebtToIncomeRatio"* contributes to its exceptional predictive power and ability to generalize well across data.