Introduction

Credit risk assessment plays a crucial role in the financial sector, guiding lending decisions and ensuring financial stability. The dataset under analysis, sourced from Kaggle's Credit Risk Dataset, provides a rich repository of information on loan applicants, including demographic details, financial characteristics, and credit history. This dataset comprises **32,581 observations** and **12 variables**, capturing diverse attributes such as applicants' age, income, homeownership status, employment length, and loan details like intent, grade, amount, interest rate, and status. It also includes indicators of past credit behavior, such as default history and credit history length.

The primary objective of this statistical analysis is to predict the **loan status** (good or bad) based on these features, which will enable lenders to better understand the risk profiles of potential borrowers. By exploring and modeling this dataset, we aim to identify key factors influencing creditworthiness and evaluate the effectiveness of various predictive models.

This report will address the following key research questions:

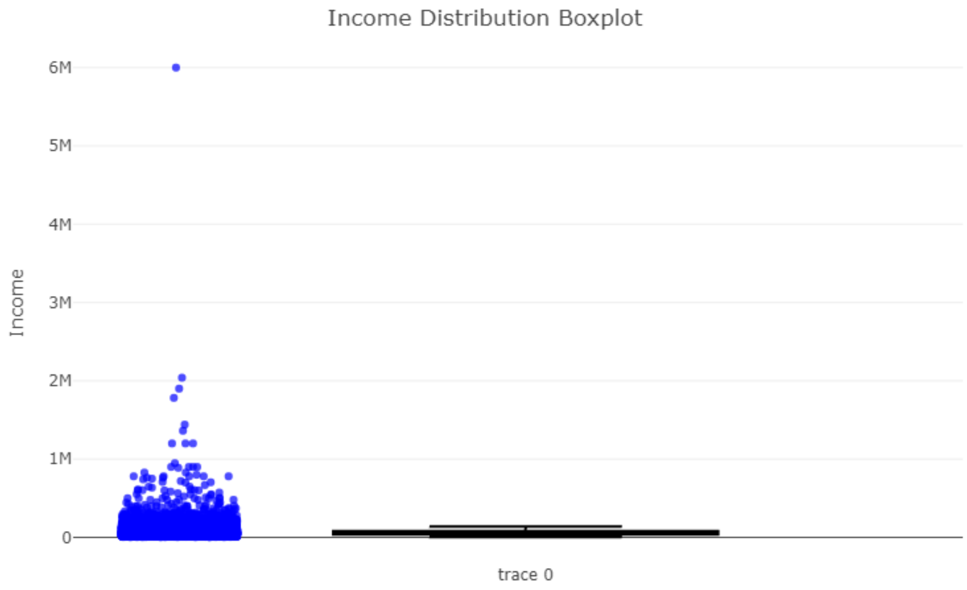
1. **What factors are most predictive of a loan being in good standing?**
2. **Which predictive modeling techniques are most effective for determining loan status?**

To achieve these objectives, the dataset will be subjected to rigorous preprocessing, exploratory analysis, and modeling using a range of statistical and machine learning techniques. The findings from this analysis will provide actionable insights into credit risk assessment, helping financial institutions make informed decisions and manage risks effectively.

Descriptive Statistics

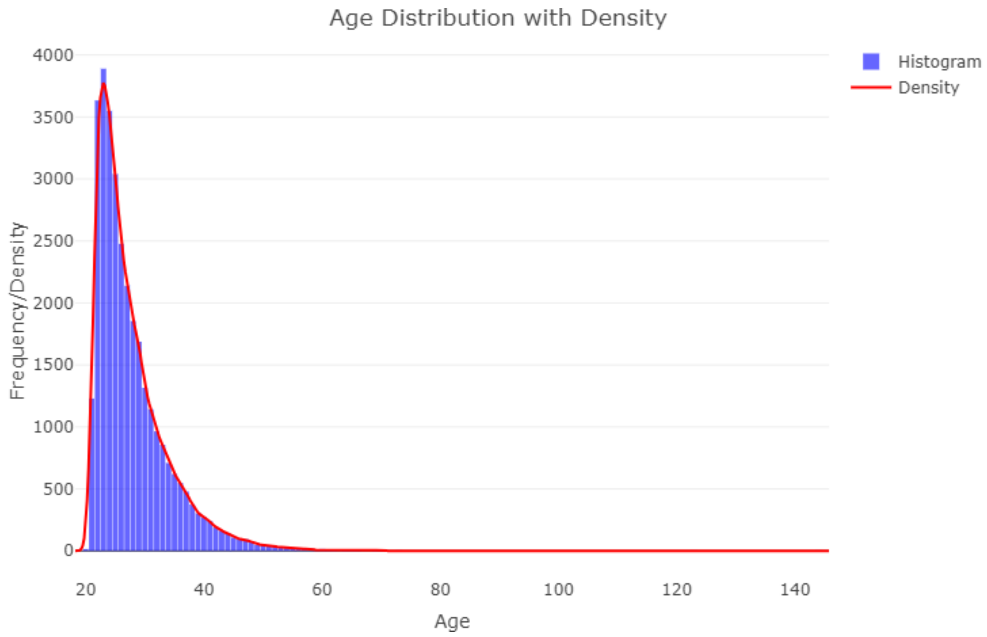
Income Summary (person_income)

Min	1st Qu	Median	Mean	3rd Qu	Max
4000	38500	55000	66075	79200	60000000



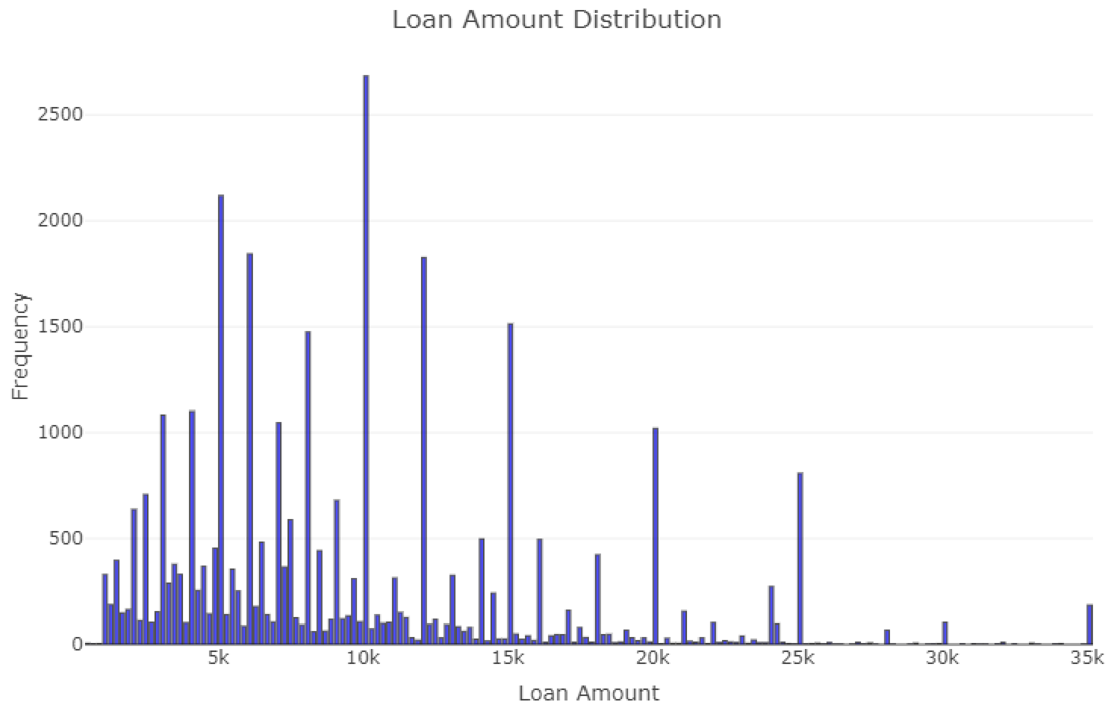
Age Summary (person_age)

Min	1st Qu	Median	Mean	3rd Qu	Max
20	23	26	27.73	30	144



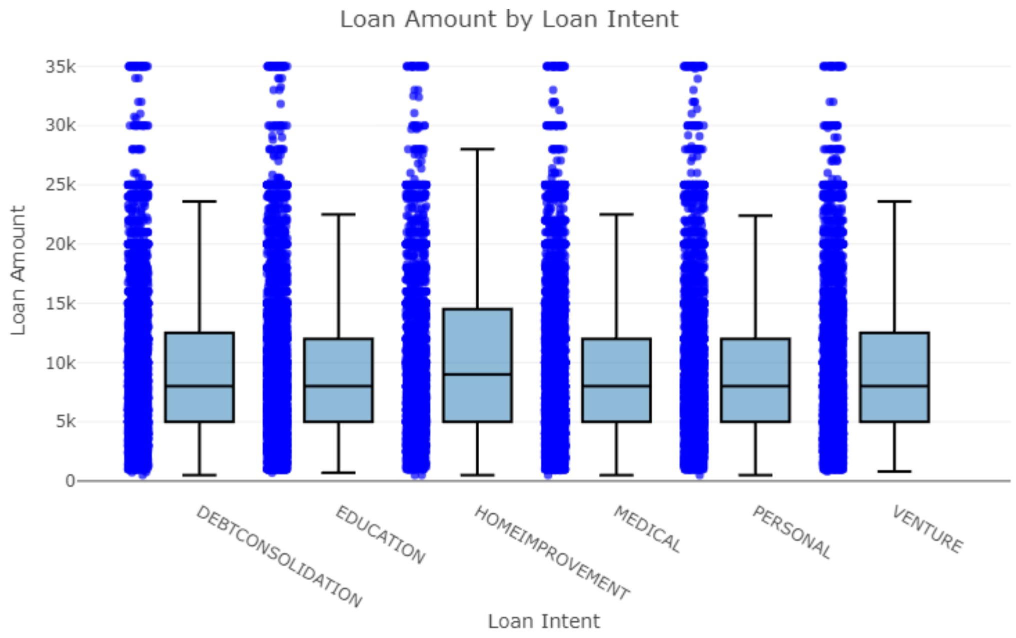
Loan Amount Summary (loan_amnt)

Min	1st Qu	Median	Mean	3rd Qu	Max
500	5000	8000	9589	12200	35000



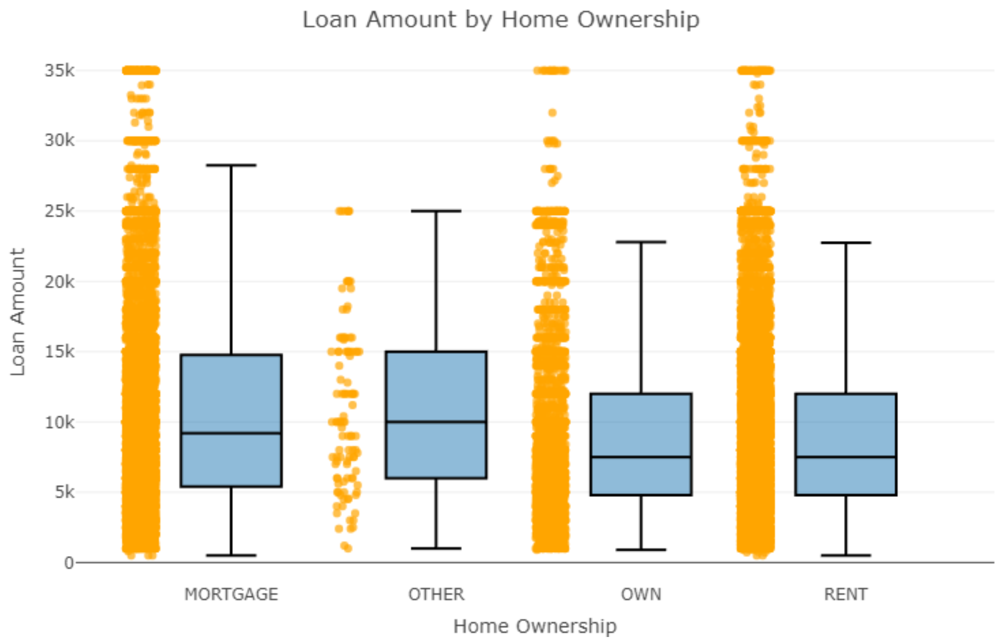
Loan Intent Summary

DEBT CONSOLIDATION	EDUCATION	HOME IMPROVEMENT	MEDICAL	PERSONAL	VENTURE
5212	6453	3605	6071	5521	5719

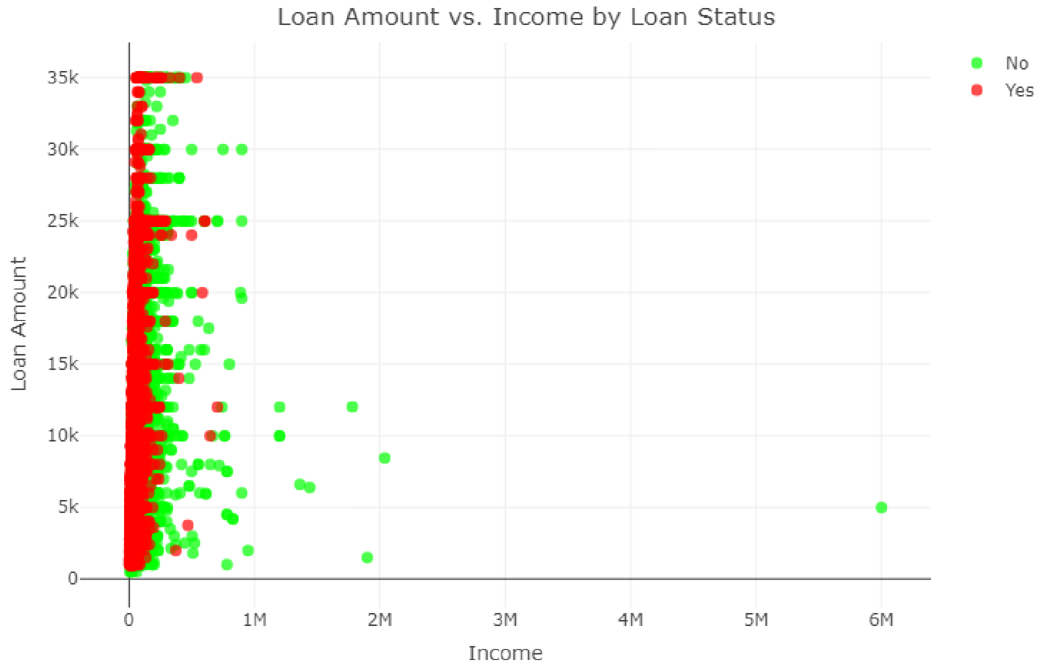


Home Ownership Summary (person_home_ownership)

MORTGAGE	OTHER	OWN	RENT
13444	107	2584	16446



Income VS Loan Amount



Statistical Methodology

The statistical methodology adopted in this study involved data preprocessing, feature engineering, and the application of multiple statistical and machine learning models to predict the target variable, **loan_status**. This approach ensured a robust framework for analyzing relationships within the data and building predictive models.

1. Feature Selection and Data Splitting

In this analysis, we decided to leverage eight variables selected for their theoretical and empirical relevance to credit risk prediction:

- **Numerical predictors:** *person_age, person_income, person_emp_length, loan_amnt,* and *loan_int_rate*.
- **Categorical predictors:** *person_home_ownership, loan_intent,* and *loan_grade*.

The dataset was divided into training (80%) and testing (20%) subsets using random sampling using ***seed(123)***, ensuring that the model training was independent of the data used for performance evaluation. This split allowed for an unbiased assessment of model generalizability.

2. Statistical and Machine Learning Models

To explore the relationships between the predictors and the target variable, loan_status, a diverse set of models was employed. These models represented a range of statistical techniques, from linear models to non-linear machine learning approaches:

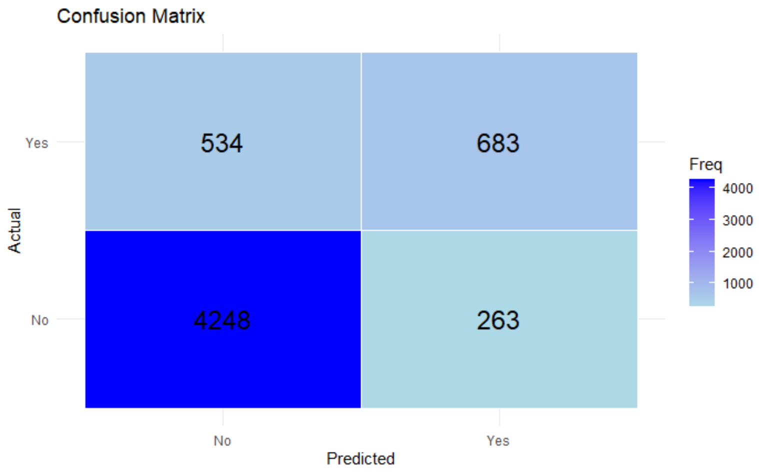
1. **Logistic Regression**:

   A probabilistic classification model was used to estimate the likelihood of a loan being in good standing. Assumed a linear relationship between the log-odds of the outcome and the predictors.

   The model was trained with **altering threshold** to find the best fit with the least error rate. With this parameter alteration, we got a model with an **error rate (13.9%)** with a threshold at **41%.**

   Error rate: 0.1391411

   Confusion Matrix

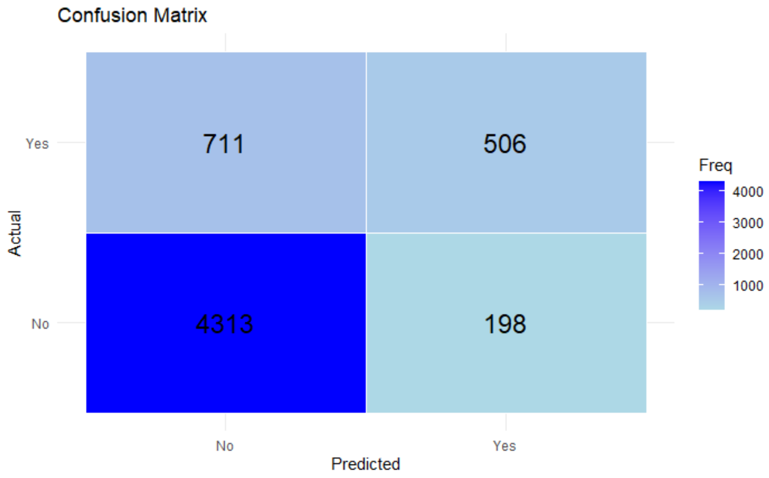


2. **K-Nearest Neighbors (KNN)**:

   A non-parametric method that classified loans based on the majority label of their k-nearest neighbors in the feature space. Dependent on the choice of **k**, which was tuned to optimize performance, we ended up with a model with an **error rate (15.9%)** on k at **31**.

   Error rate: 0.1585196

   Confusion Matrix

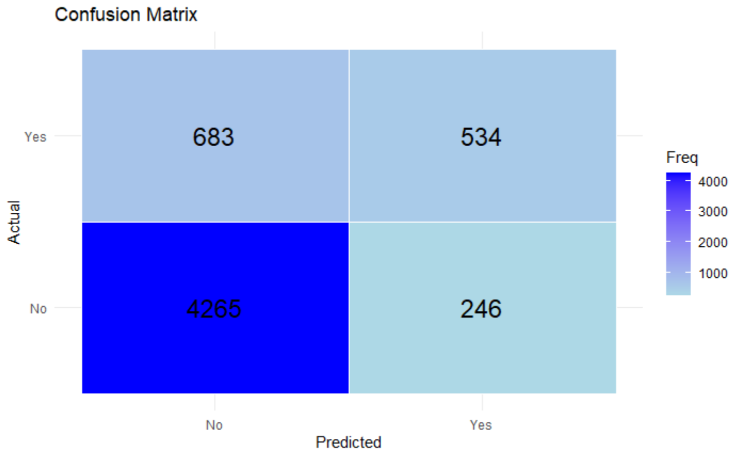


3. **Linear Discriminant Analysis (LDA)**:

   Assumed normally distributed predictors within each class and equal class covariance. Combined predictors into a single discriminant axis for classification. The LDA model achieved an **error rate (16.2%)**.

   Error rate: 0.1621858

Confusion Matrix

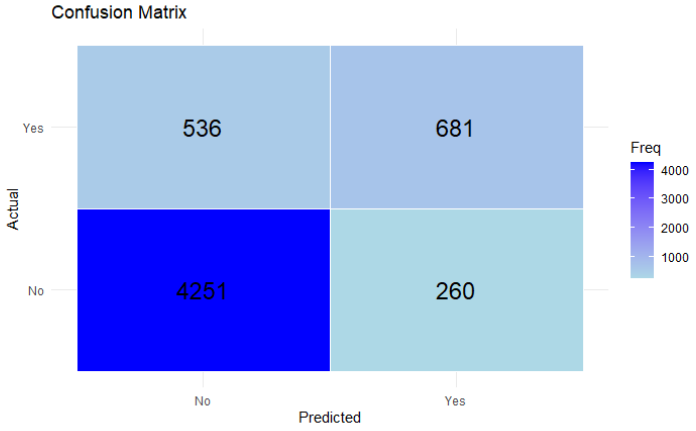


4. **Least Squares Regression**:

   A baseline linear regression model treated loan_status as a continuous variable before thresholding it for classification. The least square model identified five important variables for prediction ***person_income***, ***person_home_ownership, loan_intent***, ***loan_grade***, and ***loan_amnt*** using logistics regression. However, the model performed almost the same as compared to logistic regression with all variables. The model achieved an **error rate (13.9%)** with a threshold of **41%**.

   Error rate: 0.1389665

   Confusion Matrix



5. **Partial Least Squares Regression (PLS) and Principal Component Regression (PCR)**:

   Dimensionality-reduction techniques were applied to address multicollinearity among predictors. PLS maximized covariance between predictors and loan_status, while PCR focused on explaining predictor variance. Both models with cross-validation produces the same error rate **(21.2%)** with **8 components**.

PLS Error rate: 0.2124651

PCR Error rate: 0.2124651

6. **Regularized Regression Models**:

   LASSO (Least Absolute Shrinkage and Selection Operator): Applied L1 regularization to encourage sparsity in model coefficients. Executed with optimized lambda, the model with the least lambda has an **error rate (14.2%)**.

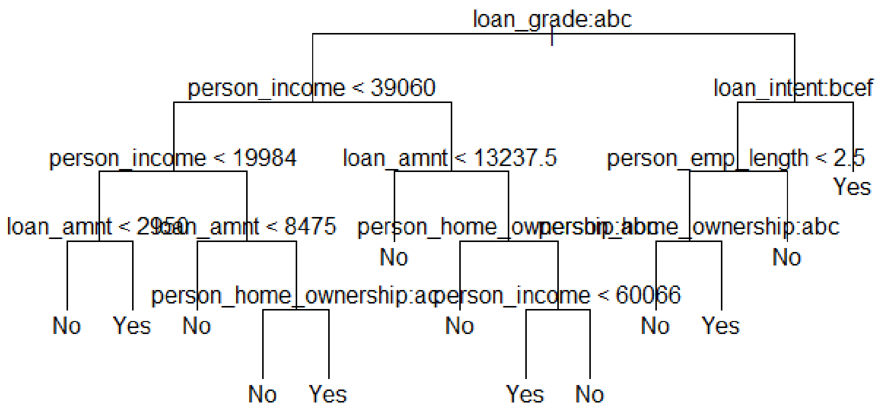
   Ridge Regression: Employed L2 regularization to penalize large coefficients, stabilizing model predictions. With optimized lambda as well, the model with best lambda has an **error rate (15.2%)**.

   LASSO Error rate: 0.1421089

   Ridge Error rate: 0.1517109

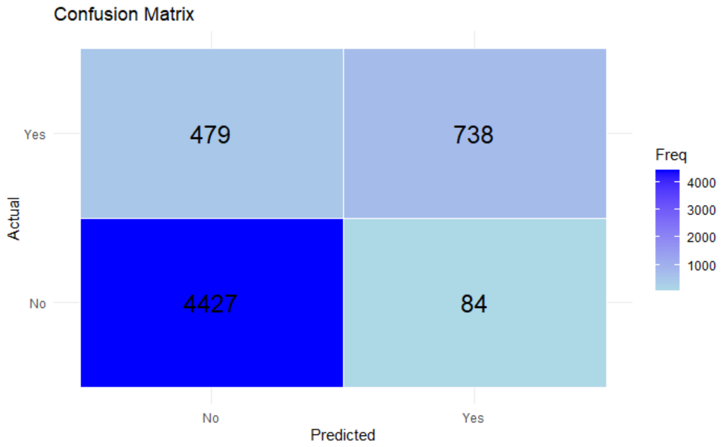
7. **Decision Tree**:

   A tree-based model created decision rules from the data to predict loan_status. The model was introduced to cross-validate with *prune.misclass* function for tree pruning. The original decision tree and the pruned performed the same with an **error rate (9.8%)**.



Error Rate: 0.09828911

Confusion Matrix

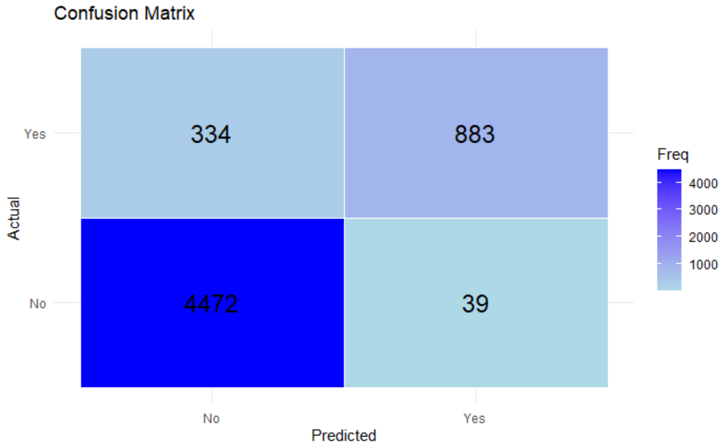


8. **Random Forest**:

   An ensemble method combined predictions from multiple decision trees, reducing variance and enhancing model robustness. Leveraged bootstrapping and feature randomness to improve generalization. Optimizing the number of trees and the number of variables to achieve best generalization. The model achieves an **error rate (6.5%).**

   Error rate: 0.06494413

   Confusion Matrix:

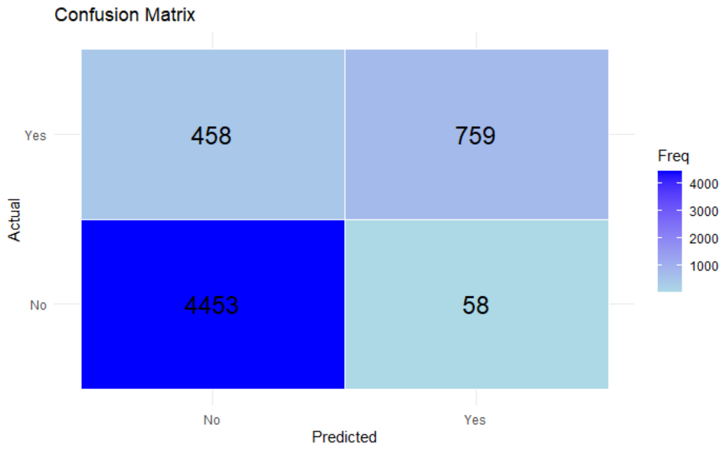


9. **Support Vector Machine (SVM)**:

   A discriminative classifier used hyperplane optimization to separate classes. With SVM analysis, we optimize the values of the cost and Kernel functions allowed for non-linear separability in high-dimensional feature spaces for better prediction. The model achieved an **error rate (10.8%)** with **kernel – radial** and **cost 10**.

Error rate: 0.1078911

Confusion Matrix



Analysis and Findings

This section provides a comprehensive analysis addressing the research questions posed at the outset of this report.

Research Question 1: What Factors Are Most Predictive of a Loan Being in Good Standing?

Through the application of multiple modeling techniques, several key variables emerged as significant predictors of loan status. These include:

- Person_income
- Loan_amnt
- Loan_int_rate
- Loan_grade

The Random Forest model, in particular, proved instrumental in identifying these factors. By leveraging multiple decision trees, Random Forest effectively handled complex, non-linear relationships among variables and mitigated overfitting—a common challenge in datasets with intricate patterns. The model's ability to construct trees on different subsets of data enhanced its generalizability and ensured a robust identification of predictive features.

Research Question 2: How Do Different Predictive Models Compare in Accuracy and Effectiveness?

To evaluate the accuracy and effectiveness of the predictive models, error rates and other performance metrics were compared. The **Random Forest** model emerged as the most accurate, achieving an error rate of **6.49%**. Its ensemble-based approach, which combined the predictions of multiple trees, allowed it to model non-linear relationships and maintain high accuracy across various subsets of data.

In comparison:

- **Decision Trees** performed well with an error rate of **9.83%**, even when pruned to simplify the model and avoid overfitting. However, they lacked the robustness of ensemble methods like Random Forest.
- **Partial Least Squares Regression (PLSR)** and **Principal Component Regression (PCR)**, while effective in addressing multicollinearity and reducing dimensionality, showed limited effectiveness in predicting loan_status for this dataset. Both models exhibited higher error rates of **21.25%**, highlighting their suboptimal alignment with the dataset's complex structure.

These findings underscore the importance of selecting models that are well-suited to the specific characteristics of the data. Models like Random Forest demonstrated the capacity to balance accuracy and complexity, making them the most effective tool for this analysis.

Summary of Model Performance

The following table provides a summary comparison of the models evaluated, stating the parameters and cross-validation used to address the optimization of the models:

Model Type	Key Parameters	Error rate
Logistic Regression	Threshold adjusted to 0.41	13.91%
K-Nearest Neighbor (KNN)	Number of neighbors (k = 31)	15.85%
Linear Discriminant Analysis (LDA)	None	16.22%
Decision Trees	Pruned for misclassification	9.83%
Random Forest	Mtry = 5, ntree = 414	6.49%
Support Vector Machine (SVM)	Cost = 10, Kernel = 'radial'	10.78%
Partial Least Squares Regression (PLSR)	Number of components = 18	21.25%
Principal Component Regression (PCR)	Number of components = 18	21.25%
LASSO	Lambda = 0.0000001	14.21%
Ridge Regression	Lambda = 0.0000001	15.17%

Conclusion

From the above results, **Random Forest** is the most suitable model for predicting loan status, offering the highest accuracy and robustness due to its ability to handle non-linear relationships and diverse predictor interactions. While **Decision Trees** and **SVM** also provided solid results, their performance was less consistent compared to the ensemble-based Random Forest. Regularized models like **LASSO** and **Ridge Regression** demonstrated moderate accuracy and utility in addressing multicollinearity. Conversely, **PLSR** and **PCR** were the least effective, underscoring the need to match model choice to the data's structure and predictive goals. This analysis highlights the critical role of ensemble techniques in predictive tasks involving complex datasets.