Loan Default Rates Prediction

Data Analysis Project

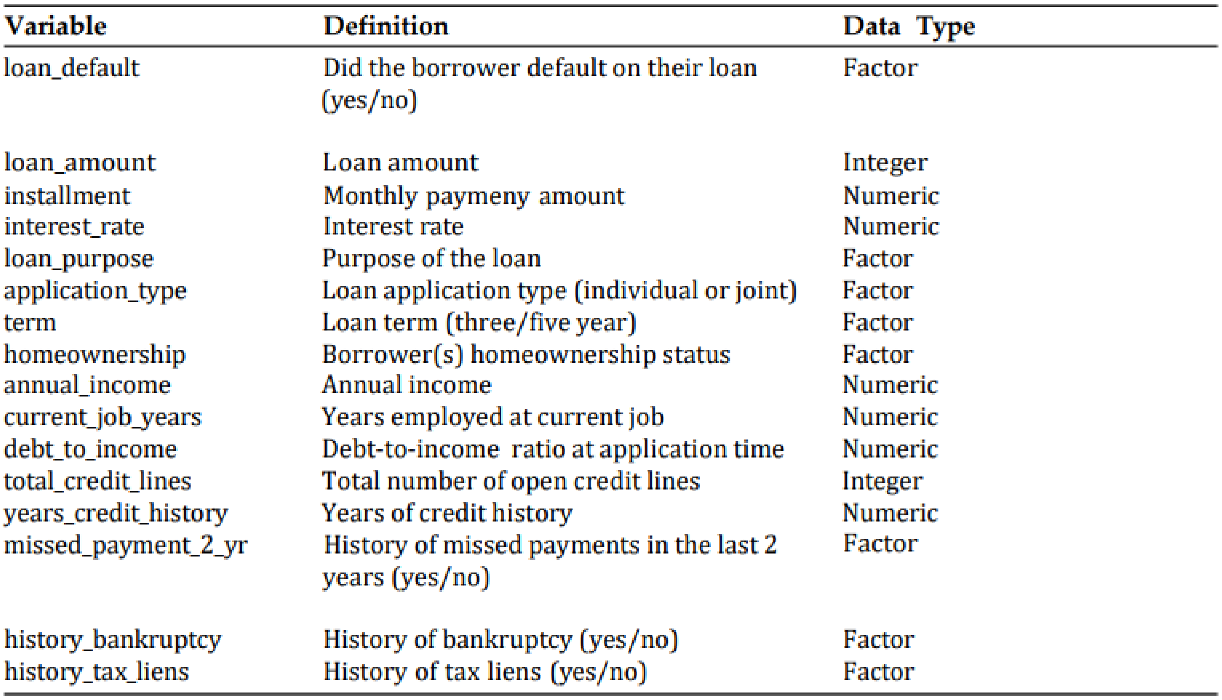
Ivan Francis

10-29-2023

1. Introduction.

The data set that is going to be analyzed is the loan default data that contains information on over 3,500 individuals who secured a personal loan in 2017 from a national bank. The objective of this project is to explore the factors that lead to loan default and develop a machine learning algorithm that will predict the likelihood of an applicant defaulting on their loan in the future. The company is looking to see if it can determine the factors that lead to loan default and whether it can predict if a customer will eventually default on their loan. For answering the former questions, we focus on 5 research questions pertaining to the dataset while utilizing machine learning algorithms to do the same.

The dataset contains information on various factors such as loan amount, interest rate, loan purpose, and applicant details.



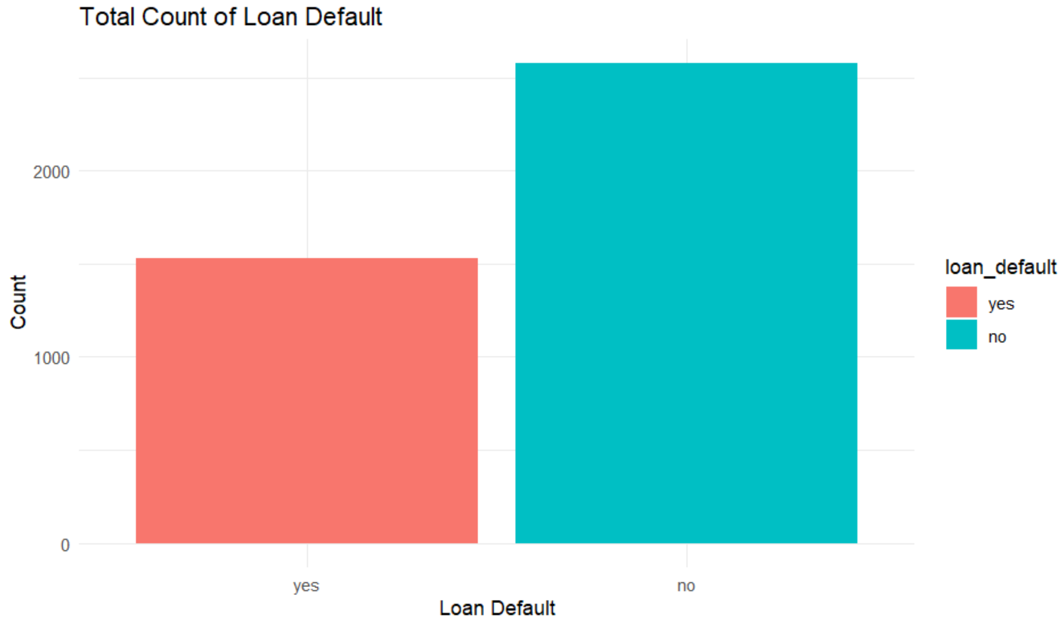
Variable	Definition	Data Type
loan_default	Did the borrower default on their loan (yes/no)	Factor
loan_amount	Loan amount	Integer
installment	Monthly paymeny amount	Numeric
interest_rate	Interest rate	Numeric
loan_purpose	Purpose of the loan	Factor
application_type	Loan application type (individual or joint)	Factor
term	Loan term (three/five year)	Factor
homeownership	Borrower(s) homeownership status	Factor
annual_income	Annual income	Numeric
current_job_years	Years employed at current job	Numeric
debt_to_income	Debt-to-income ratio at application time	Numeric
total_credit_lines	Total number of open credit lines	Integer
years_credit_history	Years of credit history	Numeric
missed_payment_2_yr	History of missed payments in the last 2 years (yes/no)	Factor
history_bankruptcy	History of bankruptcy (yes/no)	Factor
history_tax_liens	History of tax liens (yes/no)	Factor

Our primary goal is to understand the factors influencing loan defaults and answer specific research questions. The questions we aim to address are based on identifying customers at risk of defaulting on their loans to minimize financial losses as the bank in recent years has experienced record levels of customers defaulting on the past couple of years and this has led to large financial losses to the bank.

2. Data and Model

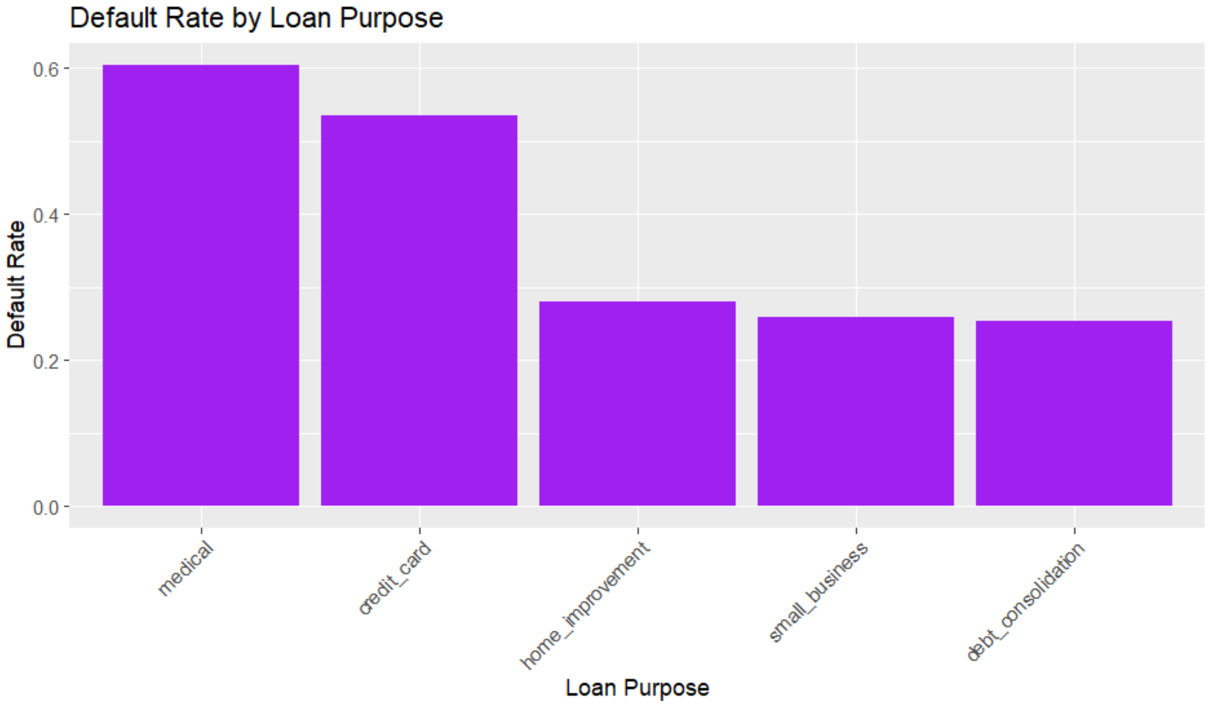
As the dataset was already clean we could straight away dive into EDA. Before answering the research question, few summary statistics and visualizations were carried out.

When we look at the percentage of customers defaulting , it is 37.2% while 62.8% have not. This means that a significant portion of the customers, roughly more than one-third, have experienced loan defaults. The reasons for this relatively high default rate could be multifaceted.



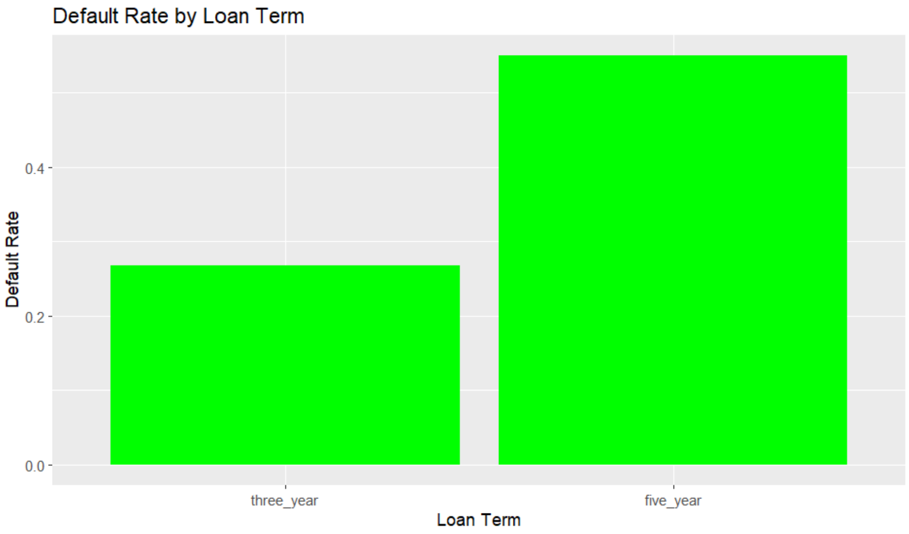
It may be associated with several factors such as the interest rates, loan terms, loan amounts, and the financial stability of the borrowers. Higher interest rates or longer loan terms, for example, could pose challenges for borrowers in meeting their repayment obligations. Additionally, borrowers with high debt-to-income ratios or those who have limited credit histories might be more prone to loan defaults.

Question 1 : What purpose of the loan has the highest rate of defaulting ?



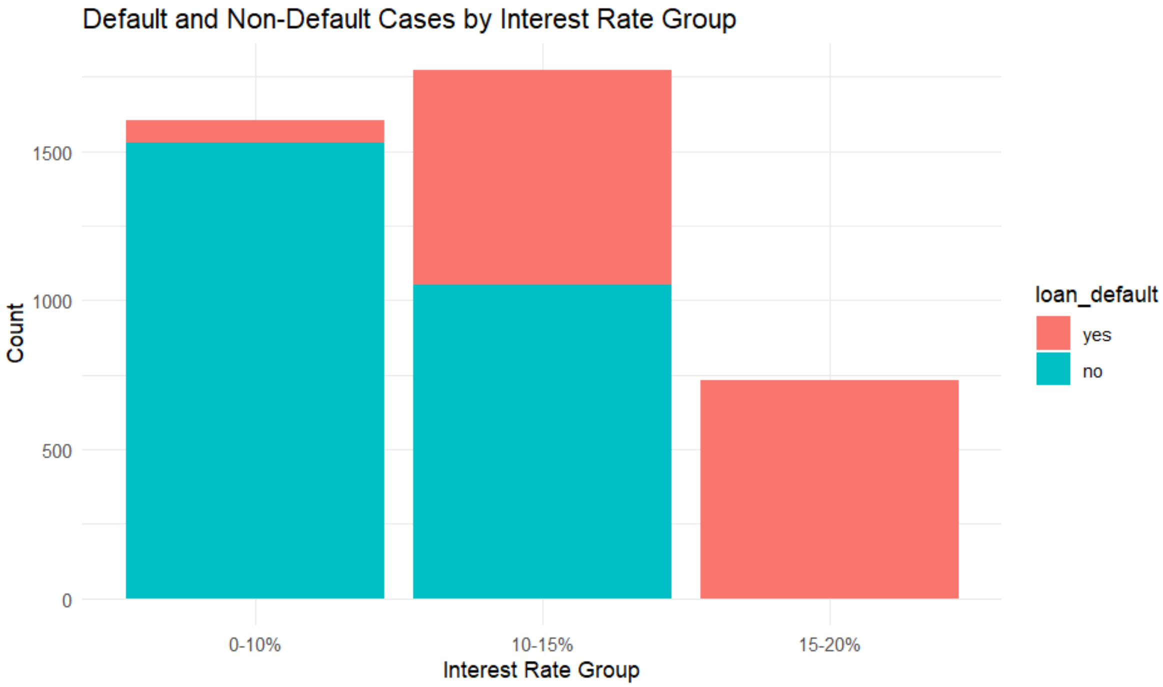
It's evident that the highest rate of defaulting among the specified loan purposes is associated with "credit_card" loans, with 470 defaults. This is followed by "medical" loans with 384 defaults and "debt_consolidation" with 308 defaults. Conversely, "home_improvement" and "small_business" loans have relatively lower default rates, with 147 and 221 defaults, respectively. Credit Card loan's default rate is at 60.5% while medical loan's default rate is at 53.5%

Question 2 : Is there a difference in default rate between loans of different terms?



The chart indicates that there is indeed a notable difference in default rates between loans with distinct terms. Specifically, for "three_year" term loans, the default rate is significantly lower at 26.8%, whereas "five_year" term loans exhibit a considerably higher default rate of 55%. This observation suggests that the loan term duration plays a crucial role in influencing default rates. While the exact factors causing this difference may require further investigation, this finding highlights the importance of considering loan terms as a significant variable in assessing and managing default risk.

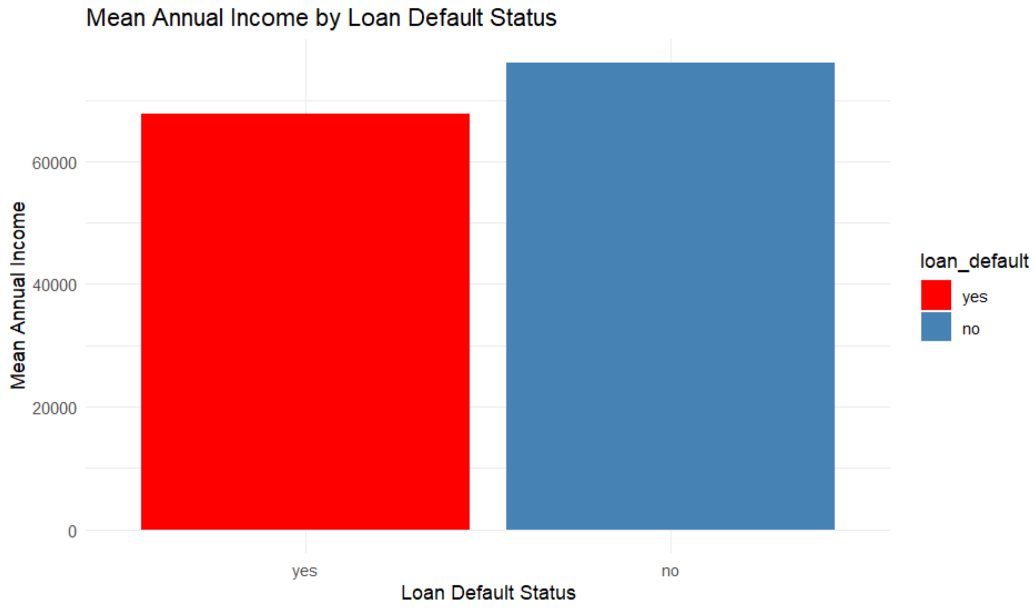
Question 3: Is there a link between default rates and with interest rates given on loans ?



We observe that there is a clear link between the two. When loans are offered with interest rates in the range of 0-10%, the default rate is quite low, with only 4.80% of loans defaulting. However, as the interest rates increase to the range of 10-15%, there is a substantial jump in the default rate to 40.67%, indicating a strong association between higher interest rates and a higher likelihood of loan defaults.

Remarkably, for loans with interest rates in the range of 15-20%, the default rate reaches 100%, signifying that all loans in this category have defaulted. This stark increase in default rate as interest rates rise reveals a significant correlation between the two variables. Borrowers might find it increasingly challenging to meet their repayment obligations as interest rates climb, leading to a higher likelihood of loan default.

Question 4 : Is there a relationship between average annual income of the borrowers and them defaulting on their loans?

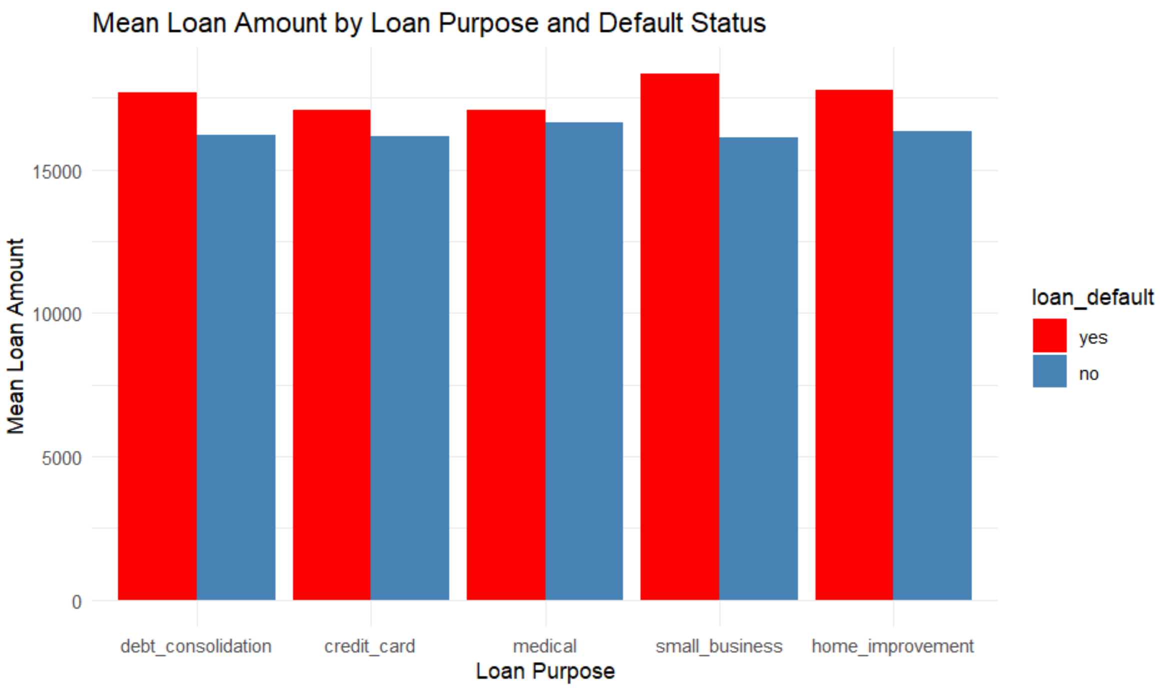


Applicants who defaulted on their loans had a lower average annual income of $67,819, while those who did not default had a relatively higher average annual income of $76,096.

This observation hints at a potential correlation between lower income levels and a higher likelihood of loan default. Borrowers with lower incomes might face financial constraints that make it more challenging for them to meet their loan repayment obligations, which could contribute to the higher default rate observed in this group.

Question 5: Is there a relationship between defaulting on the loan, loan purpose and loan amount?

When examining borrowers who defaulted on their loans ("yes" in the "loan_default" column), it is evident that the mean loan amount varies depending on the loan purpose. Among the various loan purposes analyzed, borrowers with the intent of debt consolidation had a notably higher mean loan amount, averaging around $17,704. In contrast, those who did not default on their loans had a lower mean loan amount, approximately $16,224. This discrepancy in mean loan amounts between defaulters and non-defaulters hints at a potential relationship between the loan purpose of debt consolidation and higher loan amounts contributing to a higher default rate.



Similarly, credit card-related loans exhibited a similar pattern. Borrowers who defaulted on their credit card loans had a mean loan amount of around $17,076, while non-defaulters had a slightly lower mean loan amount of about $16,173. The discrepancy suggests that the loan amount may influence the default rates for credit card loans.

For medical-related loans, those who defaulted had a mean loan amount of around $17,058, whereas non-defaulters had a mean loan amount of about $16,635. The disparity in mean loan amounts between these two groups raises questions about how the loan amount for medical purposes affects default rates.

Small business loans followed the same trend. Borrowers who defaulted on small business loans had a notably higher mean loan amount of approximately $18,351, while non-defaulters had a lower mean loan amount of around $16,116. This discrepancy highlights a potential connection between the loan purpose of small business and the loan amount, impacting the likelihood of loan default.

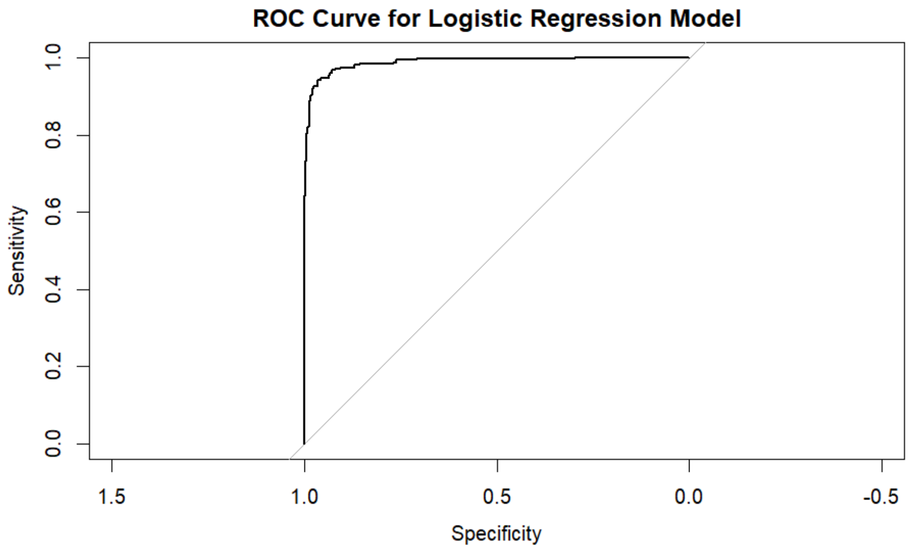
Home improvement loans displayed a similar pattern, with defaulters having a mean loan amount of about $17,755, and non-defaulters having a slightly lower mean loan amount of around $16,330. This observation suggests that the loan amount could be a contributing factor in the default rates of home improvement loans.

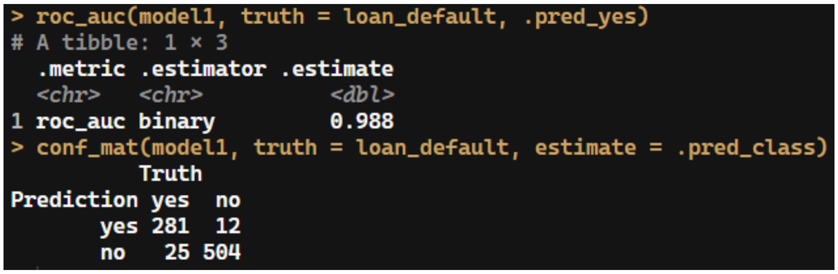
Predictive Modelling:

As the company is looking to see if it can determine the factors that lead to loan default , Two classification models are used for predicting if a customer will eventually default on their loan or not.

Logistic Regression Model:

The first prediction model created was using the logistic regression model where loan_default was the target variable. The dataset was split into 80% for training the data and 20% for testing the data. Further, cross validation was used where k=10 for the most optimum results. The logistic regression model was created using the recipe library in R which was used for preprocessing the data.



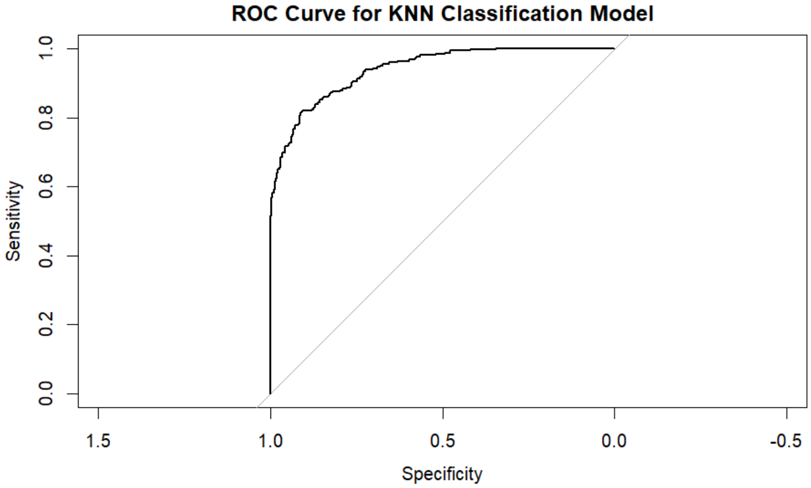


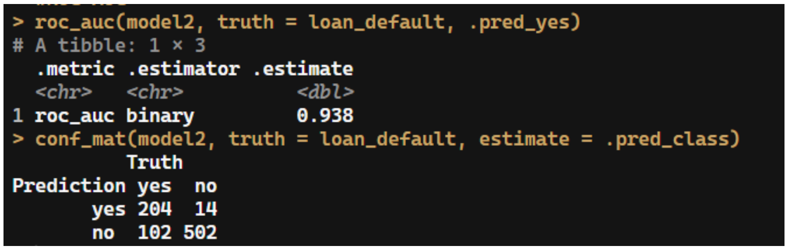
```
> roc_auc(model1, truth = loan_default, .pred_yes)
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.988
> conf_mat(model1, truth = loan_default, estimate = .pred_class)
          Truth
Prediction yes  no
       yes 281  12
       no   25 504
```

The output from this logistic regression model was insightful. The model was able to predict whether a loan would default, and it produced two key metrics to evaluate its performance. The receiver operating characteristic (ROC) area under the curve (AUC) was estimated to be approximately **0.988**, suggesting that the model demonstrated strong discriminative power in distinguishing between loan default and non-default cases. Furthermore, the confusion matrix revealed that the model correctly classified 281 instances of loan defaults and 504 instances of non-defaults, with only 12 false positives and 25 false negatives. The model gives an prediction accuracy rate of **95.28%**

K-Nearest Neighbors Classification:

Using the K-Nearest Neighbors (KNN) model for predicting loan defaults involved a series of steps in this analysis. The KNN model was set up with different values of 'k' (the number of nearest neighbors to consider) to determine the best parameter for classification. The 'kknn' engine was employed for the model, with 'classification' mode. This approach allows the model to assign new data points to classes based on the majority class among their 'k' nearest neighbors. The 'tune' function was used to tune over a grid of 'k' values, evaluating the model's performance using ROC-AUC as a metric.





```
> roc_auc(model2, truth = loan_default, .pred_yes)
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.938
> conf_mat(model2, truth = loan_default, estimate = .pred_class)
          Truth
Prediction yes  no
       yes 204  14
       no  102 502
```

The output shows that the best 'k' value for this KNN model is selected as it produced the highest ROC-AUC score. Regarding the output, the ROC-AUC score for the KNN model is estimated at **0.938**, suggesting strong predictive power in distinguishing between loan defaults and non-defaults. Based on the confusion matrix the prediction accuracy rate of this model is **85.5%** which is quite low when compared to the logistic regression model.

3. Conclusion(s)/Discussion.

This analysis was aimed at addressing the high default rate issue that the bank was facing. With over 3,500 individuals who secured personal loans in 2017, the primary goal was to understand the risks associated with loan defaults and develop machine learning algorithms to predict future defaults. One significant finding from the dataset is that approximately 37.2% of customers experienced loan defaults. This alarming rate raised questions about the underlying factors leading to such a high default rate. Factors such as interest rates, loan terms, loan amounts, and applicant financial stability were considered as potential contributors. Based on the analysis that we carried out we get an idea of the recommendations and policies the bank should consider implementing before supplying loans to their customers.

The recommendations for the bank based on our analysis is as follows:

We observed that a five-year loan resulted in 55% of loan defaults which means that more than half the applicants who took five-year loans have defaulted which is not a good sign as it indicates that people struggle to make their longer-term loan payments.

One of the variables that appears to correlate clearly is interest rate; the average interest rate for individuals who defaulted on their payments was 60% higher than for those who did not. This observation implies that loans with higher interest rates will also likely have a higher default probability. Another important observation was that applicants having interest rates of above 15% have always defaulted.

Therefore, before granting a loan, it is advised that this bank first determines the applicant's level of risk. If they are going to default on the debt, this will provide as a proper indicator. Consideration must be given to a few factors to evaluate this risk. Interest rates are the most important factor; in order to reduce the number of applications that default, the bank should strive to keep interest rates around 10% or below.

Considering that the average interest rate for applicants who did not default on their loan payments was 9.3%, should be an appropriate ceiling for the interest rates. Reducing or eliminating five-year loan issuance is another suggestion. Comparing these loans to the three-year loans, the default rate was more than twice as high. By applying the above recommendations, I think the bank will likely reduce the rate of defaults among their customers.

Coming to prediction/classification models, two models were tested and utilized for predicting future loan defaults. The logistic regression performed really well compared to the KNN classification model with a precision of 95%. Using the ROC AUC Curve we observed that logistic regression outperforms the KNN model here as well which was 0.988 leading to very accurate predictions with respect to loan defaults.