Logo, company name

Description automatically generated

**George Mason University**

5/9/2023

**Spring  
2023**

Instructor: Dr. Jie Xu

Faith Hill

Huong Thi Thu Le

Tarex Jarrouje

Priya Indu Sharma

Bloti B Teh

**SYST/OR 568  
Project Report**

**Team Versatile**

**Report Title: Vehicle Loan Default Prediction**

Table of Contents

[Abstract 2](#_Toc134450535)

[1 Introduction 2](#_Toc134450536)

[1.1 Background 2](#_Toc134450537)

[1.2 Problem Space 1](#_Toc134450538)

[1.3 Research 1](#_Toc134450539)

[1.4 Solution Space 2](#_Toc134450540)

[1.5 Project Objectives 2](#_Toc134450541)

[1.6 Definition of Terms: 2](#_Toc134450542)

[2 Data Acquisition 3](#_Toc134450543)

[2.1 Overview: 3](#_Toc134450544)

[2.2 Field Descriptions: 3](#_Toc134450545)

[2.3 Data Context: 3](#_Toc134450546)

[2.4 Data Conditioning 4](#_Toc134450547)

[2.5 Data Quality Assessment: 5](#_Toc134450548)

[3 Analytics and Algorithms 5](#_Toc134450549)

[4 Findings 10](#_Toc134450550)

[5 Summary 10](#_Toc134450551)

[6 Future Work 10](#_Toc134450552)

[7 Appendix A: Code References 10](#_Toc134450553)

[8 Appendix B: Risk Section 11](#_Toc134450554)

[9 References 12](#_Toc134450555)

Table of Figures

[Figure 1. Concerns from Consumers for Dealers to Provide Details on Pricing & Financing Automobiles 1](#_Toc134450556)

[Figure 2. Preprocessing Statistics 4](#_Toc134450557)

[Figure 3. Classification Tree 5](#_Toc134450558)

[Figure 4. Classification Tree Model 6](#_Toc134450559)

[Figure 5. Gini Index for Classification Tree 7](#_Toc134450560)

[Figure 6. ROC Curve 8](#_Toc134450561)

Table of Tables

[*Table 1. Average Auto Loan Balance from Q3 of each Year* 4](#_Toc134450562)

[*Table 2. Auto Loan Delinquency Rate from Q3 of each Year* 4](#_Toc134450563)

[*Table 3. Average Auto Loan by Generations* 0](#_Toc134450564)

[Table 4. Consumer Information 3](#_Toc134450565)

[Table 5. Model Performance (Classification Tree) 8](#_Toc134450566)

[Table 6. Model Performance (Logistic Regression) 9](#_Toc134450567)

[Table 7. Model Performance (Logistic Regression) 9](#_Toc134450568)

[Table 8. Models Comparison 9](#_Toc134450569)

[Table 9. Roadblocks - Risks & Mitigations 11](#_Toc134450570)

# Abstract

Data is ubiquitous in a way that it can be utilized anywhere and everywhere. It is almost impossible for financial institutions to underestimate of data analytics initiatives. They rely on data to validate their intuitions to aide in determining consumer behaviors. The goal and objective and objective for this project is to predict the likelihood of default or nondefault loaner.

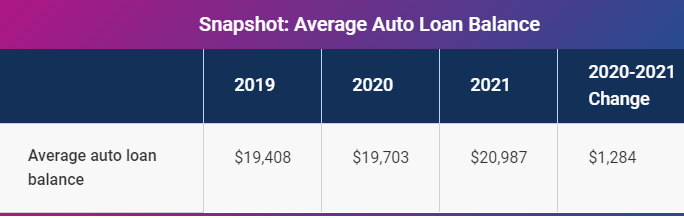
In this report, we demonstrate how we can use three different models to predict this behavior - default or nondefault. The overall findings gave promising insights.

# Introduction

## Background

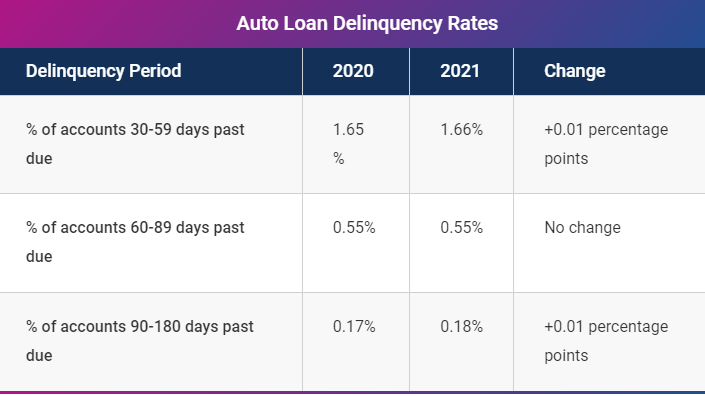
According to Experian and other credit bureau agencies, auto loan debt has reached a record high $1.43 Trillion [1]. A grand sum of $1.43 trillion is an outstanding debt owed on vehicles by customers in the third quarter of 2021. Which is an increase of $78 billion from 2020 and the average auto lean balance exceeded $20,000 for first time auto owners.

Amidst the rise in prices for several other goods and services that were hiked up in 2021, auto loan and auto prices were the most astonishing of all. The average price of vehicles increased significantly more than any other major category in the consumer price index.

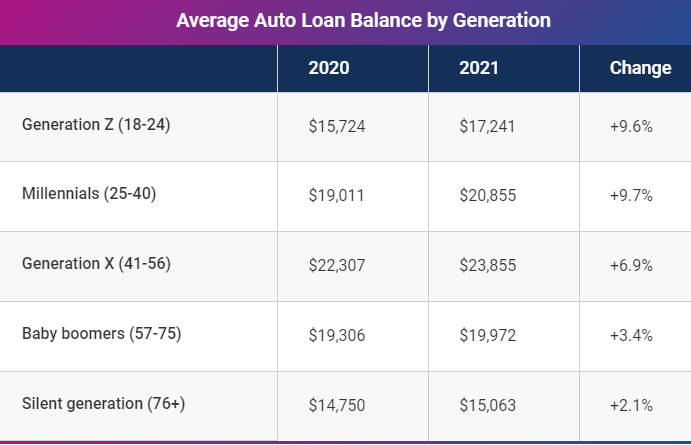
*Table 1. Average Auto Loan Balance from Q3 of each Year*

Meanwhile, auto loan delinquencies remain stable in 2021. Consumers have maintained the equilibrium with their auto loan payments in spite of changes in employment and income.

Measured in terms of percentage accounts that are pas due by 30 days or more, delinquencies have barely budged since the third quarter of 2020 [1].

*Table 2. Auto Loan Delinquency Rate from Q3 of each Year* 

Another factor that has driven auto loan debt is the young generation. Statistics show that for the second consecutive year, Millennials and Generation X (both of which travel mostly by car/road) are reported to have the largest average auto loan in 2021 [1]. The average auto loan for both Generations have exceeded $20,000 as of last year.

*Table 3. Average Auto Loan by Generations* 

* 1. Problem Space

Typically, most auto loan delinquencies have originated from with low credit scores. Thus, with fewer low-credit borrowers getting new loans, defaults have remained fairly low. Many low-credit borrowers did not finance new loans due to less demand for a vehicle with the lock down alongside stringent acceptance approval measures implemented by lenders.

Though delinquencies/defaults can damage a borrowers’ credit scores and consequently lead to repossession, the automobile industry is seeing a decline in repossession activities.

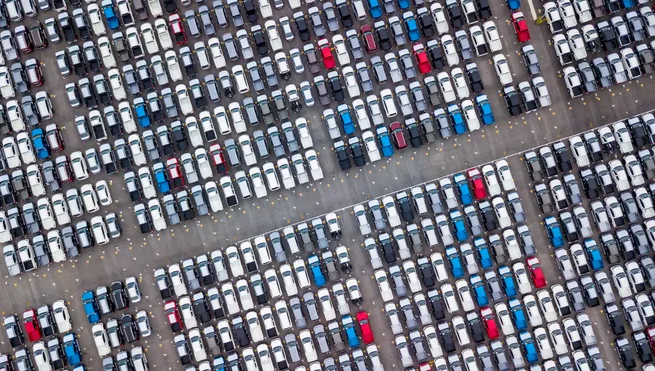


Figure 1. Concerns from Consumers for Dealers to Provide Details on Pricing & Financing Automobiles

Financial organizations suffer substantial losses from defaults on car loans, this necessitates an investigation to identify the factors that contribute to the default of vehicle loans. Doing so will ensure that consumers capable of repayment are not rejected and important determinants can be identified which can be further used for minimizing the default rates.

The objective of this project is to predict the likelihood of loan default or nondefault. Based on the information collected from the borrower, we can predict the likelihood that a loan can be returned or not.

## Research

Based on the background information vehicle loan default can be avoided by consumers upon deciding to take out a loan on a shorter term and invariably making a large down payment to avoid the risk of unmanageable monthly payments.

For this project we will focus more on the analytical aspect of defaulting on car loans and their corresponding predictors. Operation research is an approach to decision-making, which involves a set of methods to operate a system. Logistic regression is used for obtaining the most optimal solution with given predictors. In nonlinear programming, we formulate our real-life problems into a mathematical model. It involves data pre-processing applied to transform the raw data sets into a format applicable predictive modeling. This includes but not restricted to Box-Cox transformation, centering, scaling, removal of near zero variance predictors, and/or removal of bad entries and missing values.

## Solution Space

Operation research methods will most definitely deliver accurate and precise results. Beginning with the different aspects of nonlinear regression models, we will determine a pattern labeled data for classification and prediction, and unsupervised learning will determine provide insights into clustering relationships using observed unlabeled data. A taxonomy of techniques for descriptive (what is currently happening), predictive (what will happen), and prescriptive (what should be done) analyses will also provide a good insight. Data modeling and simulation using various statistical analysis techniques are highly considered and encouraged as well.

Our analyses will generate a good insight on multiple baselines in understanding the variables, and implementing the strategic methods for predicting auto loan defaults and delinquencies. This will be the most fundamental approach.

## Project Objectives

The team will learn how to perform classification models and various predictive analysis to predict default or non-defaults. By assessing multiple variables, the team can analyze using statistical techniques to provide predictive measurements via the lens of consumer debt ratio and credit bureau agencies perspective.

In the end, the goal would be to provide our analysis on the probability that a consumer will default or not on an automobile loan. In doing will ensure that consumers capable of repayment are not rejected and most important determinants identified to further minimize vehicle loan default rates.

## Definition of Terms:

**Loan**: is an act of lending, anything, especially money at an interest.

**Delinquent**: Failing in duty, often concerning bad debt or debtor.

**Default**: Failure to fulfil a financial obligation.

**Non default**: Not default.

**Consumer**: One who uses or consumes.

**Origination**: Standing as a source in relating to something.

# Data Acquisition

## Overview:

For a descriptive overview of the dataset for knowledge discovery. We considered a data set from Kaggle website which contains information on a borrower (Demographics including – age, income, date of birth, credit history, employment, among other details; disbursed loan details (amount, number of active accounts, loan status, etc.); Credit Bureau information (credit score, loan status, etc.)

## Field Descriptions:

URL: [Vehicle Loan Default Prediction | Kaggle](https://www.kaggle.com/datasets/avikpaul4u/vehicle-loan-default-prediction)

Title: Each field is titled per dataset contents based on the consumer demographic details, loan information, bureau data and history.

Authors: Avik Paul

Publication Date: 2020

Text (Type: string): XLSX

Data set Type: Open Source

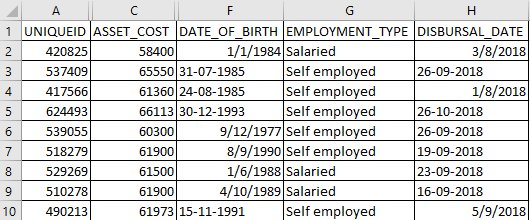
Data Size: 233,154 observations and 41 variables

## Data Context:

The descriptive context of the dataset is related to the customer and vehicle purchased. Each field has its own uniqueID, asset cost, loan amount information, consumer employment type and other peculiar information.

There are about two main categories of consumers – self-employed and salaried.

Table 4. Consumer Information



## Data Conditioning

Data conditioning was performed at the initial stage by identifying missing values and found 7661 rows missing from consumer Employment Type Predictor/Variable. We applied other pre-processing methods to transform the raw data sets into a format that enables us to predictive modeling. Formatted the date column using the Lubridate package and Library. We defined the list of date functions used in the data set, looped over the rows of the data frame and using a string format placed the output in a csv file.

Later converted the character to date columns using ymd()or mdy() or dmy() and write the output to a csv file. Other pre-processing method applied include: the removal of near zero variance predictors and checking for or removal of bad entries and missing values (in this case, missing numeric values). The code detected and removed [1] 233151 30 near-zero variance predictors and found no missing numeric data type.

The next exploratory data analysis (EDA) applied was imputing missing values within the categorical data and assigned the output to a data frame and validated the imputed data. This process detected 7661 missing values within the Employment\_Type category. Next in the EDA process was the extraction of predictors using the Data Explorer Package. The create\_report() function in the Data Explorer Package generates a detailed report that includes a summary of statistics, histograms, density plots, and scatter plots for each variable in the data set.

Next in the pre-processing analysis, we considered the correlation between all the numeric predictors. If the correlation plot shows no stars, then the variable is NOT statistically significant, while one, two and three stars mean that the corresponding variable is significant at 10%, 5% and 1% levels, respectively).

Next, was splitting the data set into training and testing sets. The training set comprised of 80% and the testing set was 20%. The Caret Function was used to skewness and estimate the transformation process of Centering, BoxCox and Scaling.



Figure 2. Preprocessing Statistics

We then apply these transformations and predict a generic function from the results of various model fitting function. the function invokes particular methods which depend on the class of the first argument.

Performed Principal Component Analysis (PCA) on the entire dataset and retrieved the numeric columns using the dplyr() function and computed the percentage of variance on the on each component.

## Data Quality Assessment:

The datasets acquired from Kaggle’s open source consist of a csv file format. The vehicle loan consumer dataset consists of data from 2018. The date of birth and disbursal date column showed inconsistencies within the date format and required for us to update the date fields with the right formatting to keep the data consistent for modeling purposes. That data set also showed a significant set of missing fields in the employment type column which needed to be imputed.

# Analytics and Algorithms

The scope of this project is focused on predicting whether a consumer will default on loan or not, therefore we incorporated three different predictive models to predict the outcome and behavior of each consumer: Classification Tree; Logistic Regression; Random Forest.

Classification Tree: Just like all regression trees, classification tree is used to partition data into smaller and more homogenous groups. The use of homogeneity here means the nodes that are being split are pure - that is to say they contain a larger portion of one class in each node. Thus, purity is by means of maximizing accuracy or minimizing misclassification errors.

Accuracy is how we measure purity by means of partitioning the data in a way that place samples in primarily in one class.

The primary package for building CART and Regression Tress for that matter is - rpart. The function takes only the formular required to specify the exact form of the model.

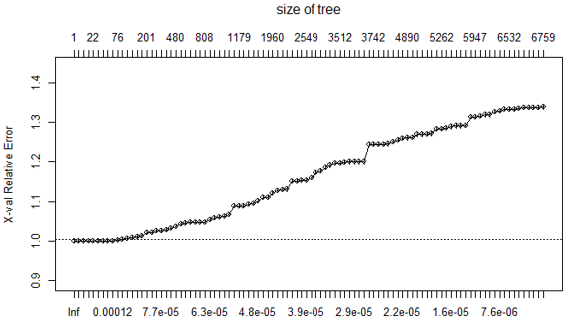


Figure 3. Classification Tree

We construction another classification tree without a different model layout showing six variables.

Trees that are constructed to have a maximum depth are notorious for over-fitting the training data, and in our case, we introduced the cp (cost complexity factor/complexity parameter) which is incorporated into the tuning process so that an optimal value can be estimated. The larger the size of the parameter the smaller number of predictors thus reducing model overfitting and we see the opposite behavior of this will result in model overfitting.

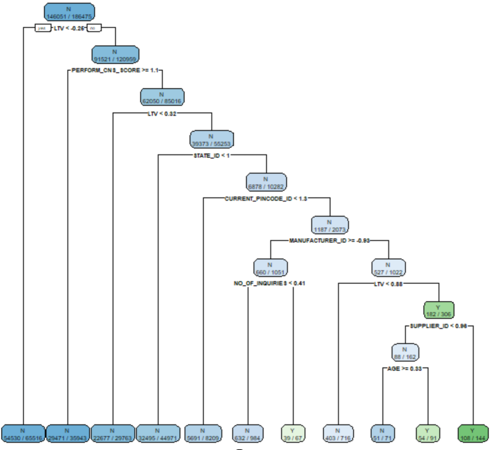


Figure 4. Classification Tree Model

There are two alternative measures of implementing the Gini Index. We will discuss Gini Index for two class problem.

Gini Index: For a two-class problem, Gini Index is defined as Class 1 & Class 2 ie. p1+p2 = 1

Usually, when working with a Categorical Response, the process for finding the optimal split point is similar to the process of Basic Regression Trees - where the model begins with the entire data set and searches every distinct value of every predictor to find the predictor and split the value that partitions the data into two groups: Class 1 & Class 2 such that the overall sums of squares errors are minimized [2].

Relative to our data set, first, the samples are sorted based on their predictor values. The split points then become the midpoints between each unique predictor value (which we see in parenthesis). Since we have a binary response, the process generates a 2X2 contingency table at each split point (all of which is calculated internally within R). The splitting process continuous within each newly created partition, thereby increasing the depth of the tree until the stopping criteria is met - where the minimum number of samples in the node or maximum tree depth [3].

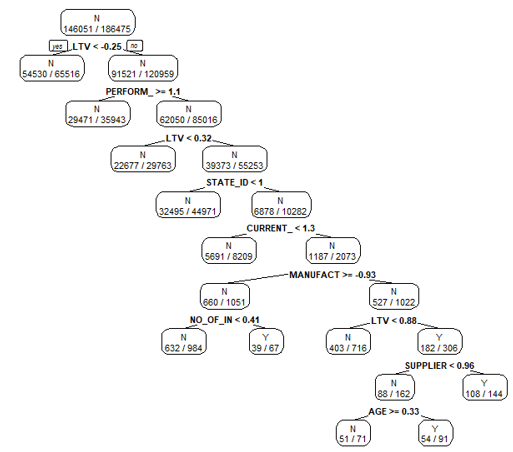


Figure 5. Gini Index for Classification Tree

There are several functions in R that are used to create the confusion matrix (which optimizes models’ performance). The caret package produces the table and all associated statistics which includes:

* Prediction: based on the class: Class1 Class2
* Accuracy from Gini Index
* CI: that work closely with ROC, AUC to determine the area under the curve
* Information Rate also from Gini Index
* P-Value [Acc > NIR] : < 2e-16
* Kappa
* Mcnemar's Test P-Value: where a P-Value less than 0.05 gives evidence that there a significant difference between the categorical variables.
* Sensitivity & Specificity: These are inversely proportional - so when one goes up the other decreases and vice versa.

**ROC Curve i**s probability curve generated from the pROC package alongside various statistics.

First, an R object is created that contains the relevant information in the pROC function. The resulting object is used to generate the ROC curve or calculate the area under the curve - where the area under the curve is the confidence interval - which expresses the degree of uncertainty around a certain effect.

In this case it tells us how much the model is capable of distinguishing between classes. The ROC is plotted with the True Positive Rate (TPR) Y-axis against False Positive Rate (FPR) X-axis

AUC represents the degree of separability. Basically, both AUC & ROC curve is a performance measurement for classification problems at various threshold setting.

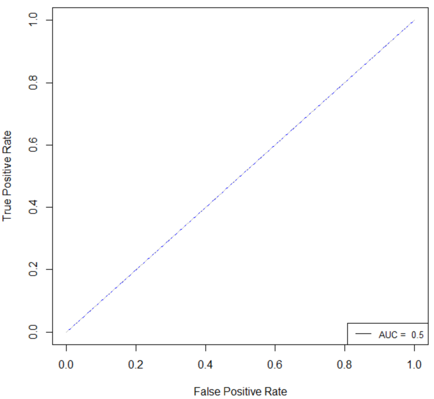


Figure 6. ROC Curve

We need the True Positive Rate (TPR)/Recall/Sensitivity. Specificity, an excellent model has AUC close to 1, and conversely, a poor model is close to 0 - making it the worst measure of separability. And when the AUC is 0.5, it means the model has no class separation capacity whatsoever, which is what our model outputs.

Table 5. Model Performance (Classification Tree)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 | AUC |
| 0.78 | 0.78 | 1.00 | 0.88 | 0.5 |

**Logistic Regression** forms a model that is linear in the parameters, and these parameters are obtained by minimizing the sum of the squared residuals. It turns out that the model that minimizes the sum of the squared residuals also produces maximum likelihood estimates of the parameters when it is reasonable to assume that the model residuals follow a normal (i.e., Gaussian) distribution. An effective logistic regression model would require an inspection of how the success rate related to each of the continuous predictors and, based on this, may parameterize the model terms to account for nonlinear effects. For our vehicle loan data, the full set of predictors was used in a logistic regression model [4].

Table 6. Model Performance (Logistic Regression)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | RMSE | R-squared |
| 0.7817 | 0.7818 | 0.9984 | 0.4044 | 0.0415 |

Table 7. Model Performance (Logistic Regression)

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
| Prediction | 0 | 1 |
| 0 | 36434 | 10128 |
| 1 | 58 | 59 |

Table 8. Models Comparison

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | N-Value | Accuracy | Precision | Recall | F1 | AUC | RMSE | R-Squared |
| LS |  | 0.78 | 0.78 | 0.99 | - | - | 0.40 | 0.04 |
| RF | 5 | 0.74 | 0.79 | 0.91 | - | - | 0.51 | 0.01 |
|  | 10 | 0.76 | 0.79 | 0.95 | - | - | 0.49 | 0.01 |
|  | 4 | 0.75 | 0.79 | 0.92 | - | - | 0.5 | 0 |
|  | 6 | 0.76 | 0.79 | 0.93 | - | - | 0.49 | 0.01 |
| CART | 186475 | 0.78 | 0.78 | 1.00 | 0.88 | 0.5 | - | - |

**Random Forest**

The primary implementation for random forest comes from the package with the same name: randomForest. The two main arguments are mtry for the number of predictors that are randomly sampled as candidates for each split and ntree for the number of bootstrap samples. The default for mtry in regression is the number of predictors divided by three.

The number of trees should be large enough to provide a stable, reproducible results. default for try in regression is the number of predictors divided by 3. The number of trees should be large enough to provide a stable, reproducible results [4].

# Findings

In general, discriminant or classification techniques seek to categorize samples into groups based on the predictor characteristics, and the route to achieving this minimization is different for each technique. While basic logistic regression has no tuning parameters, resampling can still be used to characterize the performance of the model. The summary indicates that the performance distributions are very similar.

# Summary

The goal of the project was to determine consumer behavior. Financial organizations suffer substantial losses from defaults on car loans and consequently, calling for a more effective credit risk scoring model to assess the likelihood of vehicle loan defaults. All three models achieved the results. However, the most effective models among all three are Classification Model & Logistic Regression Model.

# Future Work

Throughout this project, we learned the effective use of Exploratory Data Analysis (EDA) when developing a predictive model. While we found the importance of several related packages, we also identified the functionality and consistency of a clean data improving accuracy and performance.

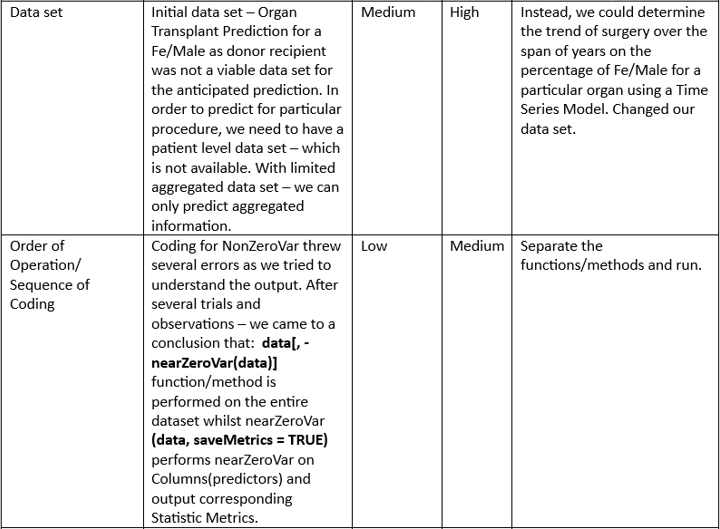
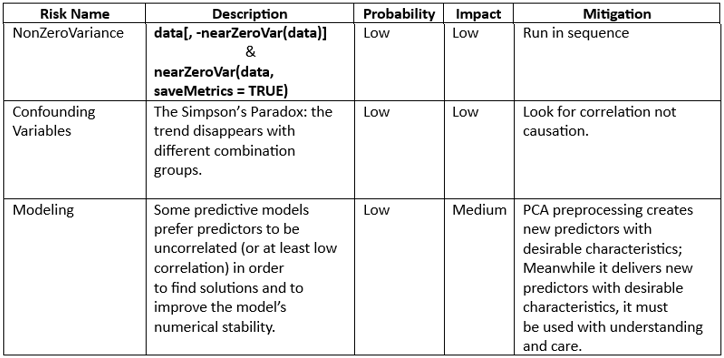
Secondly, there was one constraint variable, employment status which had several missing values requiring data imputation. Imputing the fields enhanced the determination of consumers behavior to default or not to default on a loan.

Lastly, to observe a larger sample size for precision, we can collect data from the various credit bureau agencies and comparing the trend of each consumers behavior would allow us to investigate whether our model’s accuracy would only fit that from Kaggle or are universally feasible. By doing so, we could either improve furthermore or conclude that the models are capable.

# Appendix A: Code References

Details found at: <https://github.com/fhill09/Team-Versatile>

# Appendix B: Risk Section

Table 9. Roadblocks - Risks & Mitigations

# References

|  |  |
| --- | --- |
| [1] | C. Horymski, "Experian," 2023. [Online]. Available: https://www.experian.com/blogs/ask-experian/research/auto-loan-debt-study/. [Accessed 24 04 2023]. |
| [2] | M. K. &. K. Johnson, in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 384 - 385. |
| [3] | M. K. &. K. Johnson, "Basic Regression Trees," in *Applied Predictive Modeling*, New York, Springer, 2013, p. 190. |
| [4] | M. K. &. K. Johnson, "Logistic Regression," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 48, 283, 284. |
| [5] | M. k. &. K. Johnson, "Random Forest," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 215 - 216. |