Business & Data Case Study - Report

“LoanGuard: Revealing Authenticity through AI-Powered Application Analysis”

October 2023

VINAY MITTAL

STUDENT ID: 33613877

MASTER OF DATA SCIENCE

Contents

Project Description

According to Indian Express “Banks may have to settle with some of the 16,044 willful default accounts with Rs 346,479 crore debt till end-2022“. Despite having the financial capability, the identified borrowers refused to repay loan instalments. Apparently, the loan applications were assessed based on their creditworthiness ratings. However, this evaluation ultimately led to the occurrence of financial losses described above (Indian Express, [2022]).

As a result of this, financial institutions are getting cautious and exerting in the rejection of genuine loan applications. In pursuit to counter this problem, hereby proposing a data science project to forecast loan acceptance or rejection based on consumer behavior.

Business Model

The financial institution which offers loans has always been in trouble evaluating loan applications, this led to the increase in missed opportunities and higher default rates. The model proposes an AI-Based application named “**LoanGuard: Revealing Authenticity through AI-Powered Application Analysis**”.

Additionally, LoanGuard offers “**Client inquiry management hub**” database storage, allowing for easy access to information for resolving daily client enquiries and demands. Most importantly, A crucial feature “**Smart Interest Projection Tool**” is integrated in the software to re-evaluate the interest minimization on the loan amount based on the supplied application parameters.

A. Data Roles

Data Engineer

Open-source data from platforms like Kaggle and government websites will be used to collect data based on various human behaviors. This will involve extraction, transformation, management, and conversion of raw data into valuable information.

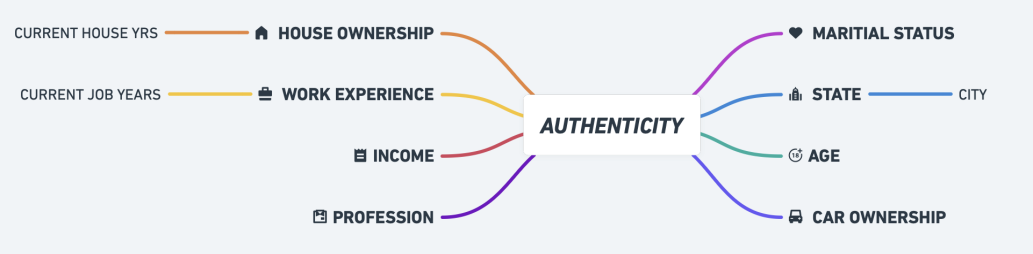

Figure 1 Human Behavior

Data Scientist

Responsible for pulling insights from clean data to understand the pattern or trend in loan defaulters.

Machine Learning Engineer

Responsible for performing an algorithm for developing a predictive model, to identify the red flag in the applied loan applicants. Based on parameters mentioned in *Figure 1*, *the model* will be developed.

Business Analyst

Responsible for understanding the process gaps and bringing valuable insights to facilitate the business.

IT Support/ Software developer

Develop an AI- Software as a Service (SaaS)-compliant. Responsible for developing and maintaining the application. I have built an initial graphical user interface (GUI). Fig 1.2 and Fig 1.3 is the dashboard built on Figma.

B. Benefits

As shown in 1.4.

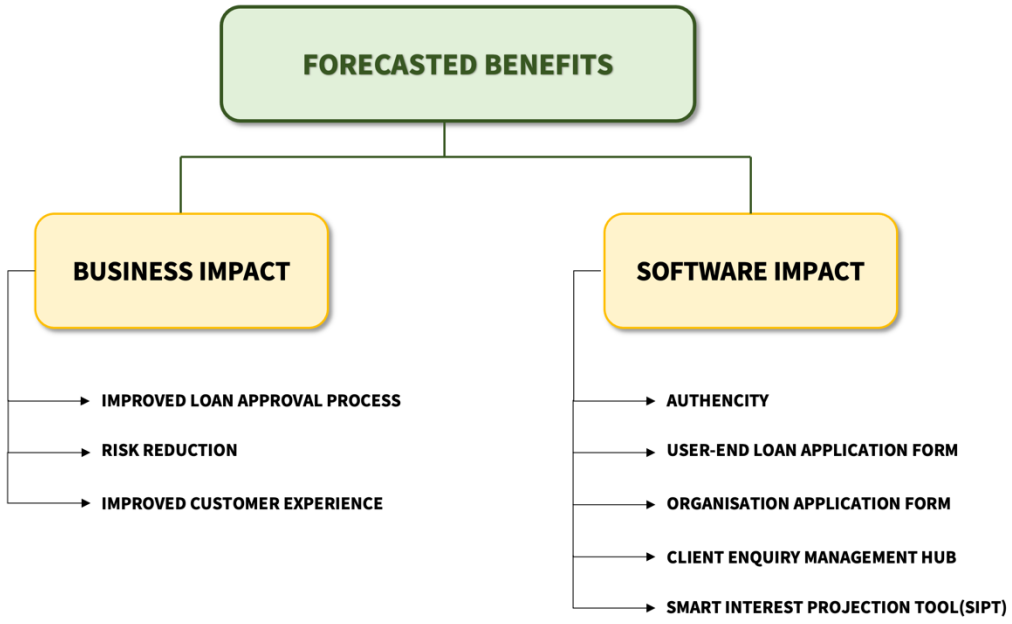

Figure 1.4 Forecasted Benefits

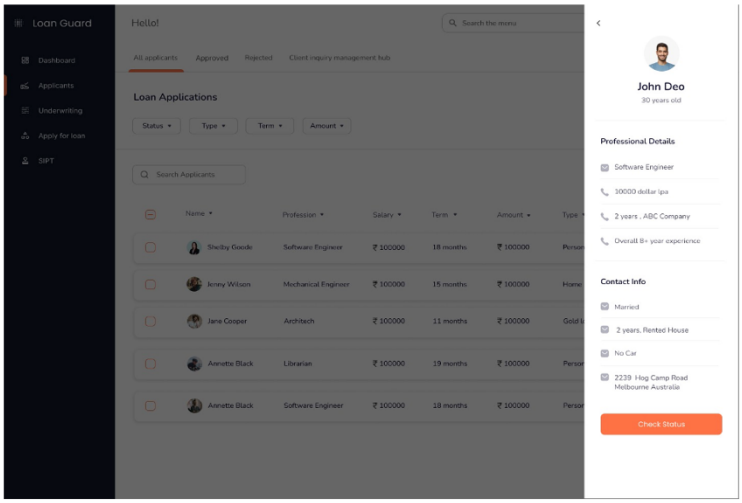

Figure 1.2 GUI Dashboard

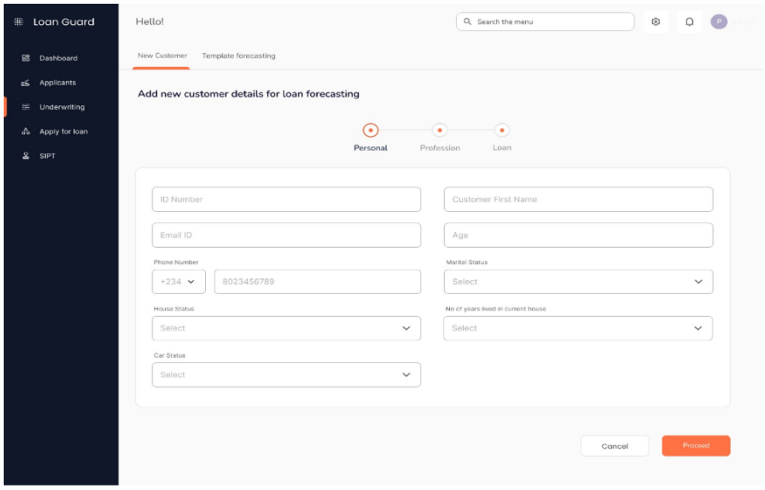

Figure 1.3 GUI Dashboard

C. Challenges

There are following challenges that we have might have while building the model.

- Limited number of parameters in the dataset. Parameters like loan amount, number of dependents, different types of loan can be a potential attribute in our model.
- Implementation of new interest on the loan application as we need to collaborate with the government.

D. Novelty

Only a few applications have been done in this field, yielding only partial results. Consequently, it is crucial that all aspects work together as a cohesive whole.

Characterising & Analysing Data

A. Potential Data Sources

It is not feasible to collect data from primary sources because of restrictions on confidentiality and the challenge of getting survey responses. Rather, secondary sources like banks, government agencies, financial institutions, and data repositories like Kaggle, will be our focus.

B. Data Characteristics

"Loan Prediction Based on Customer Behavior" from Kaggle (https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior/) using training.csv data.

This dataset contains records of people who have availed loans in India, along with a binary indication of whether the borrower defaulted on the loan indicated by risk flag column. The dataset contains diverse information which is discussed later.

1. Volume

Volume is basically the amount of data. In this case, the dataset has a large volume, containing 250,000 records and 13 columns in total. Every record is associated with a single entry of the individual and contains multiple details regarding different qualities.

2. Velocity

The second V of big data, velocity, is all about how quickly new data is generated and dispersed. Regarding velocity with the tabular dataset taken for this project, which contains information about people who have taken out loans. It reflects a static and unique collection of the data for each.

3. Variety

```
spc_tbl_ [252,000 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Id               : num [1:252000] 1 2 3 4 5 6 7 8 9 10 ...
 $ Income           : num [1:252000] 1303834 7574516 3991815 6256451 5768871 ...
 $ Age              : num [1:252000] 23 40 66 41 47 64 58 33 24 23 ...
 $ Experience       : num [1:252000] 3 10 4 2 11 0 14 2 17 12 ...
 $ Married/Single   : chr [1:252000] "single" "single" "married" "single" ...
 $ House_Ownership  : chr [1:252000] "rented" "rented" "rented" "rented" ...
 $ Car_Ownership    : chr [1:252000] "no" "no" "no" "yes" ...
 $ Profession       : chr [1:252000] "Mechanical_engineer" "Software_Developer" "Technical_writer" "Software_Developer"
 $ CITY             : chr [1:252000] "Rewa" "Parbhani" "Alappuzha" "Bhubaneswar" ...
 $ STATE            : chr [1:252000] "Madhya_Pradesh" "Maharashtra" "Kerala" "Odisha" ...
 $ CURRENT_JOB_YRS  : num [1:252000] 3 9 4 2 3 0 8 2 11 5 ...
 $ CURRENT_HOUSE_YRS: num [1:252000] 13 13 10 12 14 12 12 14 11 13 ...
 $ Risk_Flag        : num [1:252000] 0 0 0 1 1 0 0 0 0 0 ...
```

Figure 2.1 Meta Data

The idea of data diversity refers to the diverse type of information in terms of data types that can be obtained from the dataset. We find a broad range of information in the dataset that is supplied, which may be essentially divided into two categories: **Numerical and Categorical data**. 2.1 and 2.2 shows the schema the data frame created from the dataset.

Column Name	Column Description	Data Type
ID	*Unique number assigned to each record.*	*int*
INCOME	*Income of the user*	*int*
AGE	*Age of the user*	*int*
EXPERIENCE	*Professional experience of the user in years*	*int*
MARRIED/SINGLE	*Whether married or single*	*string*
HOUSE_OWNERSHIP	*Owned or rented or neither*	*string*
CAR_OWNERSHIP	*Does the person own a car*	*string*
PROFESSION	*Profession*	*string*
CITY	*CITY*	*string*
STATE	*State of residence*	*string*
CURRENT_JOB_YRS	*Years of experience in the current job*	*int*
CURRENT_HOUSE_YRS	*Number of years in the current residence*	*int*
RISK_FLAG	*Defaulted on a loan*	*int*
	1- Defaulter	
	0- Non -Defaulter	

Table 2.2 Meta Data

4. Veracity

Veracity refers to the quality and accuracy of data. We will carry out a series of tasks such as data wrangling as well as extensive data checking, to ensure the integrity of the data.

C. Data Storage

A scalable file system made for huge data storage is the **Hadoop Distributed File System**. It is also capable of handling fault tolerance and data replication. For data storage of this size, **cloud-based storage** options like Google Cloud, Amazon S3, and Azure can also be utilized.

D. Data Processing and Analysis

Tools like **MapReduce** for distributed data processing and Hive for querying are among those included in the **Hadoop ecosystem**. It will work perfectly with the way we process data. Another feature of **Apache Spark** that gives it an advantage over Hadoop Ecosystem for processing large amounts of data is its in-memory processing capability. **R-studio and PyCharm** is a potential analysis tool with diverse libraries.

E. Proposed Data Analysis

Data analysis that will implement on the dataset are as follows:

1. **Wrangling and Checking**

   In this section we will perform series of operation like dropping irreverent column, handle missing data and check data consistency and integrity. Therefore, making the data set ready for analysis.

2. **Feature Analysis**

   - ***Individual Analysis****:* Analyze the feature individually and check for pattern to understand the distribution.
     Numerical Columns
     - **Box plot**

       Reason: This will provide valuable insight into the distribution.

       High Level Output: This statistical method will provide descriptive statistics.

     - **Correlation Matrix**

       Reason: This statistical method will provide a numerical representation of the intensity and direction of these interactions, ranging from -1 to 1.

       High-Level Output: It provides a brief overview of patterns, assisting in the identification of significant positive or negative correlations.

     Categorical Columns

     - **Grouped bar chart and stacked Bar plot**

       Reason: Effective method to understanding the distribution by frequency.

       High-Level Output: This will help to bring insight between categorical columns and risk flag columns(target variable).

   - ***Combined Analysis:*** Similarly, we will combine more than one field and perform analysis.
     - **Radar Chart**

       Reason: Excellent method to showcase uniformity.

High-Level Output: To highlight income across 51 occupations, we are developing a radar graphic. The goal of highlighting the top five and worst five jobs is to simplify comprehension.

3. **Predictive Model**

Random Forest

It is an excellent predictive model because of its flexibility to handle a variety of data types and good at capturing non-linear correlations. By randomly setting the leaf node, ensemble learning reduces overfitting and improve generalization to new data. Moreover, it handles categorical variables with ease and without any preprocessing. Additionally, its feature importance measure helps to understand the main elements that affect the "Risk Flag".

Demonstration

A. Wrangling and Checking

1. Dropping Columns

The dataset's ID field is used only to provide each entry a unique identifier. Therefore, we will drop it

```
> # removing ID column from training data
> names(data_df)
 [1] "Id"               "Income"           "Age"              "Experience"       "Married/Single"
 [6] "House_Ownership"  "Car_Ownership"    "Profession"       "CITY"             "STATE"
[11] "CURRENT_JOB_YRS"  "CURRENT_HOUSE_YRS" "Risk_Flag"
> data_df1 <- data_df %>% select(-Id)
```

Figure 3.1 Dropping ID Column

2. Missing Data

Since there are no null values in the dataset as shown in 3.2, we don't need to impute any values to the cell based on the skewness of the distribution.

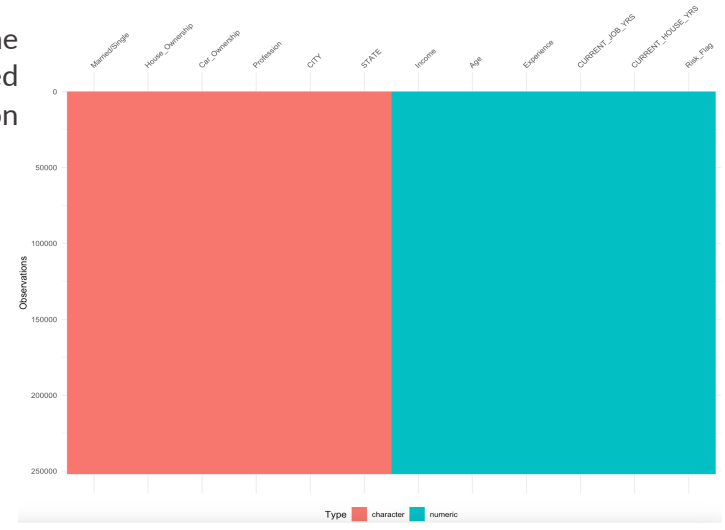

Figure 3.2 Missing data check

3. Data Consistency

It is important to ensure the logical coherence of the data.

- *Numerical Columns:* For example, would be inconsistent with logic if the existence of negative values in the Age column. As shown in 3.3 no other numerical column has negative value.

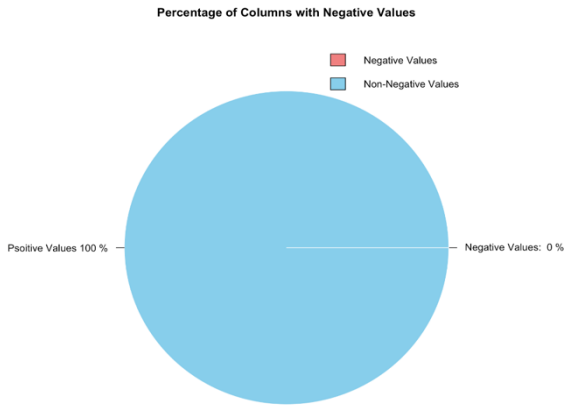

Figure 3.3 Percentage of Negative Numerical Column

- *Categorical Columns:* Found inconsistency in distinct values of State and City column. Removed special character shown in 3.4.

```
259 Nellore[14][15]      27 Uttar_Pradesh[5]
260 Visakhapatnam[4]

# Removing [] content
data_df1$CITY <- gsub("\\[.*?\\]", "", data_df1$CITY)
data_df1$STATE <- gsub("\\[.*?\\]", "", data_df1$STATE)
```

Figure 3.4 Irregularity in Categorical column

B. Feature Analysis

- **Box Plot**

A box plot showing normal distributions for all numerical columns can be observed in 3.5. Although income and other real-world variables often show a right-skewed distribution, our dataset significantly deviates from this general pattern. Rather, it displays a more regular distribution of income, indicating that people with ordinary incomes are more likely to apply for loans because they want to improve their lifestyle or achieve more.

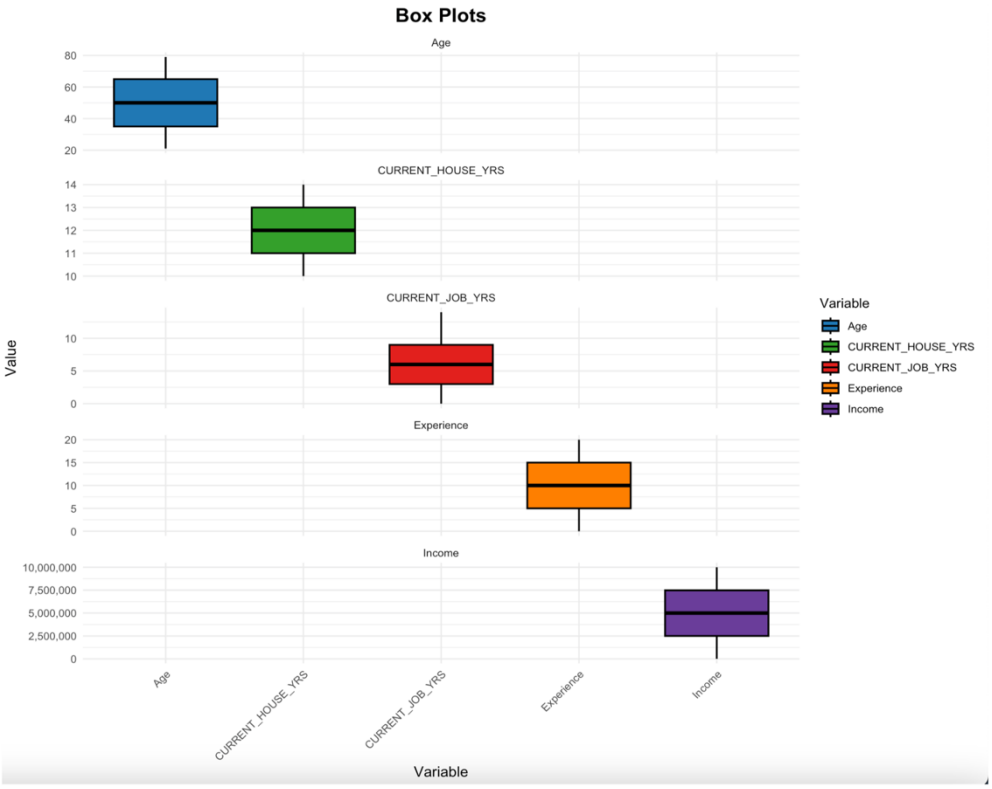

Figure 3.5 Box plot for Numerical Columns

- **Correlation Matrix**

We investigate relationships between numerical columns by using a correlation matrix. In line with the reasonable assumptions, 3.6 shows a weak positive association between Experience and Current_Job_Yrs. Other pairs have correlations that are almost equal to zero, indicating the potential of a non-linear relationship or the absence of relationship.

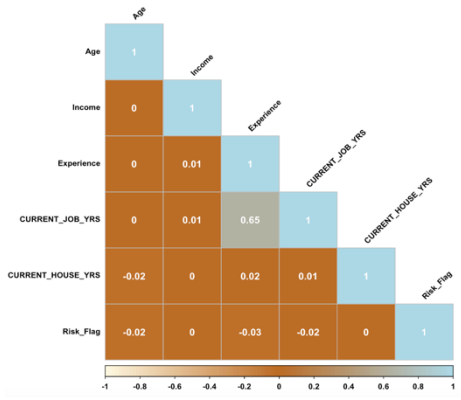

Figure 3.6 Correlation between Numerical Columns

- **Group Bar plot**

Married/Single

Bar plot showing loan defaulters according to marital status is shown in 3.7. Remarkably, the default rate for both groups is roughly 10%. Not showing much significance on its own.

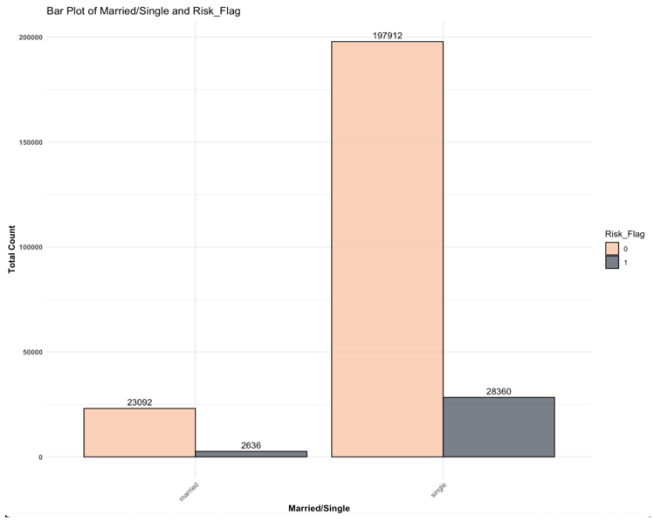

Figure 3.7 Married/Single Bar plot

House Ownership

Distribution regarding the Risk flag is shown in 3.8. Interestingly, the combined risk for Risk_Flag 1 for the "no rent" and "owned" categories is 9%, whereas the "rented" category shows a higher 13%. This implies that renting a property may be associated with a higher risk of loan default.

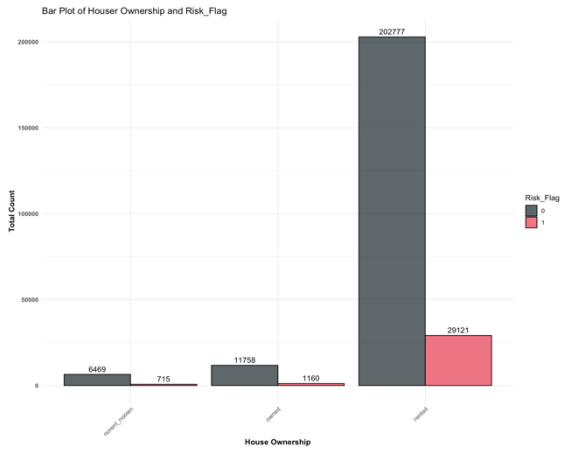

Figure 3.8 House Ownership Bar plot

Income and Profession

Average income of the top five and bottom five professions are plotted, a regular polygon shape is revealed, suggesting that there is little difference in average income across these groups.

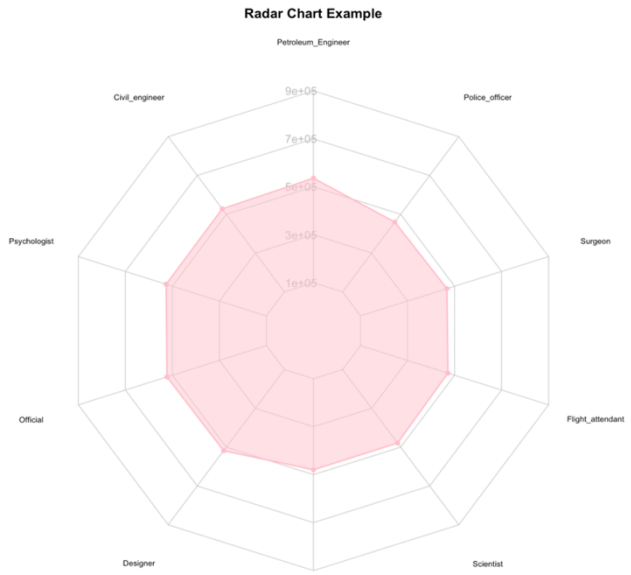

Figure 3.9 Radar chart of Average income across top 5 and Bottom 5 professions

As a result, the graph below, which shows that those in formal professions are more likely to default on loans, is not affected by income. This group has a similar average income, but there is a greater chance of non-payment.

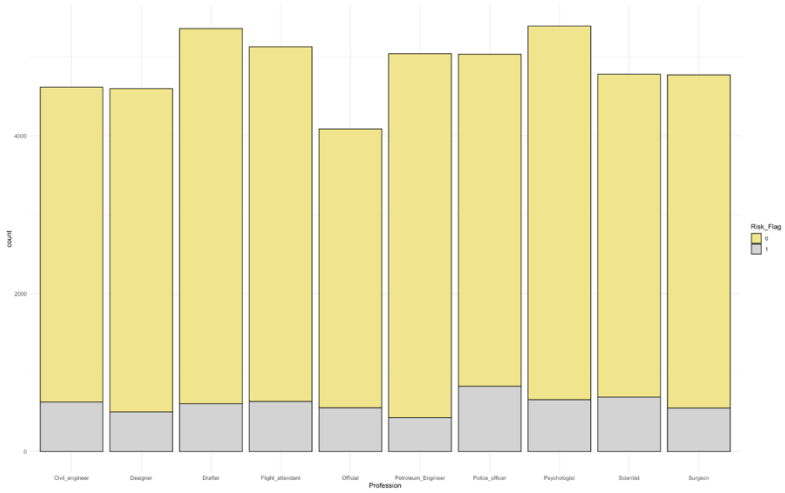

Figure 3.9 Bar plot of same top 5 and Bottom 5 Countries grouped by Risk Flag

Car Ownership

For both groups, the risk flag of 1 is roughly 11%, suggesting that it may be difficult to make meaningful judgements based only on this criterion.

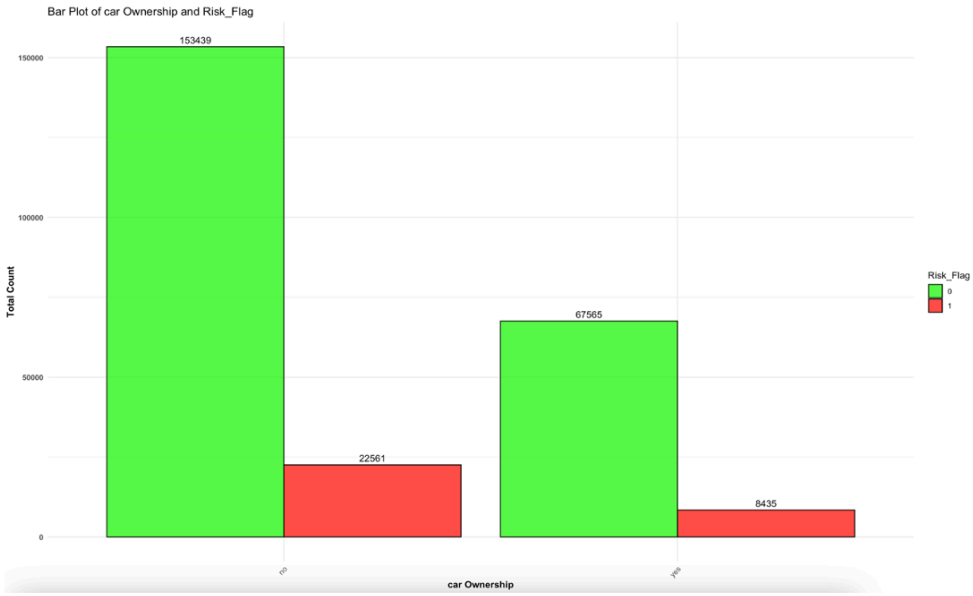

Figure 3.10 Bar plot showing loan defaulter according to Car ownership

C. Random Forest Predictive Model

We will perform following series of operation to get the best model.

Critical and Creative Novel Idea

We have employed a **creative approach** to determine the perfect number of trees to find the best Random Forest model. 4.1 presents the outcomes by methodically changing the number of trees and recording accuracy. The graph indicates the ideal hyper-parameter value and shows the peak accuracy, which occurs at two trees. It is **critical** to understand that only accuracy is not sufficient so plotted ROC curve to show that model is feasible.

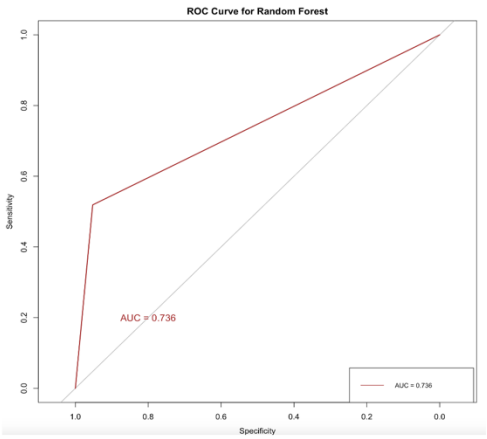

Figure 4.2 ROC Curve

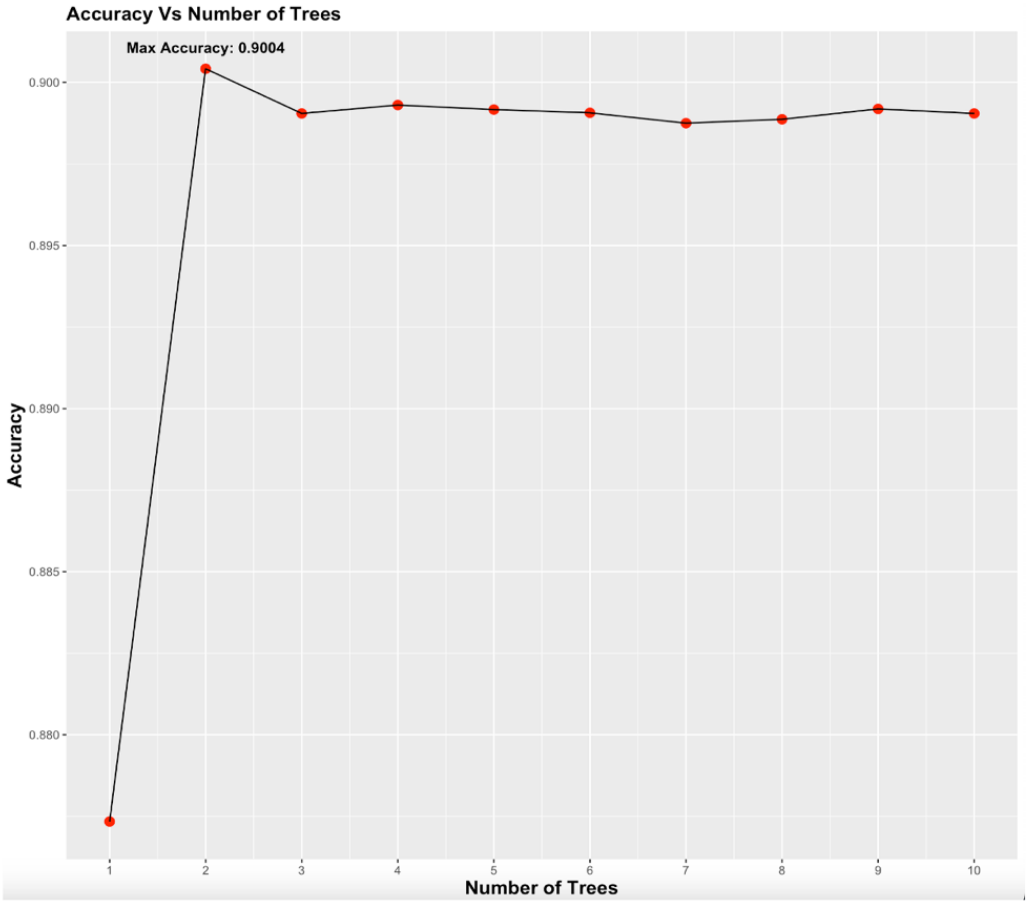

Figure 4.1 Determination of Optimal Number of Trees

We will apply **k-fold cross-validation** to mitigate overfitting and improve performance on unseen data. Using the smallest dataset, we may create a variety of combinations of training and test data using this technique. This **novel approach,** by integrating k-fold cross-validation not only mitigates possible overfitting issues but also strengthens the robustness of our model assessment.

Standard for Data Science Process, Data Governance & Management

Data Science Process Standard

The project follows Cross Industry Standard Process for Data Mining (CRISP – DM) framework, involving six stages in figure 5:

1. **Business Understanding**

   Project objective of categorizing the loan application into potential defaulters.

2. **Data Understanding**

   Data was obtained from secondary resources like online repositories (Kaggle).

3. **Data Preparation**

   Data checking, wrangling including removal of anomalies. Feature analysis was done prior to pass the data for model.

4. **Modelling**

   Random Forest model

5. **Evaluation**

   Accuracy for different hyperparameter was chosen and optimum value was selected.

6. **Deployment**
   Proposing the model's deployment computer software falling SaaS category.

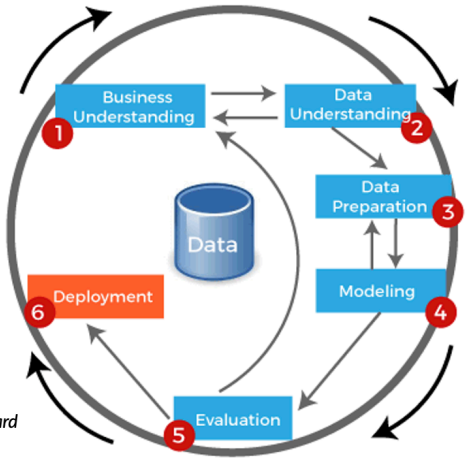

Figure 5 Data Science Process Standard

Data Governance and Management

The following are appropriate data governance and management procedures and related difficulties:

1. **Accessibility**

   It is necessary to keep a data catalogue that should be describe the potential loan defaulter data to relevant financial institution and government.

2. **Security**
   - High level of security is required to preserve the highly sensitive data of the applicants.
   - Encrypt data end to end.

3. **Confidentiality**

   Personal data is sensitive, it is imperative that information be protected utilizing anonymization techniques.

4. **Data Retention Policy**

   Establishing a precise policy for data preservation that ensures compliance and reduces administrative costs by regularly reviewing and archiving data.

5. **Ethical Concerns**
   - Make sure the model and analysis are transparent and explainable.
   - Follow ethical and legal regulations, such as data protection laws.

References

Banks may have to settle with some of the 16,044 wilful default accounts with Rs 346,479 crore debt till end-2022. (2023, June 18). The Indian Express. https://indianexpress.com/article/business/banking-and-finance/banks-may-have-to-settle-with-some-of-the-16044-wilful-default-accounts-with-rs-346479-crore-debt-till-end-2022-8670020/

What Is Data Science Process and Its Significance? (2021, September 23). Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. https://www.analytixlabs.co.in/blog/data-science-process/

Beautiful Radar Chart in R using FMSB and GGPlot Packages. (2020, December 12). Datanovia. https://www.datanovia.com/en/blog/beautiful-radar-chart-in-r-using-fmsb-and-ggplot-packages/