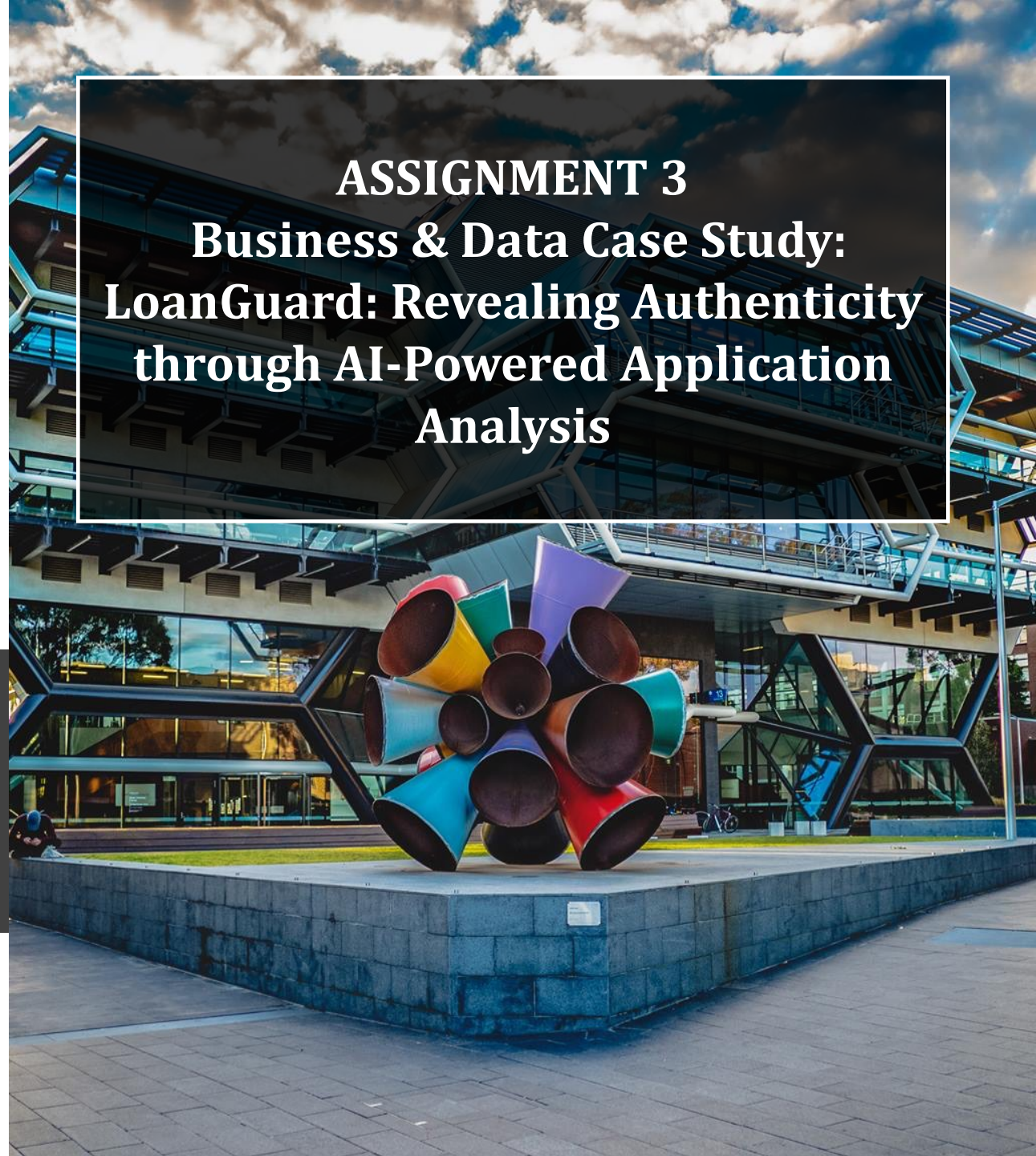
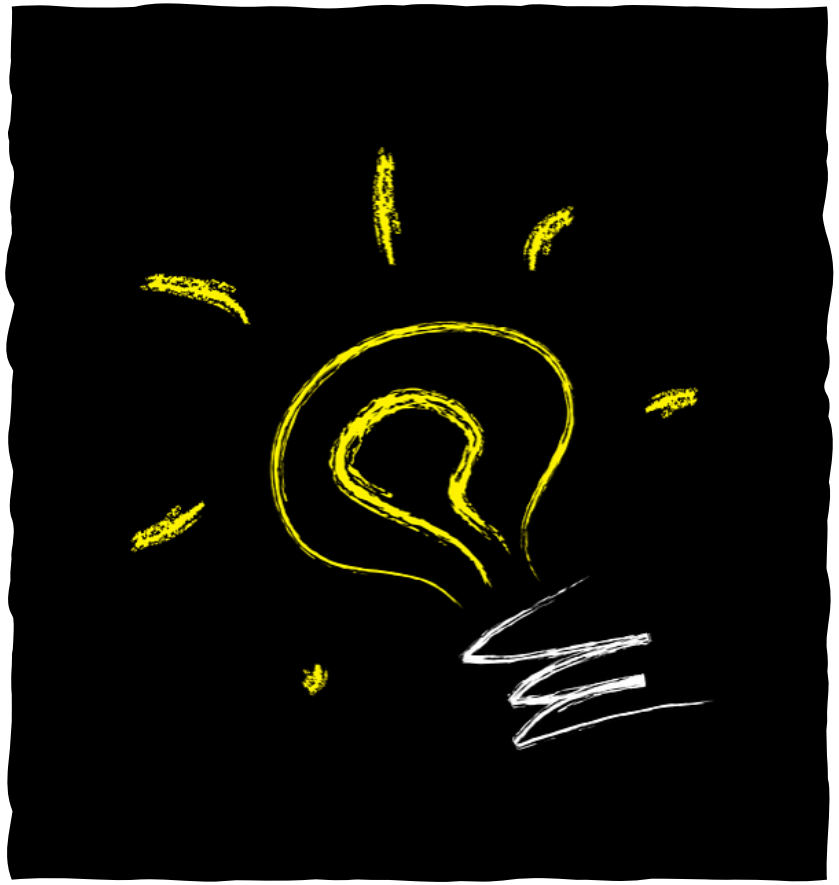
MONASH University

ASSIGNMENT 3
Business & Data Case Study: LoanGuard: Revealing Authenticity through AI-Powered Application Analysis



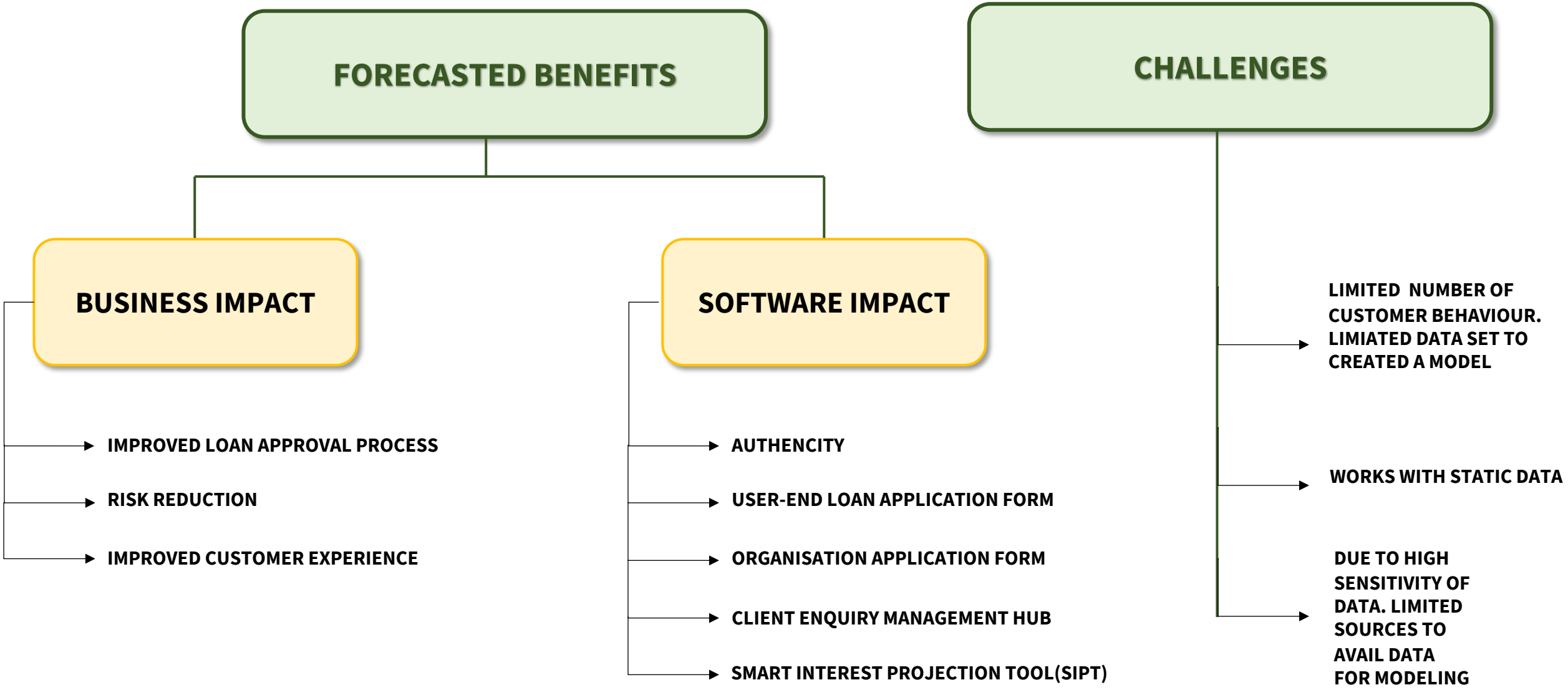
NAME: VINAY MITTAL
STUDENT ID: 33613877
COURSE: Masters Of Data Science
UNIT ID : FIT 5145

Project Description



Objective	AI-Powered Feature	Data Science Roles
• Forecast loan acceptance or rejection based on consumer behavior	• LoanGuard: Revealing Authenticity through AI-Powered Application Analysis • Client inquiry management hub • Smart Interest Projection Tool	• Data Engineer • Data Scientist • Machine Learning Engineer • Business Analyst • IT Support/ Software developer

Business Model



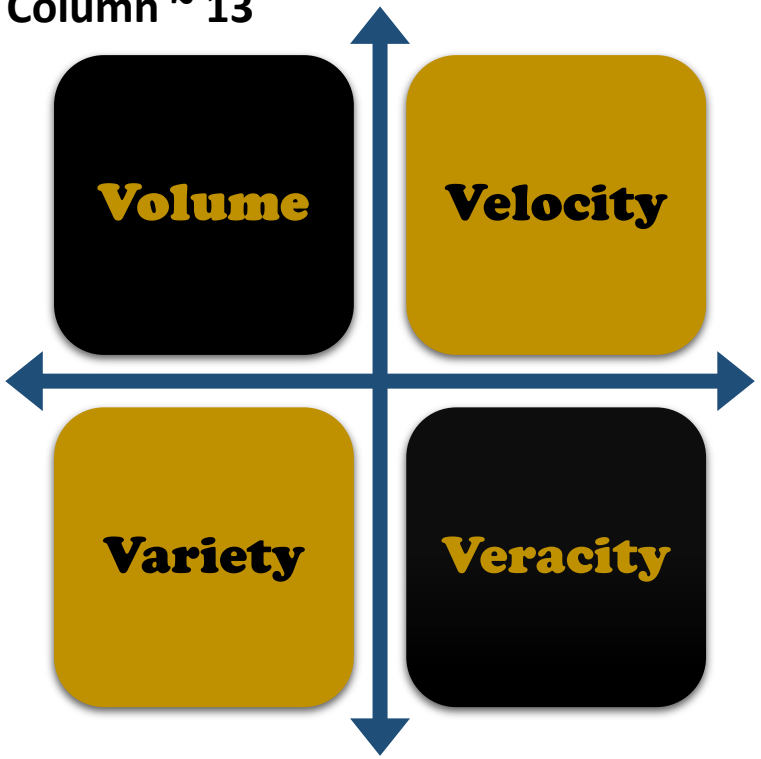
Data Characteristics

Potential Data Sources

- Sensitive Information
- Secondary Sources like Financial Institution, Government and bank records and Online Repository like Kaggle.

Rows ~ 250000
Column ~ 13

Static Tabular Data



Categorical & Numerical Data

Inconsistency

Data Storage

- Hadoop Distributed File System
- cloud-based storage

Data Processing and Analysis

- MapReduce
- Hadoop ecosystem
- Apache Spark
- R-studio
- PyCharm

Proposed Data Analysis

Wrangling and Checking

- Dropping Columns
- **Missing Data**
- **Data Consistency**

Feature Analysis

- **Individual Analysis**
- **Combined Analysis**
- **Statistical Method:** Box plot, grouped bar chart, stacked Bar plot, Correlation Matrix and Radar Chart

Predictive Model

Random Forest

Demonstration: Wrangling and Checking

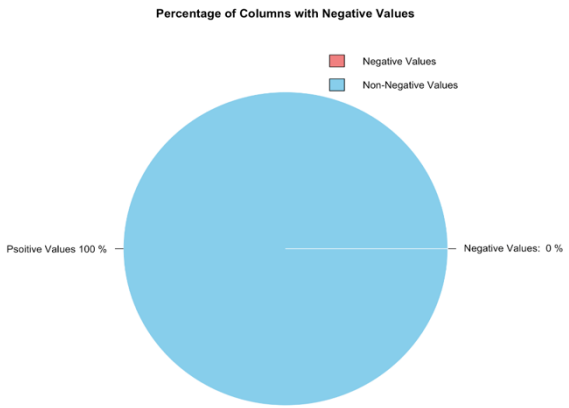
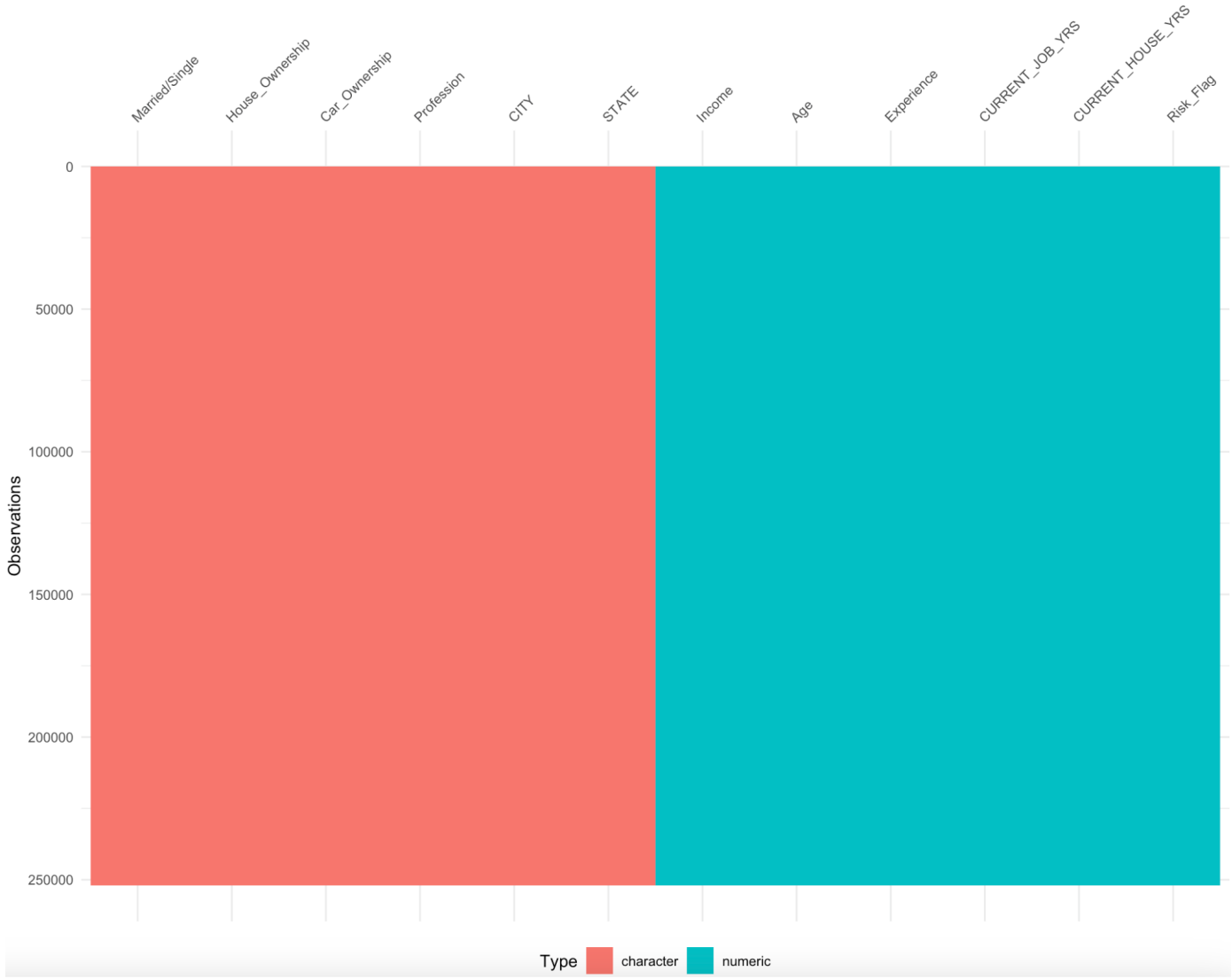


Figure 3.3 Percentage of Negative Numerical Column

```
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



Demonstration: Feature Analysis

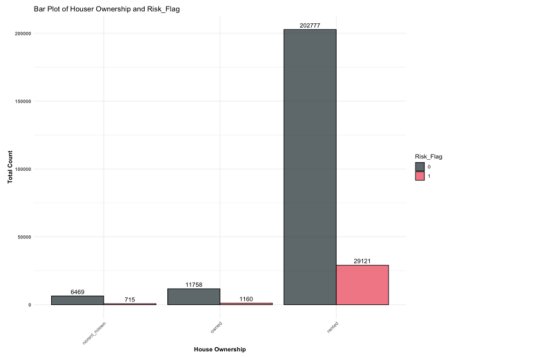


Figure 3.8 House Ownership Bar plot

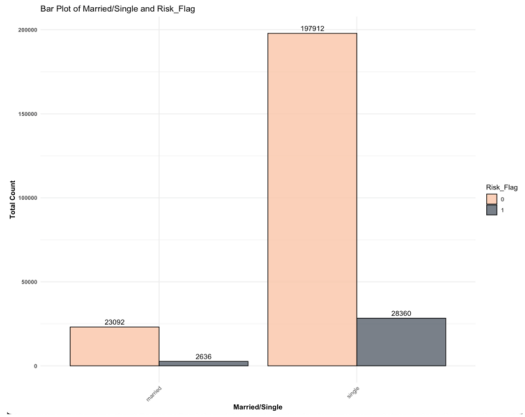


Figure 3.7 Married/Single Bar plot

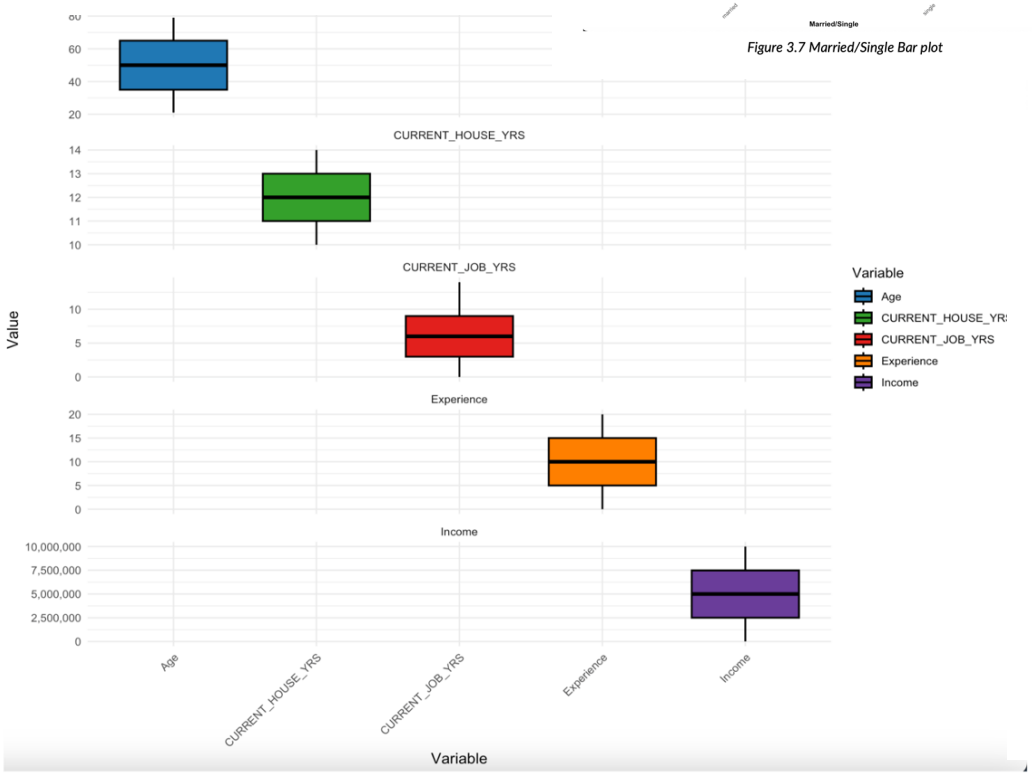


Figure 3.5 Box plot for Numerical Columns

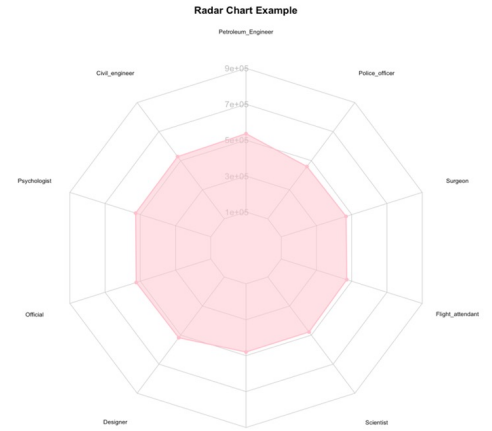


Figure 3.9 Radar chart of Average income across top 5 and Bottom 5 pro

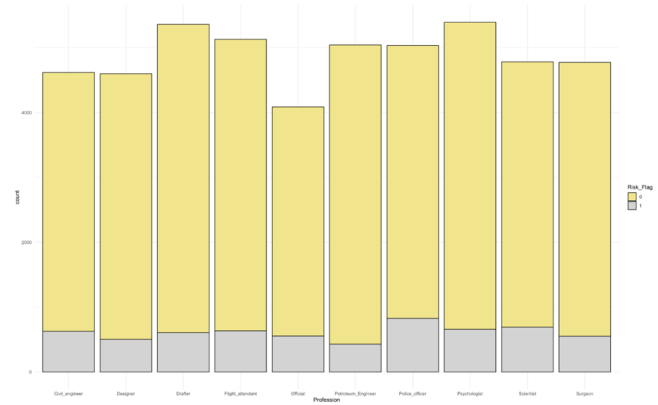


Figure 3.9 Bar plot of same top 5 and Bottom 5 Countries grouped by Risk Flag

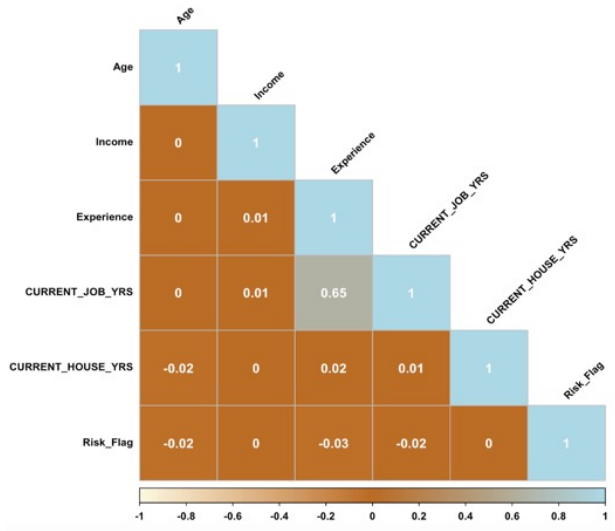


Figure 3.6 Correlation between Numerical Columns

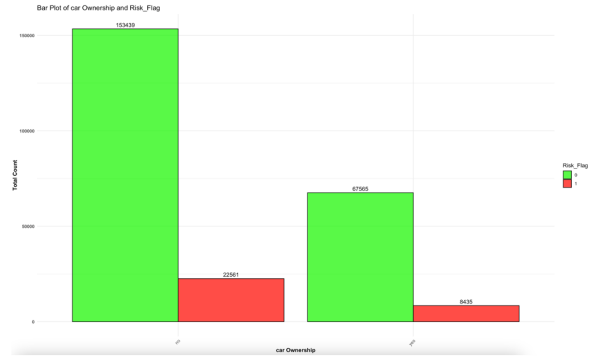


Figure 3.10 Bar plot showing loan defaulter according to Car ownership

Demonstration: Predictive Analysis

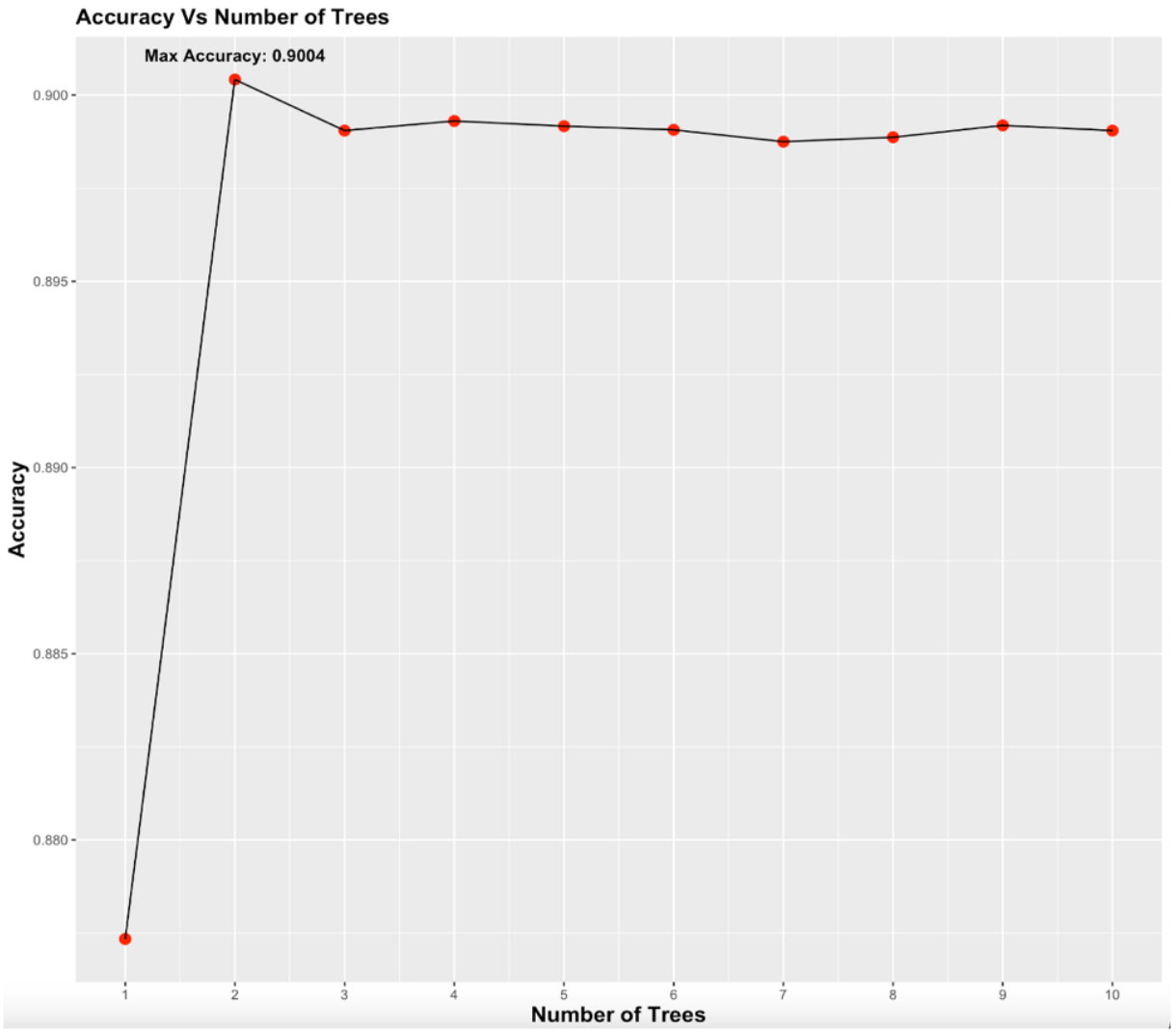


Figure 4.1 Determination of Optimal Number of Trees

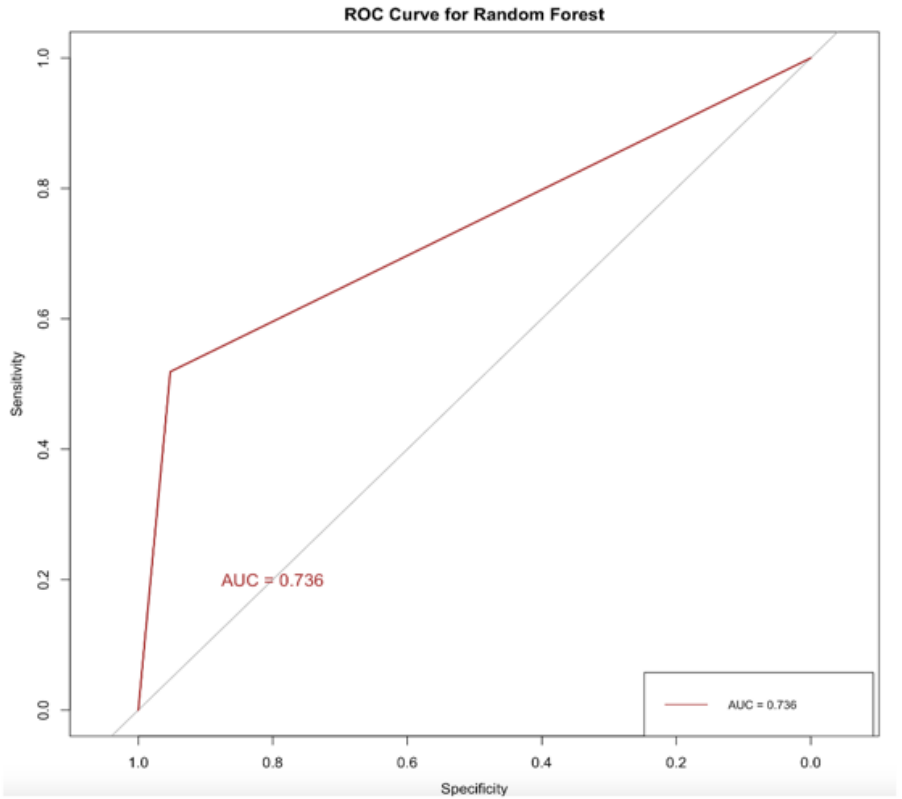
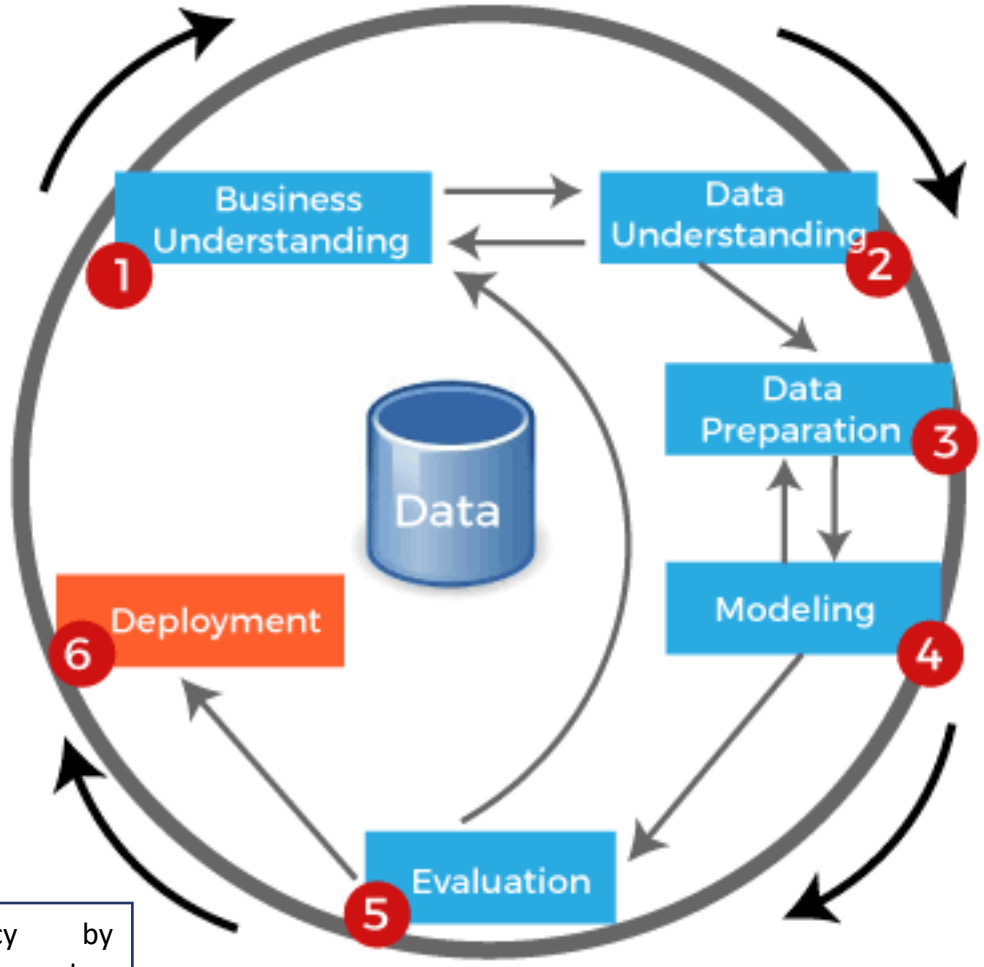


Figure 4.2 ROC Curve

Data Science Process Standard

Project follows **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP – DM) framework

Discussed Objective and Business Model.



Loan defaulters and corresponding customer behavior was obtained from secondary sources that is Online data repositories like Kaggle.

Data Wrangling and checking, Feature Analysis has been performed based on Numerical data columns and Categorical Data Columns.

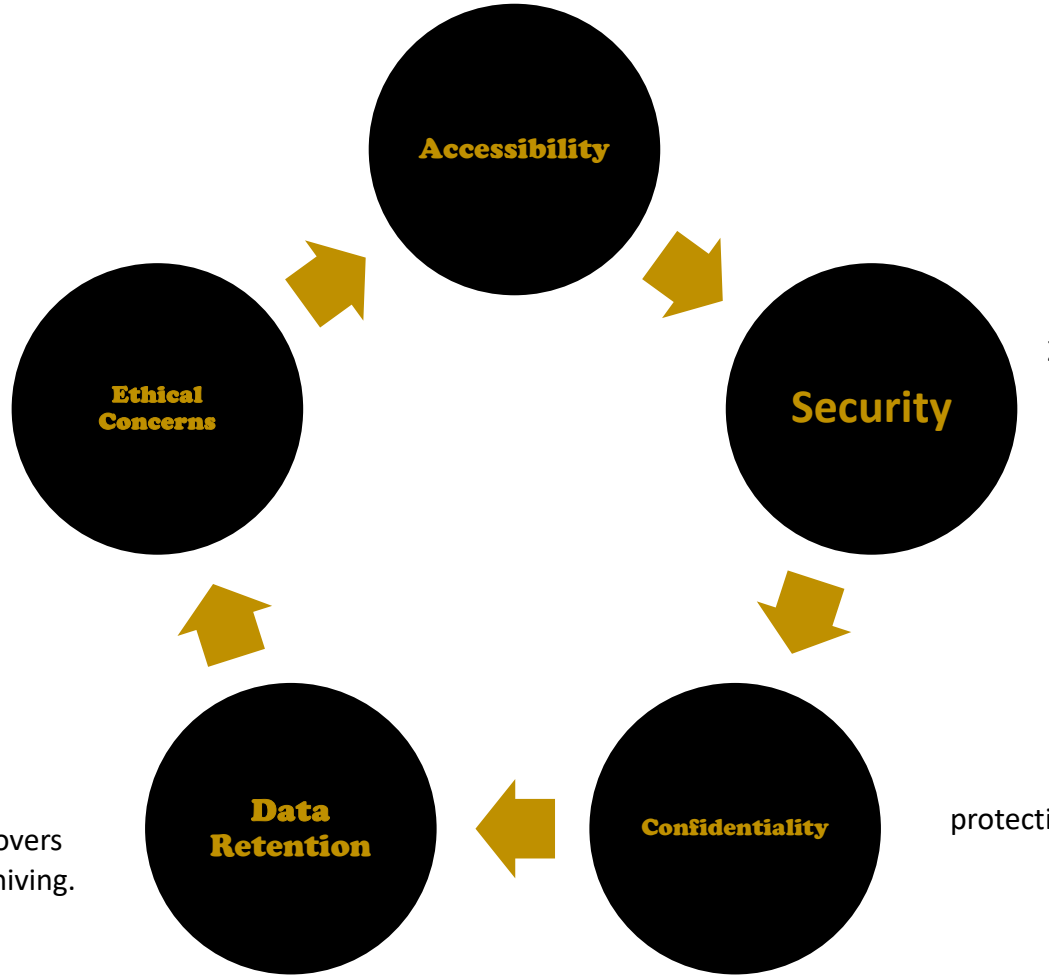
Selected Random Forest as Predictive Model.

Proposing the model's deployment as computer software Integrated with SaaS.

Calculated Accuracy by iterating for different number of trees and found an optimum best performing model.

Data Governance and Management

Maintaining Data Base and controlling accessibility for clients.



Zero tolerance and high security required due to sensitiveness of the data as it has all personal details.

protection of private information by anonymization.

A precise retention policy that covers frequent data inspection and archiving.

Follow ethical and legal regulations, like data protection laws