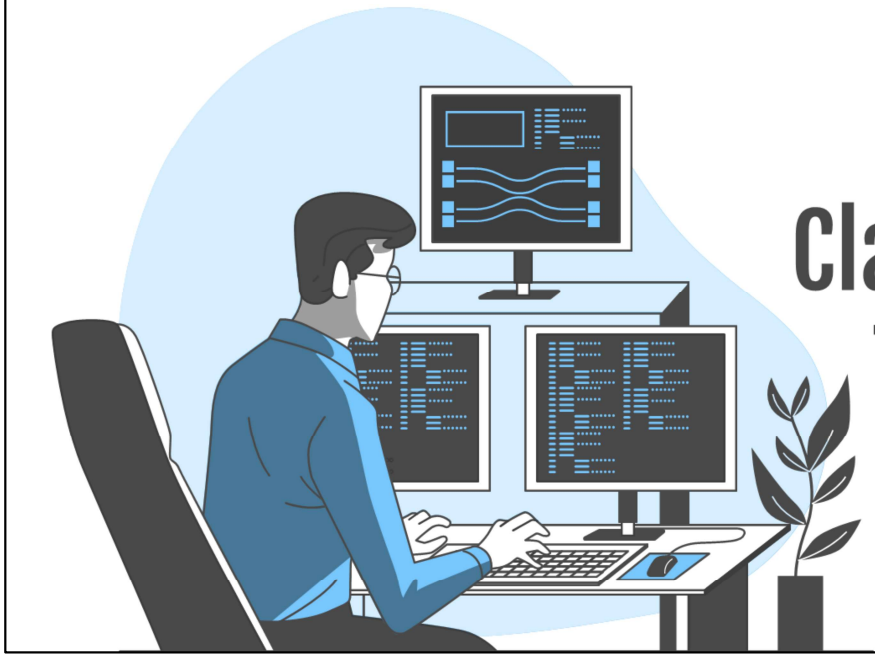


# Advanced Classification Techniques

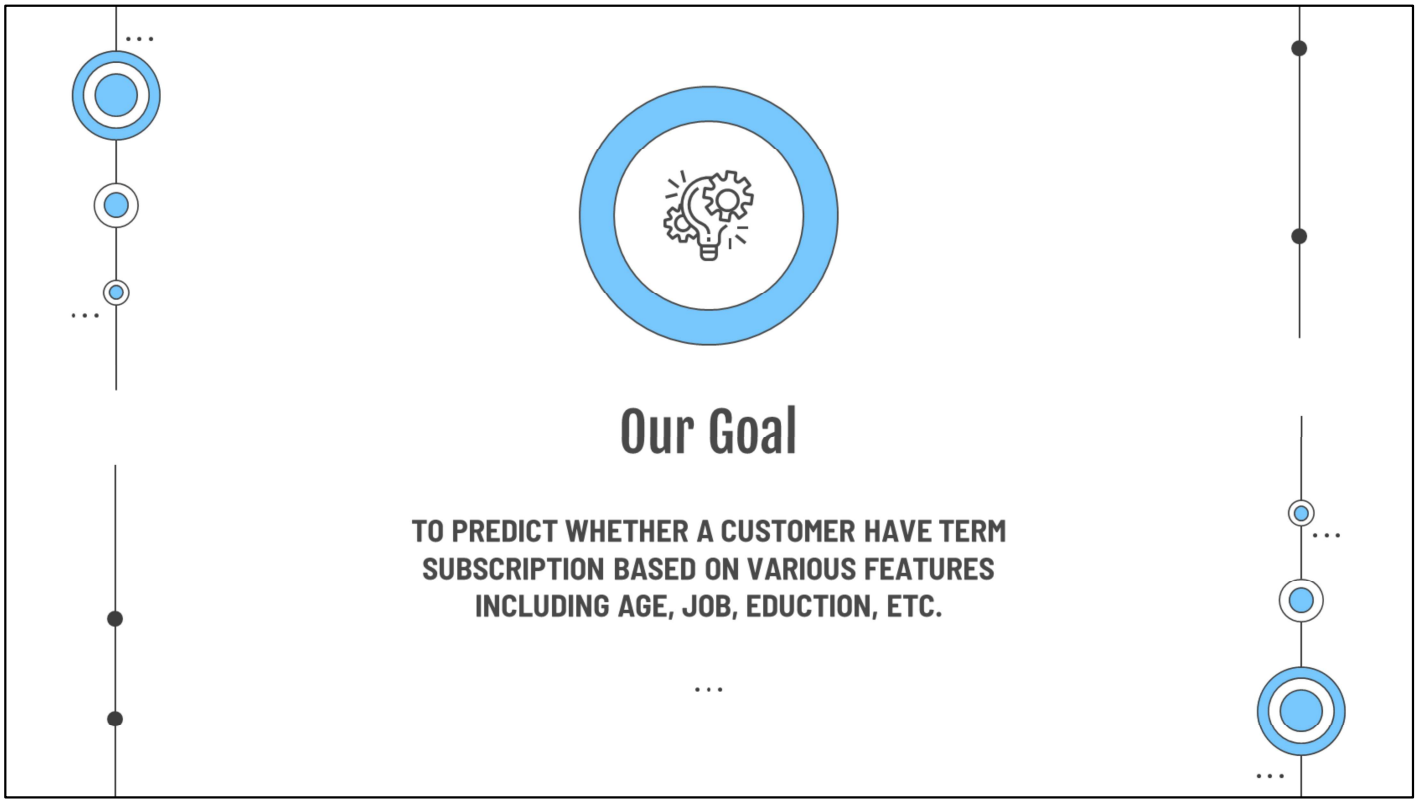
## Whether Customers Pay Term Subscription

Presented by: Yiwei Han
Will Yang

# Overview


- The dataset describes whether a customer subscribes a term deposit from a bank based on many features
- The data consists of 45211 observations
- The dataset has 17 attributes
- Binary Classification & Prediction Task
- Models will be assessed by metrics of confusion matrix

This is an overview of our dataset. The dataset consists of 45211 observations, and each observation has 7 attributes, with categorical, ordinal, binary and numeric. We will perform binary classification task, and the result will be assessed by confusion matrix, accuracy,precision, recall and f1 score.




Our Project utilizes a consumer profile dataset to predict whether a customer subscribes a term deposit from bank.

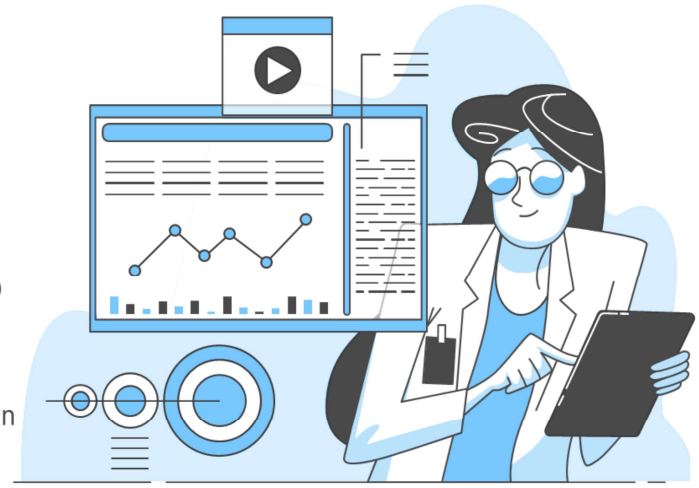
# Content Overview

01 **Exploratory Data Analysis**
Understand the Structure, Distribution of Data, and Feature Correlation

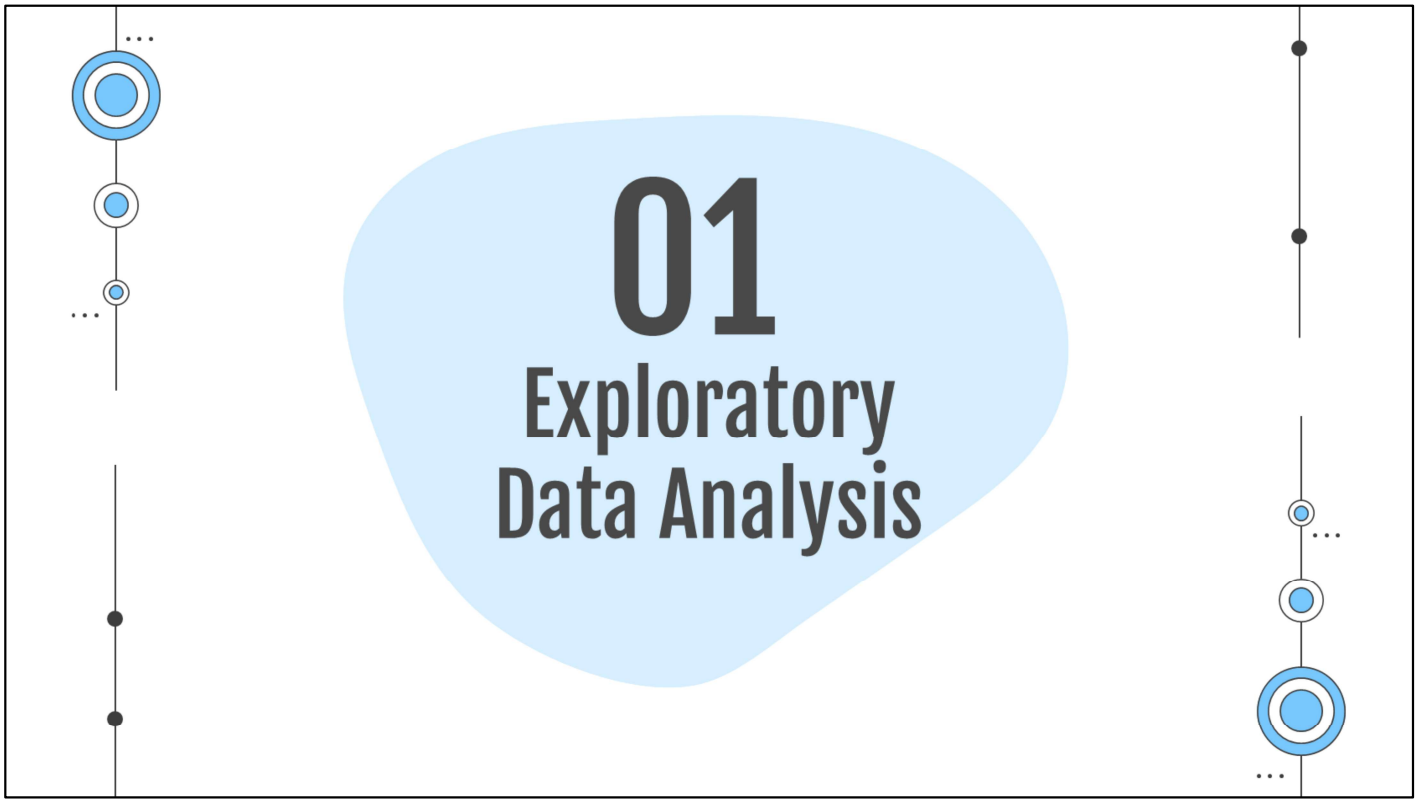
02 **Data Preprocessing**
Ordinal, Categorical and Binary Data were Encoded, Numerical Data were Normalized

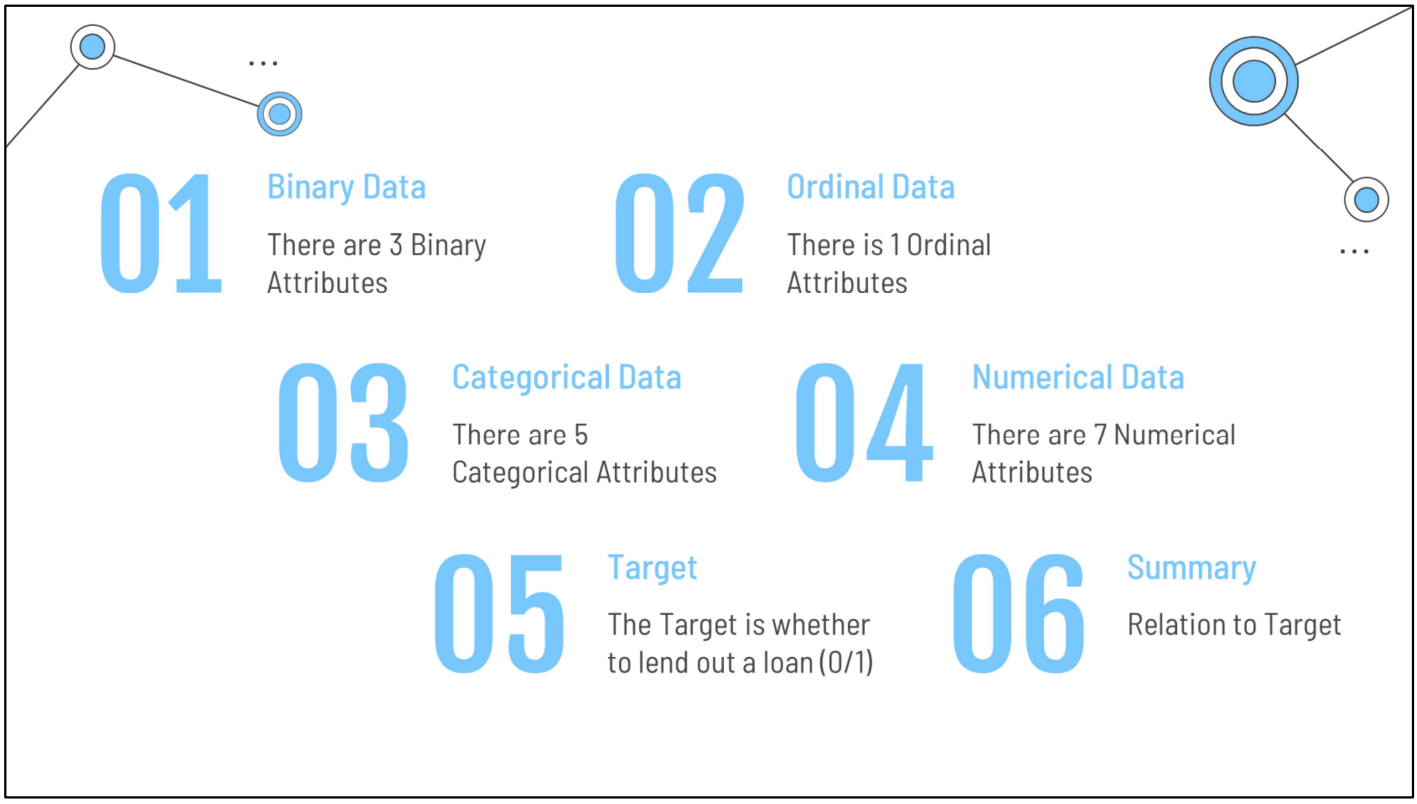
03 **Modeling**
Different Classification & Boost Techniques and Deep Learning

04 **Analysis**
Analysis of the Classification Result and Conclusion



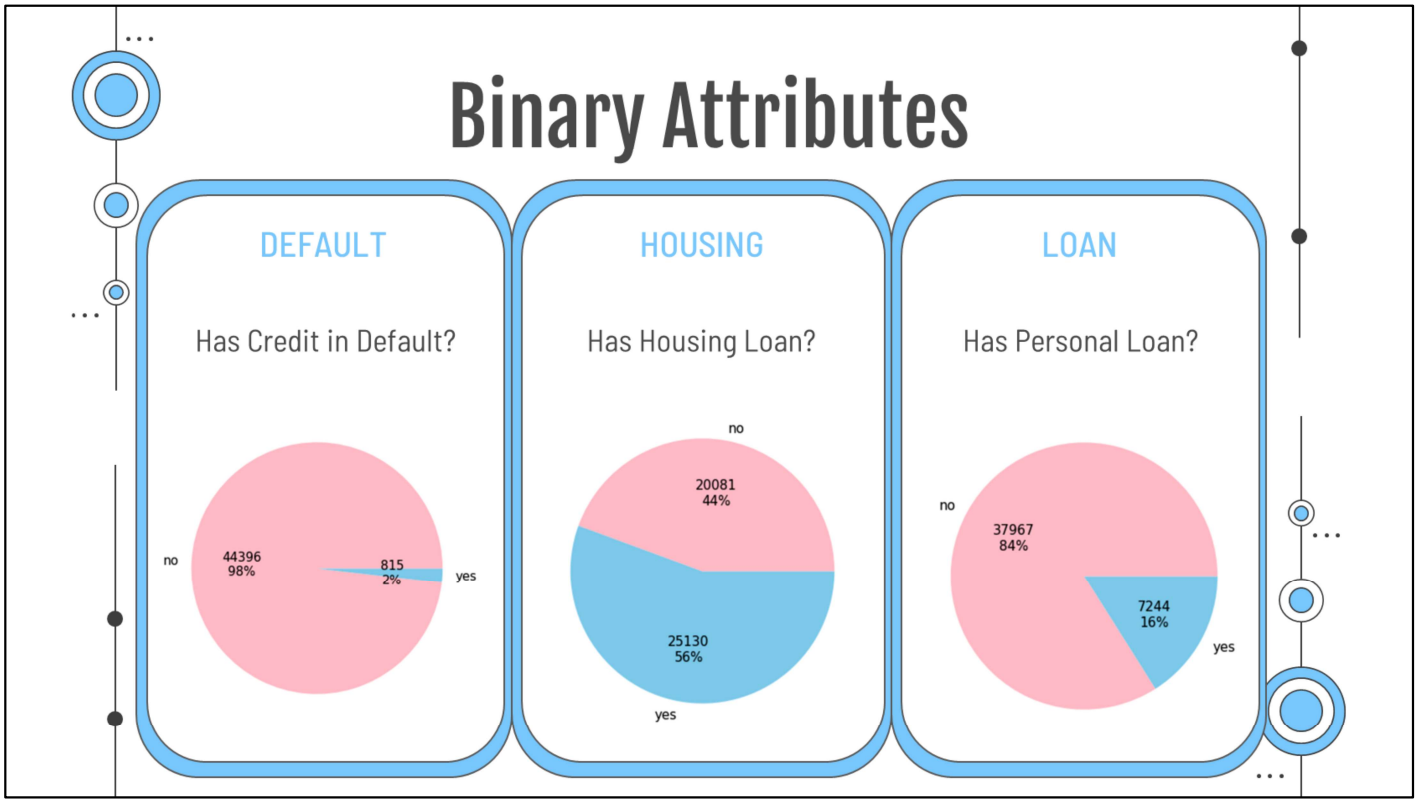
There are couple essential steps to our project: Exploratory Data Analysis, which helps us to know the dataset and features, Data Preprocessing, including feature selection and engineering, Modeling, and Analysis


01
Exploratory
Data Analysis

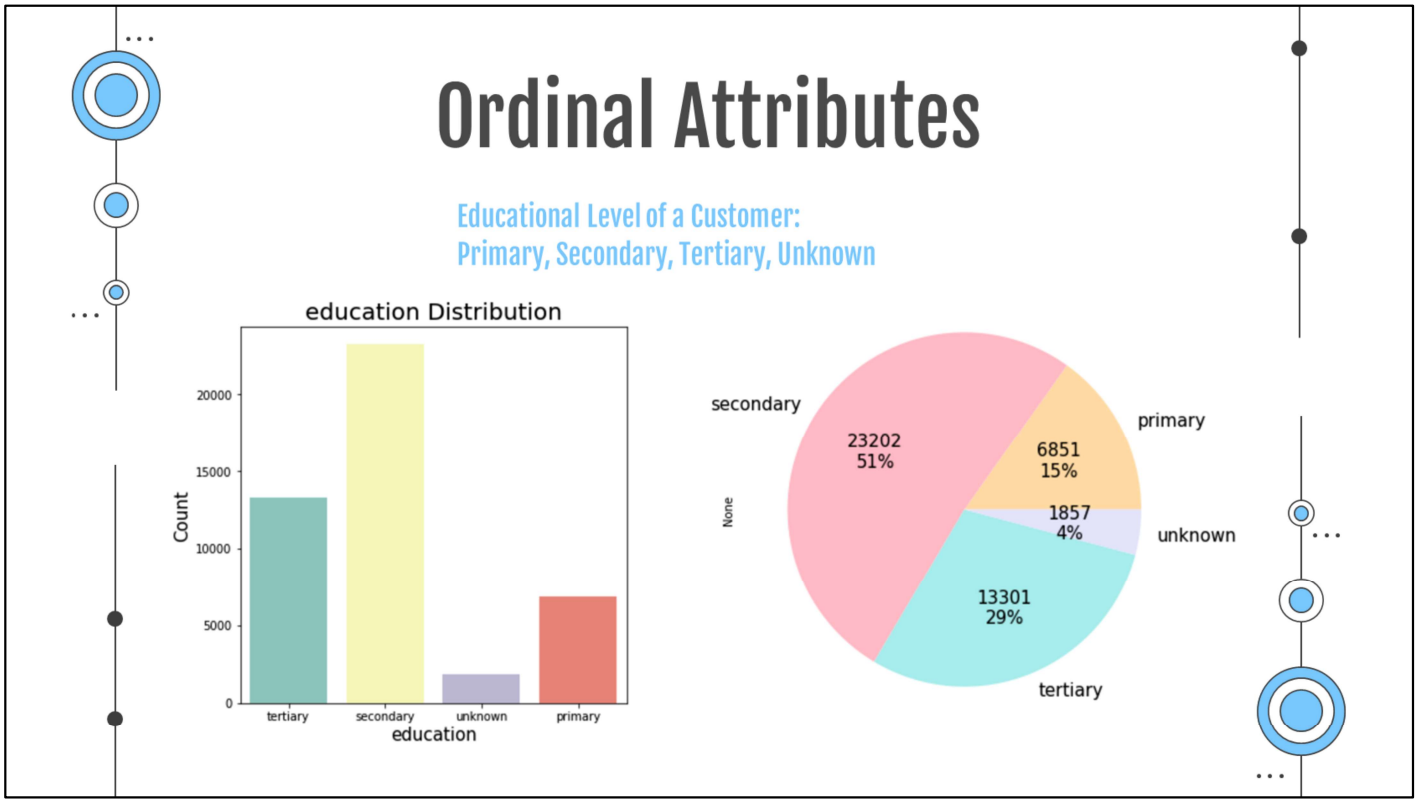



We encountered 4 different data types in dataset, and we will analysis each attribute's association with our target in the following slides

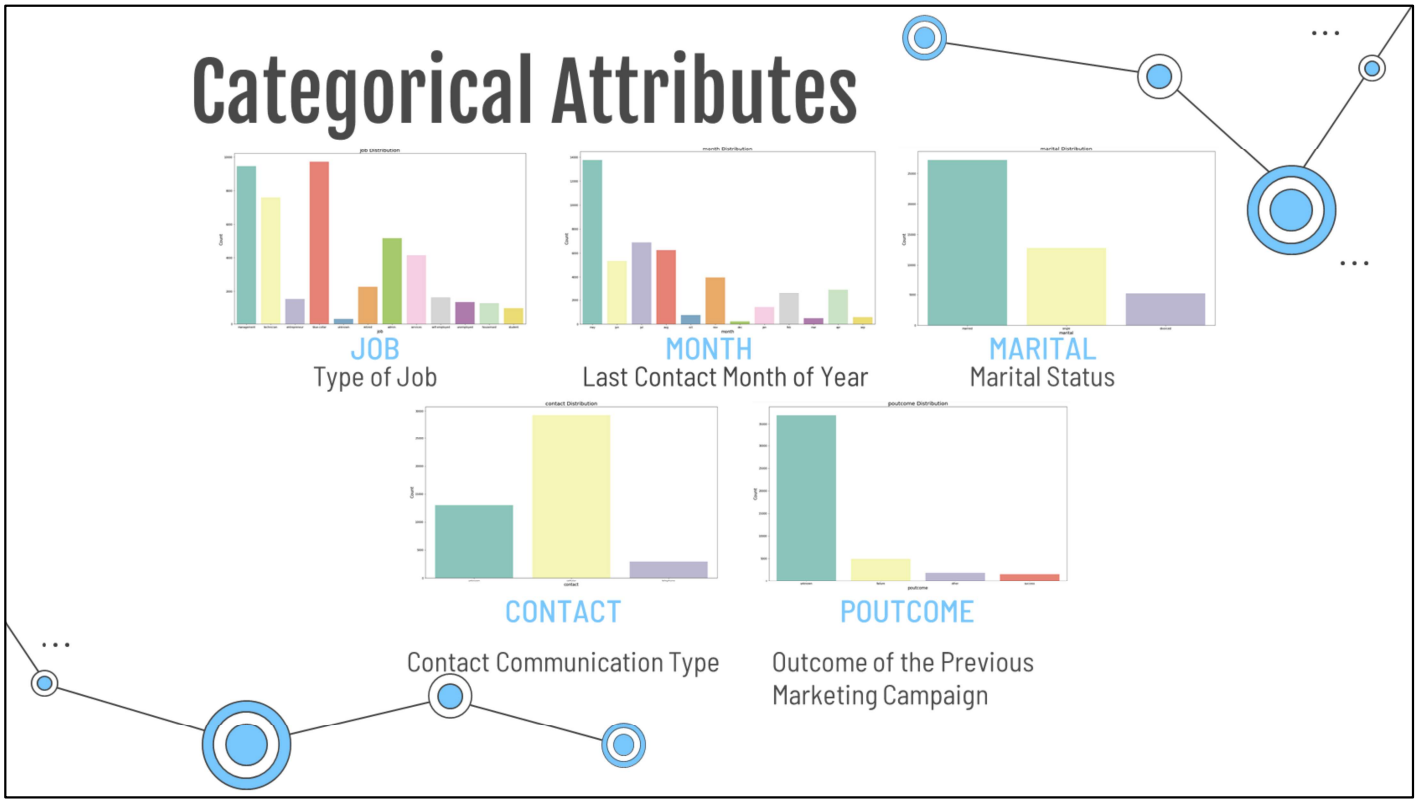

Fortunately, we don’t have missing data in column




Binary Attributes: A question Whether you have or not. Have is True; Don’t Have is False.



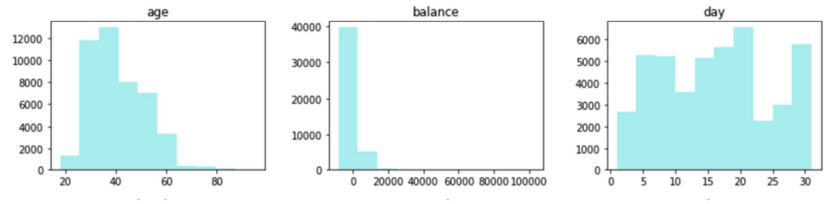

Ordinal Attributes: Education level, we denoted it with Primary, Secondary, Tertiary, or Unknown

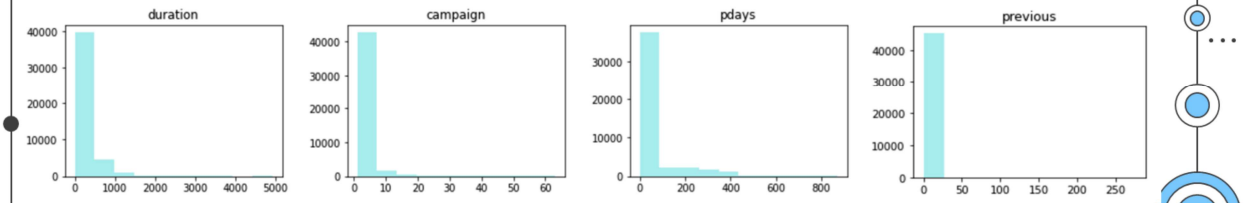



Categorical Attributes: Job type, last contact month of year, Marital info, Contact method, and Past-outcome

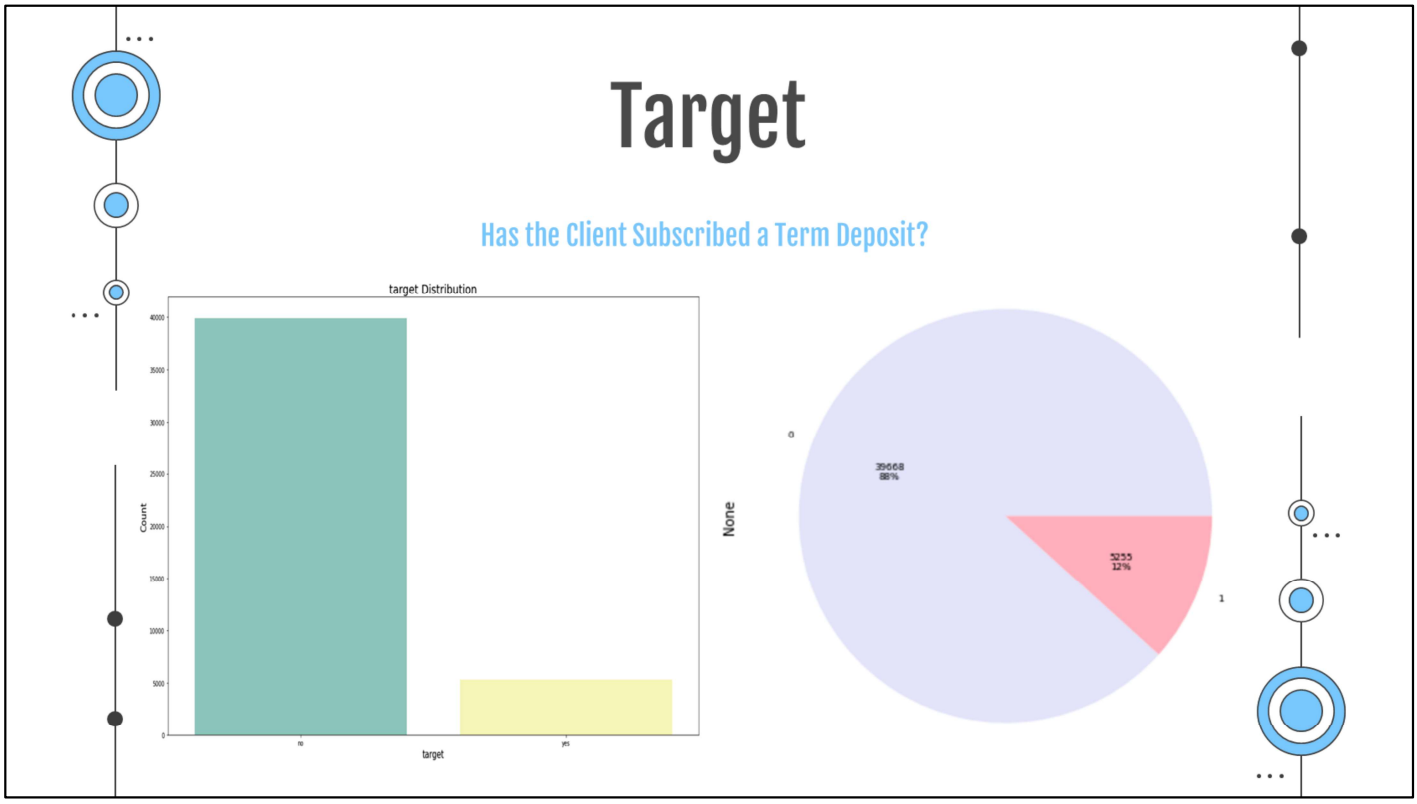
# Numerical Attributes

## — Bar Chart

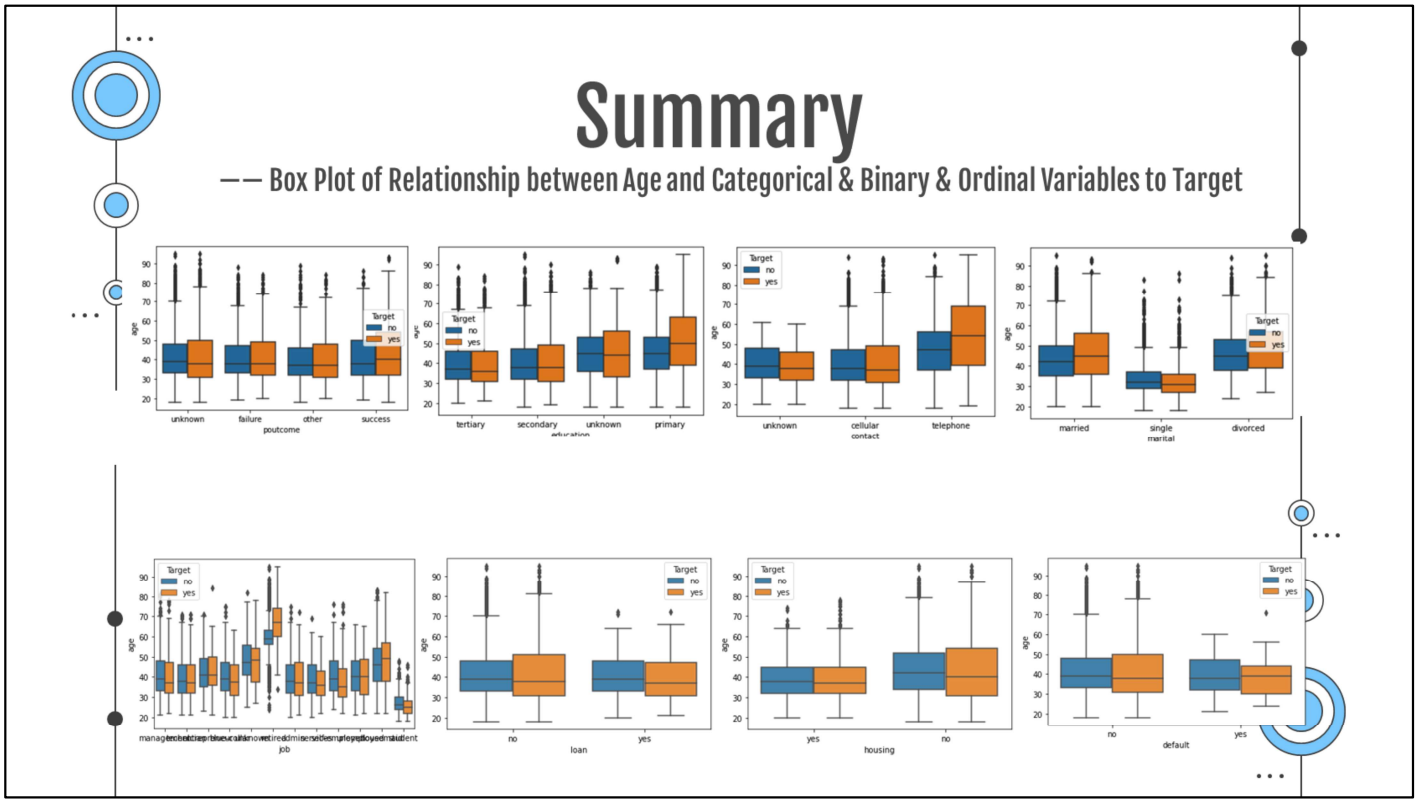


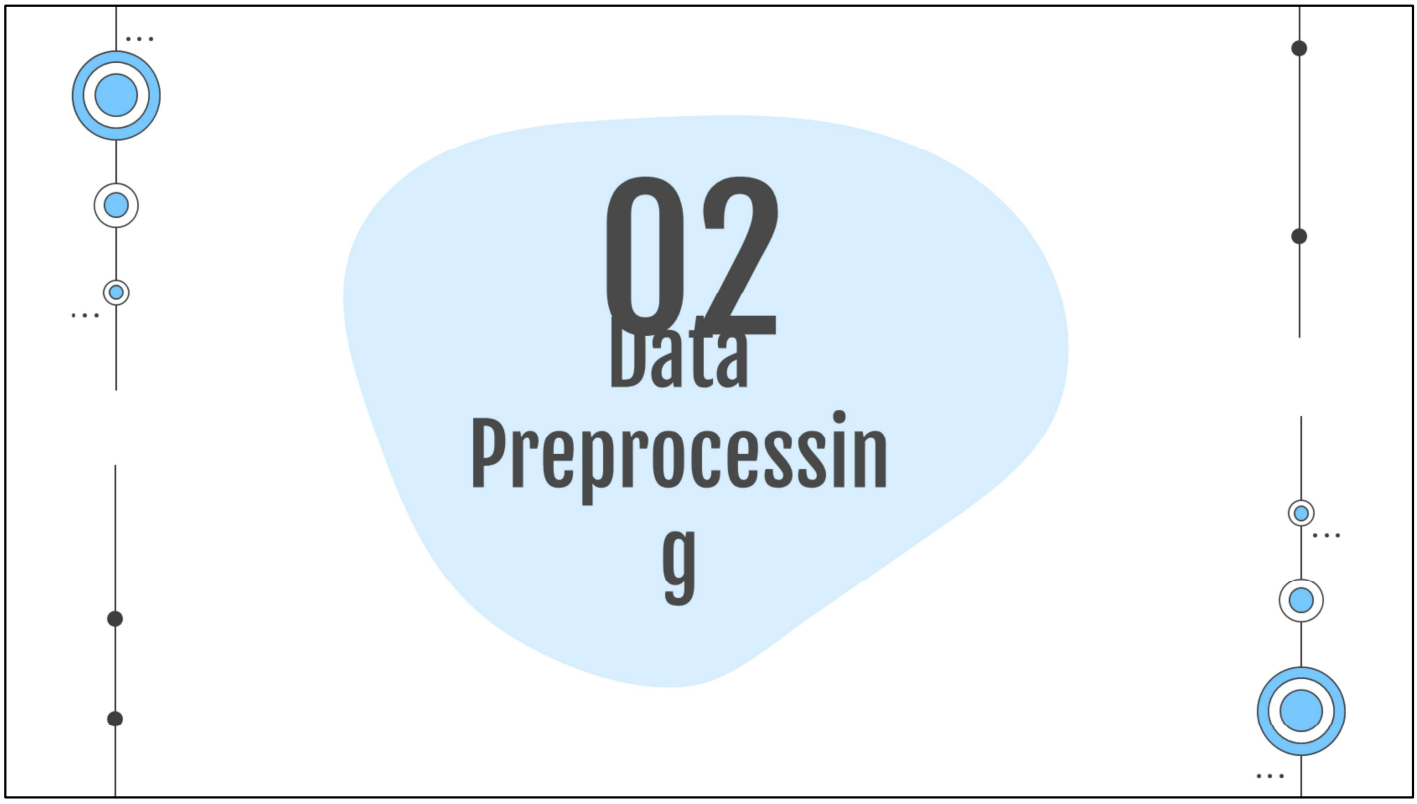

There are the Numerical Attributes: Ages, balance, days…etc

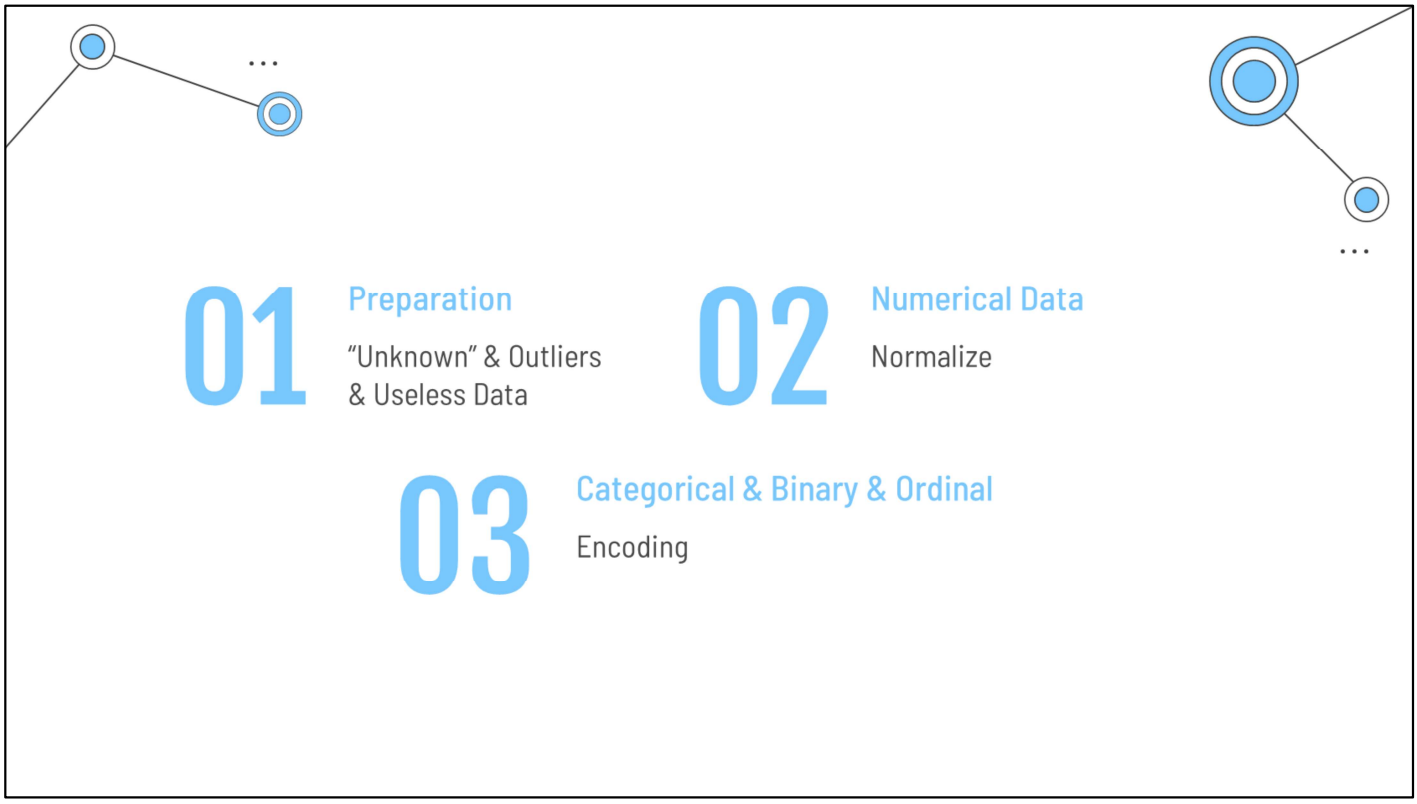


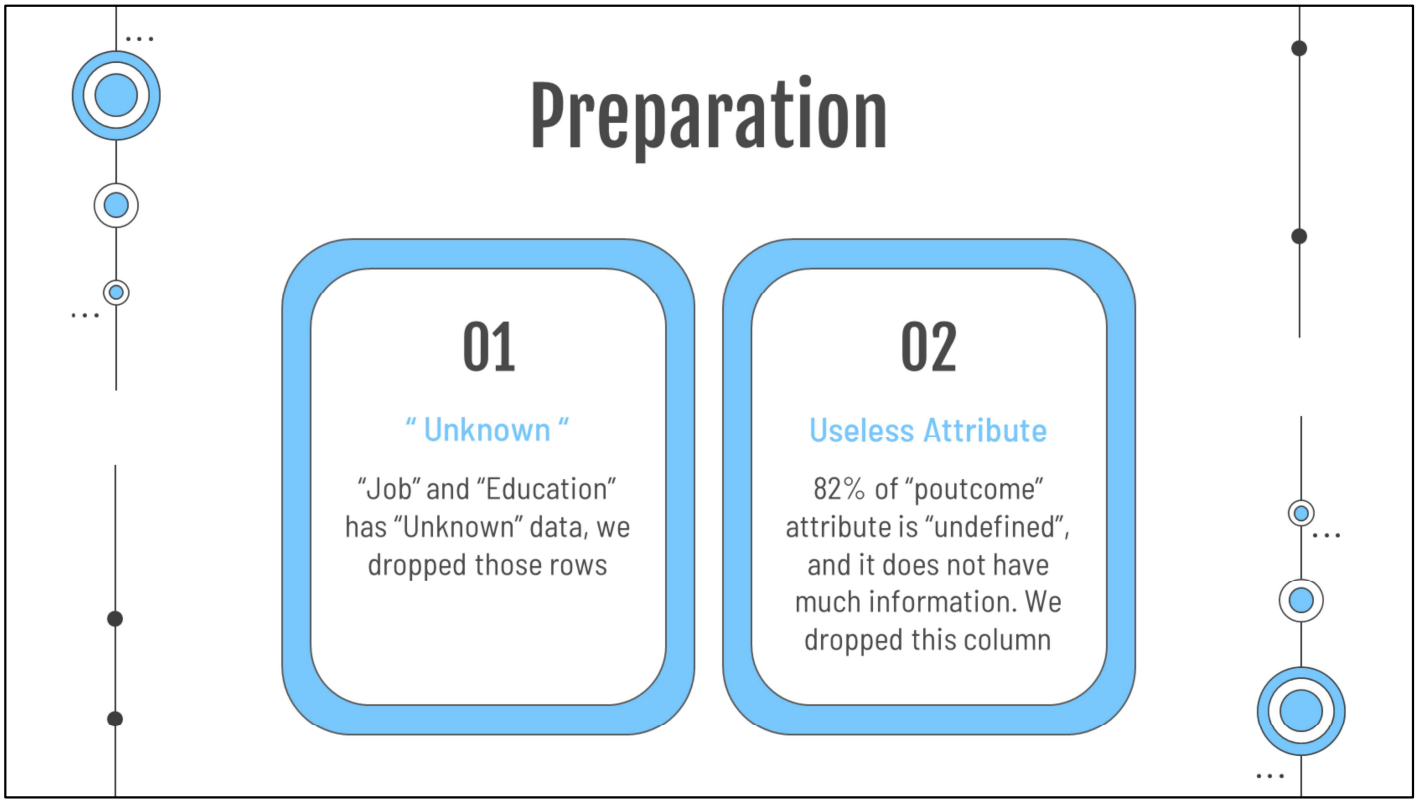

Target parameter data type is binary, which indicates whether a person has subscribed the term deposit




Associations of attribute Target parameter, they are displayed in the box plot, and with each categorical variables' relationship to age, and separated by target, which whether they have term subscription


02
Data
Preprocessin
g

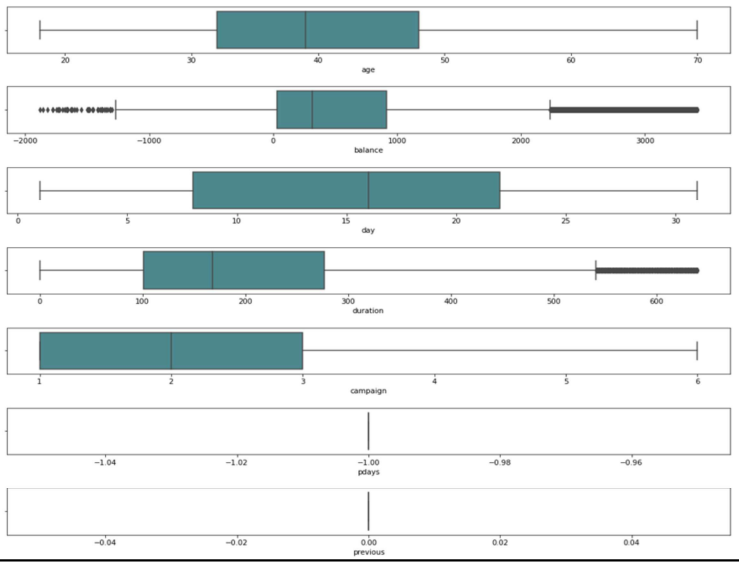

01
Preparation
"Unknown" & Outliers
& Useless Data
02
Numerical Data
Normalize
03
Categorical & Binary & Ordinal
Encoding




The categorical data job and education has some unknown entries, but they are not much, so we drop the rows with those two has unknown entries. Because the attribute poutcome has 82% of unknown, we think it is not very useful for our classification, so we dropped this attribute.

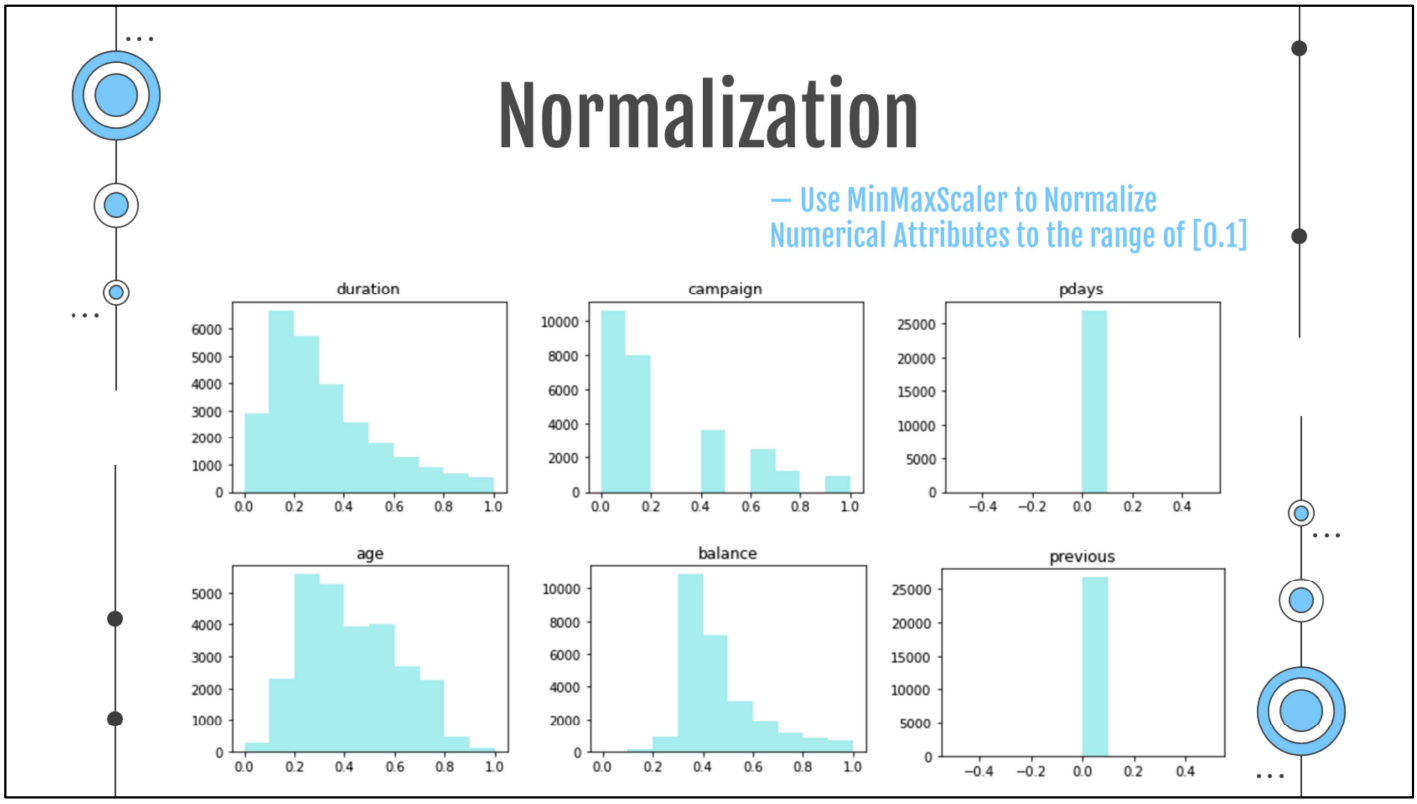
# Preparation

## — Outlier Analysis

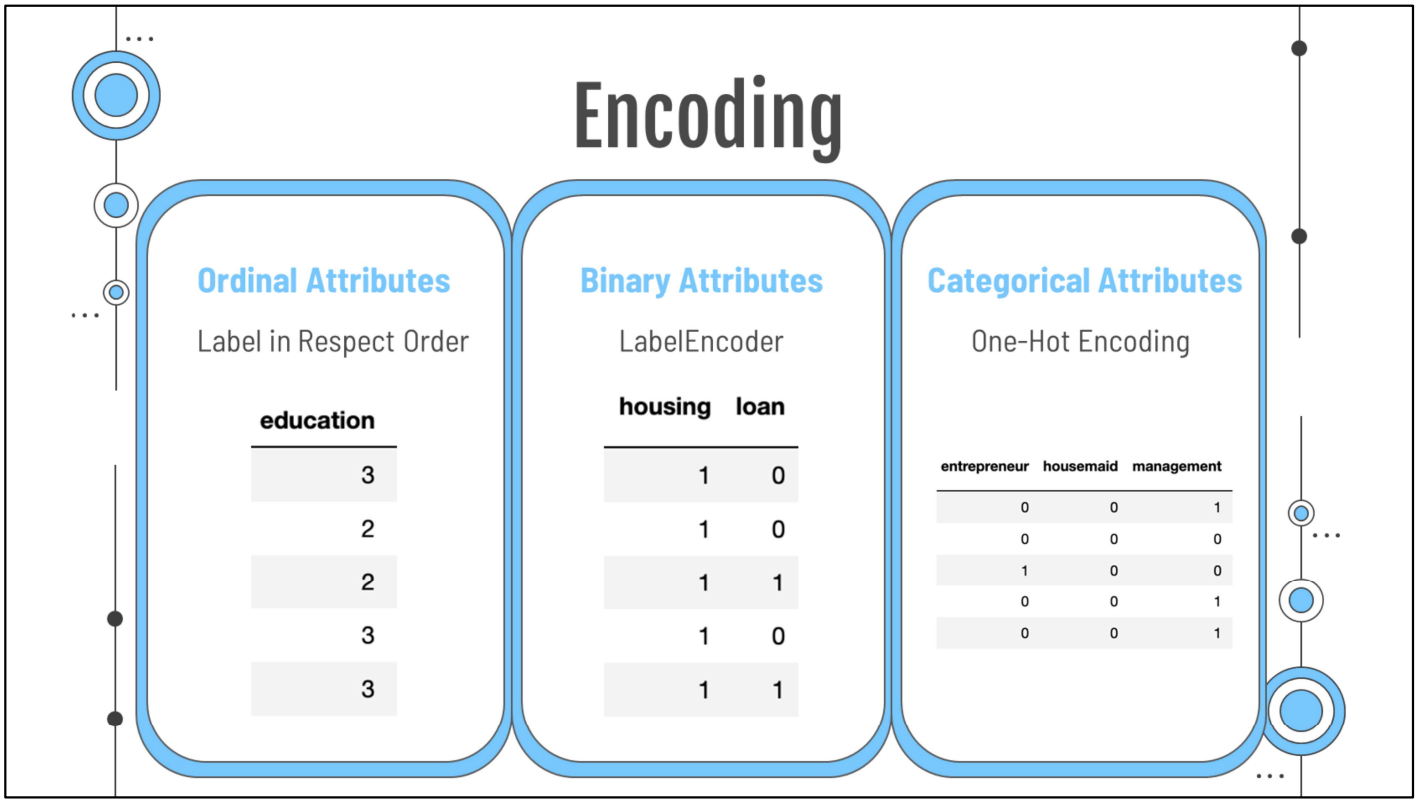



Remove all numerical values with value lower than 1.5*Quartile 1 or higher than 1.5 * Quartile 3

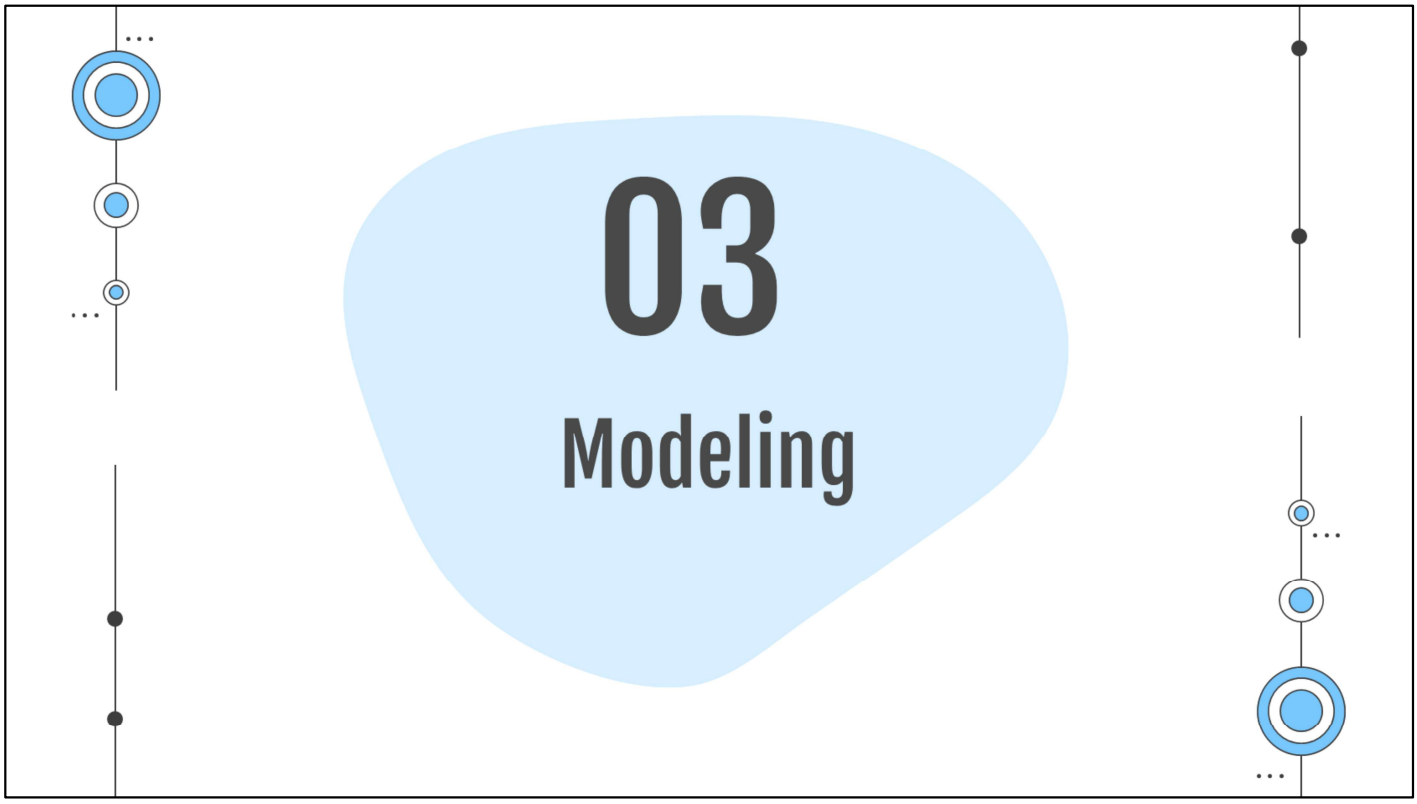
Numerical data has some outliers, which affects the accuracy of our prediction, so we remove outliers in statistical sense, the box plot after removing the outliers are shown.

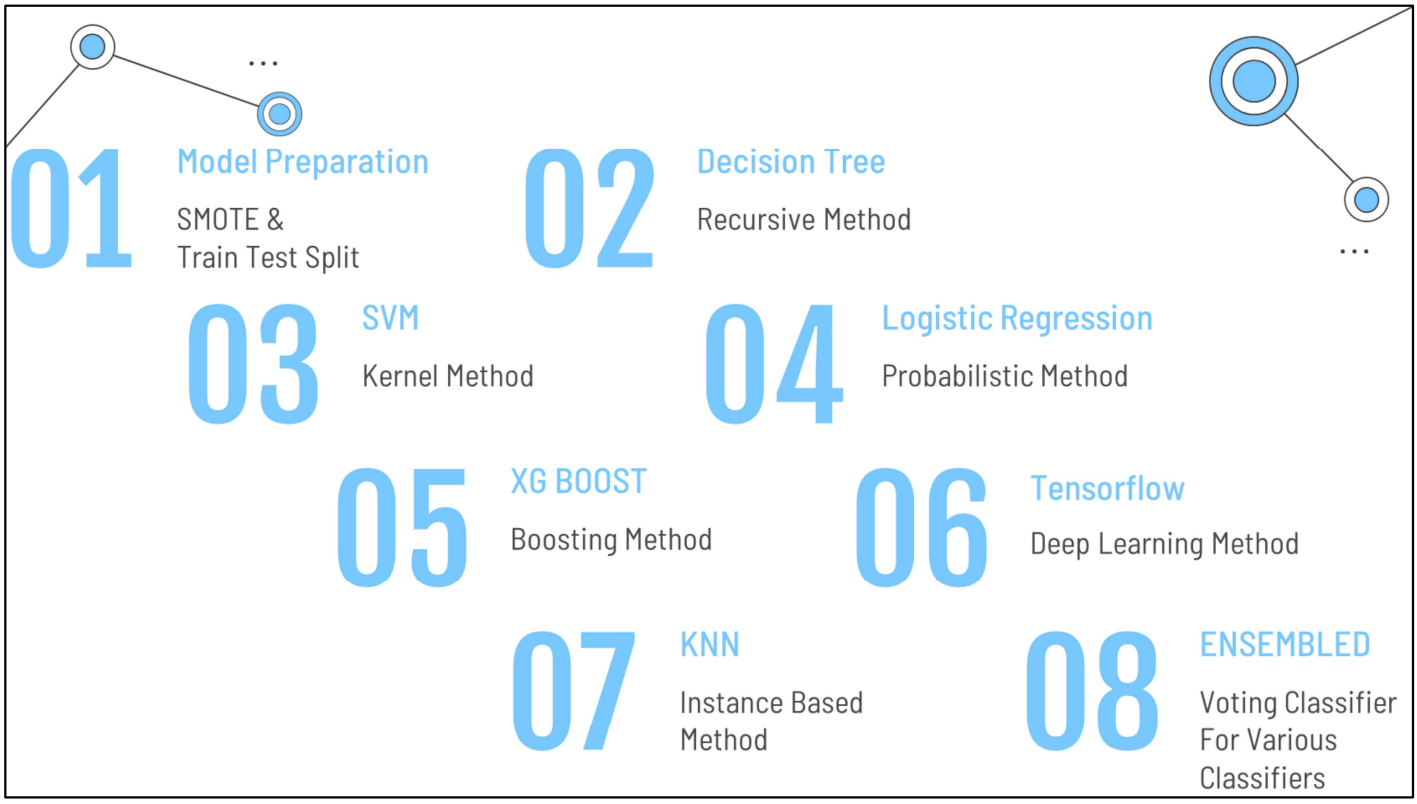


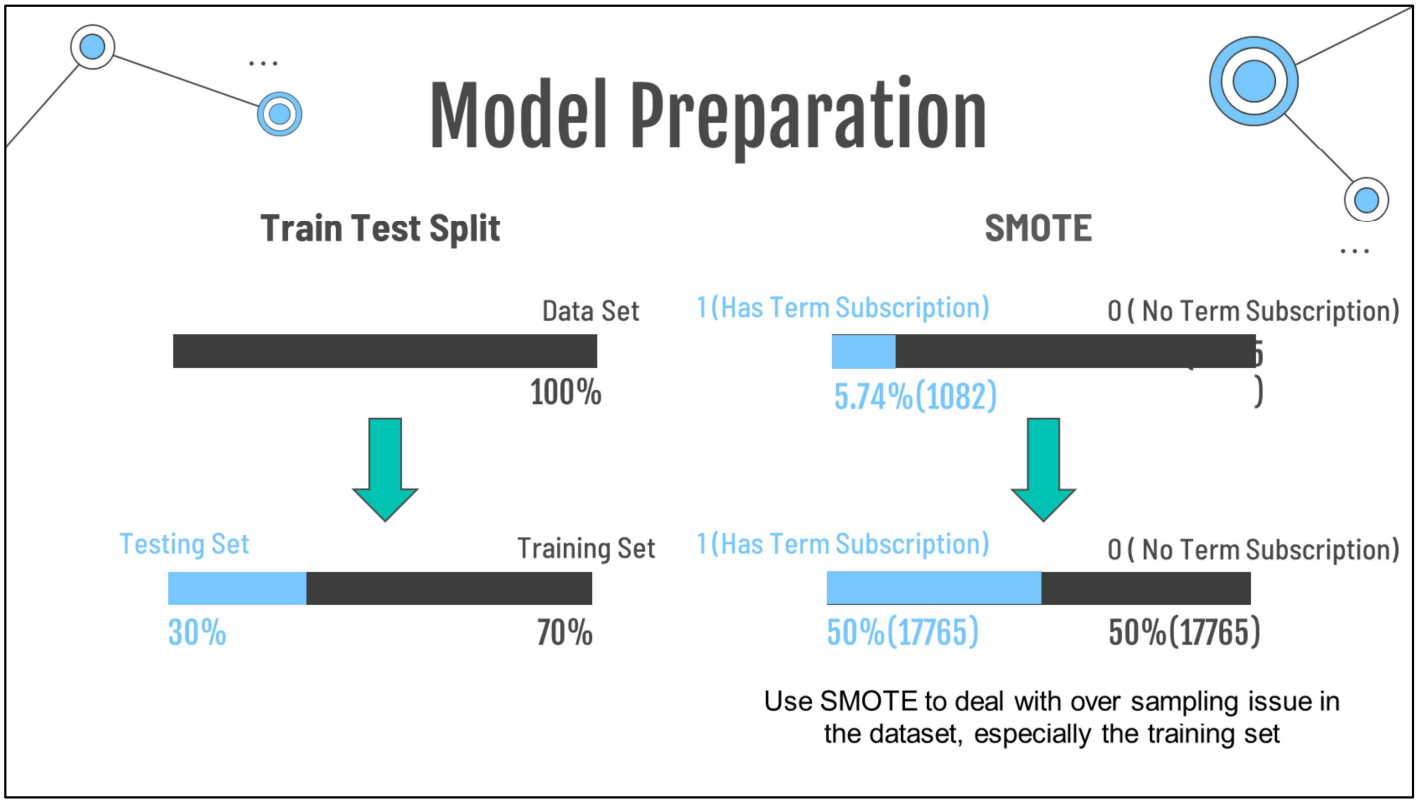

We Normalize the value range of each attribute into (0 to 1) using the minmaxscaler



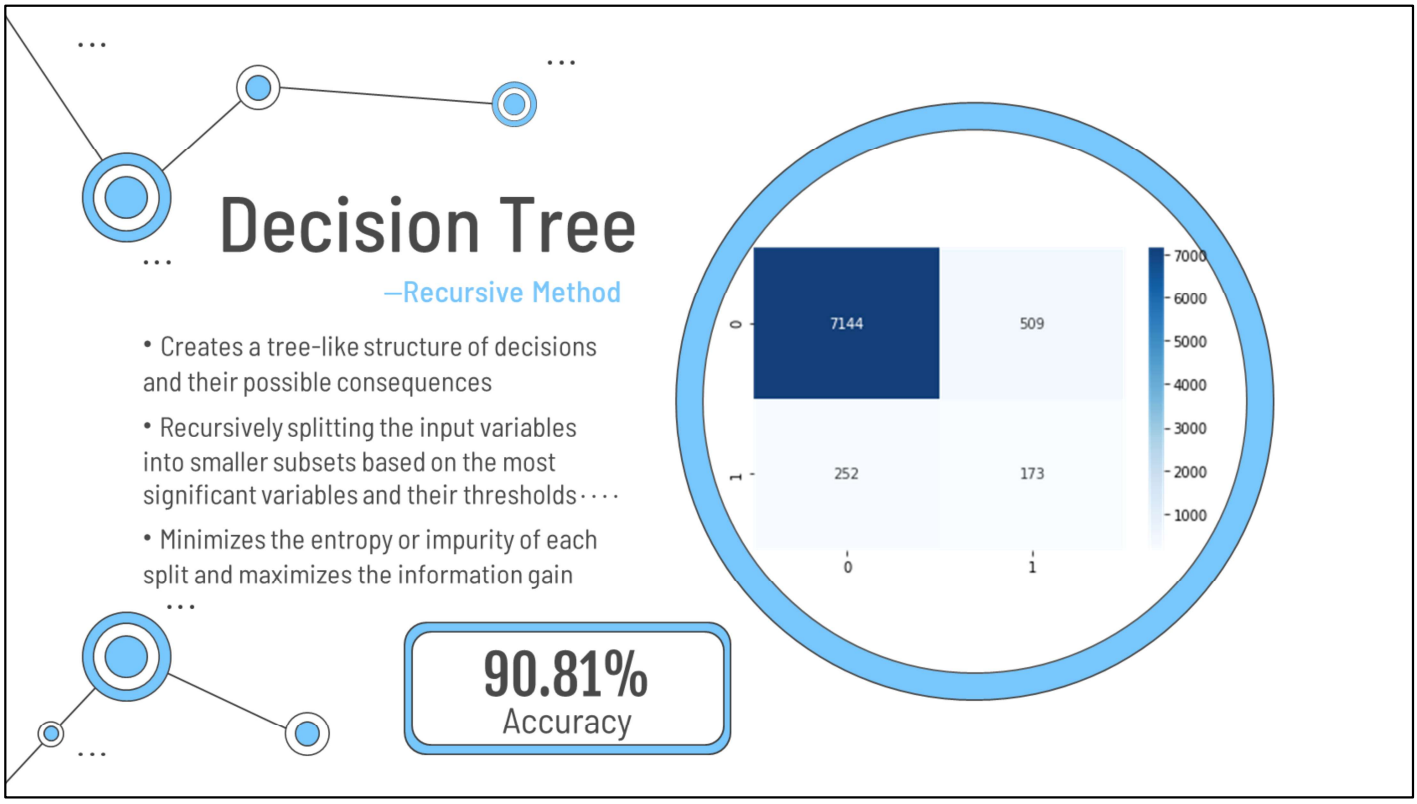

Example of Specific treatment for different data types. We used label encoder for binary attributes, which 1 represents have and 0 represent don't have. We used one hot encoding for categorical variables. We used specific order of ordinal attributes, where larger the number represents higher the eduction.


03
Modeling

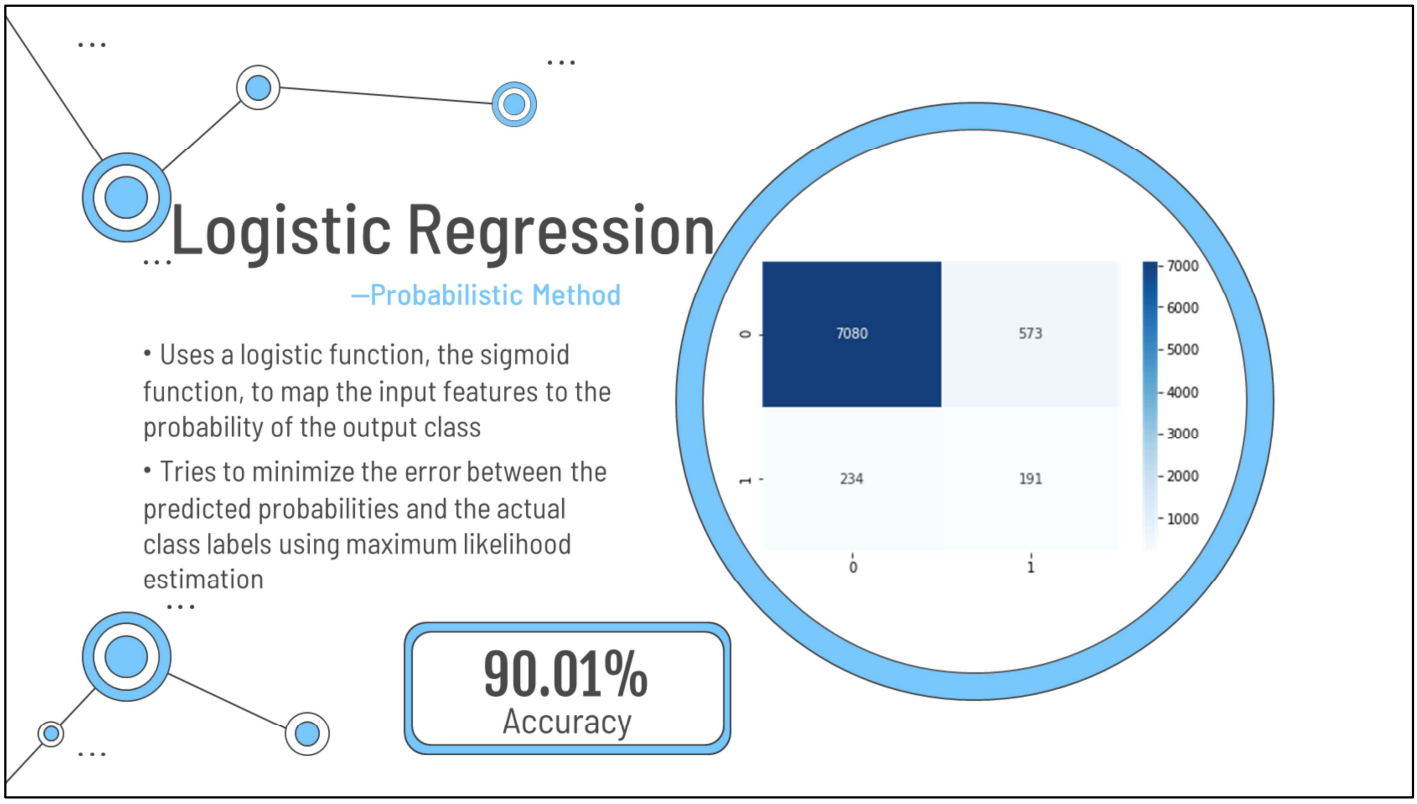

01
Model Preparation
SMOTE &
Train Test Split
02
Decision Tree
Recursive Method
03
SVM
Kernel Method
04
Logistic Regression
Probabilistic Method
05
XG BOOST
Boosting Method
06
Tensorflow
Deep Learning Method
07
KNN
Instance Based
Method
08
ENSEMBLED
Voting Classifier
For Various
Classifiers



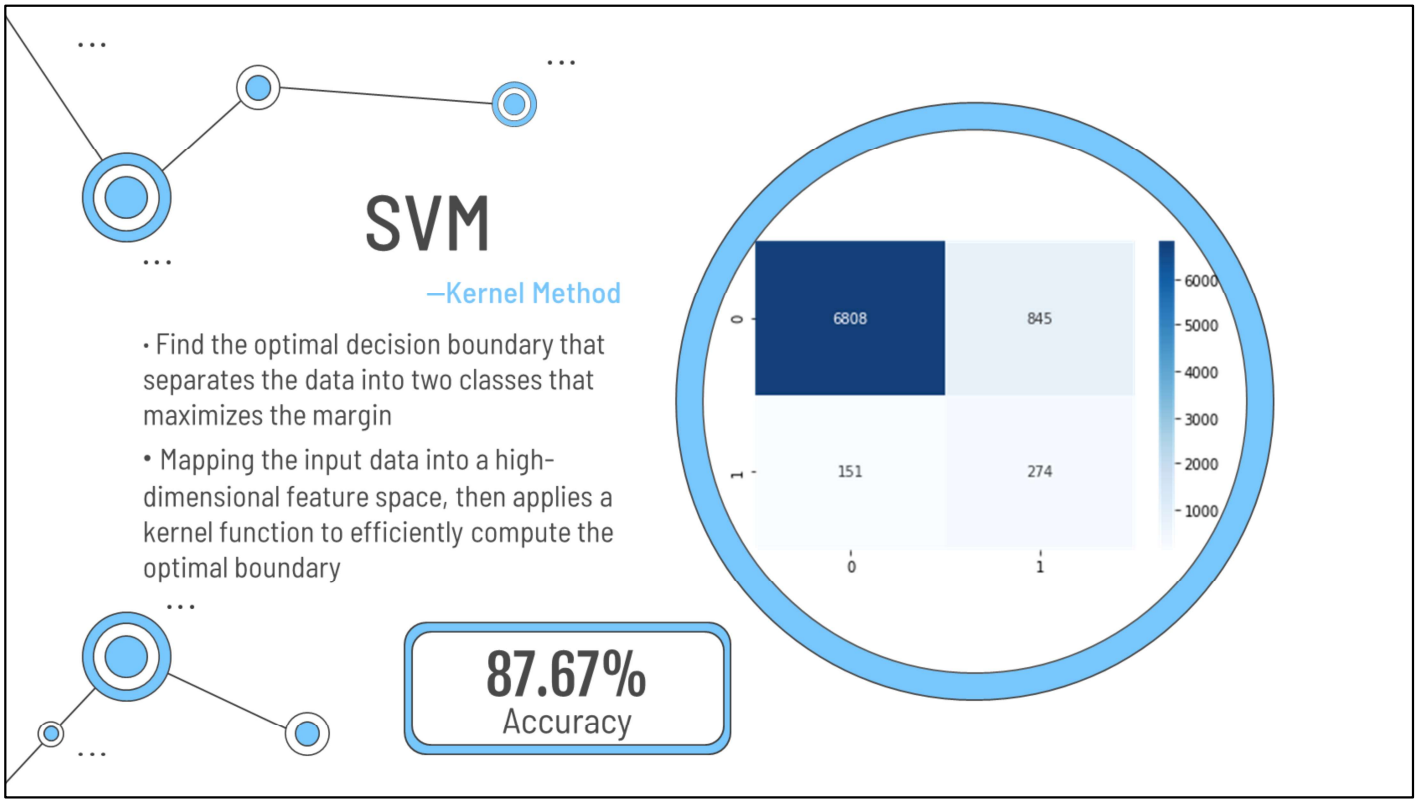

First, we perform train test split the data to 70% and 30% portion. Because our data has oversampling issue, which the target has too much 0, which is people without term subscription than people who has term subscription. As a result, we use SMOTE model, which works by creating synthetic samples of the minority class by interpolating between neighboring examples. Specifically, SMOTE selects a minority class sample and finds its k nearest neighbors in the feature space. It then creates new synthetic samples by randomly selecting one of the k neighbors and creating a linear combination of the two samples.



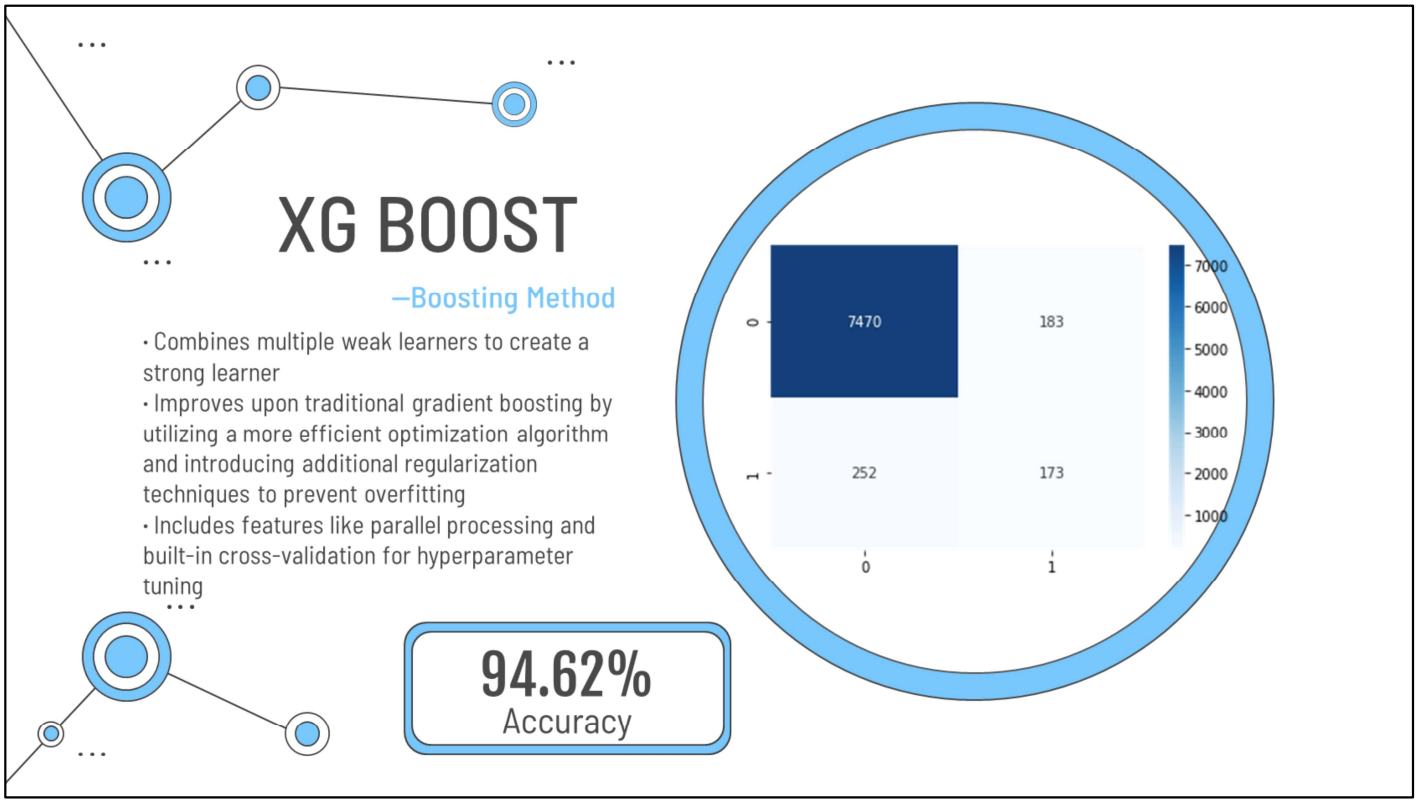

Decision Tree is our baseline model. Decision trees are easy to interpret and visualize, making them a popular choice for explaining the decision-making process to non-technical stakeholders.




Logistic regression is another baseline model. Logistic regression is robust to noise and can handle datasets with irrelevant or correlated predictor variables. It can also handle missing data by using the available observations in the model.Logistic regression is computationally efficient and can be trained on large datasets with many predictor variables. It is also less prone to overfitting compared to more complex models.




SVMs are a margin-based algorithm, which means that they aim to maximize the margin between the support vectors and the hyperplane. This can help to reduce the risk of overfitting and improve the generalization performance of the model. SVMs have a regularization parameter that helps to control the complexity of the model and prevent overfitting, which can be tuned to find the optimal balance between bias and variance in the model.

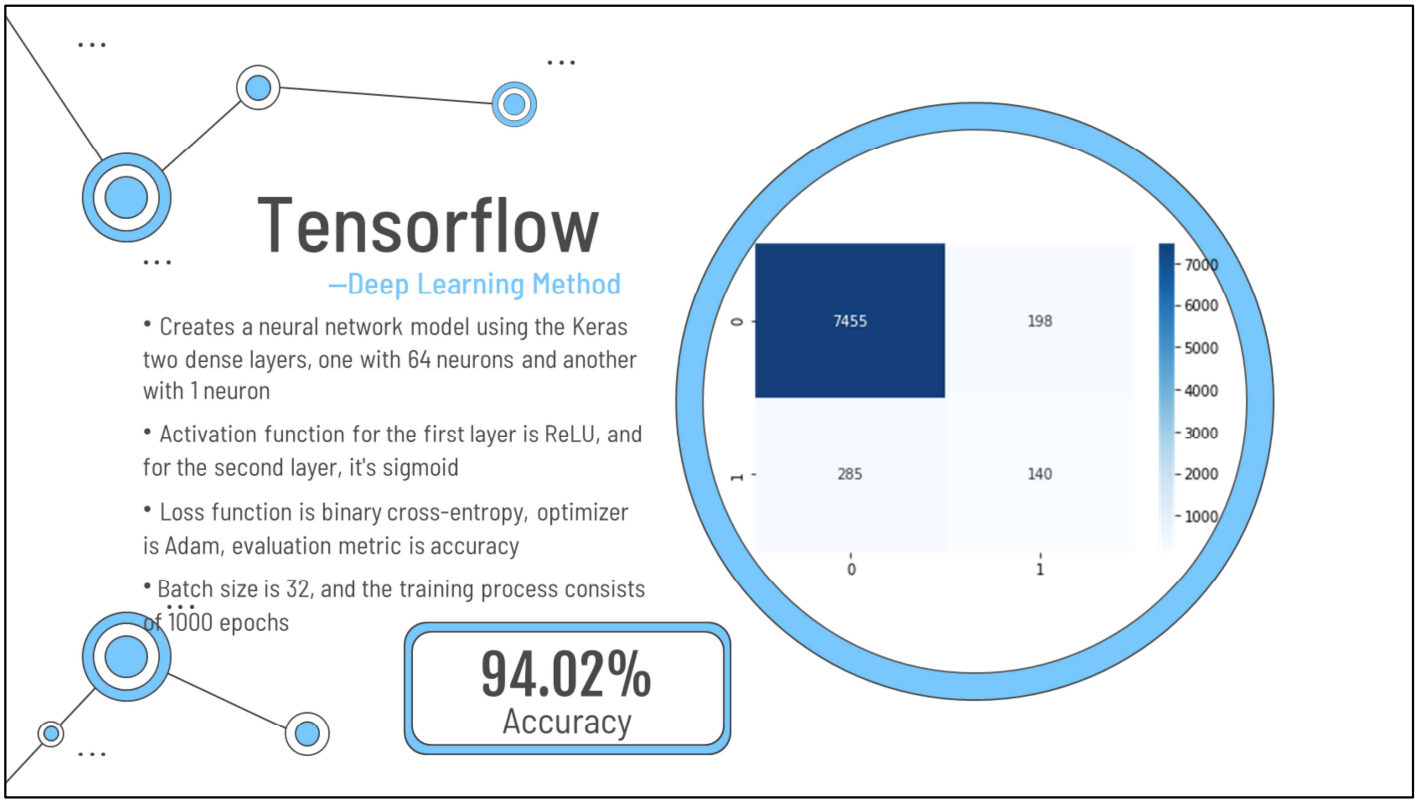



XGBoost is known to provide highly accurate predictions. It is also less prone to overfitting and can handle missing values and outliers.

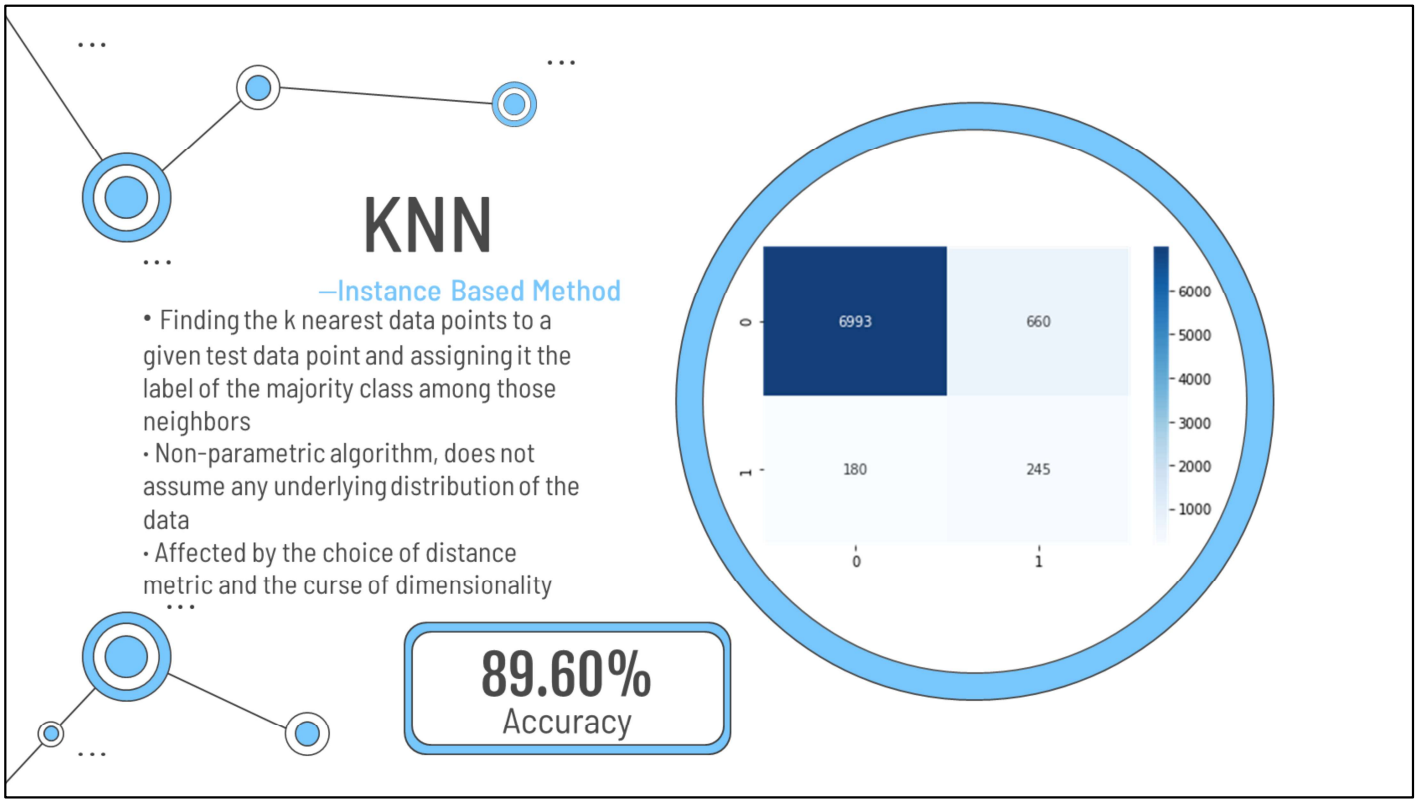
XGBoost includes regularization techniques such as L1 and L2 regularization, which can help prevent overfitting and improve the generalization performance of the model.

XGBoost provides a measure of feature importance, which can be used to identify the most important features in the dataset. This can help to reduce the dimensionality of the dataset and improve the accuracy of the model.

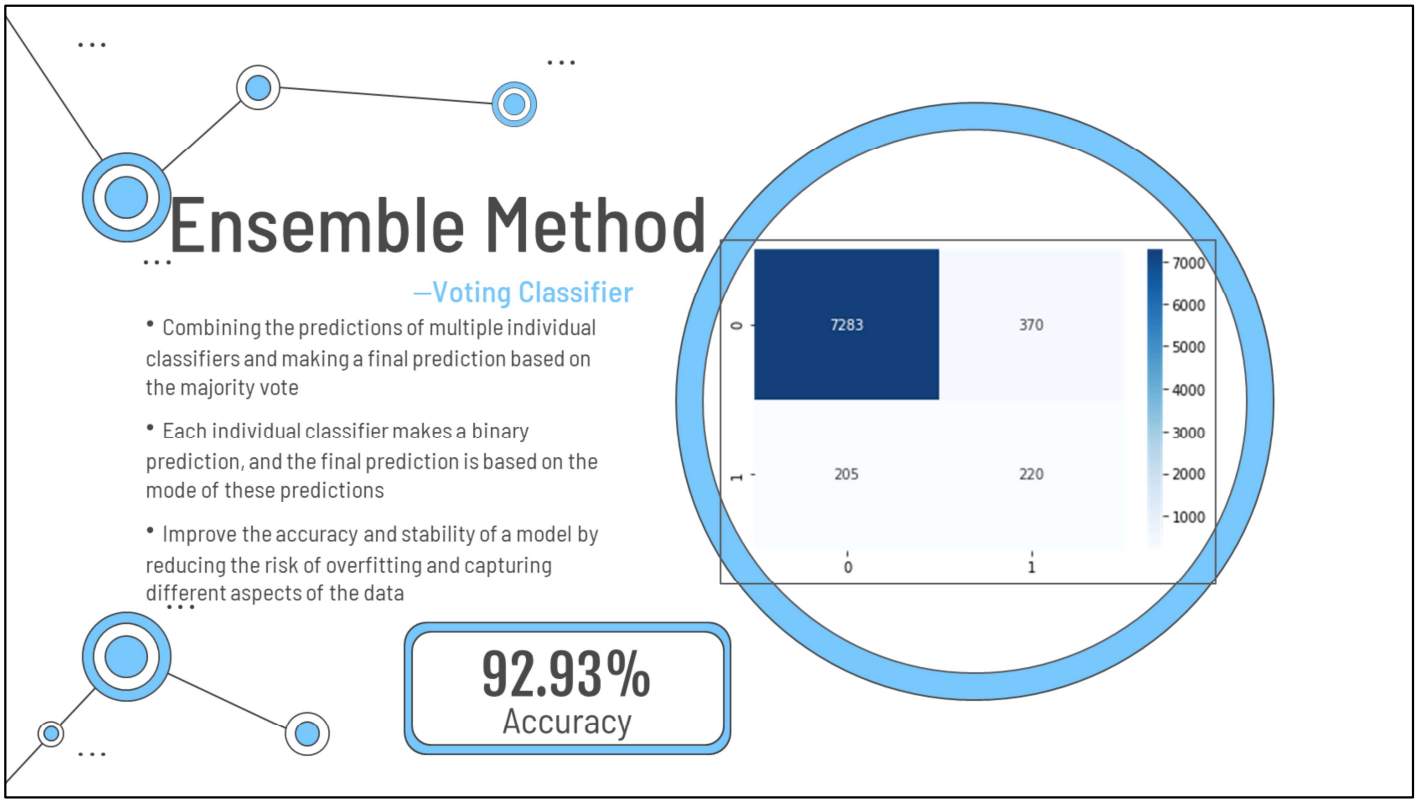
XGBoost includes techniques such as weighted and adaptive boosting, which can help to handle imbalanced datasets and improve the accuracy of the model for the minority class.




TensorFlow is optimized for high-performance computing and can leverage the power of GPUs and TPUs to accelerate the training process. This can significantly reduce the training time for large and complex models.



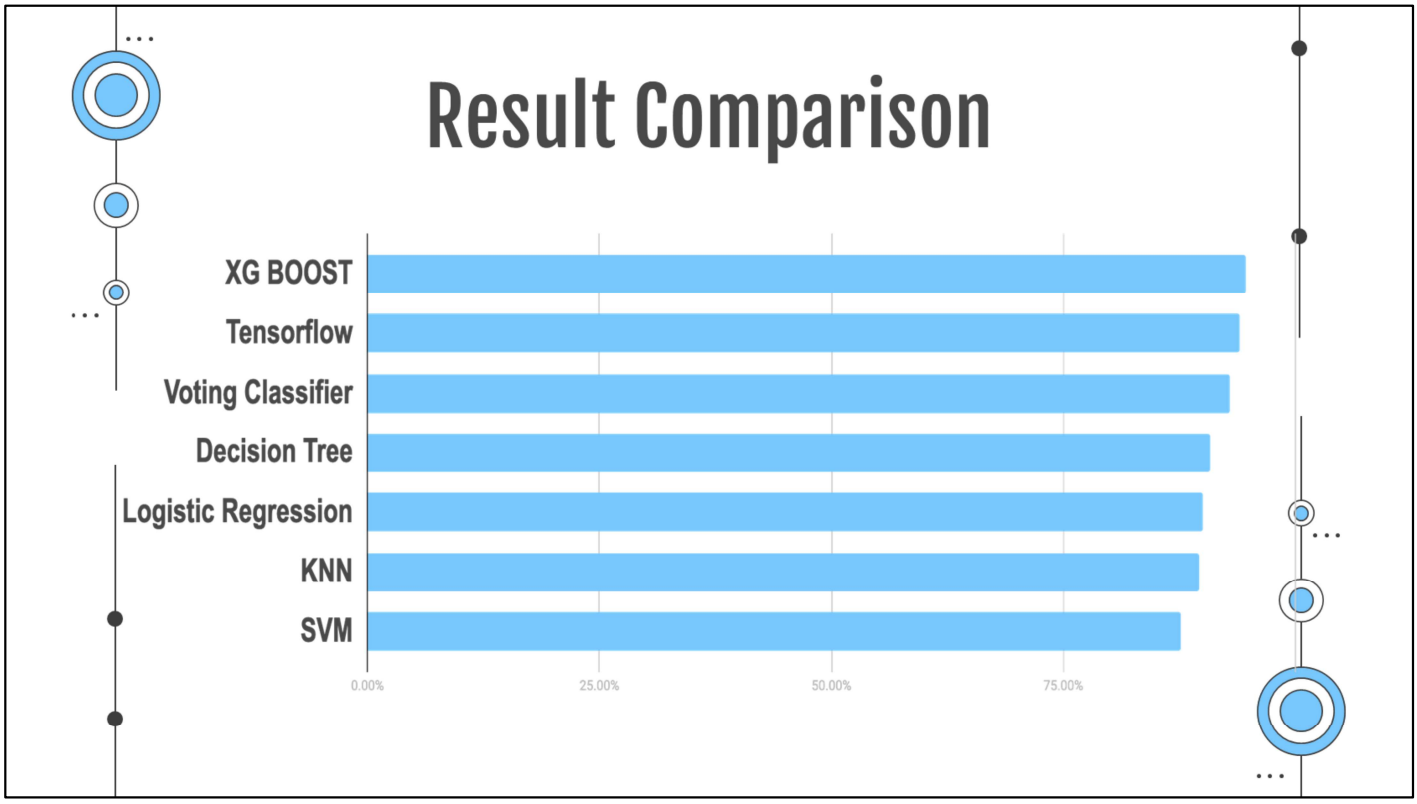

KNN does not assume anything about the distribution of the data, making it suitable for datasets with complex and nonlinear relationships between the predictors and the outcome variable.



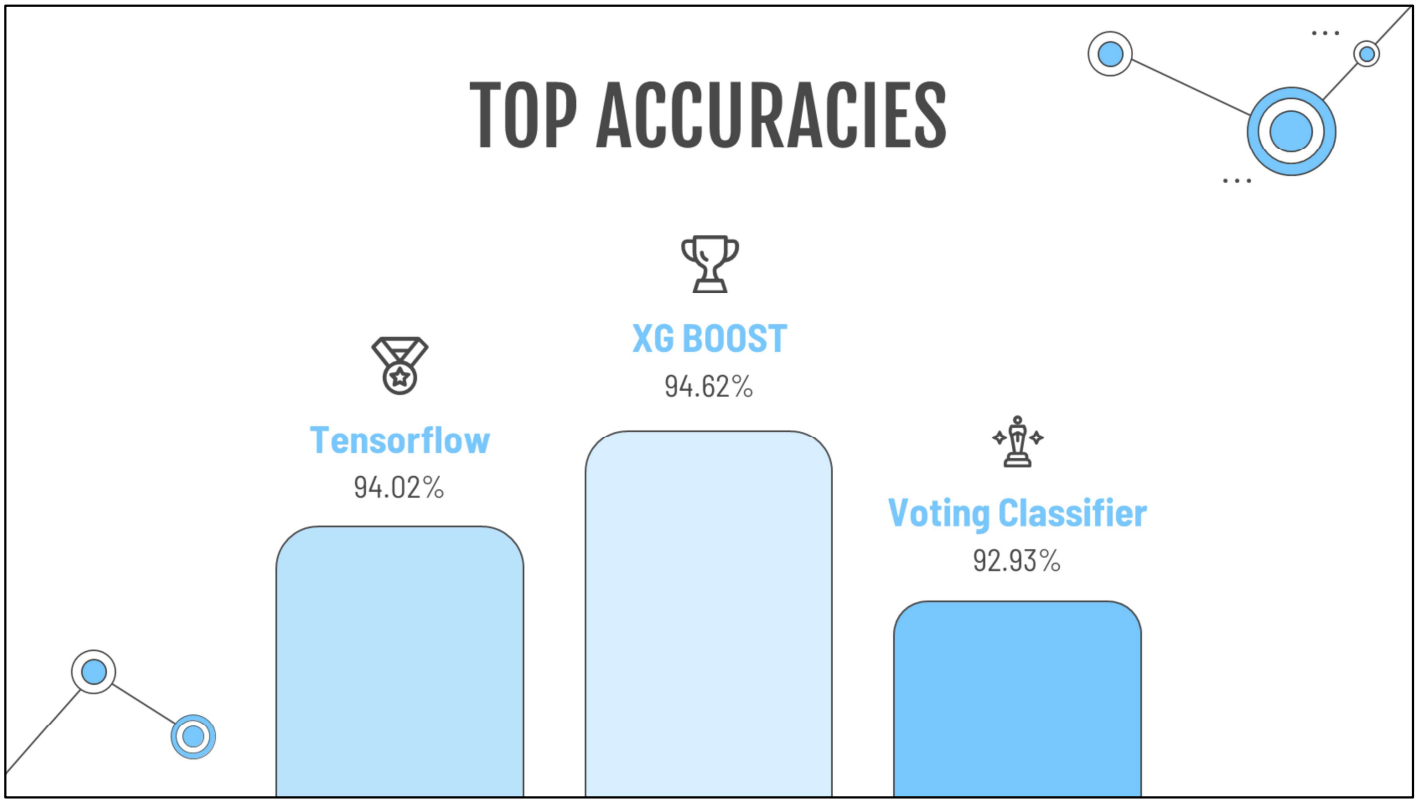

Voting Classifier combines the predictions of multiple classifiers, which can lead to better accuracy than any individual classifier. By combining the strengths of different models, the Voting Classifier can overcome the weaknesses of individual classifiers and improve the overall performance of the model. Voting Classifier is a robust algorithm that can handle noise and errors in the data. Even if some of the individual classifiers make incorrect predictions, the Voting Classifier can still make accurate predictions by taking into account the predictions of other classifiers.By combining multiple classifiers, the Voting Classifier can reduce the risk of overfitting the data. This is because individual classifiers may overfit the data in different ways, but combining their predictions can smooth out these overfitting effects and improve the generalization performance of the model.

# 04

# Analysis & Comparison




Here are the accuracy comparison of the models, we can see that the boosting method of XG Boost has the highest accuracy.




The top three accuracies are XG Boost, tensorflow and voting classifier.

# Other Metrics

| | Precision | Recall | F-1 |
|---|---|---|---|
| KNN | 27.07% | 57.65% | 36.84% |
| SVM | 24.49% | 64.47% | 35.49% |
| Logistic Regression | 25.00% | 44.94% | 32.13% |
| Voting Classifier | 37.37% | 51.53% | 43.32% |
| Decision Tree | 26.46% | 42.59% | 32.64% |
| XG Boost | 48.60% | 40.71% | 44.30% |

Here are the comparison of precision, recall and F-1 score of the models. We can see the highest precision is achieved by XG Boost, lowest recall is achieved by XG Boost and highest F-1 is achieved by XG Boost. As a result, XG boost is the best model in every metric.

# Possible Improvement

**Cluster First**

Cluster based on age gap of 10 in data preprocessing

**Reason**

Different age range typically has similar features

**Result**

The accuracy is higher, clustering provides a generalized idea, better for prediction

There are some thoughts about our possible improvement in the accuracy. For example, we could cluster the age to a gap of 10 each, so the range of the age becomes bigger, which will improve the accuracy of our prediction.


Thanks!
Do you have any questions?