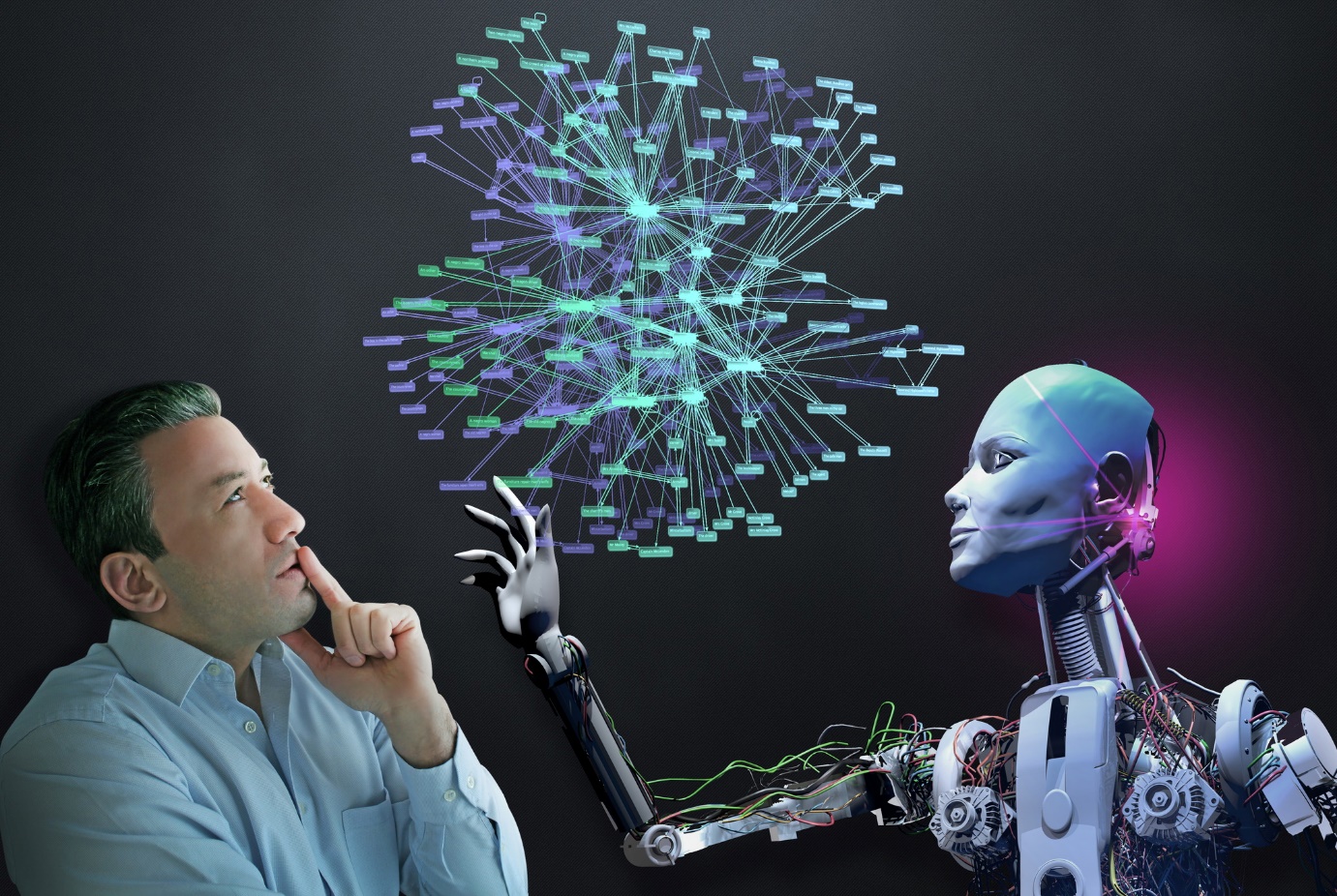
**DATA ANALYTICS AND COGNOS**

**Project: Customer Churn Production**

**Phase - 2**

**1. Introduction**



[This Photo](https://researchleap.com/fostering-open-collaborative-innovation-for-micro-and-small-technology-based-firms-in-brazil/) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

**1.1 Abstract:-**

The Problem is based on the domain of the Banking sector where the bank wants to predict the Churn of a customer depending upon the previous data of the customer. By churn it is meant that the bank wants to predict if a customer would be a defaulter in the next quarter depending upon its previous credit history.

**What is the problem?**

The main problem is to predict if a customer would be credit defaulter or not depending upon the previous data of the customer.

**Why is it important?**

It is important from a bank’s perspective in order to maintain business and customer relationship/ Apart from that if someone could be predicted as a defaulter then primitive measures can be taken in order to ensure that such violations do not happen.

**What is your basic approach?**

The basic approach of solving this problem was first studying the data , then bringing out insights from the dataset and after that I have followed a machine learning pipeline in order to solve the problem.

The ML Pipeline that I have followed is :

* Importing the necessary libraries and the dataset
* Performing Data Preprocessing (Exploratory Data Analysis and Data Manipulation)
* Modelling using Logistic Regression, KNN and Random Forest
* Performing Prediction
* Visualization in between Actual and predicted Values

**The environment used was python 3.7 and the libraries such as numpy, pandas, matplotlib , Standard Scaler and Scikit Learn module were used for Scientific computations.**

**2. Problem Definition and Algorithm**

**The Problem Statement is as follows :-**

**Customer Churn Prediction:**

A Bank wants to take care of customer retention for its product: savings accounts. The bank wants you to identify customers likely to churn balances below the minimum balance. You have the customers information such as age, gender, demographics along with their transactions with the bank.

Your task as a data scientist would be to predict the propensity to churn for each customer.

**Data Dictionary**

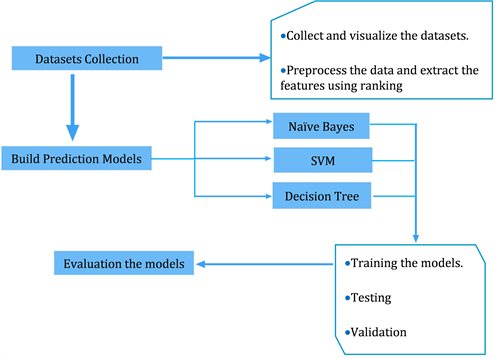
There are multiple variables in the dataset which can be cleanly divided into 3 categories:

I. Demographic information about customers

* **customer\_id** - Customer id
* **vintage** - Vintage of the customer with the bank in a number of days
* **age** - Age of customer
* **gender** - Gender of customer
* **dependents** - Number of dependents
* **occupation** - Occupation of the customer
* **city** - City of the customer (anonymized)

II. Customer Bank Relationship

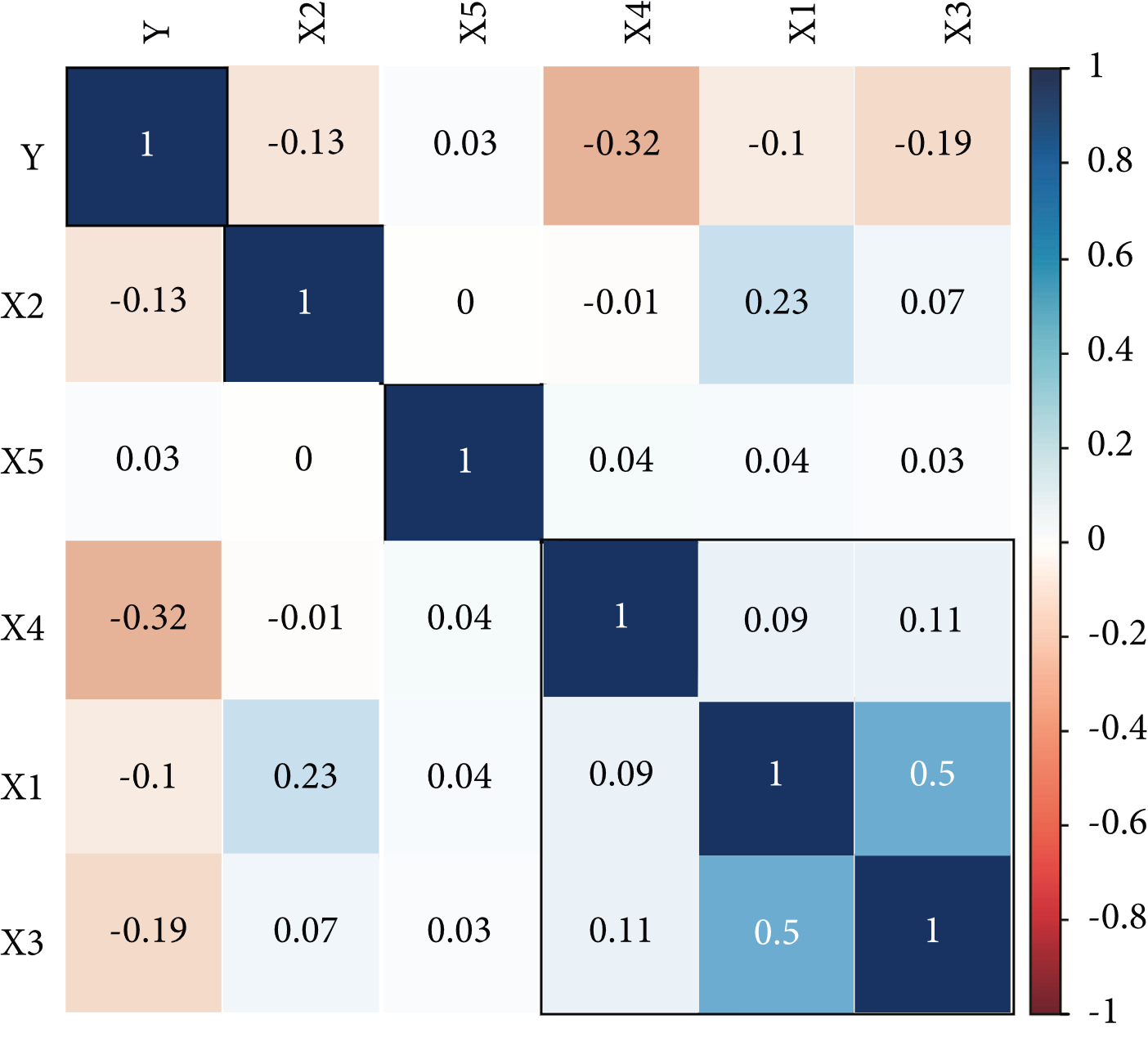
* **customer\_nw\_category** - Net worth of customer (3: Low 2: Medium 1: High)
* **branch\_code** - Branch Code for a customer account
* **days\_since\_last\_transaction** - No of Days Since Last Credit in Last 1 year



[This Photo](https://www.scirp.org/journal/paperinformation.aspx?paperid=96177) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

III. Transactional Information

* **current\_balance** - Balance as of today
* **previous\_month\_end\_balance** - End of Month Balance of previous month
* **average\_monthly\_balance\_prevQ** - Average monthly balances (AMB) in Previous Quarter
* **average\_monthly\_balance\_prevQ2** - Average monthly balances (AMB) in previous to the previous quarter
* **current\_month\_credit** - Total Credit Amount current month
* **previous\_month\_credit** - Total Credit Amount previous month
* **current\_month\_debit** - Total Debit Amount current month
* **previous\_month\_debit** - Total Debit Amount previous month
* **current\_month\_balance** - Average Balance of current month
* **previous\_month\_balance** - Average Balance of previous month
* **churn** - Average balance of customer falls below minimum balance in the next quarter (1/0)



[This Photo](https://www.hindawi.com/journals/ddns/2021/7160527/) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

* **2.1 Task Definition**
* The task was to predict those customers likely to churn balances below the minimum balance.The customers information such as age, gender, demographics along with their transactions with the bank.The task as a data scientist was to predict the propensity to churn for each customer.
* **Logistic Regression**
* Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.
* It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.
* Linear Regression Equation:
* Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.
* Sigmoid Function:
* Apply Sigmoid function on linear regressi



[This Photo](https://researchleap.com/role-service-quality-patients-customer-satisfaction-public-healthcare-institutions-ghana/) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

# Advantages:

* Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
* It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
* The algorithm can be used in both classification and regression problems.
* Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
* You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

# Disadvantages:

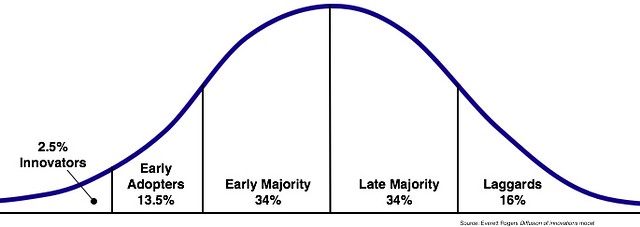
* Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
* The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

# Finding important features

Random forests also offers a good feature selection indicator. Scikit-learn provides an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction. It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.

This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.



[This Photo](http://www.flickr.com/photos/wfryer/2564440831/) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

[This Photo](https://link.springer.com/article/10.1186/s40497-018-0116-4) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)