**BANK CUSTOMER CHURN PREDICTION**

**PROJECT Report**

***Submitted by***

**Ravikumar Yadav – RA1911028010089**

**Nived N Sooraj – RA1911028010090**

**Sidharth Manoj – RA1911028010091**

**Mandla Praneeth – RA1911028010094**

**Gowthami Priya – RA1911028010140**

***in partial fulfilment for the award of the degree***

***Of***

**B. TECH**

***in***

**COMPUTER SCIENCE and ENGINEERING**

**IN**

**CLOUD COMPUTING**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**KATTANKULATHUR**

**[April, 2022]**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
| **1.** | **ABSTRACT** | **2** |
| **2.** | **DEPENDENCIES** | **3** |
| **3.** | **WORKING PROCESS** | **4** |
| **4.** | **DATA COLLECTION AND DATA PROCESSING** | **5** |
| **5.** | **TRAIN AND TEST DATA** | **6** |
| **6.** | **MODEL TRAINING** | **7** |
| **7.** | **MODEL EVALUATION** | **8** |
| **8.** | **CONCLUSION** | **9** |

**ABSTRACT**

Customer churn is a major problem of customers leaving your products/subscription and moving to another service. Due to direct effect on profit margins, businesses now are looking to identify customers who are at the risk of churning and retaining them by personalized promotional offers. In order to retain them, they need to identify the customers as well as the reason of churning so that they can provide the customers with personalized offers and products. The aim of our project is to solve this problem for banking domain, by identifying which customers are at risk of churning and what are the reasons for churning with the help of data mining and machine learning algorithms. The project focuses on 2 deliverables - Predict customers likely to churn using supervised learning classification algorithms and customer segmentation of customers using unsupervised learning to validate the similarities in the ‘likely to churn’ customer subset to come up with different segments. The reasons for a particular customer churn can vary from internal factors as well as external factors but we will try to understand the reasons of churning depending on internal factors using explainable AI, which breaks into the black box of machine learning algorithms and gives a clear explanation of the predictions.

**DEPENDENCIES**

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don't need to write the same code again and again for different programs.

**NumPy** – NumPy can be used to perform a wide variety of mathematical operations on arrays. Pandas - Pandas is mainly used for data analysis and associated manipulation of tabular data in data frames.

**Pandas** -- Pandas is mainly used for data analysis and associated manipulation of tabular data in Data Frames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel.

**Matplotlib** – Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

**Seaborn** -- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data.

**Sklearn** - Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

**WORKING PROCESS**

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

The ML Pipeline that we have followed is:

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

Importing the necessary libraries and the dataset:  
Here, the libraries such as NumPy, pandas and matplotlib were called. NumPy is known to be Numerical Python which is responsible for performing all the Numerical tasks in this project whereas Pandas would make the data frame and Matplotlib was used for visualization.

Performing Data Pre-processing (Exploratory Data Analysis and Data Manipulation): -  
In this step the data was thoroughly analysed and the steps such as Univariate Analysis and Bivariate Analysis were performed.

If the data is said to be a continuous data, then in order to do a univariate analysis, we check the distribution of the data i.e., we check if the data is normally distributed or not and if the data is said to be a categorical data, then we would check the bar plot of that date. In order to perform a Bivariate Analysis (The study of two data at an instance) we would have to plot the scatter plot in order to check the relationship in between them.

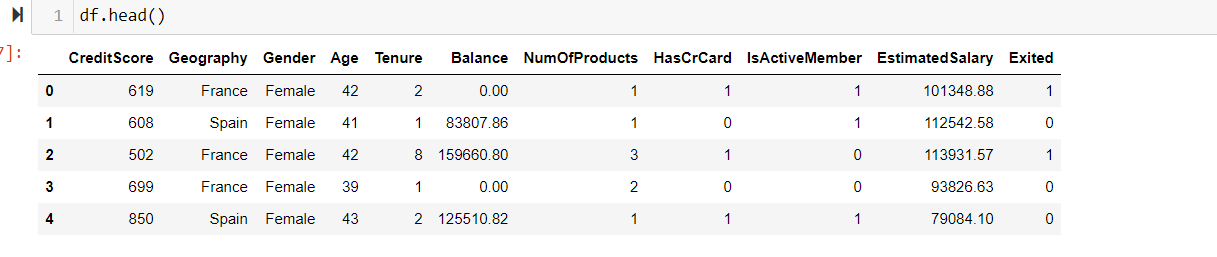
Modelling using Logistic Regression, KNN and Random Forest

The modelling was done by using Logistic Regression, Random Forest and KNN and at the end it was ensemble by using Voting Classifier.

**DATA COLLECTION AND DATA PROCESSING**

**Data collection** is the process of gathering and measuring information from countless different sources.

**Data Processing** is the task of converting data from a given form to a much more usable and desired form i.e., making it more meaningful and informative.

 *Reading the Dataset*

1.Loading the dataset and making it an pandas data frame.

2. Read.csv () to read the csv file containing the data.

3. Since there is no header file, we need to mention that as ‘none’.



4. shape () - We check the number of rows and columns present in the data set.

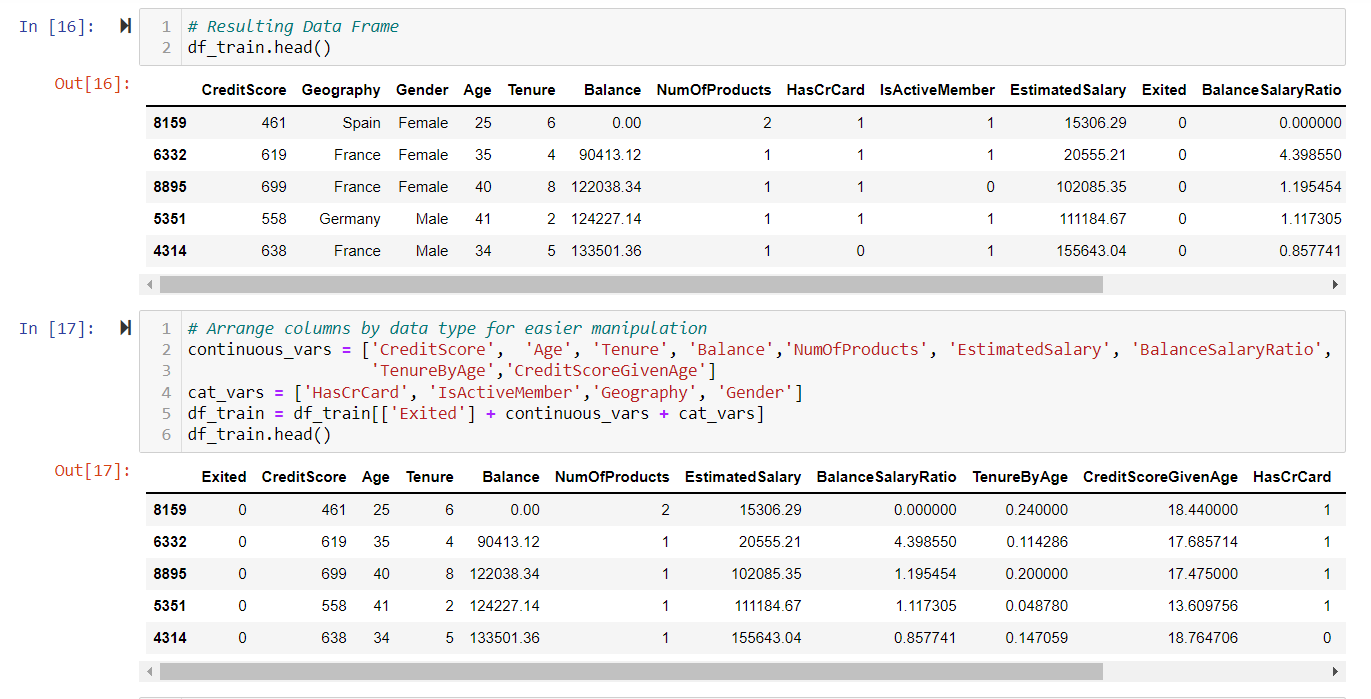
5. We get the result as 1000 rows, 14 columns.

**TRAINING AND TEST DATA**

• We need to split the data into training and test data.

**Training data** is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict.

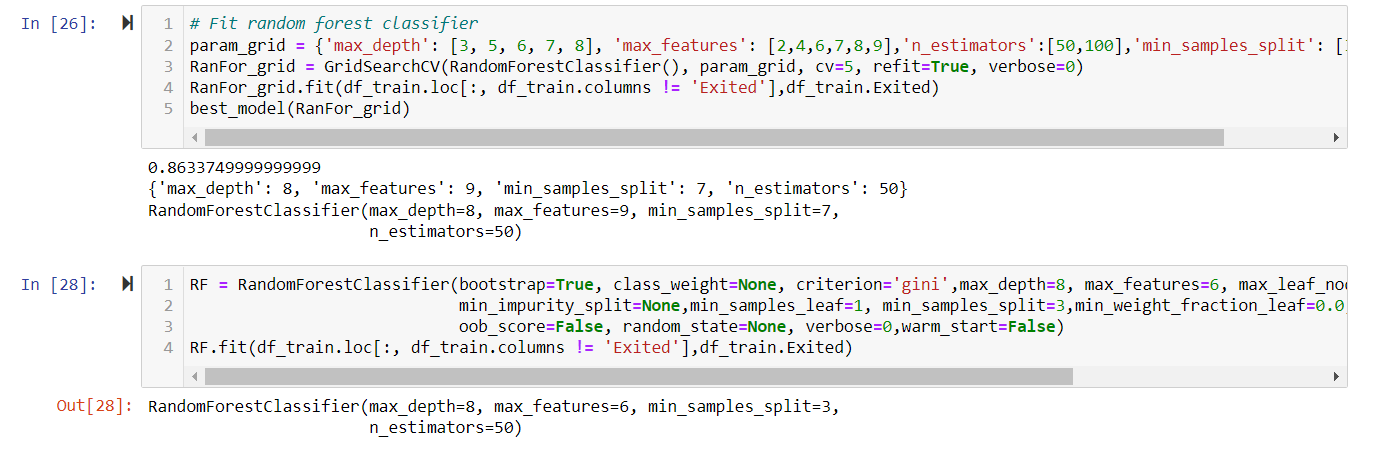
**Test data** is used to measure the performance, such as accuracy or efficiency, of the algorithm you are using to train the machine.



We need to split a dataset into train and test sets to evaluate how well our machine learning model performs.

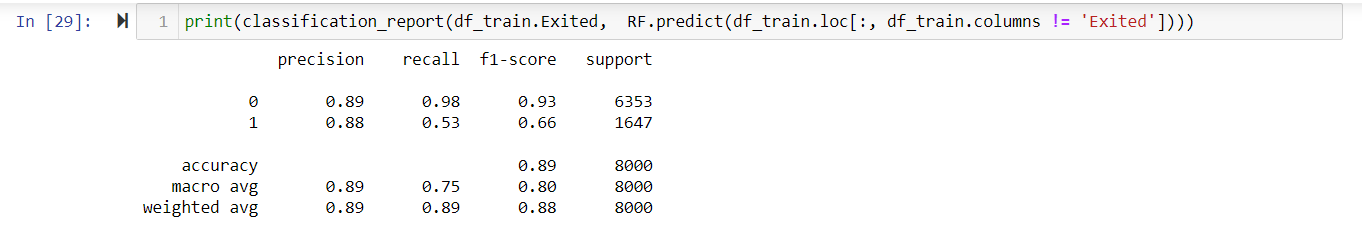
Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing

**MODEL TRAINING**



**Model training** in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. For our model training we used **Logistic Regression and Random Forest**. Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

**MODEL EVALUATION**

*Classification report of training data*

• The training data accuracy and the test data accuracy are both calculated and both are more than 80%, so the accuracy is good.

• Since we have used more data somewhat like 1000-2000 the accuracy is good.

• Model.predict() – prediction of training data accuracy score.

• >80% accuracy for training.

• Now we will check accuracy on test data.

The model has not seen this data yet.

• 79.19 % accuracy for testing.

• So out of 100 time 79 times the model can predict correctly

## **CONCLUSION**

The precision of the model on previously unseen test data is slightly higher with regard to predicting 1's i.e., those customers that churn. However, in as much as the model has a high accuracy, it still misses about quarter of those who end up churning. This could be improved by providing retraining the model with more data over time while in the meantime working with the model to save the 79% that would have churned.

.

its previous credit

y