Project Report
Customer Churn Analysis in Banking Sector
Group 58

Introduction

Customer churn is when customers discontinue doing business with a company. Maintaining a strong client base is crucial for the success of any organization because it is significantly less expensive to keep an existing customer than to find a new one. Banking consumers have several options, hence this sector has a relatively high rate of customer churn. Customers tend to have safe options, high returns, low rate of interest, and variety of benefits expected from the banks.

Motivation

The following are the general reasons for the need of Customer churn Analysis:

- ► Analysis of churn rate in customers is very important for the banks to improve their functionality and attract more customers.
- ► The cost of attracting new customers can be five to six times more than holding on to existing customers
- ► Long term customers become less costly to serve, they generate higher profits, and they may also provide new referrals
- ► Losing a customer usually leads to loss in profit for the banks
- ► During this recession, slowing sales results in a spike in the nominal churn rate

The project primarily consists of 3 key activities to execute the following:

- Import data into a structured light weight database, in this case we used SQLite.
- Exploratory Data Analysis (EDA) for data visualization and figuring out the factors responsible for churn.
- Building Predictive Model based on the factors.

Import Data into SQLite:

We converted the raw data into organized manner and imported the data into light weight database (SQLite) and created a normalized database to store the data in proper format.
We divided the attributes accordingly and stored them in four different tables containing General Information, Personal Information, Bank details, Status of the Customer using ingestion scripts

Data Definition:

General Information: This table is used to store all the general information like CustomerID, CreditScore etc.

Column Name	Data Type	Constraints
RowNumber (ID)	Integer	PK, NOT NULL
CustomerID	Integer	
CreditScore	TEXT	
Tenure	INTEGER	
EstimatedSalary	REAL	

Personal Information: This table contains the personal information of the customer.

Column Name	Data Type	Constraints
RowNumber (ID)	Integer	PK, NOT NULL
Surname	TEXT	
Geography	TEXT	
Gender	TEXT	
Age	INTEGER	

Bank Information: The table consists of Bank related data like.

Column Name	Data Type	Constraints
RowNumber (ID)	INTEGER	PK, NOT NULL
Balance	REAL	
NumOfProducts	INTEGER	
HasCrCard	INTEGER	

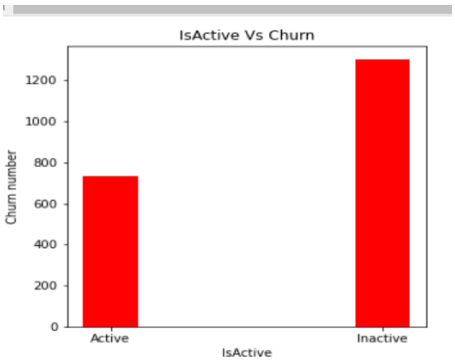
Status Information: The fourth table consists of Activity and exit status of the customer.

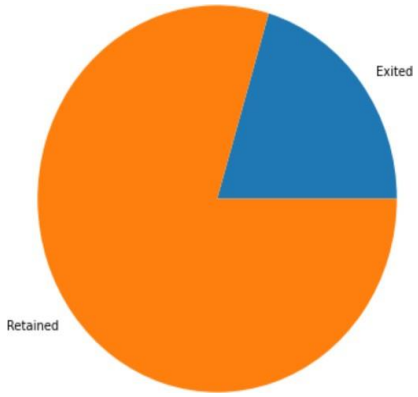
Column Name	Data Type	Constraints
RowNumber (ID)	INTEGER	PK, NOT NULL
IsActiveMember	INTEGER	
Exited	INTEGER	

After the data was entered into the database, the following step was to undertake exploratory data analysis.

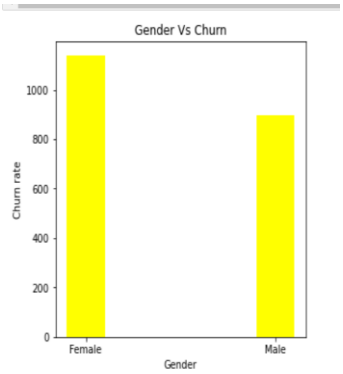
Exploratory Data Analysis:

Pie chart representing percentage of customer exit ,Bar graph representing IsActiveMember status Vs Churn

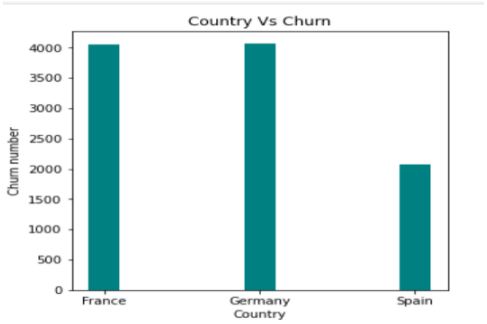




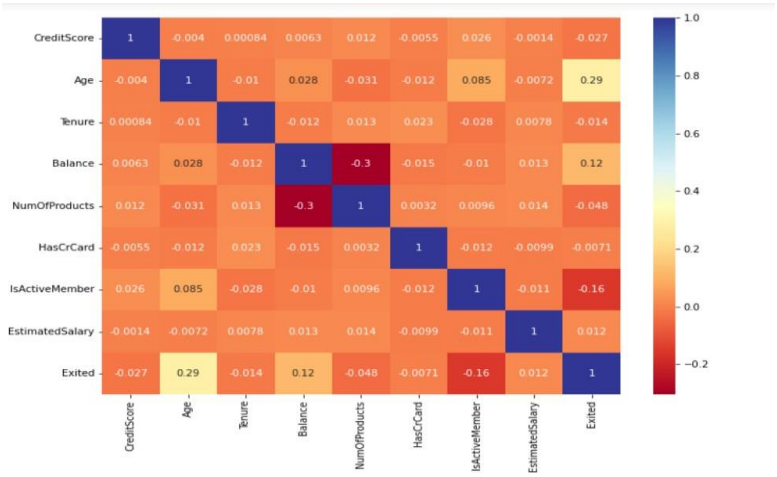
Bar graph representing Gender Vs Churn Rate



Bar graph representing Country Vs Churn rate



Correlation matrix:



Analysis:

After the factor selection, we build predictive models using Logistic Regression and Random Forest algorithms.

Logistic Regression:

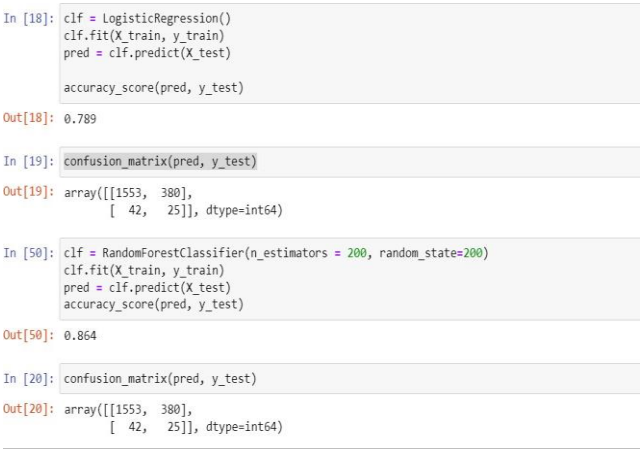
Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

Random Forest:

When using Random Forests, which is a form of ensemble learning, several decision trees are used to decide or predict a value. Regression or classification can be done using it.

Result:

Model	Accuracy
Logistic Regression	78.9%
Random Forest	86.4%



```
In [18]: clf = LogisticRegression()
         clf.fit(X_train, y_train)
         pred = clf.predict(X_test)

         accuracy_score(pred, y_test)

Out[18]: 0.789

In [19]: confusion_matrix(pred, y_test)

Out[19]: array([[1553,  380],
                [  42,   25]], dtype=int64)

In [50]: clf = RandomForestClassifier(n_estimators = 200, random_state=200)
         clf.fit(X_train, y_train)
         pred = clf.predict(X_test)
         accuracy_score(pred, y_test)

Out[50]: 0.864

In [20]: confusion_matrix(pred, y_test)

Out[20]: array([[1553,  380],
                [  42,   25]], dtype=int64)
```

Conclusion:

In this Project, Random Forest gives more accuracy than Logistic regression since data mostly comprised of categorical variables and depicts more non-linearity. With these models, Banks can have an idea of what factors affect customers leave or attract, and through this analysis , they can improve their services and introduce new schemes to attract more customers .

Dataset:https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/data
Code Link:https://buffalo.box.com/s/9ixs1cfhzzybnzrq338plvf0ddl1vwh1