

# **Customer Churn Prediction using Machine Learning in banking**



Santosh Kumar

24-01-22

## *Abstract*

The number of service providers are being increased very rapidly in every business. In these days, there is no shortage of options for customers in the banking sector when choosing where to put their money. As a result, customer churn and engagement has become one of the top issues for most of the banks. In this report, a method to predict the customer churn in a bank, using machine learning techniques, which is a branch of artificial intelligence has been proposed. This report promotes the exploration of the likelihood of churn by analysing customer behaviour. The KNN, Decision Tree, and Logistic regression classifiers have been used in this report. Also, some feature selection methods have been done to find the more relevant features.

## **1. Problem Statement**

In an era of increasingly saturated markets that have intensified competition between companies, customer defection poses a real problem. Therefore, it has become clear to companies and managers that the historical customer information, which can be used to create models, in the existing customer base is one of the most important assets to combat customer churn. The search and identification of customers who show a high inclination to abandon the company or customer churn prediction is of crucial importance.

## **2. Market/Customer/Business need assessment**

- Finding a new customer is 5 times more expensive than keeping an existing one.
- To increase profits for continuing operations and enhance the core competitiveness, commercial banks must avoid the loss of customer while acquiring new customers.
- Harvard Business Review believes that by reducing the customer defection rate by 5%, companies can increase profits by 25% to 85%, while Business Week thought the profits will increase by 140%.

## **3. Target Specification**

- We want to retain the existing customer with new one.
- To understand the needs of the customers.
- Finding the changing customer behaviour and their rising expectation.
- Delivering reliable service on time and in budget to customer while maintaining a good working partnership with them.
- This report aims to build a framework that can predict the client churn in banking sector using some machine learning techniques.

## **4. External Search (information sources)**

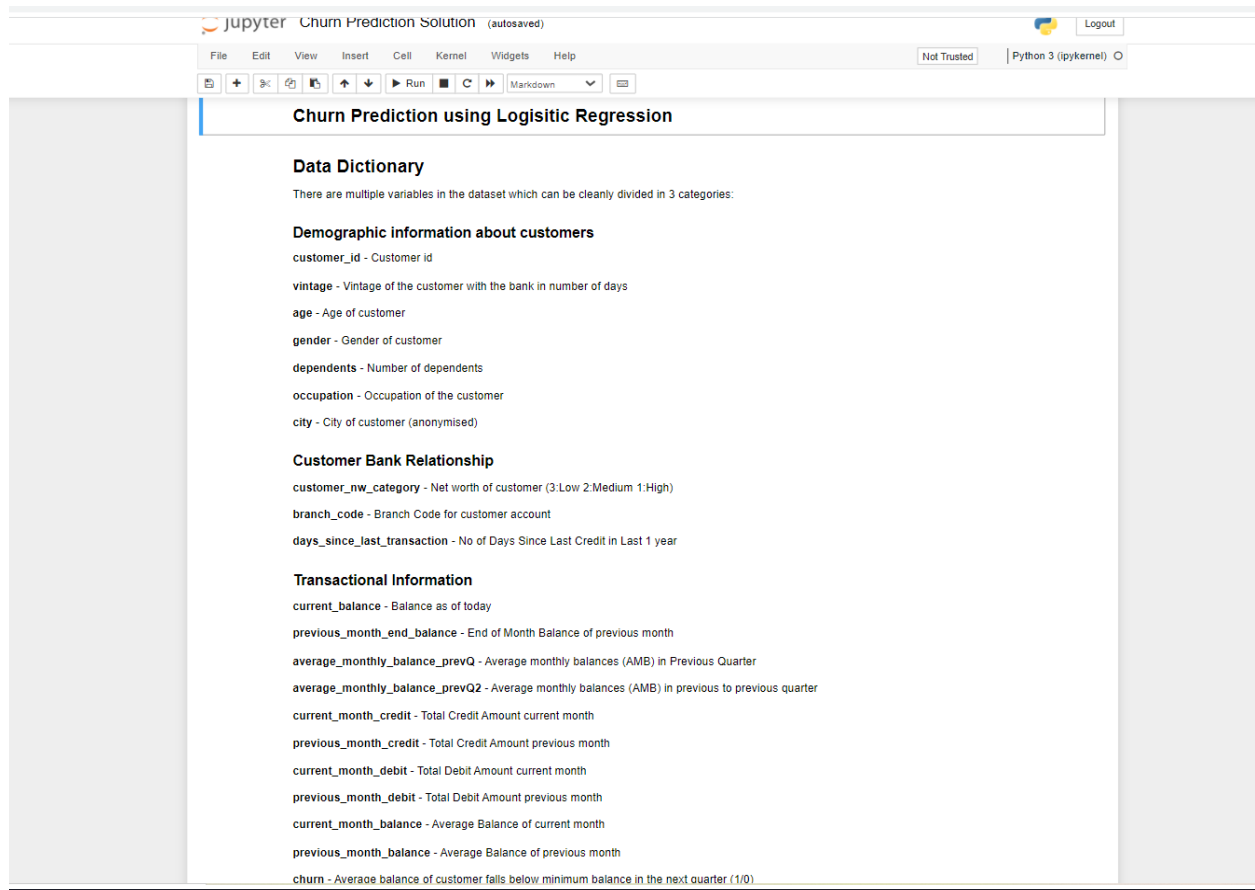
I have used customer churn prediction dataset which can be found here-[web link](#).

Our dataset has the customers information such as age, gender, demographics along with their transactions with the bank.

### **Some important papers:-**

- M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction in Banking," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1196-1201, Doi: 10.1109/ICECA49313.2020.9297529.
- M. Malyar, M. V. Mykola Robotyshyn and M. Sharkadi, "Churn Prediction Estimation Based on Machine Learning Methods," *2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC)*, 2020, pp. 1-4, doi: 10.1109/SAIC51296.2020.9239230.

## Let's have a look on my dataset:



**Churn Prediction using Logistic Regression**

**Data Dictionary**

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

**Demographic information about customers**

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

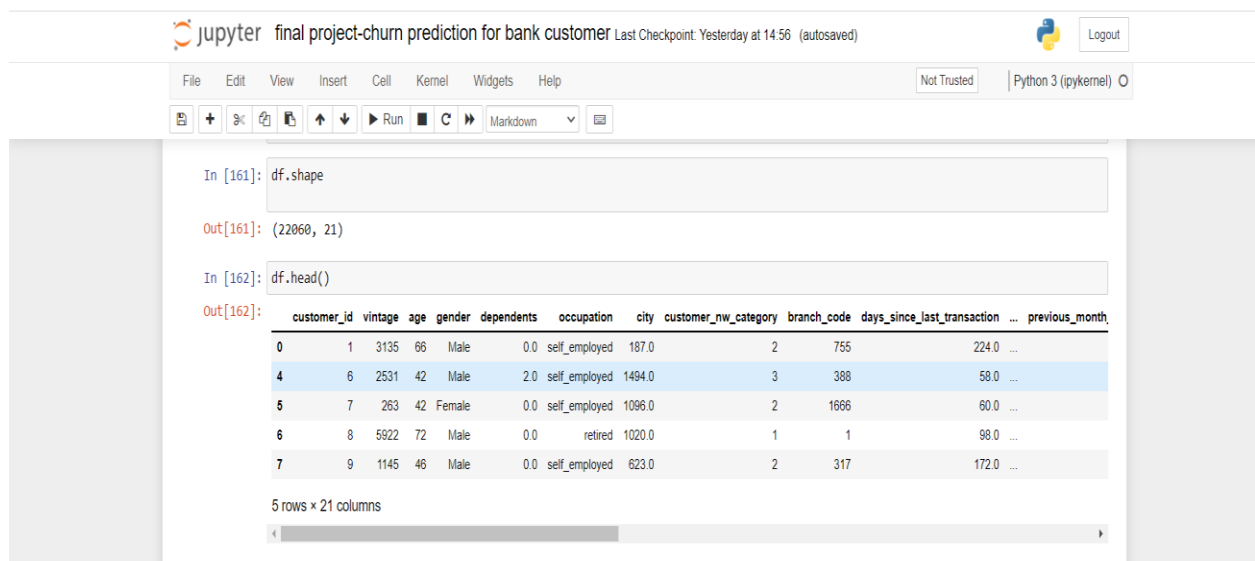
**Customer Bank Relationship**

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

**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 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)

## Important features:



**Jupyter final project-churn prediction for bank customer** Last Checkpoint: Yesterday at 14:56 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel)

In [161]: `df.shape`

Out[161]: (22060, 21)

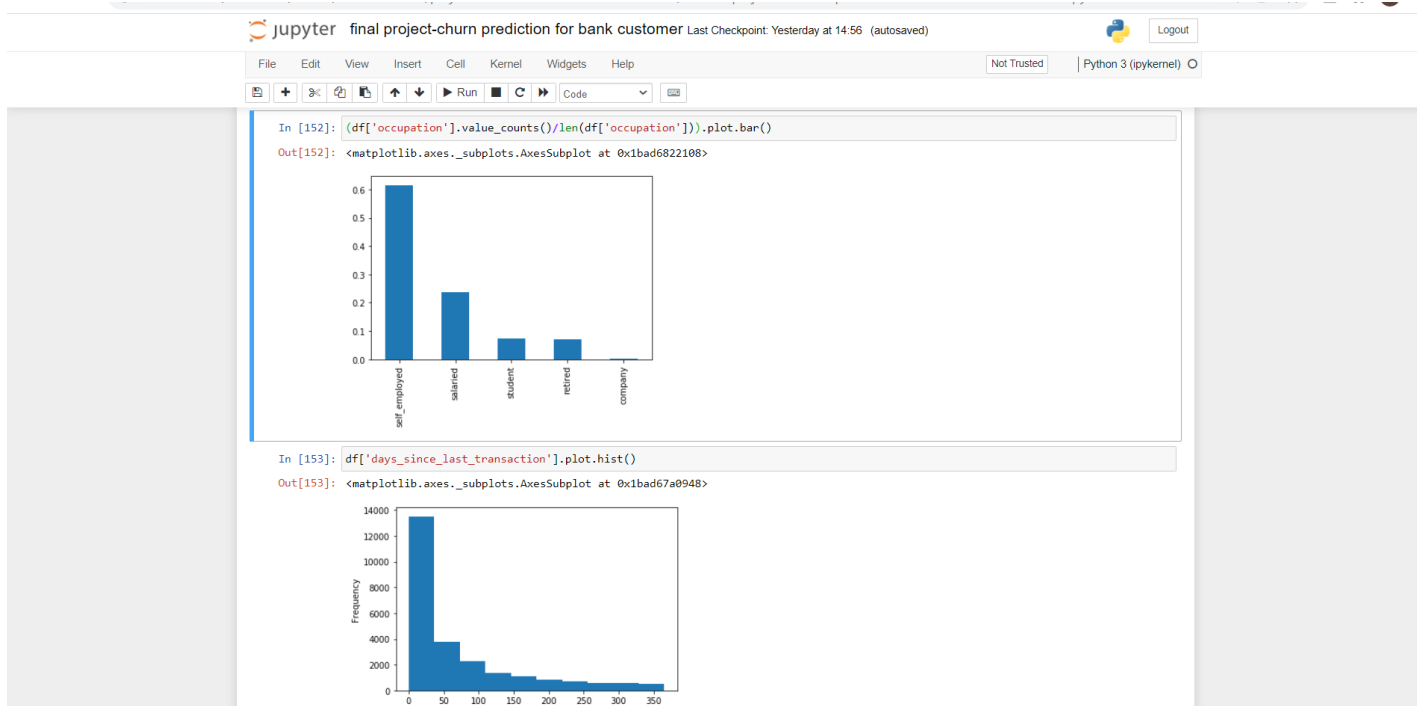
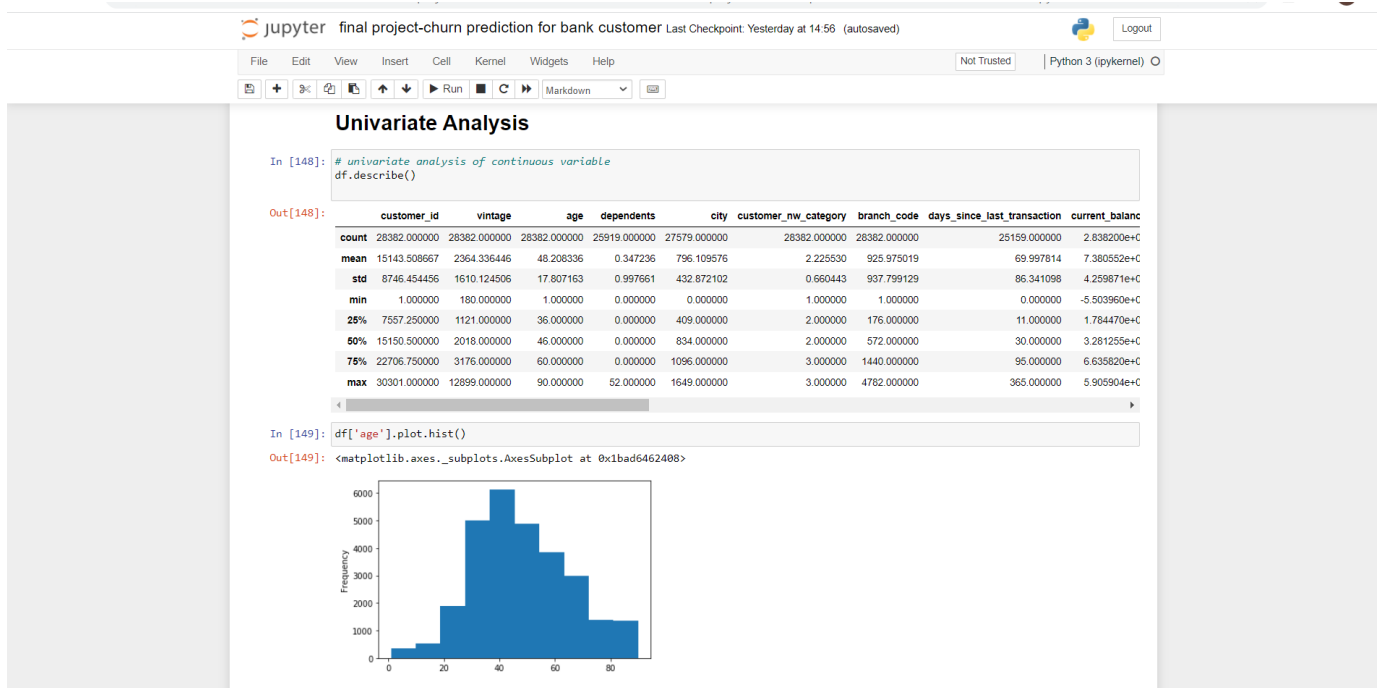
In [162]: `df.head()`

Out[162]:

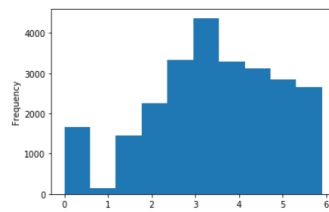
	customer_id	vintage	age	gender	dependents	occupation	city	customer_nw_category	branch_code	days_since_last_transaction	...	previous_month
0	1	3135	66	Male	0.0	self_employed	187.0	2	755	224.0	...	...
4	6	2531	42	Male	2.0	self_employed	1494.0	3	388	58.0	...	...
5	7	263	42	Female	0.0	self_employed	1096.0	2	1666	60.0	...	...
6	8	5922	72	Male	0.0	retired	1020.0	1	1	98.0	...	...
7	9	1145	46	Male	0.0	self_employed	623.0	2	317	172.0	...	...

5 rows x 21 columns

## 5. Benchmarking



```
Out[154]: <matplotlib.axes._subplots.AxesSubplot at 0x1bad26398ce>
```



## Missing Value Treatment

```
In [155]: df.dropna(how='all')
```

```
Out[155]:
```

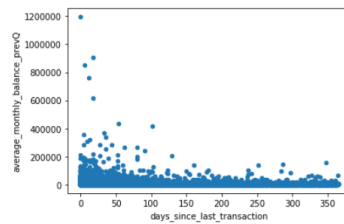
	customer_id	vintage	age	gender	dependents	occupation	city	customer_nw_category	branch_code	days_since_last_transaction	...	previous_m
0	1	3135	66	Male	0.0	self_employed	187.0	2	755	224.0	...	...
1	2	310	35	Male	0.0	self_employed	NaN	2	3214	60.0	...	...
2	4	2356	31	Male	0.0	salaried	146.0	2	41	NaN	...	...
3	5	478	90	NaN	NaN	self_employed	1020.0	2	582	147.0	...	...
4	6	2531	42	Male	2.0	self_employed	1494.0	3	388	58.0	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
28377	30297	1845	10	Female	0.0	student	1020.0	2	1207	70.0	...	...
28378	30298	4919	34	Female	0.0	self_employed	1046.0	2	223	14.0	...	...
28379	30299	297	47	Male	0.0	salaried	1096.0	2	588	0.0	...	...
28380	30300	2585	50	Male	3.0	self_employed	1219.0	3	274	NaN	...	...
28381	30301	2349	18	Male	0.0	student	1232.0	2	474	59.0	...	...

```
Out[158]: <matplotlib.axes._subplots.AxesSubplot at 0x1bad28b2b08>
```

## Outlier Treatment

```
In [158]: df.plot.scatter('days_since_last_transaction', 'average_monthly_balance_prevQ')
```

```
Out[158]: <matplotlib.axes._subplots.AxesSubplot at 0x1bad28b2b08>
```



```
In [159]: df['average_monthly_balance_prevQ']
```

```
Out[159]:
```

0	1458.71
4	1643.31
5	15211.29
6	7859.74
7	6511.82
...	...
28375	8082.48
28377	2282.19
28378	3668.83
28379	53444.81
28381	3213.44

Name: average\_monthly\_balance\_prevQ, Length: 22067, dtype: float64

```
In [160]: df=df[df['average_monthly_balance_prevQ']<400000]
```

jupyter final project-churn prediction for bank customer Last Checkpoint: Yesterday at 14:56 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel)

### Label Encoding

```
In [163]: from sklearn.preprocessing import LabelEncoder

In [164]: df['gender'].value_counts()
Out[164]: Male      13416
          Female    8644
          Name: gender, dtype: int64

In [165]: df['gender']=df['gender'].map({'Male':1,'Female':0})

In [166]: df['occupation'].value_counts()
Out[166]: self_employed    13428
          salaried         5601
          retired         1638
          student         1369
          company           24
          Name: occupation, dtype: int64

In [167]: df['occupation']=df['occupation'].map({'company':0,'student':1,'retired':2,'salaried':3,'self_employed':4})

In [168]: df.head()
Out[168]:
```

	customer_id	vintage	age	gender	dependents	occupation	city	customer_nw_category	branch_code	days_since_last_transaction	...	previous_month_e
0	1	3135	66	1	0.0	4	187.0	2	755	224.0	...	
4	6	2531	42	1	2.0	4	1494.0	3	388	58.0	...	
5	7	263	42	0	0.0	4	1098.0	2	1666	60.0	...	
6	8	5922	72	1	0.0	2	1020.0	1	1	98.0	...	
7	9	1145	46	1	0.0	4	623.0	2	317	172.0	...	

5 rows x 21 columns

## 6.Applicable Patents

- [Patent -1 Churn prediction and management system.](#)
- [Patent- 2 System and method for predicting and preventing customer churn.](#)

There are lot of patents but these two are most relevant which can be looked upon.

First patent describes the system and method for managing churn among the customers of a business. The system and method provide for an analysis of the causes of customer churn and identifies customers who are most likely to churn in the future.

Second patent describes methods for predicting and preventing customer churn.

## 7. Applicable Regulation

- Data protection and privacy regulation(Customer)
- Govt. Regulation for banks.
- Patents on ML algorithm developed.
- Ensuring open-source ,academic and research community for an audit of algorithms.
- Review of existing work authority regulations.

## 8. Applicable Constraints

- Data collection from bank customers.
- Understanding the dynamic behavior of customers.
- Regularly updating the dataset.
- Developed model should have high accuracy.

## 9. Business opportunity

Churn prediction is used in a variety of different industries and types of businesses.

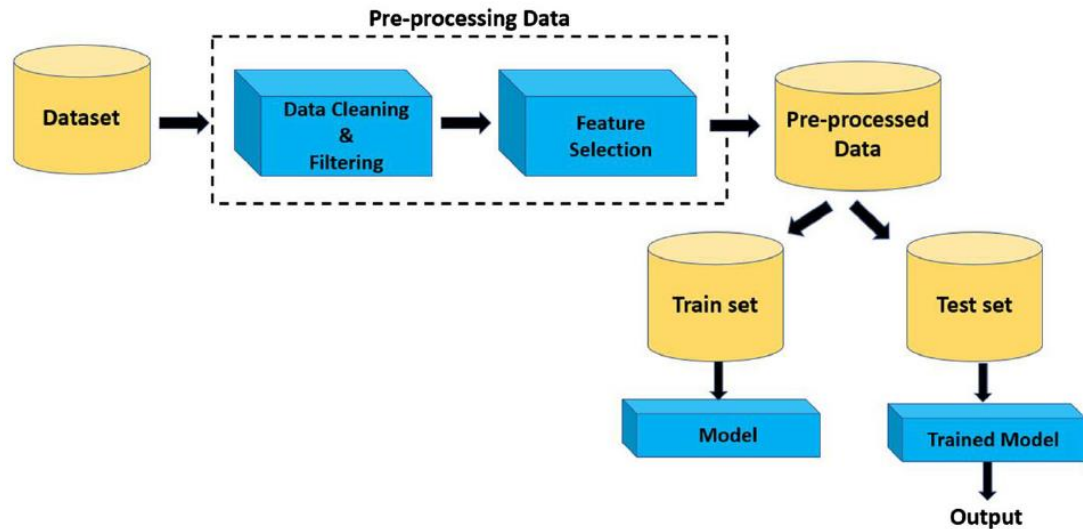
It is, however, most relevant to SaaS(software as a service) companies and membership-based businesses that charge an ongoing monthly, quarterly, or annual fee for their software or services.

As far as how churn prediction can be used within our business, it's one of the key components of [determining the lifetime value of customers](#).

## 10. Concept Generation

My proposed methodology, consists of six phases. In the first two phases, data pre-processing and feature analysis is performed. In the third phase, feature selection is taken into consideration. Next, the data has been split into two parts train and test set in the ratio of 80% and 20% respectively. In the prediction process, most popular predictive models have been applied, namely, logistic regression, K-

nearest neighbour, decision trees, etc. on train set as well as ensemble techniques are applied to see the effect on accuracy of models.



- **clean the data**
- **split the data into train- test**

```
In [170]: # importing train_test split function|
from sklearn.model_selection import train_test_split
train_x,valid_x,train_y,valid_y=train_test_split(x,y,random_state=101,stratify=y)
```

```
In [171]: # importing different ml models to use it in ensemble modelling
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
```

**We will use three different models and use ensemble technique to finalize the result.**

## 1. Logistic Regression

```
In [172]: # for logistic regression:-
model1 = LogisticRegression()
model1.fit(train_x,train_y)
pred1=model1.predict(valid_x)
pred1[:10],model1.score(valid_x,valid_y)
```

```
Out[172]: (array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64), 0.8195829555757026)
```



## 2. K-nearest neighbour

```
In [173]: # for k-nearest neighbour
model2=KNeighborsClassifier(n_neighbors=5)
model2.fit(train_x,train_y)
pred2=model2.predict(valid_x)
pred2[:10],model2.score(valid_x,valid_y)

Out[173]: (array([0, 0, 1, 0, 0, 1, 0, 0, 0, 0], dtype=int64), 0.8235720761559383)
```

## 3. Decision Tree

```
In [174]: # for Decision Tree classifier
model3= DecisionTreeClassifier(max_depth=7)
model3.fit(train_x,train_y)
pred3=model3.predict(valid_x)
pred3[:10],model3.score(valid_x,valid_y)

Out[174]: (array([0, 0, 1, 0, 0, 0, 0, 0, 0, 0], dtype=int64), 0.8417044424297371)
```

**By analysing all the models, we can say that Decision tree is giving highest accuracy.**

**Thus, we would use Decision tree as our final model for model deployment.**

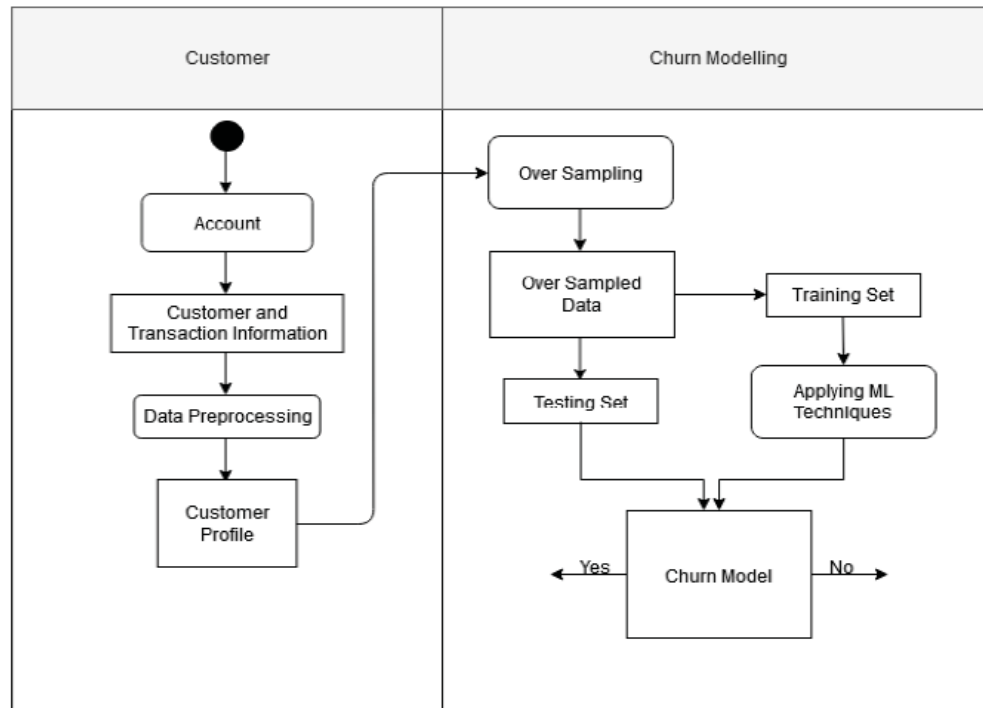
## 11. Concept development

The concept can be developed by using the appropriate API (flask in this case) and using Django as framework for the same and for its deployment , The cloud services have to be chosen according to the need.



## 12. Final Product Prototype

The final product is a service which enables small, medium, as well as large banks to predict the customer who is going to stop taking their service in nearby future. According to need, a web application or mobile application can be built on top of developed model and can be hosted on any one of available cloud services.



## 13. Code Implementation

This is GitHub Link: - <https://github.com/Santosh175/ML-project>

## 14. Conclusion

While the banking sector is considered, like any other organization, customer engagement has become one of the primary concerns. To resolve this crisis, banks need to identify customer churn possibilities as quickly as possible. The purpose of this report is to build the most appropriate model to predict client churn in a bank in the early stages.

## 15. References

- [1] Hend Sayed, Manal A. Abdel-Fattah, Sherif Kholief “Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study”.
- [2] Ketut Gde, Manik Karvana, Setiadi Yazid, Amril Syalim, and Petrus Mursanto “Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry”.
- [3] Manas Rahman, V Kumar “MACHINE LEARNING BASED CUSTOMER CHURN PREDICTION IN BANKING”.
- [4] Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi “Customer churn prediction system: a machine learning Approach”
- [5] Seyed Hossein Iranmanesh, Mahdi Hamid, Mahdi Bastan, Hamed Shakouri G., Mohammad Mahdi Nasiri “Customer Churn Prediction Using Artificial Neural Network: An Analytical CRM Application”.
- [6] Yaya Xie , Xiu Li , E.W.T. Ngai , Weiyun Ying “Customer churn prediction using improved balanced random forests”.