**CHURN MODELLING**

**Mini Project Report**

Submitted in partial fulfillment of the requirements for the degree of

**Bachelor of Engineering (Computer Engineering)**

by:

|  |  |
| --- | --- |
| VedantHariom Agrawal | TU3F1819078 |
| Swaroop Singh Rajpurohit | TU3F1819080 |
| Vinit Sanjay Sonawane | TU3F1819082 |

**Under the Guidance of**

**Prof. ReshmaKoli**



**Department of Computer Engineering**

**TERNA ENGINEERING COLLEGE**

Nerul (W), Navi Mumbai 400706

**(University of Mumbai)**

(2020-21)

**Internal Approval Sheet**

****

**TERNA ENGINEERING COLLEGE, NERUL**

**Department of Computer Engineering**

Academic Year 2020-21

**CERTIFICATE**

This is to certify that the mini project entitled **“Churn modelling”** is a bonafide work of

|  |  |
| --- | --- |
| VedantHariom Agrawal | TU3F1819078 |
| Swaroop Singh Rajpurohit | TU3F1819080 |
| Vinit SanjaySonawane | TU3F1819082 |

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the Bachelor of Engineering (Computer Engineering).

**Guide Head of Department Principal**

**Approval Sheet**

**Project Report Approval**

This Mini Project Report – entitled “**Churn Modelling**” by following students is approved for the degree of ***B.E. in "Computer Engineering"***.

**Submitted by:**

|  |  |
| --- | --- |
| VedantHariom Agrawal | TU3F1819078 |
| Swaroop Singh Rajpurohit | TU3F1819080 |
| Vinit Sanjay Sonawane | TU3F1819082 |

Examiners Name & Signature:

1.---------------------------------------------------------

2.----------------------------------------------------------

Date: ---------------------------------

Place: ---------------------------------

**Declaration**

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

|  |  |  |
| --- | --- | --- |
| VedantHariom Agrawal | TU3F1819078 |  |
| Swaroop Singh Rajpurohit | TU3F1819080 |  |
| Vinit Sanjay Sonawane | TU3F1819082 |  |



Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Place: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Acknowledgement**

We would like to express our sincere gratitude towards our guide **Prof. Priyanka Sherkhane**, Mini Project Coordinators **Prof. D.K.Chitre, Prof. D. V. Thombre, Dr. Kiran Bhandari**, **Prof. Rohini Palve, Prof. Nayana Vaity, Prof. Pravin Hole, Prof. A. B. Umbare, Prof. Reshma Koli, Prof.Saylee Narkhede,** for their help, guidance and encouragement, they provided during the project development. This work would have not been possible without their valuable time, patience and motivation. We thank them for making our stint thoroughly pleasant and enriching. It was great learning and an honor being their student.

We are deeply thankful to **Dr. Archana Mire (H.O.D Computer Department)** and entire team in the Computer Department. They supported us with scientific guidance, advice and encouragement, they were always helpful and enthusiastic and this inspired us in our work.

We take the privilege to express our sincere thanks **to Dr. L. K. Ragha** our Principal for providing the encouragement and much support throughout our work.

|  |  |  |
| --- | --- | --- |
| VedantHariom Agrawal | TU3F1819078 |  |
| Swaroop Singh Rajpurohit | TU3F1819080 |  |
| Vinit Sanjay Sonawane | TU3F1819082 |  |

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Place: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Abstract**

In the current challenging era, there is prominent competition in bank industry. To improve quality and level of service, bank concentrates on customer retention as well as customer churning. This paper discusses the classification problem of banking industry. It focuses on the customers of a bank concerns towards churning, predicting the departing customers from potential customers. Machine learning is the cutting edge technology that is practical and handy to solve such problems. Using supervised machine learning, a proprietary algorithm (a typical machine learning model) is created to forecast and inform the bank about the customers who are at the highest risk in leaving the bank. A customer churn prediction can be used here as churn and nonchurn customers are to be defined. Using ML, gap is to be resolved between churn and nonchurn customers. Different accuracy levels are achieved by classifiers using different data sheets. A novel approach K-nearest neighbor algorithm (KNN) is presented in which dataset is suitably grouped into training and testing models depending on weighted scales along with XGBooster algorithm for high and improved accuracy

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| Chapter 1 | Introduction | 08 |
|  | 1.1 Aim and Objectives of Project | 09 |
|  | 1.2 Scope | 09 |
|  |  |  |
| Chapter 2 | Literature Survey | 11 |
|  | 2.1 Data Acquisition  2.2 Data Preprocessing | 12  12 |
| Chapter 3 | Design | 13 |
|  | 3.1 Flowchart Diagram | 13 |
|  | 3.2 Hardware and Software Requirements | 14 |
| Chapter 4 | Methodology | 15 |
|  | 4.1 Project Module | 16 |
| Chapter 5 | Implementation Details | 17 |
|  | 5.1 Screenshots | 19-22 |
| Chapter 6 | Problem Timeline | 23 |
|  | 6.1 Timeline Chart | 23 |
| Chapter 7 | Applications | 24 |
| Chapter 8 | Conclusion  8.1 Future Scope | 25  25 |
| References | | 26 |

**CHAPTER - 1**

**Introduction**

Customers are surrounded by number of resources of information in today’s digitized environment, and they have all resources at the tip of their fingers. Smartphones, e.g., provide instant access to various branded products, mbanking, comparative information. And customer perspectives demand based on advanced technology, convenience reasons, price sensitivity, service quality reasons and socio factors Due to availability of different options as a boon of advanced technologies, customers keep on changing from one service provider to another. That is why for companies, it is difficult to retain and attract the customers and loses their wealth due to switching action by their customers. The procedure of customers leaving their service providers is called churn For banking sector, this customer churn prediction is the serious issue and gargantuan impact on the profit line of bankers. Thus, customer retention scheme can be targeted on high-risk customers who wish to discontinue their custom and switch to another competitor. To minimize the cost of bank sectors customer retention marketing scheme an accurate and prior identification of these customers is hypercritical. Customer churning is the estimate or analysis of degree of customers who turn to shift to an alternative. It is the most common problem witnessed in any industry. Banking is one such industry that focuses a lot on customer’s behavior by tracking their activities. It is very extortionate to add a new customer to the bank when compared to retention Companies can raise their profits by handling these customers. Hence, there is a need to keep up the existing customers, which will be achieved only by understanding the customer’s grievances of changing the bank. The paper presents a model to churn the bank customers using k-nearest neighbor (KNN) algorithm. This simple KNN algorithm is used to classify the customers into two classes, those who will leave the bank and those who will not leave. To enhance the accuracy, XGBooster algorithm is applied, whereas many research papers are available from various journals based on bank customer churn prediction, but techniques applied are decision tree logistic regression random forest unsupervised learning artificial neural network (ANN), data mining, neurocomputing

**1.1 Aim and Objectives of Project**

This is a classification project, since the variable to be predicted is binary (churn or loyal).The goal here is to model the probability of churn, conditioned on the customer features.

It is advantageous for banks to know what leads a client towards the decision to leave the company.Churn prevention allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible.

**1.2 Scope**

It is much more expensive to sign in a new client than keeping an existing one.It is advantageous for banks to know what leads a client towards the decision to leave the company.

[Churn prevention](https://www.neuraldesigner.com/solutions/churn-prevention) allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible. In this example, we use customer data from a bank to construct a predictive model for the likely churn clients.

**CHAPTER - 2**

**Literature Review**

From the above discussion, it is clear that customer retention is important for a company and for its business strategy. Customer churning becomes business intelligence to know which customers will shift or who will get retained. To achieve customer churning, companies started adapting machine learning techniques for customer churn prediction models. In this section, a few techniques are compared considering churn prediction. Data mining by author ‘Sen K’ aims to analyze large dataset by converting the sets of data into useful data. And a customer churn prediction model is developed and is measured using accuracy, sensitivity and specificity and Kappa’s statistics

The support vector machine (SVM) is the popular technique providing guide to the bank for customer strategy. SVM has larger probability of customer churns in the samples. With good number of plenty vectors, SVM provides good precision in predicting technique models. SVM gives high fitting accuracy rate of 0.59 by the author ‘Zhao Jing’ ‘Guoxun Wang’ focuses on the comparison of all techniques used to build credit card holder churn model for the banks in China based on multi-criteria decision algorithm and constructing techniques using PROMETHEE and TOPSIS methods . In MCDM algorithm, decision tree methods are implemented. ‘Shaoying Cui’ presents improved FCM algorithm as data mining algorithm to facilitate the banks with a new idea for predicting customer churn. It achieved accuracy rate of 80% for high-value customers and 83% for low-value customers. ‘Pradeep B’ proposed to construct a model for churn prediction for a company using logistic regression and decision trees techniques. In Pradeep’s approach there is a trial to retrieve the important factors of the customer churn that provides additional and useful knowledge which supports decision making . Alisa Bilal Zori´c applied a data mining technique ‘neural network’ in the software package ANN to predict churn in bank customer. Using this model, the reason of customer leaving the bank can be easily acquainted by entering the parameters. ‘Abinash Mishra’ proposed methodology of ensemble classifiers comprising bagging, boosting and random forest to predict customer churn for telecom industry. Random forest achieves high accuracy of 96% with low specificity and high sensitivity and low error rate ‘Ning Lu’ presented a paper in which an experimental evaluation proves that the boosting provides a good source of churn data, efficiently providing the customer churn model. The measures for churn prediction are calculated using a training set of customers over a period of six months ‘Hend Sayed’ presented a methodology of decision tree in which two packages ML and MLib were conducted, to evaluate accuracy, model training and model evaluation. They got effective result with ML package

**2.1 Data Acquisition**

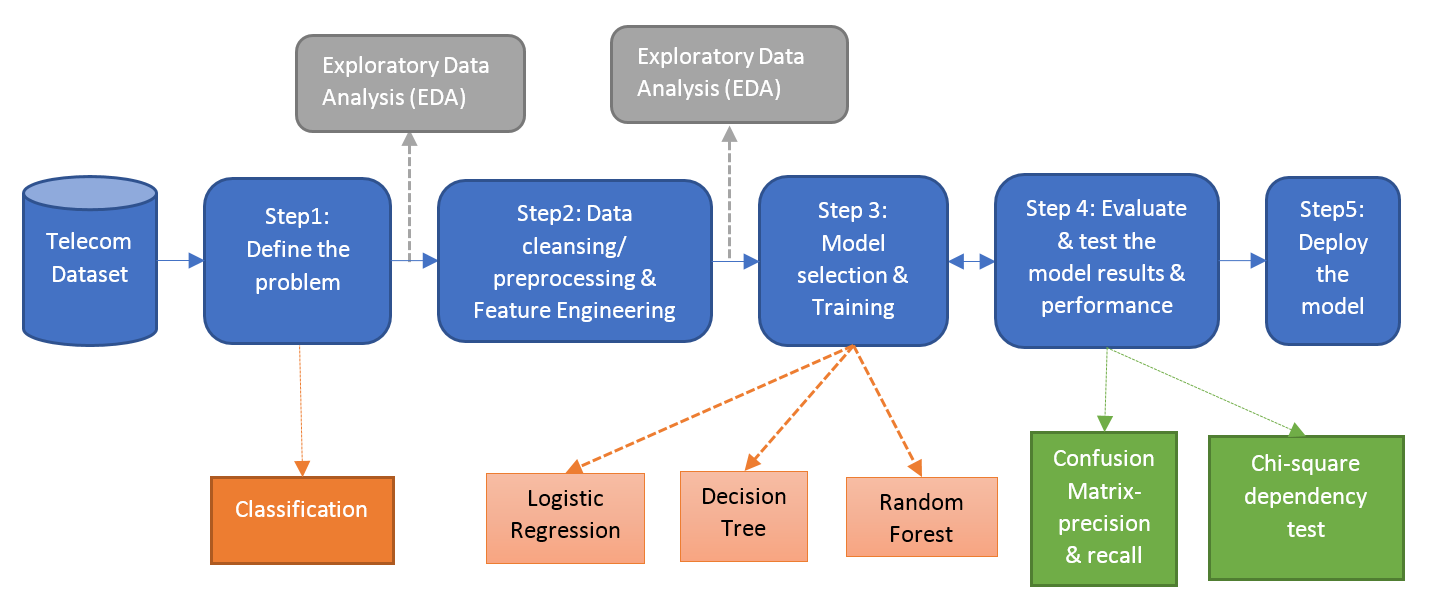
Dataset used for this supervised prediction is acquired from an online source. The target dataset is subjected to churning of customers of bank containing information about 10,000 customers with 14 features for each customer. The customers of the bank are identified as churn or loyal based on the potential features like credit score, age, gender, estimated salary, etc. A user of the bank is classified as loyal if he/she is active and remains with the bank. Customers are classified as churners if they switch to another bank. The variable exited in the dataset gives the actual status of the customer if he/she had switched to another bank

**2.2 Data Preprocessing**

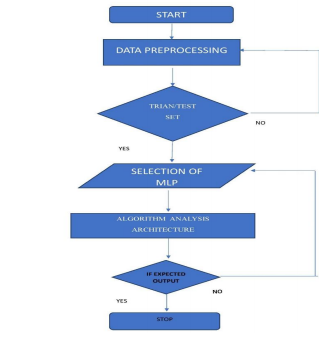
The process to identify the required independent variables for predicting the exit status of a customer and to predict the binary dependent variable ‘EXITED’ using the independent variables is data preprocessing. The dataset used for predicting churning of customers of a bank contains information about 10,000 customers with 14 features for each customer. These features include row number, customer id, surname, credit score, geography, gender, age, tenure, balance, number of products, has cr card, is active member, estimated salary, exited To predict the churning of customers, dataset is split suitable for training and testing. At this instance, splitting has 80% training rate and 20% testing rate . The value of this attribute will be 1 if the customer has left the bank and 0 if remained there. Feature scaling or data normalization is a technique used to standardize the range of independent variables in the dataset.

**CHAPTER - 3**

**Design**

****

**3.1Flowchart Diagram**



**3.2Hardware & software requirements**

* **Software Requirements**

1. Jupyter Notebook

2.Web Browser

* **Hardware Requirements**

1.Desktop

* **Languages Requirements**

1.Python

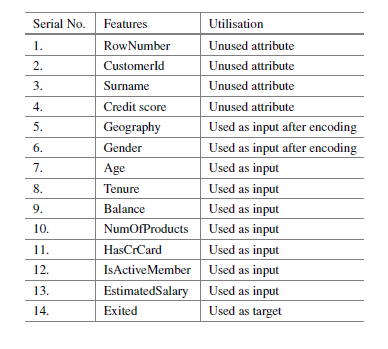
**CHAPTER - 4**

**Methodology**

In this project,whole focus is using flexible technique to boost the accuracy in customer churning process. So, along with K-nearest neighbors (KNN) algorithm, XGBoost algorithm is implemented. The block diagram is represented below to describe the whole process

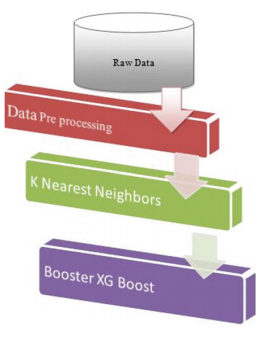
Utilization of

features of dataset



Block diagram of

build customer churn model



**4.1Project modules**

## 1. Application type

This is a [classification](https://www.neuraldesigner.com/learning/tutorials/neural-networks-applications#Classification) project, since the variable to be predicted is binary (churn or loyal).

The goal here is to model the probability of churn, conditioned on the customer features.

## 2. Data set

The data set contains information for creating our model. We need to configure three things here:

* Data source.
* Variables.
* Instances.

The data file [bank\_churn.csv](https://www.neuraldesigner.com/files/datasets/bank_churn.csv) contains 12 features about 10000 clients of the bank.

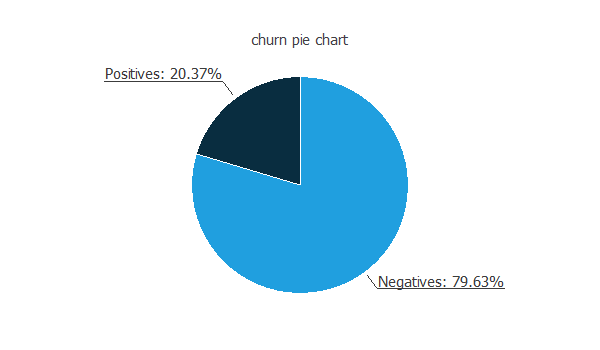
The features or [variables](https://www.neuraldesigner.com/learning/tutorials/data-set#Variables) are the following:

* **customer\_id**, unused variable.
* **credit\_score**, used as input.
* **country**, used as input.
* **gender**, used as input.
* **age**, used as input.
* **tenure**, used as input.
* **balance**, used as input.
* **products\_number**, used as input.
* **credit\_card**, used as input.
* **active\_member**, used as input.
* **estimated\_salary**, used as input.
* **churn**, used as the target. 1 if the client has left the bank during some period or 0 if he/she has not.

On the other hand, the [instances](https://www.neuraldesigner.com/learning/tutorials/data-set#Instances) are split at random into training (60%), selection (20%), and testing (20%) subsets.

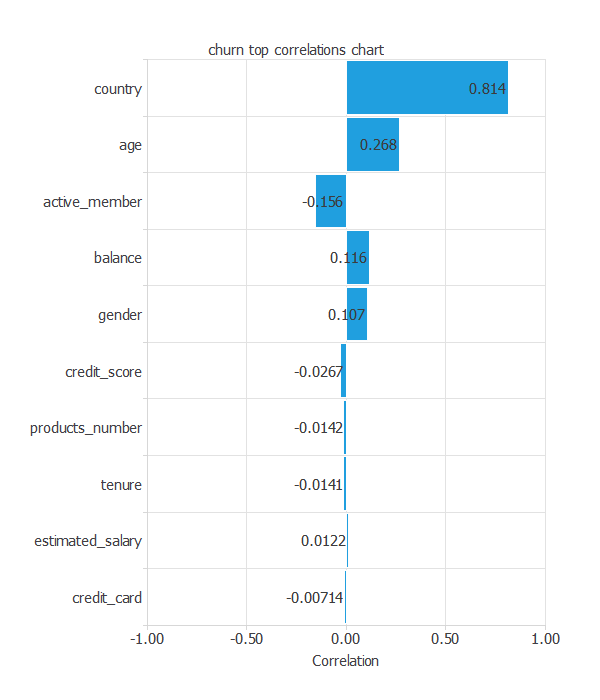
Once the variables and instances are configured, we can perform some analytics on the data.

The data [distributions](https://www.neuraldesigner.com/learning/tutorials/data-set#Distributions) tell us the percentages of churn and loyal customers.



In this data set, the percentage of churn customers is about 20%.

The [inputs-targets correlations](https://www.neuraldesigner.com/learning/tutorials/data-set#InputsTargetsCorrelations) might indicate which variables might be causing attrition.



From the above chart, we can see that the country has a great influence and that older customers have more probability of leaving the bank.

## 3. Training strategy

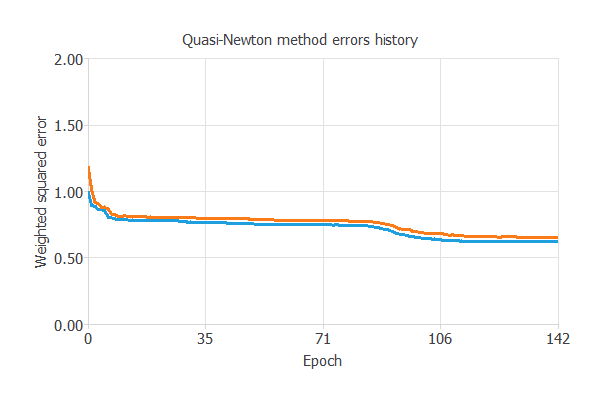
The [training strategy](https://www.neuraldesigner.com/learning/tutorials/training-strategy) is applied to the neural network to obtain the best possible performance. It is composed of two things:

* A loss index.
* An optimization algorithm.

The selected [loss index](https://www.neuraldesigner.com/learning/tutorials/training-strategy#LossIndex) is the [weighted squared error](https://www.neuraldesigner.com/learning/tutorials/training-strategy#WeightedSquaredError) with [L2 regularization](https://www.neuraldesigner.com/learning/tutorials/training-strategy#L2Regularization). The weighted squared error is very useful in applications where the targets are unbalanced. It gives a weight of 3.91 to churn customers and a weight of 1 to loyal customers.

The selected [optimization algorithm](https://www.neuraldesigner.com/learning/tutorials/training-strategy#OptimizationAlgorithm) is the [quasi-Newton method](https://www.neuraldesigner.com/learning/tutorials/training-strategy#QuasiNewtonMethod).

The following chart shows how the training (blue) and selection (orange) errors error decrease with the training epochs.

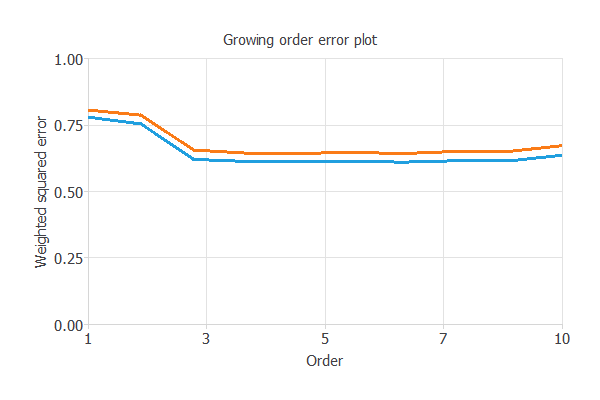


The final training and selection errors are **training error = 0.621 WSE** and **selection error = 0.656 WSE**, respectively. In the next section, we will try to improve the generalization performance by reducing the selection error.

## 4. Model selection

[Order selection](https://www.neuraldesigner.com/learning/tutorials/model-selection#OrderSelection) is used to find the complexity of the neural network that optimizes the generalization performance. That is the number of neurons that minimize the error in the [selection instances](https://www.neuraldesigner.com/learning/tutorials/data-set#SelectionInstances).

The following chart shows the training and selection errors for each different order after performing the [incremental order](https://www.neuraldesigner.com/learning/tutorials/model-selection#IncrementalOrder) method.



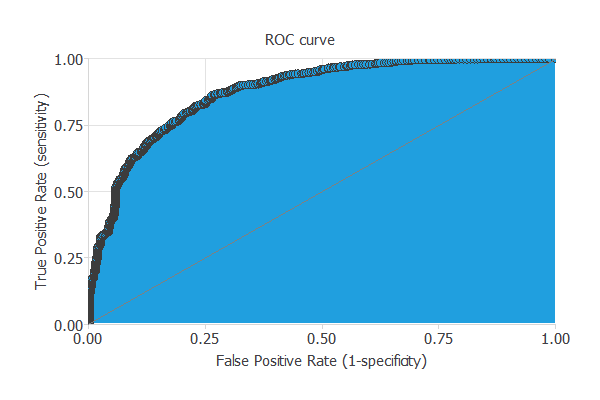
As the chart shows, the optimal number of neurons is 6, with **selection error = 0.643**.

The following figure shows the final network architecture for this application.

## 5. Testing analysis

The next step is to perform an exhaustive [testing analysis](https://www.neuraldesigner.com/learning/tutorials/testing-analysis) to validate the neural network's predictive capabilities.

A good measure for the precision of a binary classification model is the [ROC curve](https://www.neuraldesigner.com/learning/tutorials/testing-analysis#RocCurve).



We are interested in the area under the curve (AUC). A perfect classifier would have an AUC=1, and a random one would have AUC=0.5. Our model has an **AUC = 0.874**, which is great.

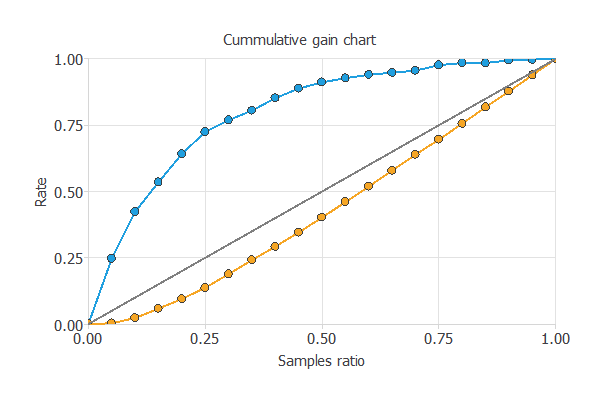
We can also look at the [confusion matrix](https://www.neuraldesigner.com/learning/tutorials/testing-analysis#ConfusionMatrix). Next, we show the elements of this matrix for a **decision threshold = 0.5**.

|  |  |  |
| --- | --- | --- |
|  | **Predicted positive** | **Predicted negative** |
| **Real positive** | 305 (15%) | 80 (4%) |
| **Real negative** | 344 (17%) | 1271 (63%) |

From the above confusion matrix, we can calculate the following [binary classification tests](https://www.neuraldesigner.com/learning/tutorials/testing-analysis#BinaryClassificationTests):

* **Classification accuracy: 78.8%** (ratio of correctly classified samples).
* **Error rate: 21.2%** (ratio of misclassified samples).
* **Sensitivity: 79.2%** (percentage of actual positive classified as positive).
* **Specificity: 78.7%** (percentage of actual negative classified as negative).

Now, we can simulate the performance of a retention campaign. For that, we use the [cumulative gain](https://www.neuraldesigner.com/learning/tutorials/testing-analysis#CumulativeGain) chart.



The above chart tells us that if we contact 25% of the customers with the highest probability of churn, we will reach 75% of the customers who will leave the bank.

### 6. Model deployment

Once we have tested the churn model, we can use it to evaluate the probability of churn of our customers.

For instance, consider a customer with the following features:

* credit\_score: 650
* country: France
* gender: Female
* age: 39
* tenure: 5
* balance: 76485
* products\_number: 2
* credit\_card: Yes
* active\_member: No
* estimated\_salary: 100000

The probability of churn for that customer is 38%.

**CHAPTER - 5**

**Implementation**

**Code:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

import pickle

df=pd.read\_csv('Churn\_Modelling.csv')

X=df.iloc[:,3:13]

Y=df.iloc[:,13]

X=X.drop(['Geography','Gender'],axis=1)

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.3,random\_state=0)

sc=StandardScaler()

x\_train=sc.fit\_transform(x\_train)

x\_test=sc.transform(x\_test)

lr = LogisticRegression()

lr.fit(x\_train,y\_train)

#using pickel to dump the codde

file = open('model.pkl','wb')

pickle.dump(lr,file)

file.close()

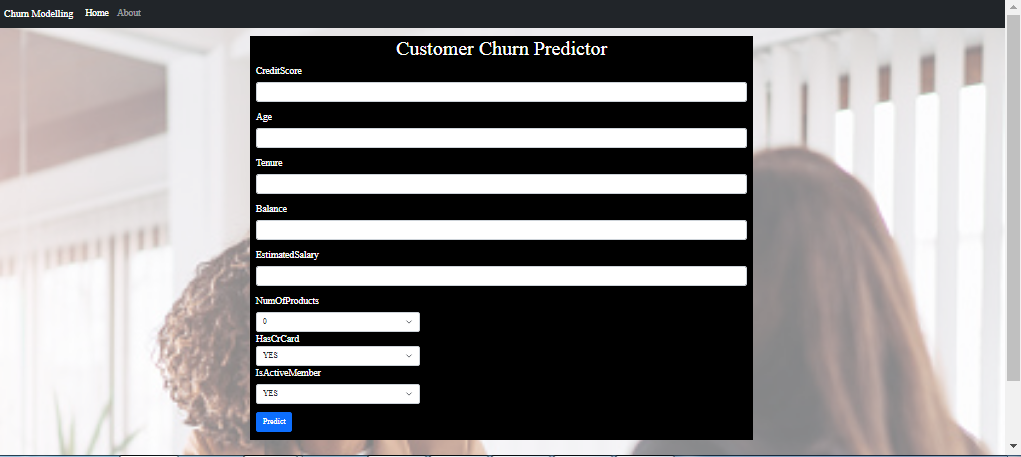
#using pickel to dump the codde

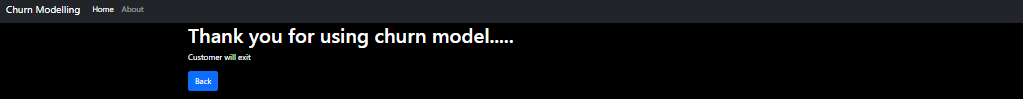
file1 = open('sc.pkl','wb')

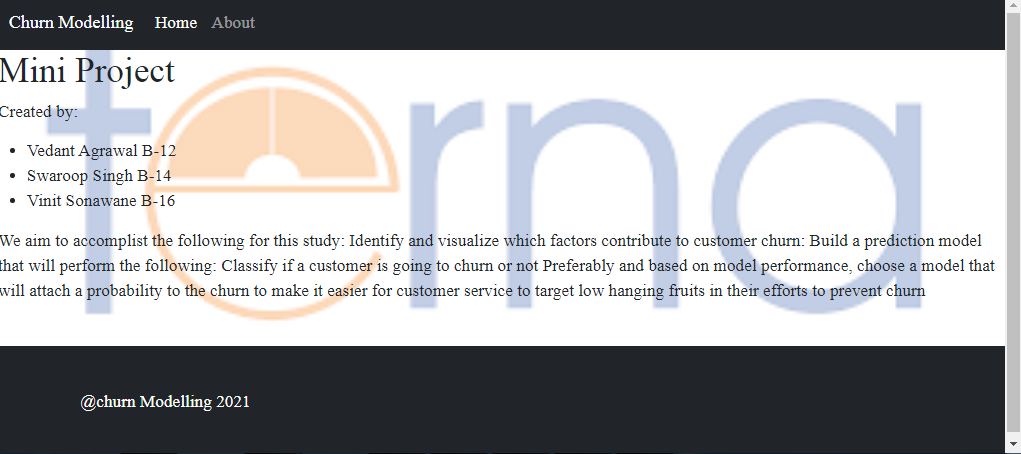
pickle.dump(sc,file1)

file1.close()

**5.1 Screenshots**

****

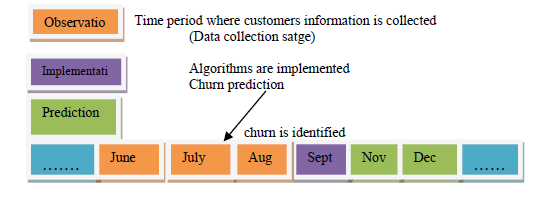




**CHAPTER - 6**

**Project Timeline**

**6.1 Time line chart**

****

**CHAPTER - 7**

**Applications**

In many applications, the recognition of colors is essential in order to separate or sort different models / components from one another. The color values are just as important for quality control. Color recognition is a standard task for which numerous machine vision solutions are implemented.

* It can recognize and detect colors and has many good new features in comparison with other color sensors. It is adequate for colorimeter measurement applications, such as medical diagnosis, color printing, computer color monitor calibration, and cosmetics, paint, textile and the process control of printing materials.
* Mounting of Car Antennas – Detecting signal cables in different colors

Nowadays, car antennas are very small and yet offer a variety of functions(10). In order to meet the requirements of e-mobility, a car antenna contains all sorts of technology whose installation is checked. Particularly important is the distinction of the signal lines to correct color affiliation. The camera examination takes place in two color matching steps:

(1) Distinction between the signal lines and the background and

(2) Delimitation of the signal lines red / brown to each other.

* Color recognition in the production of toothbrushes

Everyone knows the colorful variety of commercially available toothbrushes. The color may be indicative of the different degree of strength of a toothbrush. Within the production process, different toothbrushes are automatically visually inspected in a very short time in order to ensure, for example, the correct sorting according to hardness.

* Color recognition for caps and spray heads of deodorants

There are usually different varieties of a deodorant spray. The consumer recognizes his personal choice when buying the color of the cap. Each product type is therefore assigned a color for the closure cap. In the production of these caps, the color detection which ensures that the manufactured products correspond exactly to the color requirements of the customer.

**CHAPTER - 8**

**Conclusion**

In this project,we propose an effectivemodel of churn in bank industry. It combines the KNN with XGBoost algorithm to enhance the accuracy of the model; this proves the advantage of the technique used. XGBoost gives the best result in terms of accuracy, sensitivity and specificity. Boosting has given the increased accuracy of 86.85 with low error, high sensitivity and specificity. Organizations periodically calculate customer churn inmultiple aspects. Churning can be the number of customers lost, ratio or percentage of customers lost compared with total customers in bank. Churn can be calculated on quarter or annual basis. An accurate forecast can give insights on future using which a strategy can be formulated.

**8.1Future Scope**

Churn analysis is useful to any business with many customers, or to businesses with few, high-value customers. Which is to say, nearly every company. Companies in different industries use customer churn analytics for a variety of reasons:

* Financial services: Measure account holder lifecycle, detect users thinking of switching banks
* Consumer packaged goods: Develop a support model that encourages loyalty
* Consumer tech: Measure app churn
* Energy: Measure how much revenue is at risk of being lost to other providers
* Healthcare: Calculate the value of patients lost to other providers
* Insurance: Predict a user’s likelihood to close a policy
* Life sciences: Measure churn for device or equipment buyers
* Manufacturing: Measure churn for direct and downstream buyers
* Media and entertainment: Measure subscriber churn
* Retail and e-commerce: Predict when shoppers pose a high churn risk
* Telecommunications: Detect when customers are shopping other carriers
* Travel: Measure churn among repeat web visitors

**References**

1. N. Hashmi, N.A. Butt, M. Iqbal, Customer churn prediction in telecommunication in a decade review and classification. Int. J. Comput. Sci. Issues (IJCSI) **10**(5), 271–281 (September 2013)

2. V. Mahajan, R. Mishra, R. Mahajan, Review of data mining techniques for churn prediction in telecom. JIOS **37**(2), 183–197 (2015)

3. L. Yan, R.H. Wolniewicz, R. Dodier, Predicting customer behavior in telecommunications. 1094-7167/04© 2004 IEEE Published by the IEEE Computer Society (2004)

4. B. Kaderabkora, P. Malecek, Churning and labour market flows in the new EU member states. Int. Inst. Soc. Econ. Sci. 372–378 (2015)

5. S.A. Qureshi, A.S. Rehman, A.M. Qamar, A. Kamal, *Telecommunication Subscribers’ Churn Prediction Model Using Machine Learning* (IEEE, 2013), pp. 131–136

6. B. Mishachandar, K.A. Kumar, Predicting customer churn using targeted proactive retention. Int. J. Eng. Technol. **7**(2.27), 69–76 (2018)

7. K. Mishra, R. Rani, Churn prediction in telecommunication using machine learning, in *International Conference on Energy, Communication, Data Analytics and Soft Computing* *(ICECDS-2017)* (IEEE, 2017), pp. 2252–2257

8. E.M.L. Peters, G. Dedene, J. Poelmans, Understanding service quality and customer churn by process discovery for a multi-national banking contact center, in *Proceedings—IEEE 13th* *International Conference on Data MiningWorkshops*, *ICDMW 2013* (2013), pp. 228–233. Art. no. 6753925

9. N. Wang, D.X. Niu, Credit card customer churn prediction based on the RST and LS-SVM, in *Proceedings of the 2009 6th International Conference on Service Systems and Service* *Management*, *ICSSSM ’09* (2009), pp. 275–279. Art. no. 5174892

10. Y. Xie, X. Li, E.W.T. Ngai, W. Ying, *Customer Churn Prediction Using Improved Balanced Random Forests*. (Elsevier, Amsterdam, 2008), pp. 5445–5449

11. P. Spanoudes, T. Nguyen, Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors.