Mini Project Report on

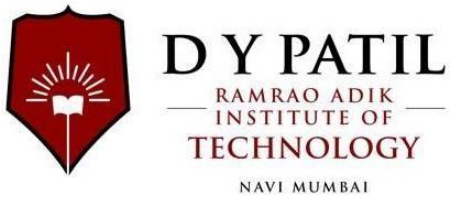
Customer Churn Prediction in Banking

by

Sanskruti Tikone 18CE1100
Jay Doshi 18CE1086
Pramey Dongre 18CE1045

Under the guidance of

Mrs Rajashree Shedge



Department of Computer Engineering

Ramrao Adik Institute of Technology

Dr. D. Y. Patil Vidyanagar, Nerul, Navi MumbaiUniversity of Mumbai

May 2021

Ramrao Adik Institute of Technology

Dr. D. Y. Patil Vidyanagar, Nerul,
Navi Mumbai

CERTIFICATE

This is to certify that Mini Project report entitled

Customer Churn Prediction in Banking

by

Sanskruti Tikone 18CE1100

Jay Doshi 18CE1086

Pramey Dongre 18CE1045

is successfully completed for Third Year Computer Engineering as prescribedby University of Mumbai.

Supervisor	Project Coordinator
(Mrs. Rajashree Shedge)	**(Dr. Bharti Joshi)**
Head of Department	Principal
(Dr. Leena Ragha)	**(Dr. Mukesh D. Patil)**

Mini Project Report Approval

This is to certify that the Mini Project entitled ***"Customer Churn Prediction in Banking"*** is a bonafide work done by ***Sanskruti Tikone, Jay Doshi*** and ***Pramey Dongre*** under the supervisionof ***Mrs. Rajashree Shedge***. This Mini Project has been approved for Third Year Computer Engineering.

Internal Examiner:

1……..
2……..

External Examiner:

1…….
2…….

Date:
Date:

DECLARATION

I declare that this written submission represents my ideas and does not involve plagiarism. I have adequately cited and referenced the original sources wherever others' ideas or words havebeen included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in mysubmission. I understand that any violation of the above will be cause for disciplinary action against me 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.

Date: _____

Sanskruti Tikone 18CE1100
Jay Doshi 18CE1086
Pramey Dongre 18CE1045

Abstract:

Customer Churn has become a major issue in numerous banks since it costs significantly more to get a new customer than retaining existing ones. With the utilization of a Customer Churn prediction model probable churners in a bank can be identified, and therefore the bank can make some move to keep them from leaving. Customer relationship and management is greatly affected by Customer Churn analysis and prediction as it improves benefits for an organization. This report presents a Random Forest machine learning model to predict Customer Churn, as the bank's customer data is enormous and largely imbalanced. The method was compared with Stochastic Gradient Descent, Logistic Regression and Decision Tree classifier regarding customer churn prediction for a commercial bank's customers. It was found out that the Random Forest model has the best precision, accuracy rate and gives a compelling estimation to bank's Customer Churn prediction.

Contents:

List of Tables:

List of Figures:

Introduction

1.1 Overview

With the growing competition in banking industry, banks are required to follow customer retention strategies while they are trying to increase their market share by acquiring new customers. Additionally, attracting new customers costs more to any company rather than retaining the old ones who are likely to produce more profit. Thus, banks should maintain their competitive advantage by taking the advantage of machine learning models to predict customer churn. Given a randomly sampled population of 10000 customers from three European-based banks, this project intends to propose an efficient predictive model for customer churn in banking industry, using different supervised classification techniques. Model performance, goodness of fit, feature selection, class imbalance and dealing with outliers will be discussed in the following sections.

1.2 Objectives

- Study the importance of data preprocessing, data normalization and feature selection.

  Information pre-handling is utilized for addressing complex structures with attributes, discretization of continuous attributes, binarization of attributes, changing discrete characteristics over to consistent, and managing absent and obscure trait esteems. Different representation procedures give important assistance in data pre-processing. Data pre-processing, like standardization, include extraction, and dimension reduction ease, is important to more readily achieve the characterization of information. The point of pre-preparing is to track down the most educational arrangement of highlights to improve the presentation of the classifier. Moreover, feature extraction is used to separate feature from the crude information to accomplish a reliable characterization. Feature extraction is the most basic piece of the sign characterization since the arrangement execution may be degraded if the features are not assigned well.

  Carefully analyze and assess six-month aggregate credit card usage volumes for active and churned users given by an anonymous financial provider.

  Our first client informational collection is from a credit card company, where we are capable to survey client gender, age, tenure, balance, number of products they are subscribed to, their estimated salary and on the off chance that they halted the membership or not.
  We can see our dataset yet we likewise need to ensure the data is spotless, so as a feature of the cleaning interaction, we take a gander at missing qualities and information types.
  At the point when we take a gander at the statistical insights, we see that the normal age of our clients, the average month client has been a part is and the estimated average salary.
  At the point when we take a gander at the gender and geographic distribution, we see that male client assessed average salary is higher than females in France and Spain, anyway in Germany female clients average salary is higher.

  At the point when we take a gander at the connection among age and credit score, the direct relationship is very weak in order to clearly define correlation.

- Carry out statistical analyses for various datasets; one from financial provider and others from accompanying theses.

  Mean: The principal technique that is utilized to play out the factual investigation is mean, which is more commonly referred to as the average. At the point when you're hoping to ascertain the mean, you include a list and afterward partition that number by the things on the list. At the point when this strategy is utilized it takes into account deciding the general pattern of an informational index, just as the capacity to get a quick and succinct perspective on the information. Clients of this strategy additionally advantage from the oversimplified and fast estimation. The factual mean is concocting the main issue of the information that is being handled. The outcome is alluded to as the mean of the information gave.

Standard deviation: Standard deviation is a strategy for factual examination that actions the spread of information around the mean. At the point when you're managing an exclusive requirement deviation, this focuses to information that is spread generally from the mean. Likewise, a low deviation shows that most information is in accordance with the mean and can likewise be known as the normal worth of a set. Standard deviation is for the most part utilized when you need to decide the scattering of information focuses (regardless of whether they're clustered).

Regression: With regards to statistics, regression is the connection between a dependent variable (the information you're hoping to quantify) and a free factor (the information used to anticipate the ward variable).It can likewise be clarified by how one variable influences another, or changes in a variable that trigger changes in another, basically circumstances and logical results. It infers that the result is reliant upon at least one factors. The line utilized in relapse investigation diagrams and outlines mean whether the connections between the factors are solid or frail, as well as showing patterns throughout a particular measure of time.

- Correlate and compare the results to know to which extent only data usage volumes could be used to predict churn.

1.3 Motivation

Customer churn has become a major issue in numerous banks since it costs significantly more to get another customer than holding existing ones. With the utilization of a customer churn expectation model conceivable churners in a bank can be recognized, and subsequently the bank can make some move to keep them from leaving. To set up a particularly model in a bank not many things must be thought of. How a churner in a bank is characterized, and which factors and techniques to utilize. We recommend that a churner for that c bank ought to be characterized as a customer who has not been dynamic throughout the previous three months dependent on the bank meaning of a functioning customer. Conduct and segment factors ought to be utilized as a contribution for the model, and either choice tree or calculated relapse utilized as a strategy.

1.4 Organization of Report

Chapter 1:
Chapter 1 gives the overall introduction regarding the Customer churn in banking. The objective of the project is to study the importance of data processing, data normalization and feature selection. Analysis of the data six-month aggregate credit card usage volumes for active and churned users given by an anonymous financial provider. Carrying out statistical analyses for various datasets; one from financial provider and others from accompanying theses. Which includes the Mean, Standard deviation and regression?

Chapter 2:
Chapter 2 is the Literature survey. The existing systems consists of the SVM-REF, ANTMiner+ALBA and the Logistic Regression. However, every model has its own limitations in order to overcome these limitations the following method has been implied.

Chapter 3:
Chapter 3 gives a complete overview of our Proposed System which contains the problem statement , methodology/techniques we have used till now to build our system, basic design of our system and finally the implementation details.

Chapter 4:
Chapter 4 provides the result and analysis of the implemented model along with its accuracy , precision etc.

Chapter 5:
Chapter 5 is the conclusion regarding the customer churn prediction in the Banking System

Chapter 2

Literature Survey

2.1 Existing Systems

M.A.H. Farquad [1] proposed a hybrid approach to overcome the drawbacks of general SVM model which generates a black box model (i.e., it does not reveal the knowledge gained during training in human understandable form). The hybrid approach contains three phases:

1) SVM-RFE (SVM-recursive feature elimination) is employed to reduce the feature set.

2) dataset with reduced features is then used to obtain SVM model and support vectors are extracted.

3) rules are then generated using Naive Bayes Tree . The dataset used here is bank credit card customer dataset (Business Intelligence Cup 2004) which is highly unbalanced with 93.24% loyal and 6.76% churned customers. The experimental showed that the model does not scalable to large datasets.

- Wouter Verbeke [2] proposed the application of Ant-Miner+ and ALBA algorithms on publicly available churn prediction dataset in order to build accurate as well as comprehensible classification rule-sets churn prediction models.
- Ant-Miner+ is a high performing data mining method based on the principles of Ant Colony Optimization which allows to include domain knowledge by imposing monotonicity constraints on the final rule-set.
- The advantages of Ant-Miner+ are high accuracy, comprehensibility of the generated models and the possibility to demand intuitive predictive models. Ant-Miner+ results in less sensitive rule-sets, but allows to include domain knowledge, and results in comprehensible rule-sets

Ning Lu [3] proposed the use of boosting algorithms to enhance a customer churn prediction. Logistic regression is used as a basis learner, and a churn prediction model is built on each cluster, respectively. The experimental results suggest that boosting algorithm provides a good separation of churn data when compared with a single logistic regression model.

In general, the performance of rotation-based ensemble classifier depends upon:

(i) The performance criteria used to measure classification performance and

(ii) The implemented feature extraction algorithm.

Our literature study concludes that following algorithms are ideal for customer churn problem :

a) Decision Tree. b) Logistic Regression. c)Neural Networks.

2.2 Limitations of Existing System

Table 2.1 (Limitations of existing system in customer churn)

SVM-RFE	AntMiner+ALBA	Logistic Regression
The results of using SVM-RFE model are very good, and this method is quickly taken as a benchmark feature selection algorithm.	AntMiner + and ALBA provide a relatively low number of rules.	Logistic Regression model works best when combined with other boosting algorithms.
However, it does not take into account the correlation probably hidden between features during the feature selection process.	This algorithm runs slower for huge datasets.	Customer data in banking sector is highly skewed so LR model has to be combined with lot many models.

Chapter 3

Proposed System

3.1 Problem Statement

Context: -

- "Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [Kaggle Data set]

Content: -

- Each row represents a customer, each column contains customer's attributes described on the column Metadata.
- The data set includes information about: RowNumber , CustomerID, Surname ,CreditScore, Geography ,Gender ,Age, Tenure, Balance ,Number of Products, HasCrCard ,IsActiveMember ,Estimated Salary and Exited[Target]
- Customer account information — how long they've been a customer, Credit Score, Estimated Salary and Demographic info about customers — location[France ,Spain ,Germany] , gender, age range, and Tenure.

3.1 Proposed Methodology/Techniques

- These are the models that we will try to use now:
  - Stochastic Gradient Descent (SGD) classifier
  - Logistic Regression
  - Random Forest Classifier

a) Stochastic Gradient Descent (SGD) classifier: -

- The word 'stochastic 'signifies a framework or a cycle that is connected with an random likelihood. Subsequently, in Stochastic Gradient Descent, a few test entries are chosen randomly rather than the entire data for every cycle.
- One data point is selected at random by SGD from the entire data set at each iterative step.This reduces the computations significantly.
- Generally "Mini- batch" gradient descent, or a small number of data-points are sampled rather than a single point at each iteration. It attempts to maintain a balance between computational speed of SGD and Gradient Descent goodness.

Advantages of Stochastic Gradient Descent

1. It easily accommodates in the memory as it is a solitary training model being handled by the computer network.
2. It is computationally quick as just one example is prepared in each turn.
3. It converges quicker for bigger datasets because of frequent updates in the parameters.
4. Due to constant refreshes, the means taken towards the minima of the loss function have fluctuations that can assist with escaping the local minimums of the loss function (in the event that the calculated position ends up being the local minimum).

Disadvantages of Stochastic Gradient Descent

1. The steps taken to reach minimum are very noisy because of frequent updates. Time and again, this can lean the gradient descent into different directions.
2. Because of this, it may take more time for the minima to achieve convergence.
3. Frequent updates are costly due to utilizing all assets for computing each test sample in turn.
4. It forfeits the benefit of vectorized operations as it manages just a single case at a time.

b) Logistic Regression: -

- Logistic regression is perhaps the most well-known Machine Learning algorithms, which goes under the Supervised Learning procedure. It is utilized for foreseeing a dependent variable utilizing a given arrangement of independent variables.
- Logistic regression predicts the yield of a dependent variable. Hence, the result should be an absolute or discrete value. It very well may be either Yes or No, 0 or 1, valid or False, and so forth. Yet as opposed to giving the specific value as 0 and 1, it gives the probabilistic qualities which lie somewhere in the range of 0 and 1.
- Logistic Regression is very much like the Linear Regression with the exception of that how they are utilized. Linear Regression is utilized for Regression problems, while Logistic Regression is utilized for tackling classification problems.
- In Logistic regression, rather than fitting a regression line, we fit an "S" shaped function(logistic), which predicts two greatest qualities(0 or 1).

Advantages of Logistic Regression: -

1. Logistic Regression is one of the easiest ML algorithms and is not difficult to carry out yet gives incredible preparing productivity sometimes. Additionally, because of these reasons, preparing a model with this calculation doesn't need high calculation power.
2. The anticipated boundaries (trained weights) give derivation about the significance of each component. The direction of association for example positive or negative is additionally given. So, we can utilize logistic regression to discover the connection between the components.
3. Logistic Regression yields all around adjusted probabilities alongside classification results. This is a benefit over models that only give the final classification as results. On the off

chance that a training model has a 95% likelihood(probability) for a class, and another has a 55% likelihood(probability) for a similar class, we get a conclusion about which preparing models are more exact for the defined problem.

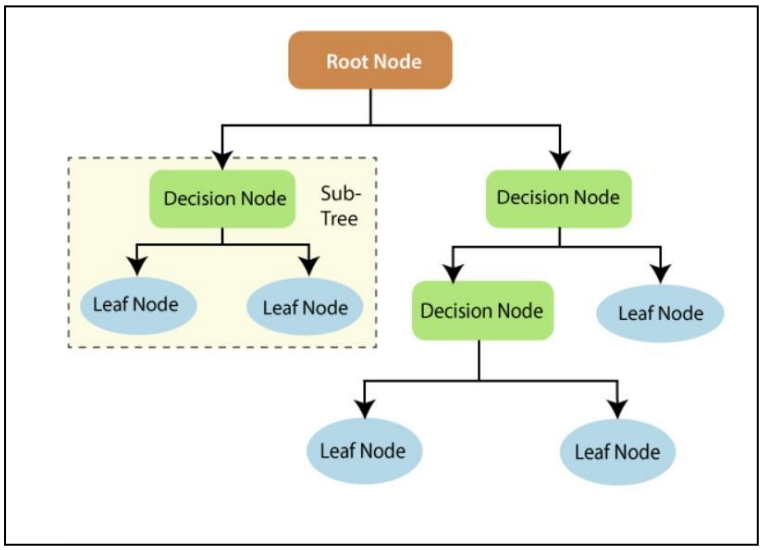
4. Maybe than straight away beginning with an intricate model, logistic regression is now and again utilized as a benchmark model to compute performance, as it is comparatively fast and simple to carry out.

Disadvantages of Logistic Regression: -

1. Non-Linear issues can't be settled with logistic regression since it possesses a linear decision surface. Linearly separable data is infrequently found in genuine situations. Thus, the change of non-linear highlights is required which should be possible by expanding the quantity of highlights to such an extent that the data turns out to be linearly separable in larger dimensions.
2. It is hard to catch complex connections utilizing logistic regression. All the more remarkable and complex methodology, for example, Neural Networks can undoubtedly beat this methodology.
3. Just significant and important highlights ought to be utilized to fabricate a model in any case the probabilistic expectations made by the model might be erroneous and the model's prescient value may corrupt.

c) Random Forest Classifier: -

- Random forest, as its name infers, comprises of huge number of individual decision trees that work as a group. Every individual tree in the random forest lets out a class prediction and the class with the most votes become our model's output (see figure underneath)



Working of Random Forest Classifier

Figure 3.1

- The principal idea driving random forest is a straightforward yet amazing one — the wisdom of crowd. In information science talk, the reason that the random forest model functions nicely is: huge number of moderately uncorrelated models (trees) working as a council will beat any

of the solitary constituent models.

- The little correlation within models is the key. Very much like how investments with low relation index (like stocks and securities) meet up to frame a portfolio that is more noteworthy than the total amount of its parts, uncorrelated models can deliver troupe forecasts that are more precise than any of the individual expectations.

- The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

Advantages of Random Forest Classifier: -

1. Random forest can take care of both sort of issues that is classification and regression and does a fair assessment at the two fronts.
2. One of advantages of Random Forest which excites us more than anything is, the ability to handle huge informational datasets with higher dimensionality. It can deal with a large number of input factors and so it is known to be as one of the dimensionality reduction method. Further, the model yields significance of variable, which can be a convenient component.
3. It has a powerful technique for predicting missing data and keeps up the accuracy when enormous extent of the data is absent.
4. It has techniques for adjusting blunders in data sets where classes are imbalanced.
5. The capacity of the above can be stretched out to unlabelled data, prompting outlier detection, data views and unsupervised clustering.

Disadvantages of Random Forest Classifier: -

1. Model interpretability: Random forest models cannot be interpreted; they are like black boxes.
2. For extensively large data sets, the size of the trees can take up a lot of memory.
3. It is generally prone to overfitting, so you must tweak the hyperparameters.

3.4 Hardware and Software required

3.4.1 Hardware Requirements

- Intel Core i3 , AMD Ryzen 3 1300Xor higher, (i5 10^{th} gen 3.5GHz recommended)
- 8GB RAM
- 100 GB hard free drive space

3.4.2 Software Requirements

- Google Chrome (or any other browser).
- Python 3.7 or higher.

3.5 Implementation Details

Implementation Environment: -

GOOGLE COLABORATORY

- Colaboratory, or “Colab” for short, is a product from Google Research.
- Colab permits anyone to compose and execute self-assertive python code through the program, and is particularly appropriate to AI, information investigation and training.
- All the more actually, Colab is a facilitated Jupyter notebook service that requires no arrangement to utilize, while giving free admittance to processing assets including GPUs.
- Colab is an open source tool.

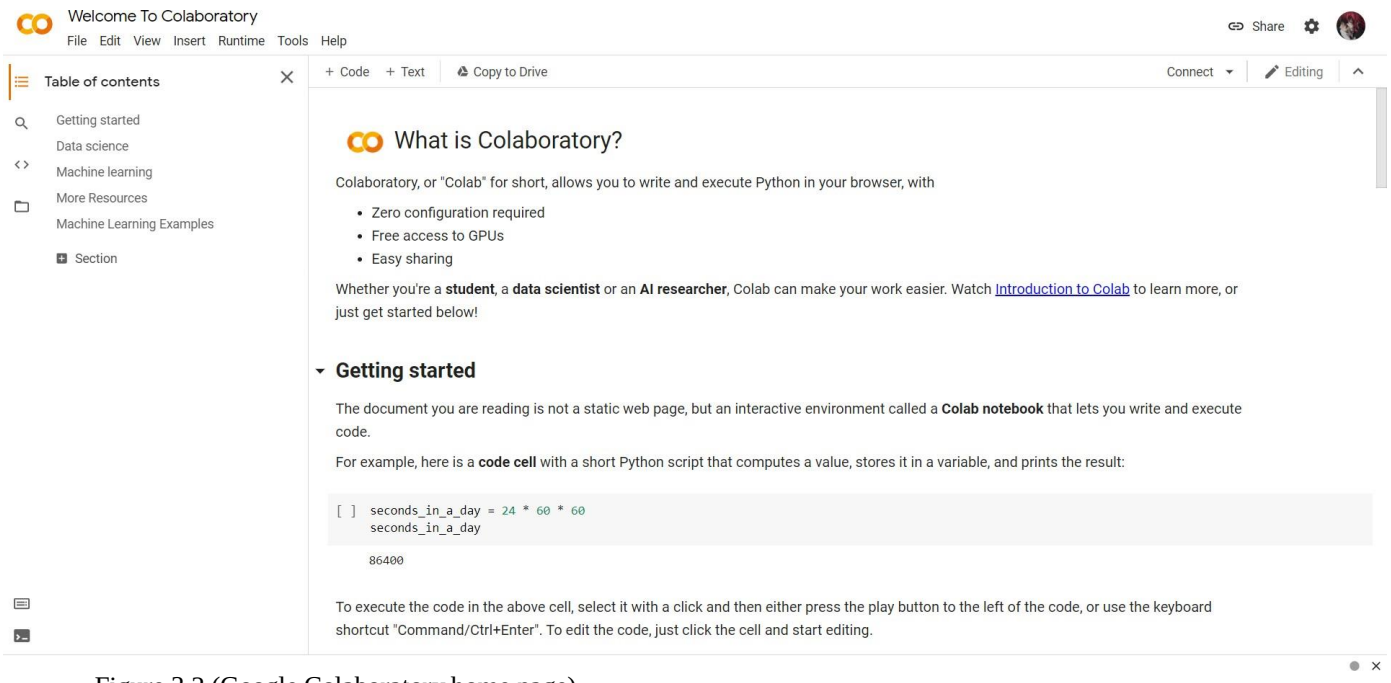


Figure 3.2 (Google Colaboratory home page)

- **STEP 1: Data selection.**

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CustomerI	Surname	CreditScor	Geography	Gender	Age	Tenure	Balance	NumOfPrc	HasCrCard	IsActiveMe	Estimated	Exited
2	1.6E+07	Hargrave	619	France	Female	42	2	0	1	1	1	101349	1
3	1.6E+07	Hill	608	Spain	Female	41	1	83807.9	1	0	1	112543	0
4	1.6E+07	Onio	502	France	Female	42	8	159661	3	1	0	113932	1
5	1.6E+07	Boni	699	France	Female	39	1	0	2	0	0	93826.6	0
6	1.6E+07	Mitchell	850	Spain	Female	43	2	125511	1	1	1	79084.1	0
7	1.6E+07	Chu	645	Spain	Male	44	8	113756	2	1	0	149757	1
8	1.6E+07	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
9	1.6E+07	Obinna	376	Germany	Female	29	4	115047	4	1	0	119347	1
10	1.6E+07	He	501	France	Male	44	4	142051	2	0	1	74940.5	0
11	1.6E+07	H?	684	France	Male	27	2	134604	1	1	1	71725.7	0
12	1.6E+07	Bearce	528	France	Male	31	6	102017	2	0	0	80181.1	0
13	1.6E+07	Andrews	497	Spain	Male	24	3	0	2	1	0	76390	0
14	1.6E+07	Kay	476	France	Female	34	10	0	2	1	0	26261	0
15	1.6E+07	Chin	549	France	Female	25	5	0	2	0	0	190858	0
16	1.6E+07	Scott	635	Spain	Female	35	7	0	2	1	1	65951.7	0

Dataset contains of 13 columns and 10,000 rows.
Dataset Source : www.kaggle.com

Figure 3.3

- **STEP 2: Data preparation / pre-processing.**

2.1 - Dropping useless features.
For example: CustomerID , Surname.

```
data.head()
```

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Figure 3.4

2.2 – Checking for and removing null values (if any) .

```
# We can check if the pandas dataframe 'data' has any null values in each of its column using the isnull() function.
# Furthermore, the sum() function tells us the total null values in each column.

data.isnull().sum()
```

```
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

Surprisingly, our dataset contains no null values.
Therefore there is no need for pre-processing step.
Figure 3.5

- **STEP 3 : Exploratory Data Analysis**

3.1 – Finding percentage of customers churned and retained.

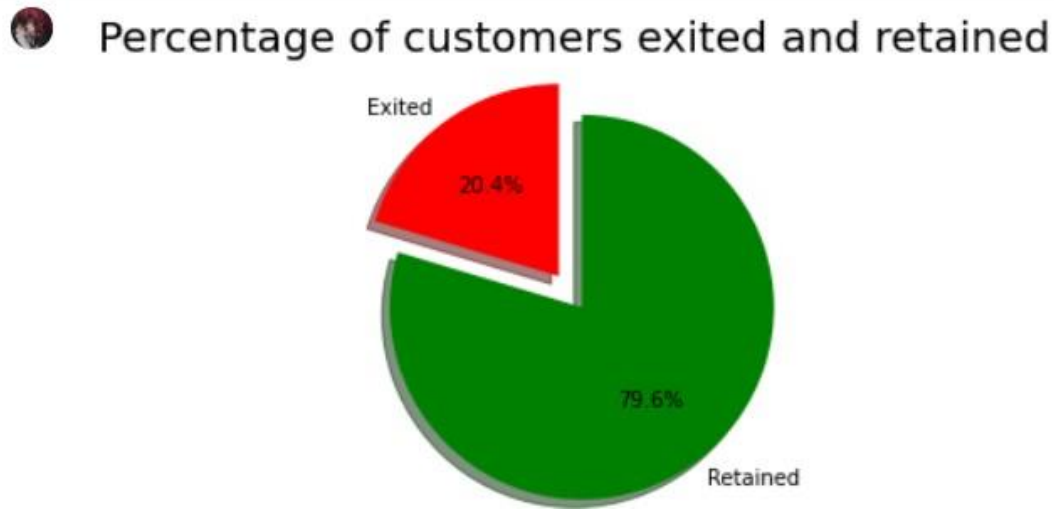


Figure 3.6

3.2 - Count column plots to map the dependence of 'Exited' column on categorical features

a) Dependence on geography:

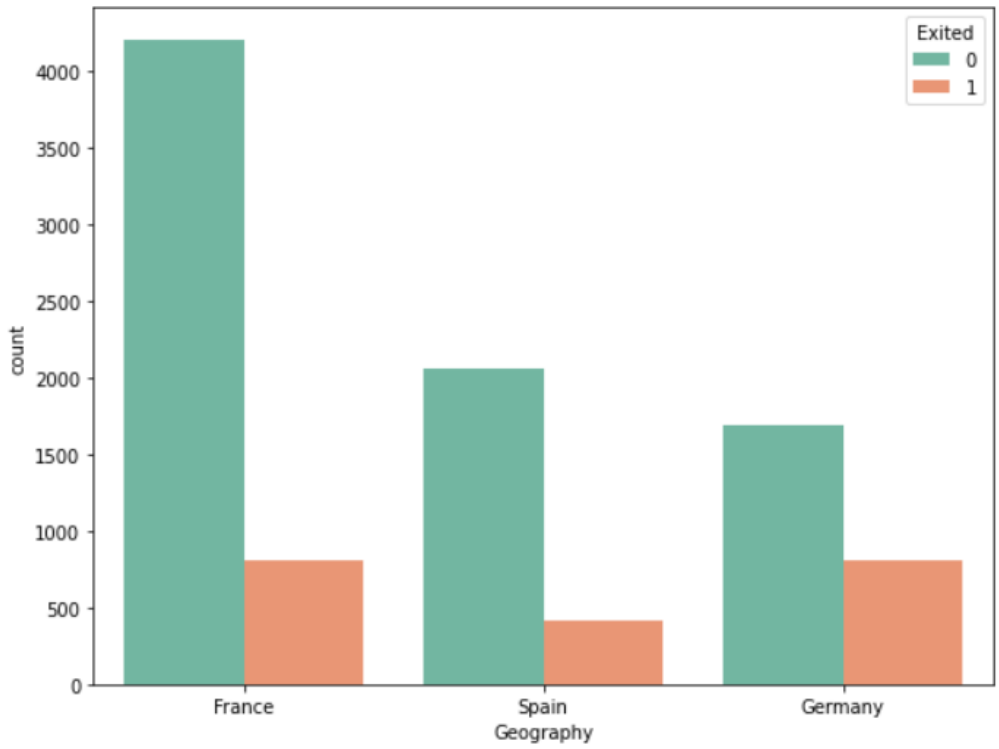


Figure 3.7

b) Dependence on Gender:

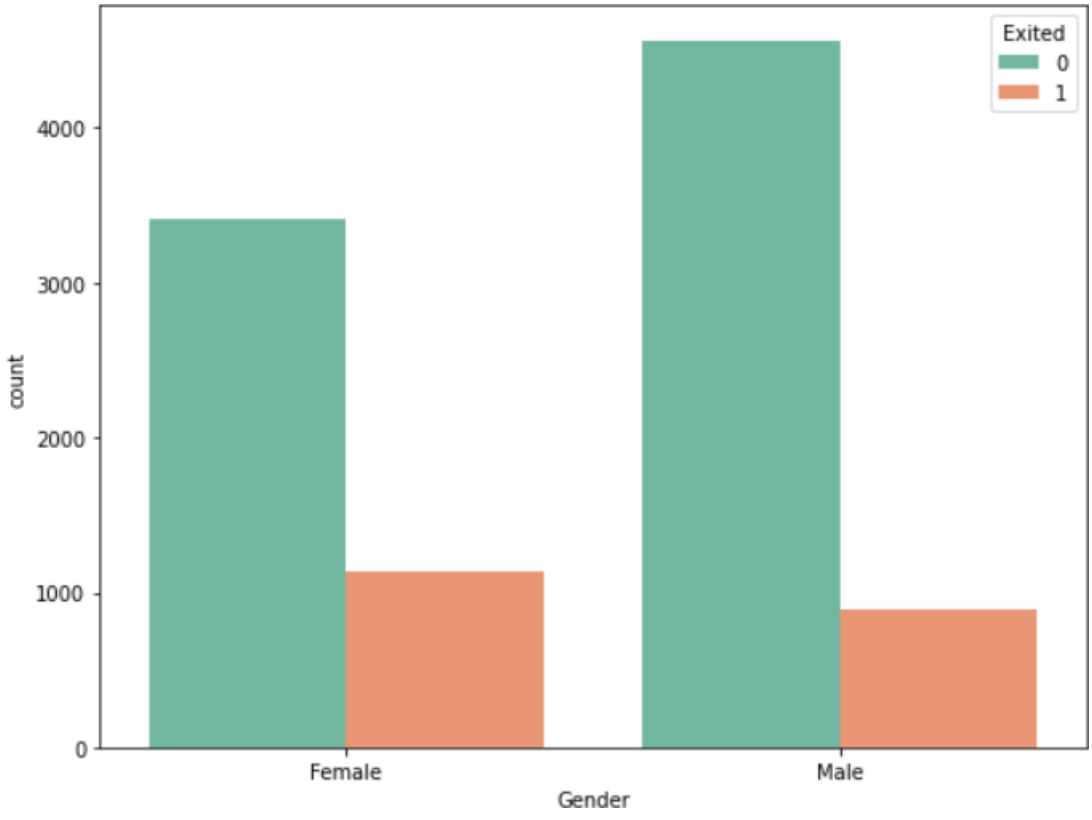
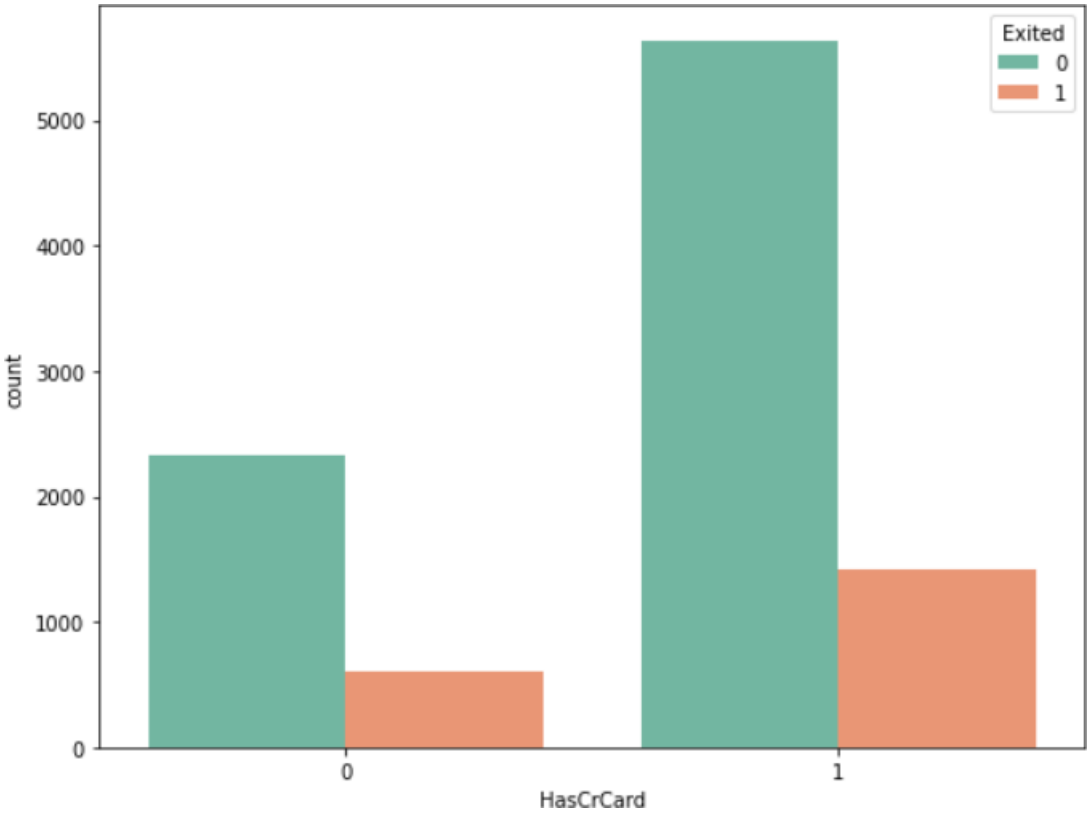


Figure 3.8

c) Dependence on owning of Credit card Figure 3.9



d) Dependence on whether an Active member or not:

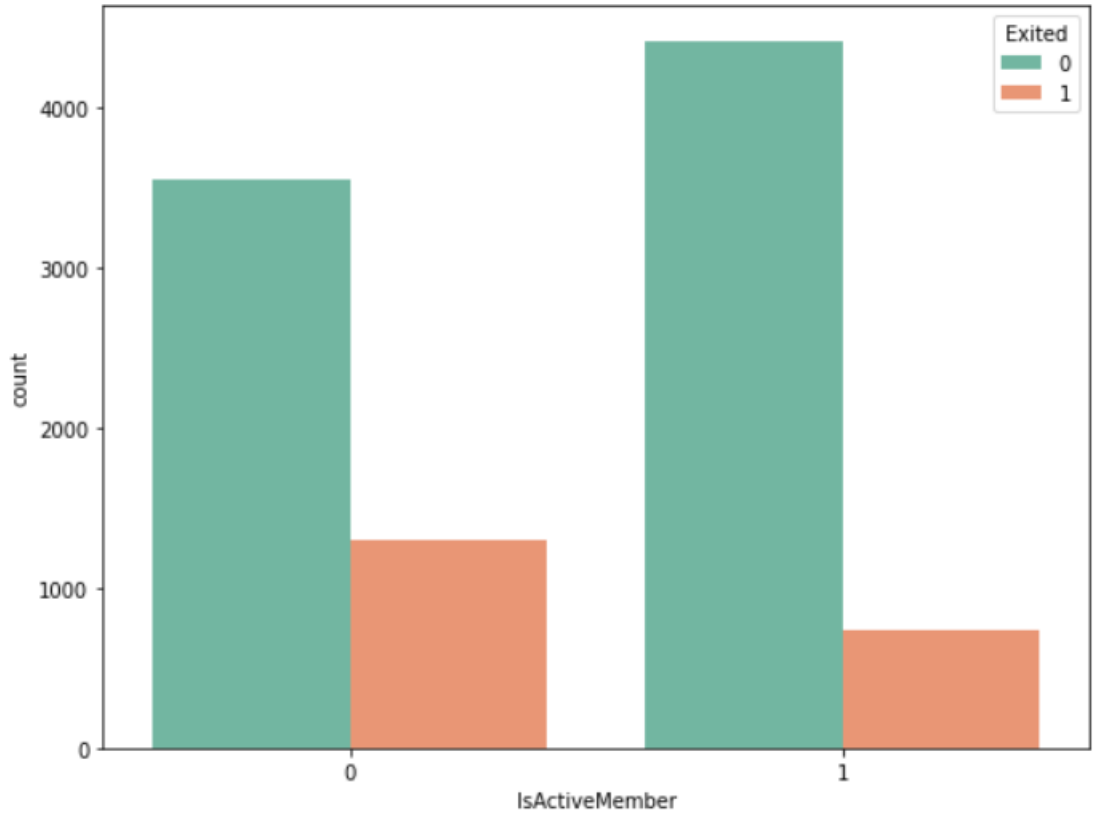


Figure 3.10

e) Dependence on Number of products owned

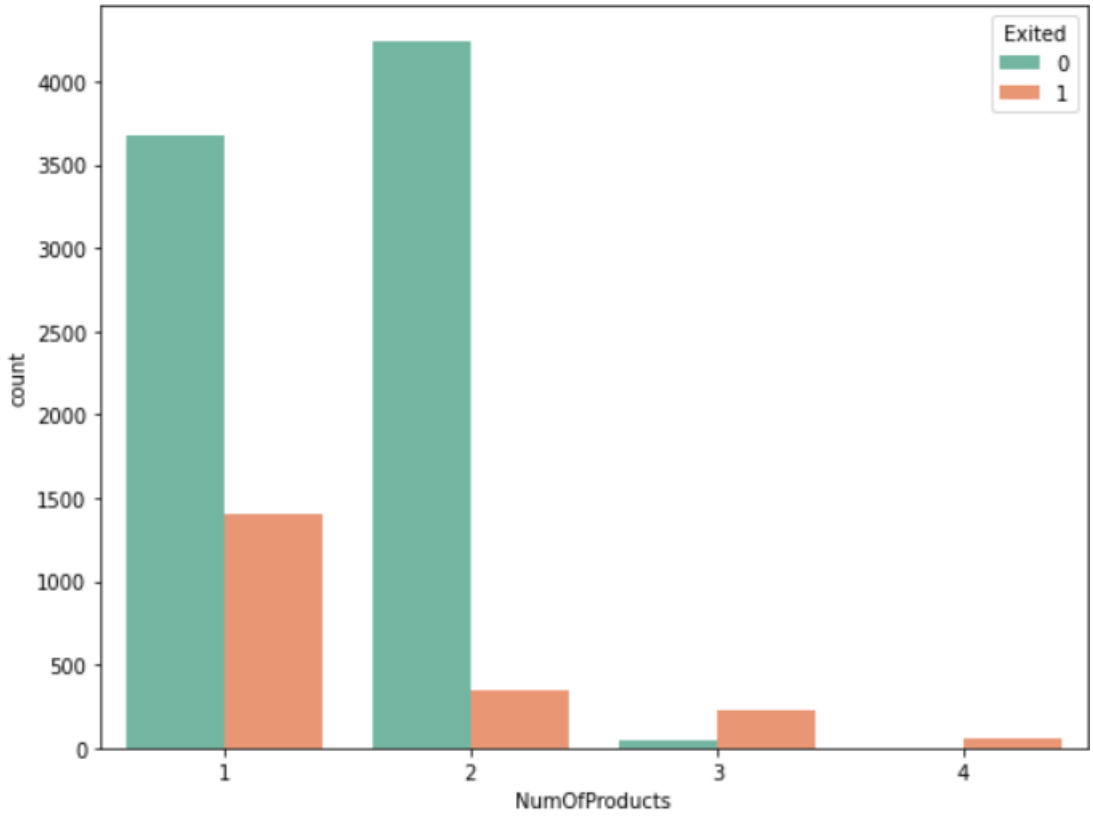


Figure 3.11

f) Dependence on Tenure:

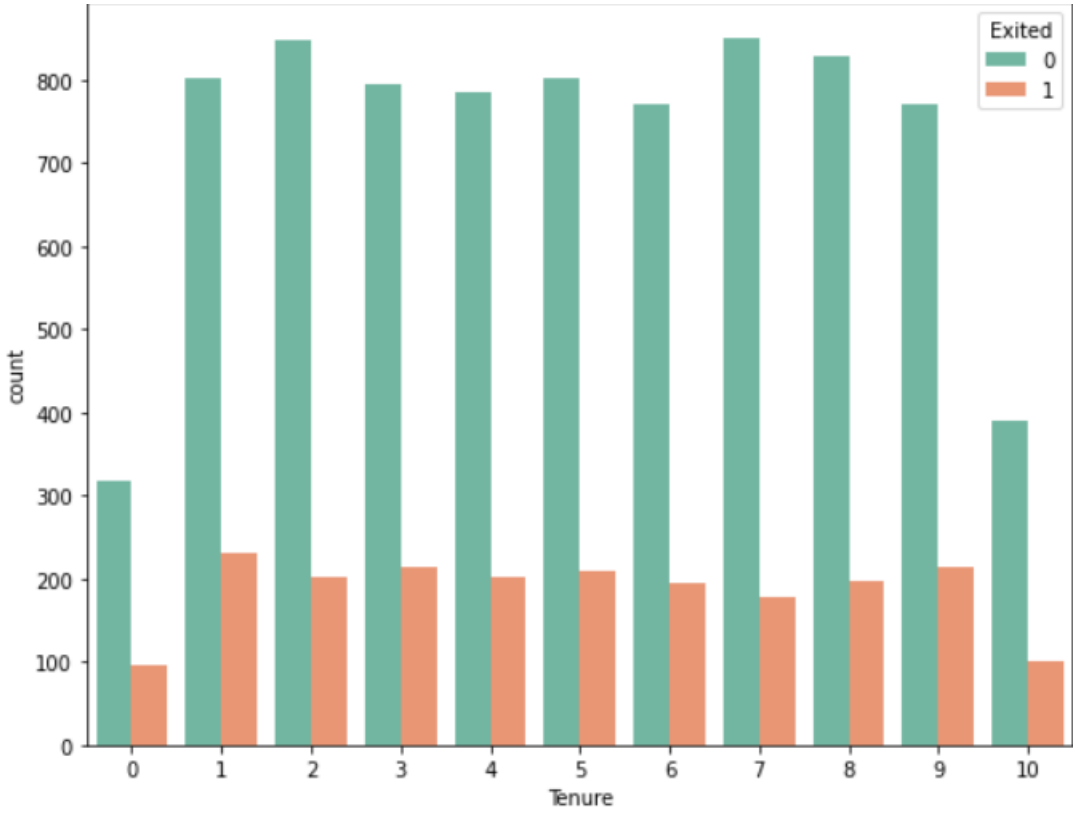


Figure 3.12

Observations

- **Geography:** Ideally for an evenly-distributed data, if the amount of people from a place is the majority, then the majority of churning should also be within that group. However, it is not so in this case as we see that number of exited people who belong to Germany is almost equal to the number of exits from France.
- **Gender:** We can clearly see the **Female customers** had more exits than the male customers.
- **Credit cards:** It is generally expected that people who have more interactions and products of the bank, would likely be retained for a longer time. However, we can see that people who have credit cards have more exits than those who do not own credit cards.
- **Active Member:** This is an expected observation. We can see that inactive members have been churned more than members who are active.
- **Number of Products:** This is also an expected observation, where we see that customers who own more products from the bank are likely to be retained for a longer time than those who own less products.
- **Tenure:** We see that the tenure of a customer does not really tell us much if that customer is likely to be churned or not. Initially, it looks like new joinees and older people (10 years) have been churned less. However, on a closer analysis we can see that the overall number of retained customer are significantly less in both these cases. As a result, we can probably conclude that new joinees and older customers may be more likely to be churned as their churn rate (percentage) is likely to be higher than other tenure rates.

3.3 - Count column plots to map the dependence of 'Exited' column continuous and numerical feature

a) Dependence on Credit Score:

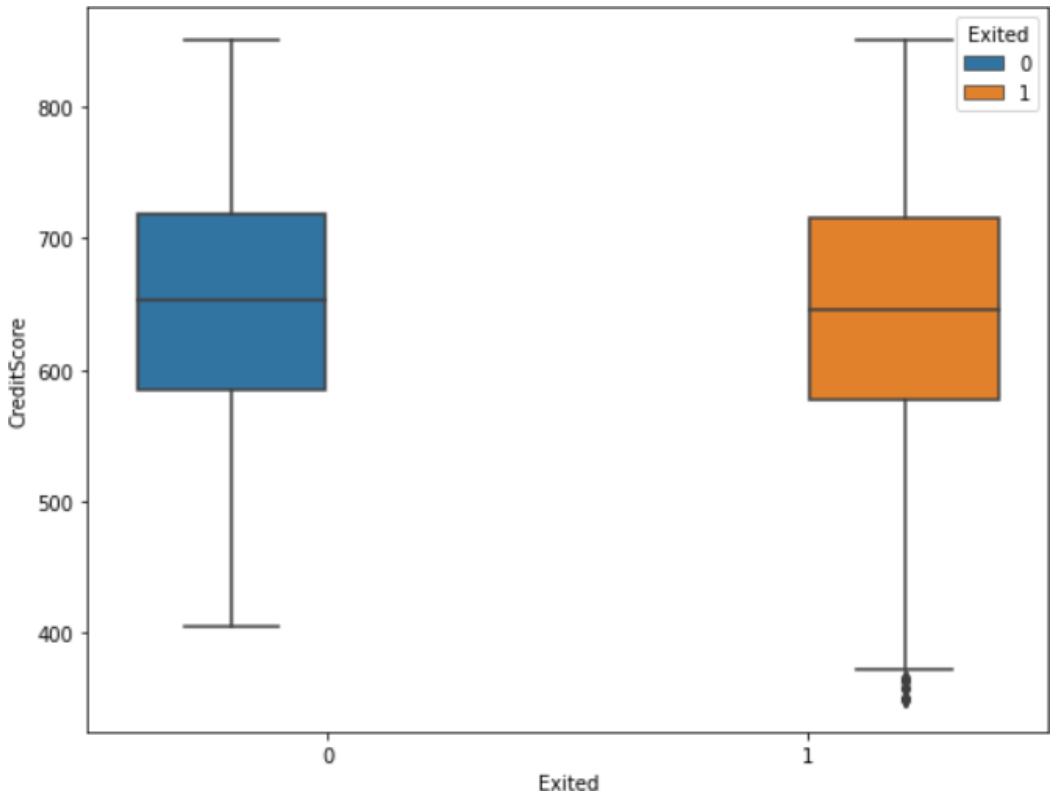
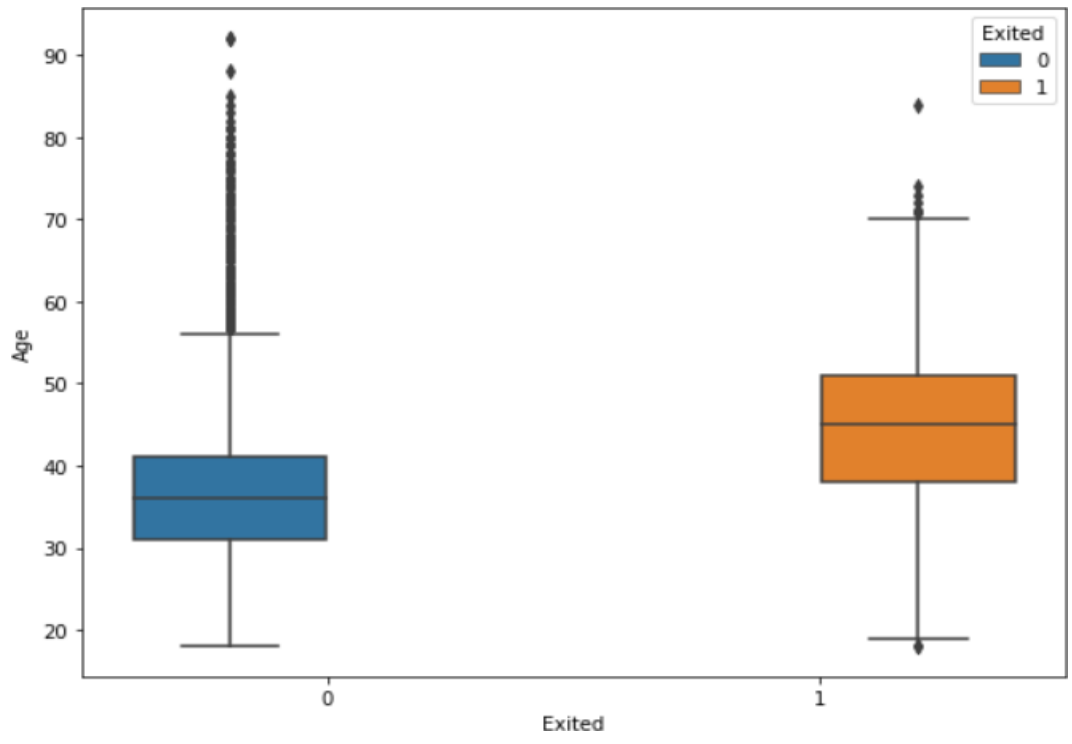


Figure 3.13

b) Dependence on Age: Figure 3.14



c) Dependence on Balance:

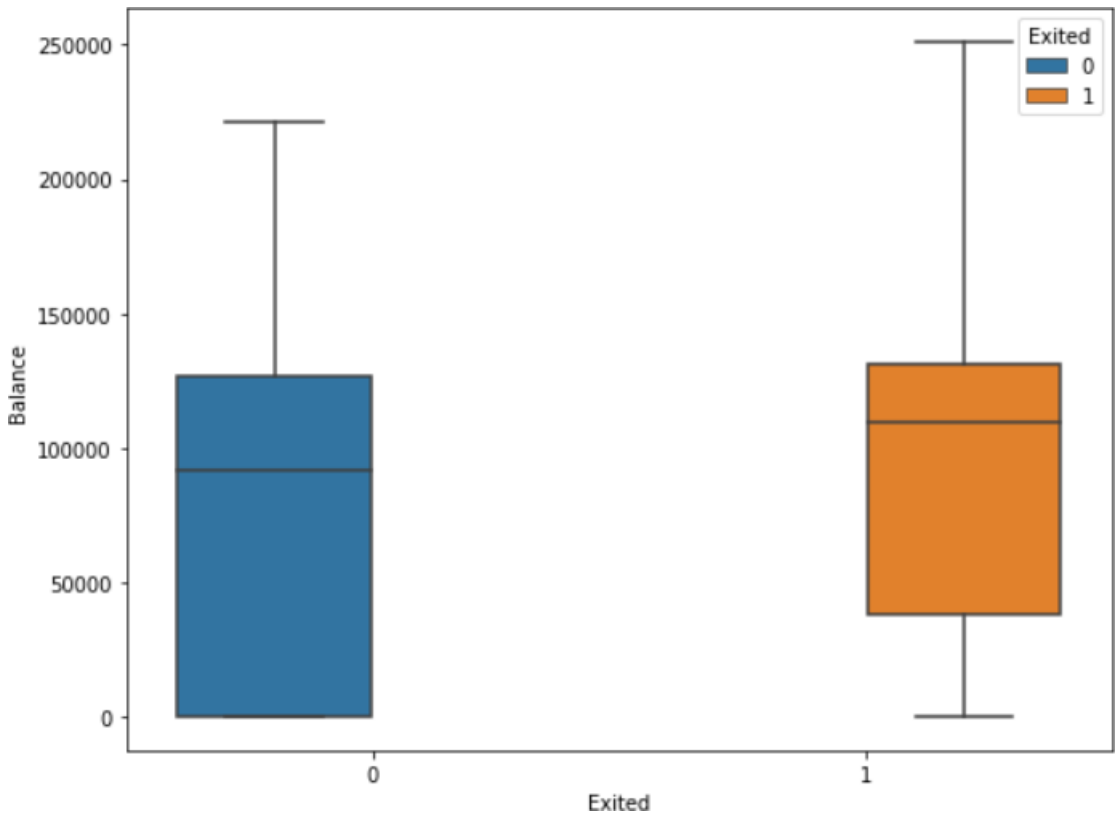


Figure 3.15 d) Dependence on Estimated Salary:

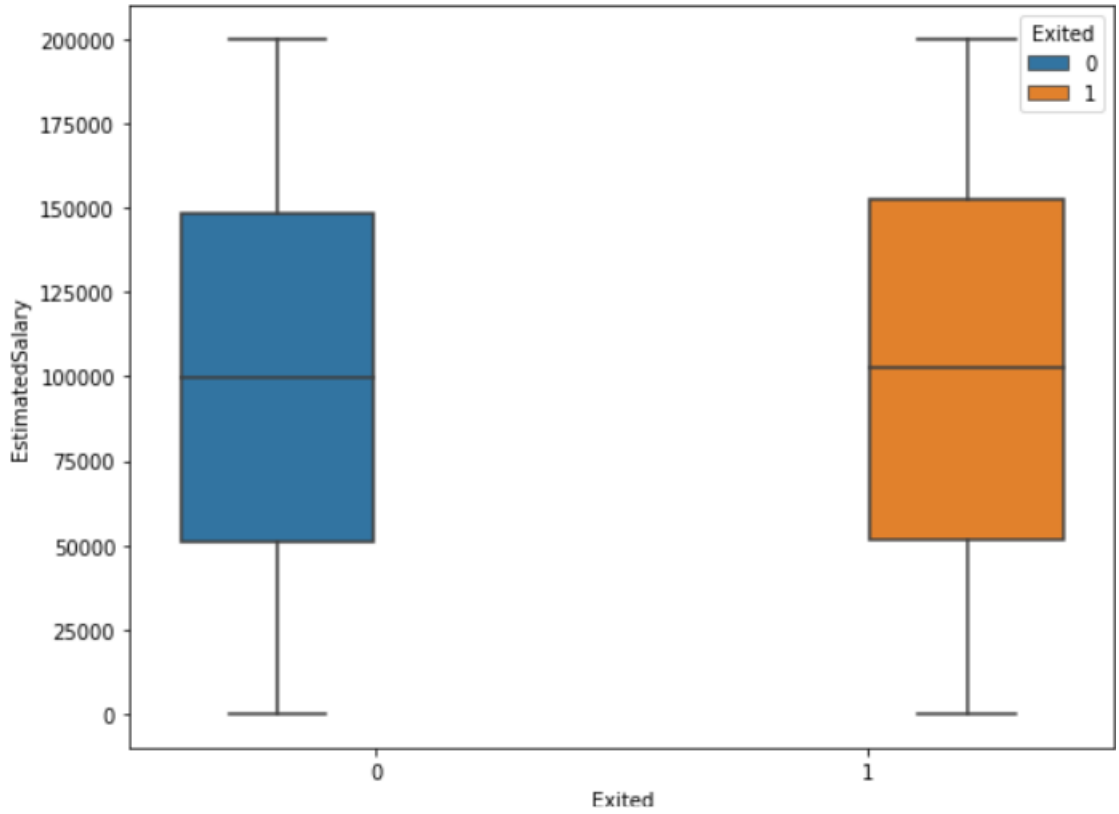


Figure 3.16

e) Dependence on Number of products:

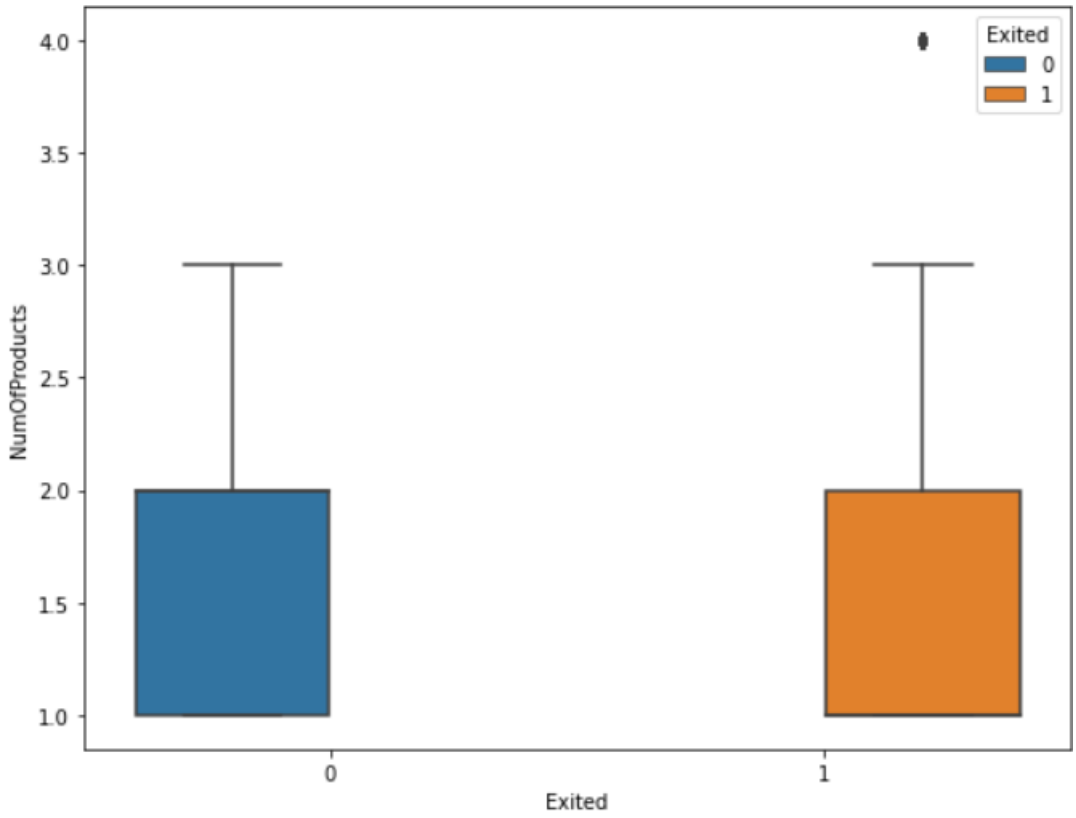
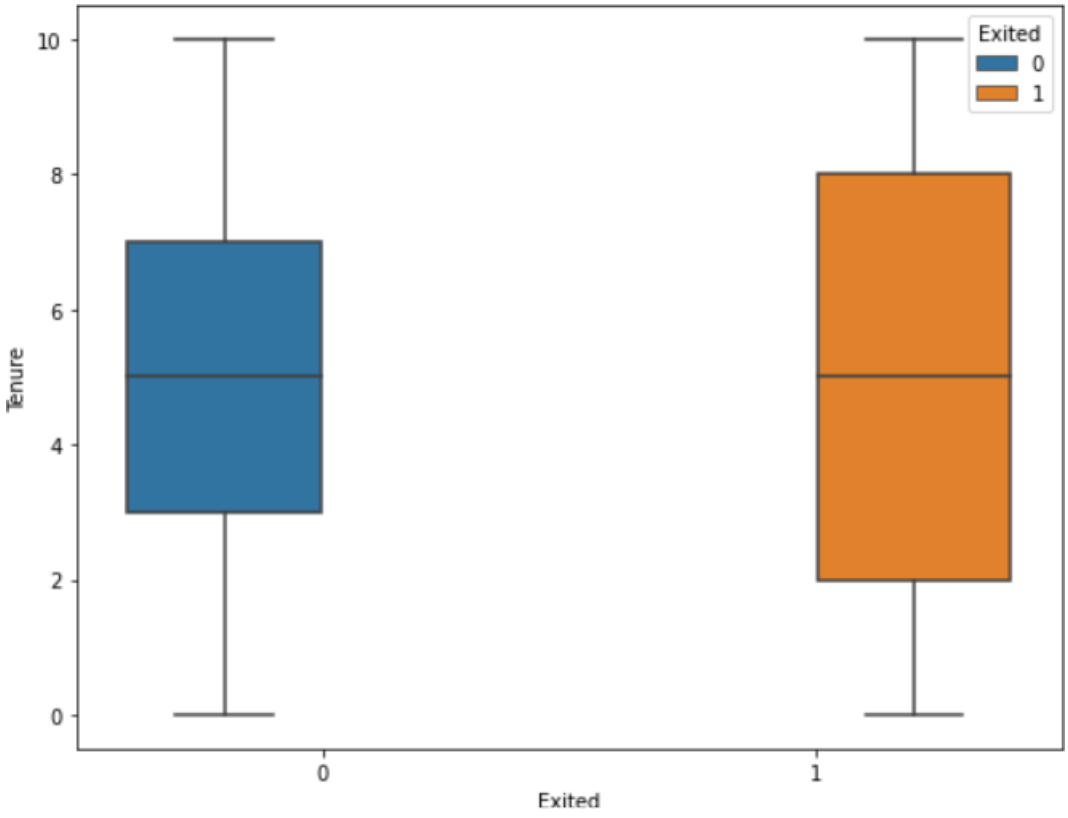


Figure 3.17

f) Dependence on Tenure: Figure 3.18



Observations

- **Credit Score:** We can see that Credit Score does not have much effect on the customer churn.
- **Age:** Here we can see that the older customers are more likely to be churned from the bank. This is most probably to keep a younger manpower in the organization.
- **Balance:** When it comes to Balance, we see that the bank is losing a significant number of customers with high balance in their accounts. This is likely to affect the bank's capital as well.
- **Estimated Salary:** Estimated Salary does not seem to affect the customer churn much.
- **Number of Products:** We see that the number of products also does not seem to affect the customer churn.
- **Tenure:** For tenure, as we can see here too, customer belonging more to the two extreme tenure groups (new joinees and older ones) are more likely to be churned.

3.4 - Split the data into training and testing set:

- The data set is split into data_train and data_test.
- data_train consists of 80% data (8000 rows) chosen randomly.
- data_test consists of remaining 20% data. (2000 rows)

3.5 - Introducing negative relation between numerical categorical values:

	Exited	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	HasCrCard	IsActiveMember	Geography	Gender
8018	1	632	23	3	122478.51	1	147230.77	1	-1	Germany	Male
9225	0	594	32	4	120074.97	2	162961.79	1	1	Germany	Female
3854	0	687	33	9	135962.40	2	121747.96	1	-1	Germany	Male
2029	0	520	33	4	156297.58	2	166102.61	1	1	France	Male
3539	0	667	42	6	0.00	1	88890.05	1	-1	France	Male

Figure 3.19

3.6 – Converting categorical object data into one-hot vectors:

Geography_Germany	Geography_France	Geography_Spain	Gender_Male	Gender_Female
1	-1	-1	1	-1
1	-1	-1	-1	1
1	-1	-1	1	-1
-1	1	-1	1	-1
-1	1	-1	1	-1

Figure 3.20

3.7 – Normalizing the discreet values using Min-max scaling:

Formula: Xsc= (X−Xmin) / (Xmax−Xmin)

CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
0.564	0.067568	0.3	0.488160	0.000000	0.736166
0.488	0.189189	0.4	0.478581	0.333333	0.814829
0.674	0.202703	0.9	0.541903	0.333333	0.608740
0.340	0.202703	0.4	0.622952	0.333333	0.830534
0.634	0.324324	0.6	0.000000	0.000000	0.444435

Figure 3.21

STEP 4 : Modelling and Testing

4.1 – Importing the necessary libraries:

Important libraries
- sklearn.model_selection : GridSearchCV

Models
- sklearn.ensemble : RandomForestClassifier

Metrics
- sklearn.metrics : accuracy_score
- sklearn.metrics : confusion_matrix
- sklearn.metrics : classification_report
- sklearn.metrics : roc_auc_score
- sklearn.metrics : roc_curve

4.2 – Finding the best parameters to apply to Random Forest Classifier using **GridSearchCV**:

GridSearchCV : -
- It belongs to the sklearn.model_classifier package.
- It performs exhaustive search over specified parameter values for an estimator.
- Important members are fit, predict.
- GridSearchCV implements a “fit” and a “score” method. It also implements “score_samples”, “predict”, “predict_proba”, “decision_function”, “transform” and “inverse_transform” if they are implemented in the estimator used.
- The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

Output of grid search model selection:

```
[INFO] Time taken: 1644.6 seconds.

0.8644999999999999
{'max_depth': 9, 'max_features': 8, 'min_samples_split': 6, 'n_estimators': 50}
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=9, max_features=8,
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=6,
                       min_weight_fraction_leaf=0.0, n_estimators=50,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Figure 3.22

4.3 - Training the Random Forest model using optimal parameters:
Accuracy = (TP+TN)/total
Precision = TP/predicted yes

Confusion Matrix :

```
array([[6269,  113],
       [ 709,  909]])
```

Figure 3.23

Output:-

```
[INFO] Random Forest classifier:

              precision    recall  f1-score   support

           0       0.90      0.98      0.94      6382
           1       0.88      0.56      0.69      1618

    accuracy                           0.90      8000
   macro avg       0.89      0.77      0.81      8000
weighted avg       0.89      0.90      0.89      8000
```

Figure 3.24

4.4 – Implementing the RF model on test data:

- Output of this stage is the ROC (**R**eceiver **O**perating **C**haracteristic) curve.
- When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (**Area Under The Curve**) ROC (**Receiver Operating Characteristics**) curve.
- It is one of the most important evaluation metrics for checking any classification model's performance.
- It tells how much the model is capable of distinguishing between classes.

The Higher the AUC, the better the model is at distinguishing between *Churned Customers* from those *Customers who are Retained*.

Chapter 4

Results and Discussion

4.1 Result and Analysis

We have used the following formulae for calculation: -

$$\text{Precision} = \frac{tp}{tp + fp}$$

When it predicts yes, the person has Exited, how often is it actually correct?

$$\text{Recall} = \frac{tp}{tp + fn}$$

When it is actually yes, the person has Exited, how often does it predict correctly?

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Weighted average between precision and recall. Useful when dealing with unbalanced samples.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

The sum of true positives and true negatives divided by the total number of samples. This is only accurate if the model is balanced. It will give inaccurate results if there is a class imbalance.

- We get the following confusion matrix with respect to the Y_train and y_pred values.

```
[ ]  confusion_matrix(Y_train, y_pred)

     array([[6269,  113],
            [ 709,  909]])
```

Figure 4.1

- Further we generate a classification report: -

```
[ ]  print('[INFO] Random Forest classifier:\n')
     print(classification_report(data_train.Exited, rf_model.predict(data_train.loc[:, data_train.columns != 'Exited'])))

     [INFO] Random Forest classifier:

                   precision    recall  f1-score   support

                0       0.90      0.98      0.94      6382
                1       0.88      0.56      0.69      1618

         accuracy                           0.90      8000
        macro avg       0.89      0.77      0.81      8000
     weighted avg       0.89      0.90      0.89      8000
```

Figure 4.2

Here we have achieved an accuracy of 90%.
We have also plotted an **ROC curve** (**receiver operating characteristic curve**) which is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

ROC curve : -

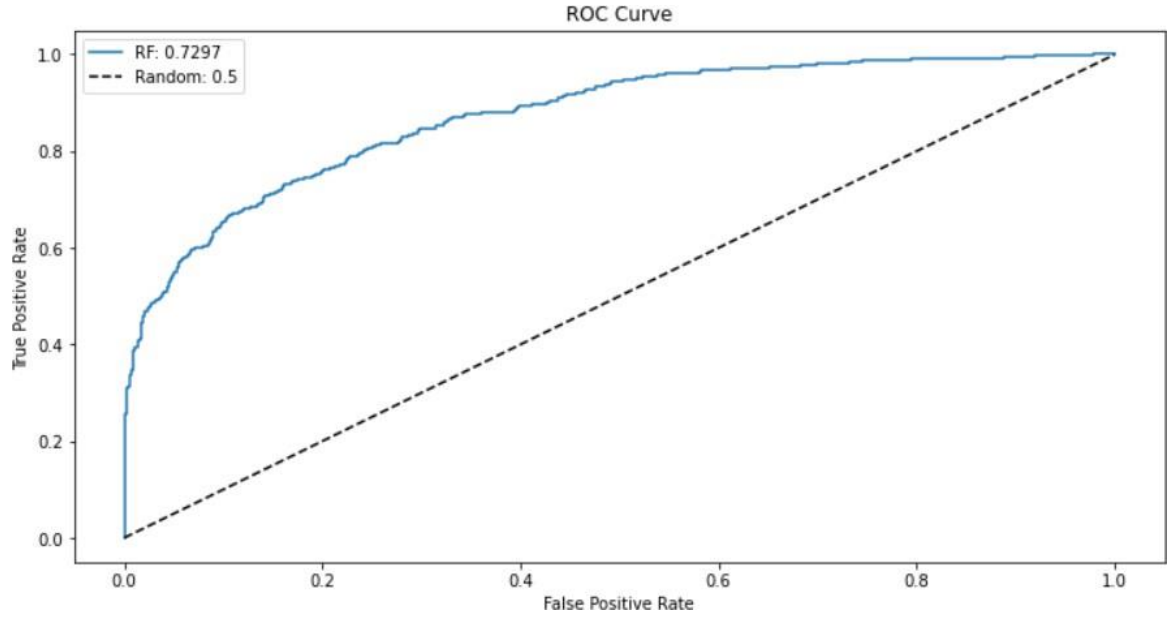


Figure 5.1

Chapter 5

Conclusion

We have implemented Customer churn prediction in Banking System, by using Random Forest Classifier. The model that we have developed is 89% accurate and provides relatively good results. This is the mean cross-validated accuracy of the model. We have improved the accuracy of out model by using GridSearchCV. We will later try to improve the accuracy of this model. We can also use different and better models while also varying the parameters slightly and obtaining more data.

Bibliography: -

- [1] M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju "Churn prediction using comprehensible support vector machine: An analytical CRM application", Applied Soft Computing 19 (2014) 31–40.
- [2] Wouter Verbeke, David Martens, Christophe Mues, Bart Baesens "Building comprehensible customer churn prediction models with advanced rule induction techniques", Expert Systems with Applications 38 (2011) 2354–2364.
- [3] Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang "A Customer Churn Prediction Model in Telecom IndustryUsing Boosting", IEEE Transactions on Industrial Informatics, vol. 10, no. 2, may 2014
- [4] Ammara Ahmed, D. Maheswari Linen "A review and analysis of churn prediction methods for customer retention in telecom industries" 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) 6-7 Jan. 2017
- **https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6**
- **http://www.diva-portal.org/smash/get/diva2:1049992/FULLTEXT02.pdf**
- **https://rstudiopubsstatic.s3.amazonaws.com/565148_6e82a5c320f14869bf63e23bcf59ce9b.html**
- **https://www.neuraldesigner.com/learning/examples/bank-churn**
- **https://medium.com/@rohitlal/customer-churn-prediction-model-using-logistic-regression-490525a78074**

Acknowledgements

We express our sincere thanks to Dr. M.D. Patil, Principal, Ramrao Adik Institute of Technology, Nerul. We pay our deep sense of gratitude to Dr. Leena Ragha (HOD) of COMPUTER department, Ramrao Adik Institute of Technology, Nerul to encourage us to the highest peak and to provide us the opportunity to prepare the project. We are immensely obliged to our friends for their elevating inspiration, encouraging guidance and kind supervision in the completion of our project. We feel to acknowledge our indebtedness and deep sense of gratitude to our guide Mrs. Rajashree Shedge whose valuable guidance and kind supervision given to us throughout the course which shaped the present work as its show. Last, but not the least, our parents are also an important inspiration for us. So, with due regards We express our gratitude to them.

Customer Churn

ORIGINALITY REPORT

35%
SIMILARITY INDEX

PRIMARY SOURCES

1	towardsdatascience.com Internet	374 words — 7%
2	mafiadoc.com Internet	367 words — 7%
3	medium.com Internet	193 words — 4%
4	www.slideshare.net Internet	169 words — 3%
5	www.javatpoint.com Internet	105 words — 2%
6	www.ijert.org Internet	93 words — 2%
7	ieeexplore.ieee.org Internet	84 words — 2%
8	content.yudu.com Internet	74 words — 1%
9	github.com Internet	65 words — 1%
10	www.ijcaonline.org Internet	

		54 words — 1%
11	www.researchgate.net Internet	46 words — 1%
12	www.coursehero.com Internet	39 words — 1%
13	journals.plos.org Internet	33 words — 1%
14	Jaehuyn Ahn, Junsik Hwang, Doyoung Kim, HyukGeun Choi, Shinjin Kang. "A Survey on Churn Analysis in Various Business Domains", IEEE Access, 2020 Crossref	28 words — 1%
15	Xiaoli Qin, Francis M. Bui, Ha H. Nguyen. "Learning from an Imbalanced and Limited Dataset and an Application to Medical Imaging", 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2019 Crossref	21 words — < 1%
16	code.i-harness.com Internet	16 words — < 1%
17	"Advanced Computing", Springer Science and Business Media LLC, 2021 Crossref	15 words — < 1%
18	Kononenko, Igor, and Matjaž Kukar. "Data Preprocessing", Machine learning and data mining, 2007. Crossref	15 words — < 1%
19	docshare.tips Internet	13 words — < 1%

20	www.e3s-conferences.org Internet	13 words — < 1%
21	repository.unika.ac.id Internet	12 words — < 1%
22	www.treasury.govt.nz Internet	11 words — < 1%
23	acadpubl.eu Internet	10 words — < 1%
24	redstatz.com Internet	10 words — < 1%
25	Cano, Alberto, Amelia Zafra, and Sebastián Ventura. "An interpretable classification rule mining algorithm", Information Sciences, 2013. Crossref	9 words — < 1%
26	Dorenda Slof, Flavius Frasincar, Vladyslav Matsiiako. "A competing risks model based on latent Dirichlet Allocation for predicting churn reasons", Decision Support Systems, 2021 Crossref	9 words — < 1%
27	pdfs.semanticscholar.org Internet	9 words — < 1%
28	Hoss Belyadi, Alireza Haghighat. "Supervised learning", Elsevier BV, 2021 Crossref	8 words — < 1%

EXCLUDE QUOTES OFF EXCLUDE MATCHES OFF

EXCLUDE BIBLIOGRAPHY OFF