

# **CUSTOMER CHURN PREDICTION**

SUBMITTED BY

**S. Madhu Mitha 19MX115**

**G. Aishwarya 19MX201**

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF  
THE REQUIRMENT FOR THE DEGREE OF

**MASTERS OF COMPUTER APPLICATIONS**

ANNA UNIVERSITY



**April 2021**

DEPARTMENT OF COMPUTER APPLICATIONS

**PSG COLLEGE OF TECHNOLOGY**

COIMBATORE – 641 004

**PSG COLLEGE OF TECHNOLOGY  
COIMBATORE – 641 004**

**18MX48 MINI PROJECT I  
CUSTOMER CHURN PREDICTION**

Bonafide record of work done by  
**S.Madhu Mitha 19MX115**  
**G. Aishwarya 19MX201**

Dissertation submitted in partial fulfillment of the requirements  
for the degree of

**MASTER OF COMPUTER APPLICATIONS**  
**ANNA UNIVERSITY**  
**April 2021**

**Faculty Guide**

.....

**Dr.G.A. Vijayalakshmi Pai**

# ACKNOWLEDGEMENT

I take this opportunity to express my sincere thanks to **Dr.K.Prakasan**, Principal, PSG College of Technology, for providing me all the necessary facilities and motivation to carry out the project.

I profoundly thank **Dr.A.Chitra**, Professor and Head, Department of Computer Applications, for providing me with the necessary library and lab facilities to complete the project.

I am grateful to thank **Dr.R.Manavalan**, Associate Professor, Program Coordinator, Department of Computer Applications, for his full-fledged support and guidance.

I also express my deepest and sincere thanks to my Project Guide **Dr.G.A.Vijayalakshmi Pai**, Professor(CAS), Department of Computer Applications, for her priceless suggestions, and unrelenting support in all my efforts to improve my project and for piloting in the right way for the successful completion of the project.

I also thank my tutor **Dr.N.Geetha**, Assistant Professor, Department of Computer Applications, for encouraging me to complete the project successfully.

Finally I express my sincere thanks to all staff members of the Department of Computer Applications, for their motivation and encouragement. Above all I thank the almighty for his support in my endeavor

# **Abstract**

Customer Churn is a focal concern of most companies which are active industries with low switching cost. There are majorly two categories of customer churn, voluntary churn and non-voluntary churn. Non-voluntary churn is initiated by the company in which a company withdraws its service from a customer. Whereas, voluntary churn is initiated by the customers when he/she decides to terminate from the service of the company. Hence voluntary churn is more difficult to determine.

Though banking and finance sectors exhibit low voluntary churn rates as compared to other sectors, the impact on profitability by losing a customer is comparatively high. Customer churn management plays a vital role for an organization to enhance long term profitability.

In order to tackle this problem we must recognize the voluntary churners before they churn. So developing a model which predicts the future churners seems to be vital. With Data analytics and machine learning, we can identify factors that lead to customer turnover, create customer retention plans, and predict which customers are likely to churn. This model has to be able to recognize the customers which tend to churn in close future.

# Table of Contents

<b>1. Introduction</b>	<b>1</b>
1.1 Churn Management	1
1.2 Problem definition	2
1.3 Project Purpose	2
1.4 System Specification	3
<b>2. Customer Churn Data Processing</b>	<b>4</b>
2.1 Data Collection	5
2.2 Data Selection	5
2.3 Data Pre-Processing	6
2.4 Data Transformation	6
2.5 Data Visualization & Interpretation	7
<b>3. Churn Prediction Models</b>	<b>14</b>
3.1 Machine Learning Models	14
3.2 Logistic Regression	16
3.3 Support Vector Machine	18
3.4 Random Forest Classifier	20
3.5 XGBoost Classifier	21

3.6 Comparison of prediction models	22
<b>4. Web Development</b>	<b>26</b>
<b>5. Conclusion</b>	<b>30</b>
<b>References</b>	<b>31</b>

# 1. Introduction

## 1.1 Churn Management

Churn is defined as the movement of customers from one provider to another and churn management is the process for holding profitable customers with the company through appropriate marketing campaign and retention strategies. There are majorly two categories of customer churn, Voluntary churn and Non-voluntary churn. Non-voluntary churn is initiated by the company in which a company withdraws its service from a customer. On the other hand voluntary churn is initiated by the customer when he/she decides to terminate his/her service from the provider, hence voluntary churn is more difficult to determine. Voluntary churn has been a challenge for all the companies and it constitutes of major portion of company's total churn. A voluntary churn can be either incidental or deliberate. Incidental churn happens due to circumstances which prevents customer from continuing his service with the provider. Deliberate churn occurs when a customer decides to switch to another service provider. Some reasons for this type churn include bad service quality or low priced offers by competitors.

Customer churn has become a massive problem that affects other aspects of Customer Relationship Management (CRM). For banks and financial organizations maintaining relationship with the customer is of highest priority. Although these sectors exhibit a low churn rate, the impact of losing a single potential customer can have a drastic effect on company's profitability. Hence it's essential for companies to efficiently manage customer churn for long term profitability and survival in the market.

## **1.2 Problem Definition**

Banking is one of the highly competitive sectors where customer relation is the utmost importance for any bank. Though banking and finance sectors exhibit low voluntary churn rates as compared to other sectors, the impact on profitability by losing a customer is comparatively high. Customer churn management plays a vital role for an organization to enhance long term profitability.

In order to tackle this problem we must recognize the voluntary churners before they churn. So developing a model which predicts the future churners seems to be vital. With Data analytics and machine learning, we can identify factors that lead to customer turnover, create customer retention plans, and predict which customers are likely to churn. This model has to be able to recognize the customers which tend to churn in close future.

## **1.3 Project Purpose**

The purpose of this project is to develop and design an effective and efficient model for customer churn prediction in banking sectors. The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for the banks. Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.



## 1.4 System Specifications

Python 3.8 is used as the base programming language for building machine learning model to predict Customer retention. Backend is completely built in Python (Anaconda). Python has a powerful set of packages for wide range of data computing and prediction. The main packages and libraries that are to be used in this project are as follows:-

- Pandas
- Matplotlib
- Seaborn

### **Pandas**

All the preprocessing work on the data will be done with the help of this extremely powerful library. It helps for manipulating numerical data for data analysis.

### **Matplotlib**

It is a comprehensive library for creating static, animated, and interactive visualizations in Python.

### **Seaborn**

It is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphs.

## 2. Customer Churn Data Processing

A Data science lifecycle is an iterative set of steps we take to deliver a data science project or product. Because every data science project and team is different, every specific data life cycle is different. The flowchart of our data science process is illustrated in the Fig. 2.1

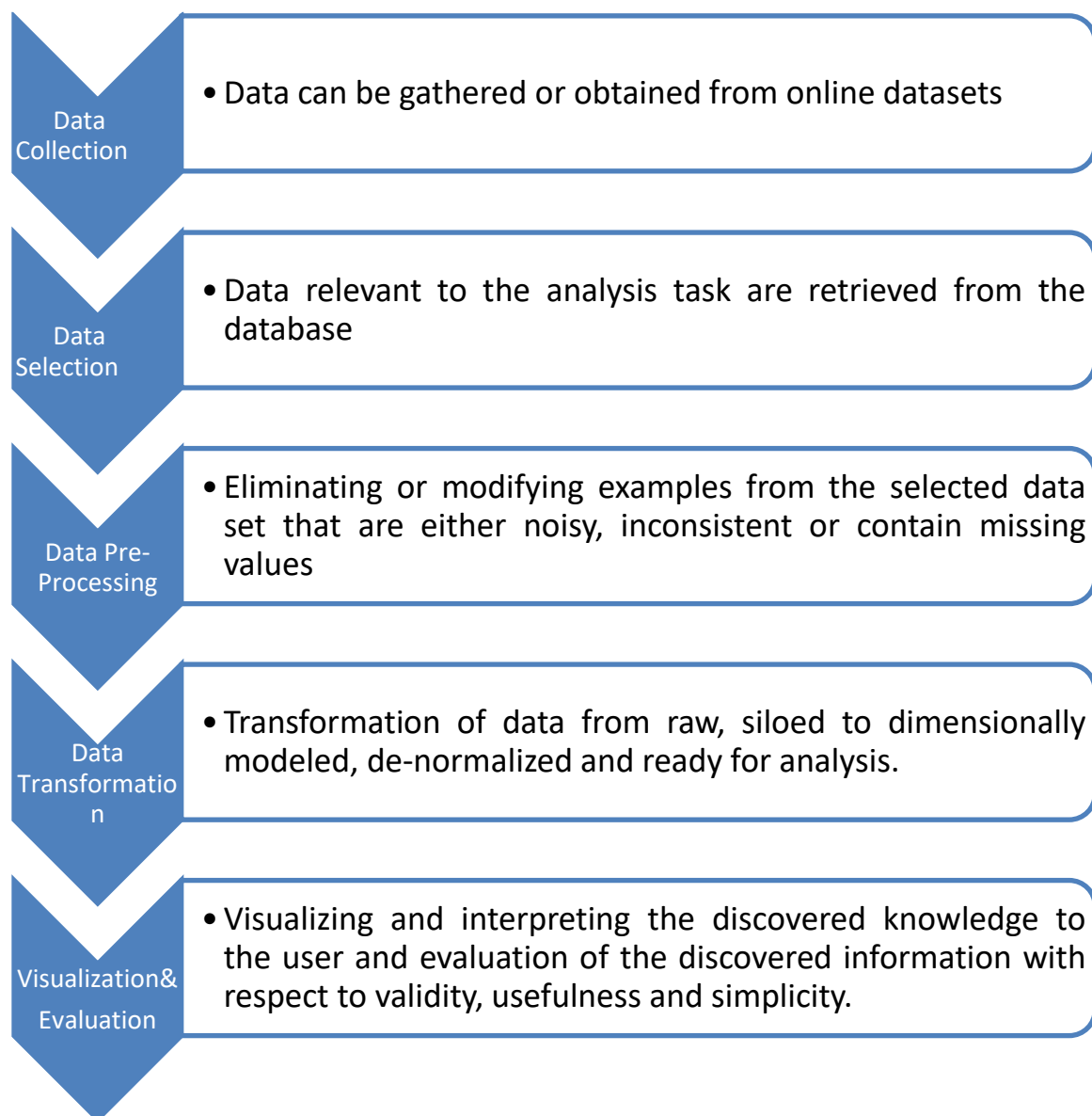


Fig. 2.1 Flowchart of Data science process

## 2.1 Data Collection

The data file Churn\_modelling.csv containing 15 features about 10000 clients of the bank obtained from Kaggle is chosen as the dataset for this project. Fig. 2.2 illustrates a snapshot of dataset

data - DataFrame

Index	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumberOfProducts	HasCreditCard	ActiveMember	EstimatedSalary	Exited	Reason
0	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	1e+05	1	High Service Charges/Rate of Interest
1	15647311	Hill	608	Spain	Female	41	1	8.4e+04	1	0	1	1.1e+05	0	Nil
2	15619304	Onio	502	France	Female	42	8	1.6e+05	3	1	0	1.1e+05	1	Long Response Times
3	15701354	Boni	699	France	Female	39	1	0	2	0	0	9.4e+04	0	Nil
4	15737888	Mitchell	850	Spain	Female	43	2	1.3e+05	1	1	1	7.9e+04	0	Nil
5	15574012	Chu	645	Spain	Male	44	8	1.1e+05	2	1	0	1.5e+05	1	High Service Charges/Rate of Interest
6	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	1e+04	0	Nil
7	15656148	Obinna	376	Germany	Female	29	4	1.2e+05	4	1	0	1.2e+05	1	Inexperienced Staff / Bad customer service
8	15792365	He	501	France	Male	44	4	1.4e+05	2	0	1	7.5e+04	0	Nil
9	15592389	H?	684	France	Male	27	2	1.3e+05	1	1	1	7.2e+04	0	Nil
10	15767821	Bearce	528	France	Male	31	6	1e+05	2	0	0	8e+04	0	Nil

Fig. 2.2 Snapshot of the customer churn data set

## 2.2 Data Selection

We first identify and extract the most relevant attributes for this project. The initial dataset consists of 15 attributes, which can be grouped in the following three categories:

- **Demographic Attributes:** contain the primary features of the customer such as sex, age, gender, geography, etc.
- **Contract Attributes:** contain the attributes associated with the customer contract for a particular service such as type of service, date of conclusion of the contract, price of the service, etc.
- **Customer behavior attribute:** describe the customer activities.

A total of 10000 customers are included, of which 2037 customers are churns, while 7963 customers still use the services from the bank.

## 2.3 Data Pre-Processing

Columns related to personal data of the customers are removed, since these columns do not have any quantitative impact on any calculations. Often, we may find missing data or NULL data in certain columns of a dataset. Such NULL data values not only pose problems in the analysis, but also terminate any mathematical calculations that are carried out on the dataset without dealing with them. Therefore, we need to make sure that there are no such data in our dataset. Fortunately, we did not find any NULL or missing values in this dataset. This is really very good for analysis.

## 2.4 Data Transformation

Data Transformation techniques can significantly improve the overall performance of the churn prediction. The prediction produces best results when data attributes are normalized (in the [0, 1] range). This step is implemented so that there is no overflow of values during the calculations. Fig. 2.3 illustrates the snapshot of a normalized customer churn data set.

Index	Exited	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	HasCrCard	IsActiveMember
8018	1	0.56	0.068	0.3	0.49	0	0.74	1	-1
9225	0	0.49	0.19	0.4	0.48	0.33	0.81	1	1
3854	0	0.67	0.2	0.9	0.54	0.33	0.61	1	-1
2029	0	0.34	0.2	0.4	0.62	0.33	0.83	1	1
3539	0	0.63	0.32	0.6	0	0	0.44	1	-1

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

Fig. 2.3 Snapshot of a normalized customer churn data set.

## 2.5 Data Visualization

Now that we have preprocessed the data by removing unnecessary information from it, we can start an exploratory analysis to find possible correlations between features (columns) of the data and resulting outcomes. We have analyzed each and every possible column in the data to measure its eligibility to be a valuable feature for the exit criteria of any customer. Fig. 2.4 illustrates the pie chart for churned customers and reason for churning

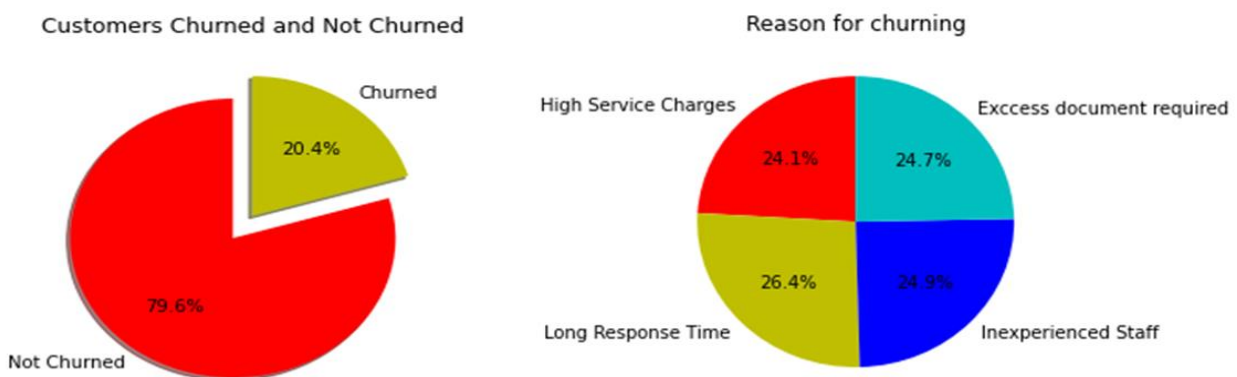


Fig. 2.4 Pie chart for churned customers and reason for churning

From the given data, it is visible that around 20% of the people have exited or churned and the reasons for churning have same ratios.

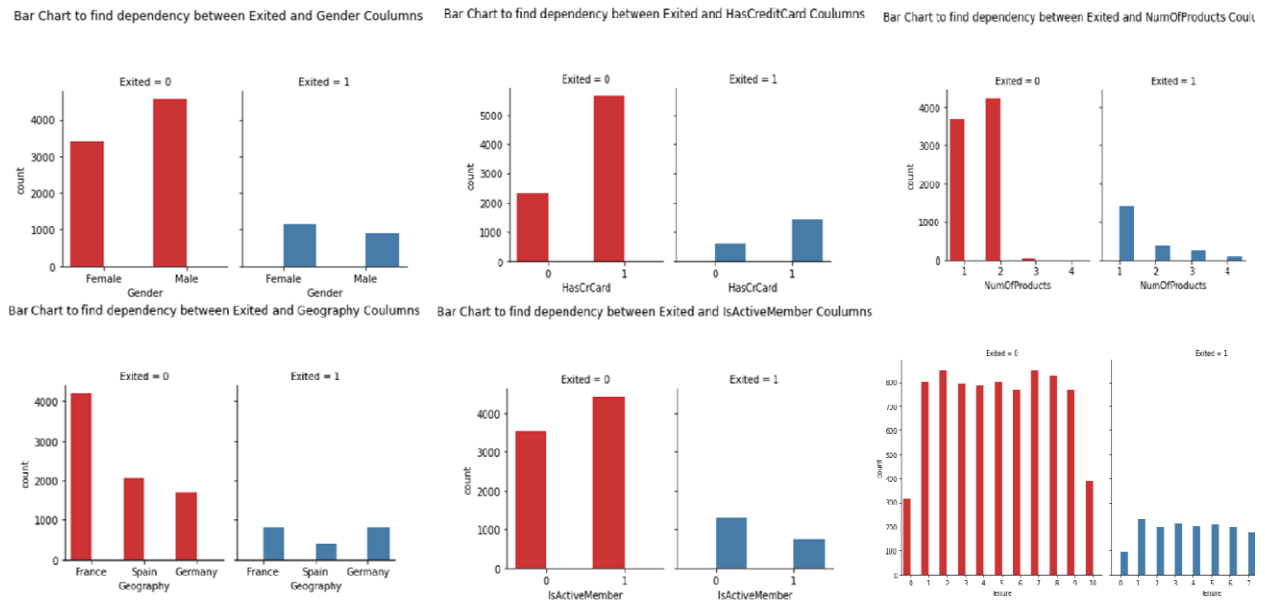


Fig 2.5 Bar chart to map the dependence of 'Exited' column on categorical features

Fig. 2.5 describes the following,

- **Geography:** We can see that majority of the data is about people France. 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 as 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 fewer 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 joiners and older people (10 years) have been churned less. However, on a closer analysis we can see that the overall numbers of retained customer are significantly less in both these cases. As a result, we can probably conclude that new joiners and older customers may be more likely to be churned as their churn rate (percentage) is likely to be higher than other tenure rates.

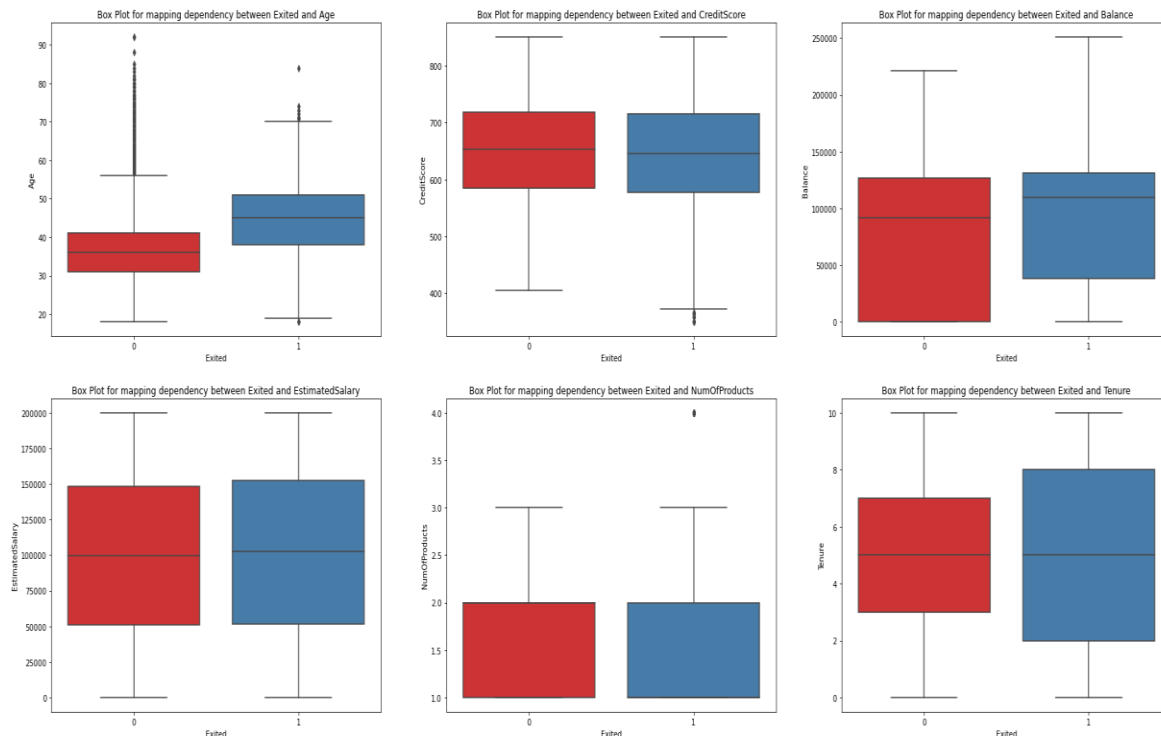


Fig.2.6 Box plot to map the dependence of 'Exited' column on continuous and numerical features

Fig. 2.6 describes the following,

- **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, customer belonging more to the two extreme tenure groups (new joiners and older ones) are more likely to be churned.

## Swarm Plot

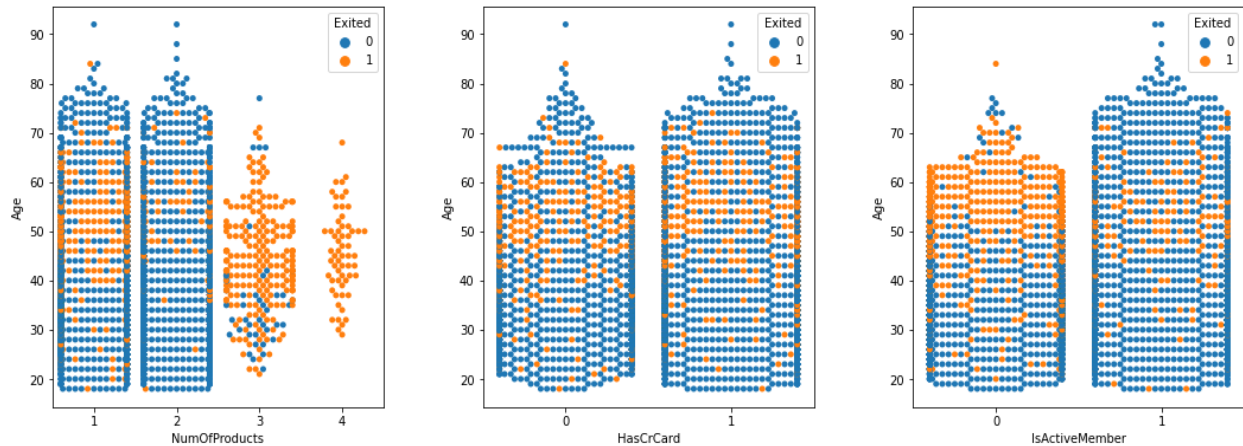


Fig. 2.7 Swarm plot to map the dependency between 'age' and other variables

From the Fig. 2.7, we visualized the dependency between *age* & *numofproducts*, *age* & *hascrcredit*, *age* & *isActivemember*. We are mapping with these three variables because these are the variables which mainly involve with Customer bank Relationship and to see whether *age* affects the customer churn.

### **NumofProducts**

Customer with 3 or 4 products have higher chances to Churn 50 to 60 years old customers having 2 products have a major churn when compared to other aged customers.

### **HasCrCard**

The plot shows that mostly 50 to 60 aged customers are the major churn who do not have creditcard and even if they have creditcard they are the major churn when compared to other aged customers who have creditcard.

### ***Isactivemember***

The plot shows that mostly 50 to 70 aged customers are the major churns who are not activemembers and even if they are active they are the major churn when compared to other aged customers who are activemembers. From the Fig. 2.7, we can conclude that 40 to 70 years old customers have higher chances to churn.

### **Heat Map:**

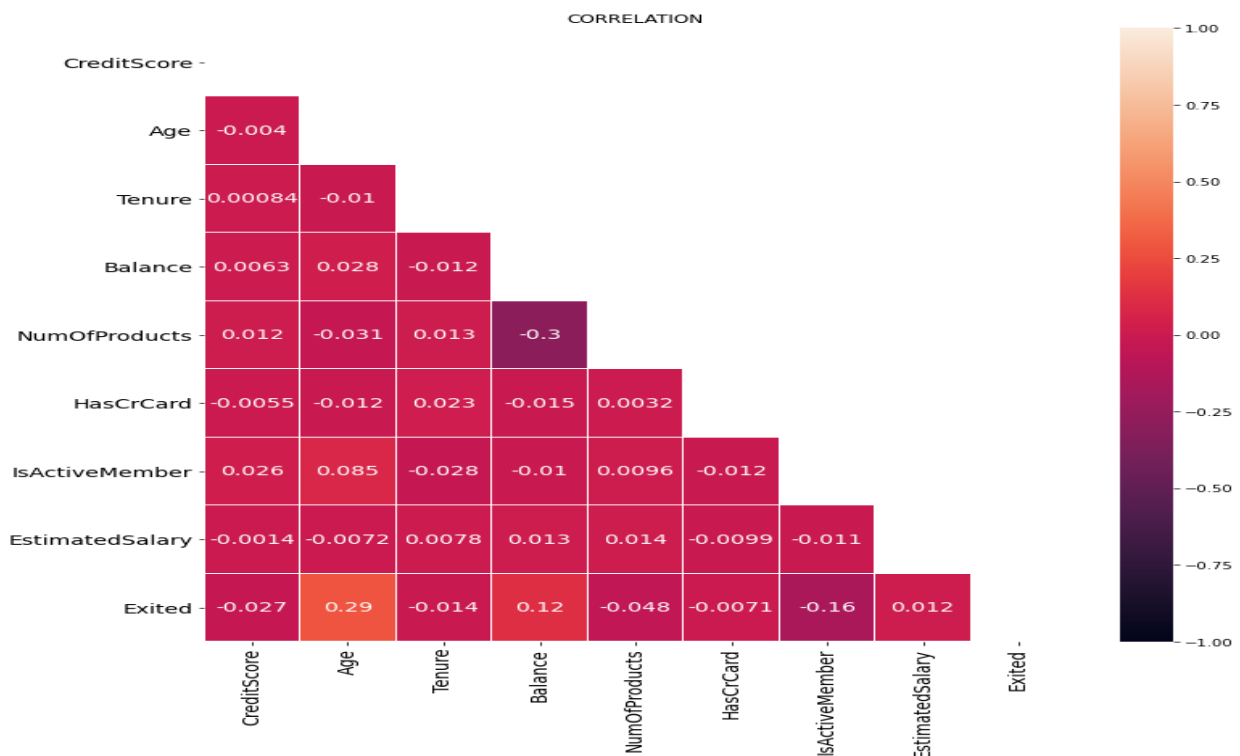


Fig. 2.8 Heat Map to find the correlation between the variables

The following facts can be derived from the Fig. 2.8:

Cells with 1.0 values are highly correlated with each other

Each attribute has a very high correlation with itself, so all the diagonal values are 1.0

Balance attribute is negatively correlated with *numberofproducts* attribute. It means one attribute increases as the other decreases, and vice versa.

### 3. Churn Prediction Models

As the first step in data preparation, the data is separated into training and test dataset in 80:20 ratio. Therefore, the number of rows in train data is 8000 and the number of rows in test data is 2000.

- The categorical values 0 are converted into -1 for ***HasCrCard*** and ***IsActiveMember*** columns. This will allow us to include a negative relation in the modeling.
- One-hot encoding is done for the remaining text categorical variables ***Geography*** and ***Gender***.
- Finally, we normalize the continuous variables between 0 and 1. This step is implemented so that there is no overflow of values during the calculations.

#### 3.1 Machine Learning Models

A “model” in machine learning is the output of a machine learning algorithm run on data. Here, we try to train different machine learning classification models to our data. Once we get the model details for each of the models, we can select the best model from them for our training and testing purposes.

**The five algorithms chosen for modeling are**

- Logistic Regression
- Support Vector Machine with RBF Kernel
- Support Vector Machine with Poly Kernel
- Random Forest Classifier
- XG Boost Classifier

A classification report is generated for all the models. The classification report visualizer displays the precision, recall, F1, and support scores for the model.

There are four ways to check if the predictions are right or wrong:

1. **TN / True Negative**: the case was negative and predicted negative
2. **TP / True Positive**: the case was positive and predicted positive
3. **FN / False Negative**: the case was positive but predicted negative
4. **FP / False Positive**: the case was negative but predicted positive

## Precision

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.

Precision:- Accuracy of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

## Recall

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall:- Fraction of positives that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

## **F1 score**

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

## **Support**

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

## **3.2 Logistic Regression**

The logistic regression is a classification model and it was developed in order to extend the linear regression model which is, instead, a regression model (i.e. the outcome is continuous). Although linear regression can be adapted for the classification task, it gives poor results. Therefore, researchers have developed the logistic regression model to study the relationships between a

binary outcome (0 or 1) and features. The intuition behind logistic regression is pretty simple. Since we need a binary outcome, we could do the following:

- Map the linear regression predictions in  $[0, 1]$
- Interpret the new result as the probability of having 1 as outcome
- Predict 1 if the probability is greater than a chosen threshold, otherwise predict 0

This makes logistic regression a powerful classifier, because it allows to specify the algorithm precision through the probability threshold. In particular, setting a high threshold leads to predict 1 only if we are very confident. Conversely, a low threshold decreases the precision, but increases the recall. Typically, if the problem has equally relevant classes, the threshold is set to 0.5. Usually, the function used to map linear regression predictions in  $[0, 1]$  is the logistic function (also called the sigmoid). This function was chosen mainly because of its peculiar characteristics, which fit the problem requirements well. The sigmoid function is the following:

$$S(x) = \frac{1}{1+e^{-x}}$$

The sigmoid transformation changes the model representation therefore it is necessary to define a particular loss function in order to train the logistic regression. In particular, the loss function returns a small error if the sigmoid shows the probability of being 1 for an instance, coherently with the real record class. Otherwise it returns a high error. Fig. 3.1 illustrates the classification report for Logistic Regression model.

Classification Report for logistic regression					
	precision	recall	f1-score	support	
0	0.87	0.97	0.92	6382	
1	0.76	0.44	0.56	1618	
accuracy			0.86	8000	
macro avg	0.82	0.70	0.74	8000	
weighted avg	0.85	0.86	0.84	8000	

Fig. 3.1 Classification report for logistic regression model

### 3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

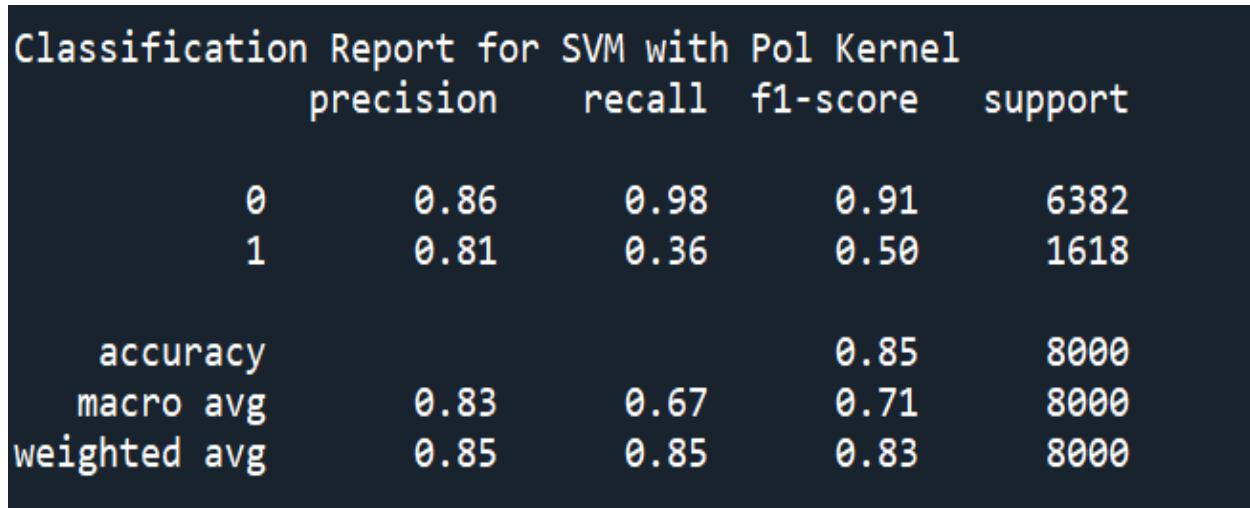
Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line)



## SVM with Poly Kernel

Tuning the parameters values for machine learning algorithms effectively improves model performance. So, here the kernel is tuned as Poly Kernel and modeled to improve the performance.

Fig. 3.2 illustrates the classification report of SVM with poly kernel



	precision	recall	f1-score	support
0	0.86	0.98	0.91	6382
1	0.81	0.36	0.50	1618
accuracy			0.85	8000
macro avg	0.83	0.67	0.71	8000
weighted avg	0.85	0.85	0.83	8000

Fig 3.2 Classification report of the SVM with poly kernel

## SVM with RBF Kernel

The kernel is tuned to RBF and modeled to improve the performance. Fig. 3.3 illustrates the classification report of the SVM with RBF kernel.

From the Fig. 3.3, we can see that the accuracy varies by tuning the parameters. It can be seen that the SVM with RBF Kernel provides more accuracy than Poly Kernel.

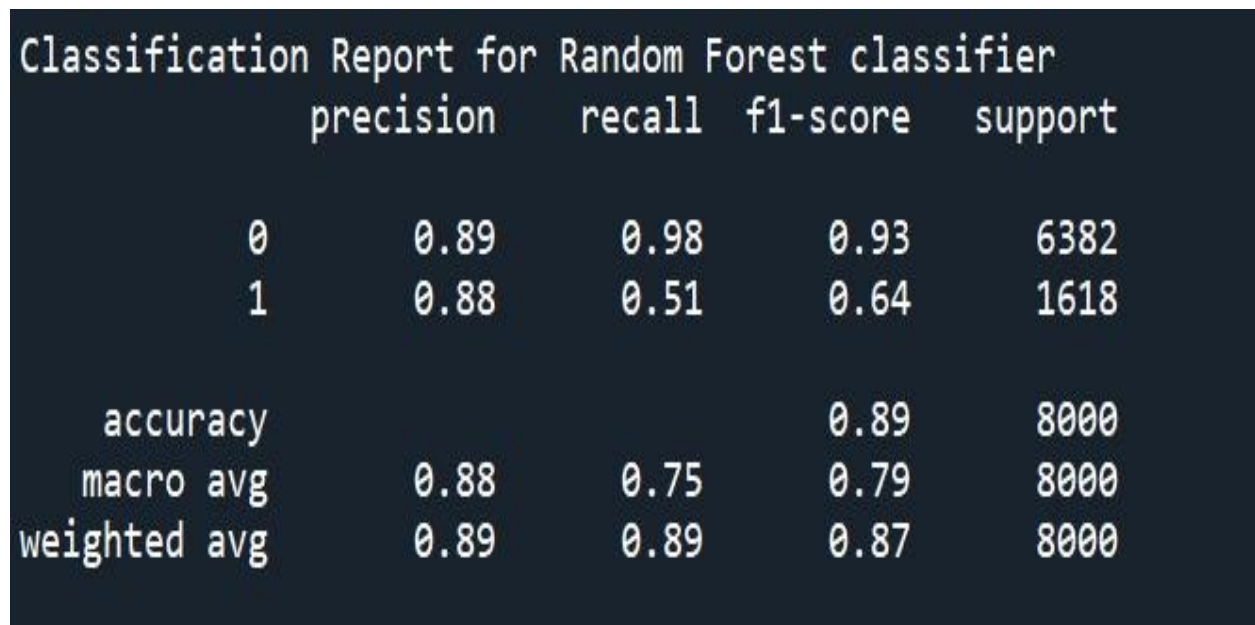
Classification Report for SVM with RBF Kernel				
	precision	recall	f1-score	support
0	0.86	0.98	0.92	6382
1	0.84	0.38	0.53	1618
accuracy			0.86	8000
macro avg	0.85	0.68	0.72	8000
weighted avg	0.86	0.86	0.84	8000

Fig. 3.3 Classification report of the SVM with RBF kernel

### 3.4 Random Forest Classifier

Random forest is an ensemble learning algorithm. An ensemble is a set of different classifiers merged together to make a more powerful model. In particular, random forest is based on the bagging technique, which is a statistical method to create some different training sets starting from a single one. Random forest builds on each training set a decision tree, using just a subset of random features for each classifier. Since decision trees are very unstable, building them on different training sets with random features will lead to very diverse classifiers. This allows lowering the correlation among the models, hence increasing the overall performance. In fact, to classify a record, random forest merges all the decision trees with a voting system: each tree votes a class for the record and then random forest chooses the most voted one as the final outcome. The voting system averages decision trees predictions, which are affected by high variance, therefore the results reflect in an increased accuracy, even with large sets of data. This also means that random forest is less prone to overfitting than a single decision tree, because it averages the tree predictions,

leading to more robust results. Moreover, it has only few parameters to set up, which is always desirable. The main disadvantage of this approach is the computational time, which increases proportionally with the number of trees. Fig. 3.4 illustrates the classification report for Random Forest Classifier



	precision	recall	f1-score	support
0	0.89	0.98	0.93	6382
1	0.88	0.51	0.64	1618
accuracy			0.89	8000
macro avg	0.88	0.75	0.79	8000
weighted avg	0.89	0.89	0.87	8000

Fig. 3.4 Classification report for Random Forest Classifier

### 3.5 XGBoost Classifier

XGBoost is an optimized Gradient Boosting Machine Learning library. The core XGBoost algorithm is parallelizable i.e. it does parallelization within a single tree. It is one of the most powerful algorithms with high speed and performance. It can harness all the processing power of modern multicore computers. It is feasible to train on large datasets. Consistently outperform all single algorithm methods. XGBoost is usually used with a tree as the base learner, that decision tree is composed of the series of binary questions and the final

predictions happens at the leaf. XGBoost is itself an ensemble method. The trees are constructed iteratively until a stopping criterion is met.

XGBoost uses CART(Classification and Regression Trees) Decision trees. CART is the trees that contain real-valued score in each leaf, regardless of whether they are used for classification or regression. Real-valued scores can then be converted to categories for classification, if necessary. Fig. 3.5 illustrates the classification report for XGBoost classifier.

Classification Report for Extreme Gradient Boost Classifier				
	precision	recall	f1-score	support
0	0.89	0.97	0.93	6382
1	0.82	0.50	0.62	1618
accuracy			0.88	8000
macro avg	0.85	0.74	0.77	8000
weighted avg	0.87	0.88	0.86	8000

Fig. 3.5 Classification report for XGBoost classifier

### 3.6 Comparison of prediction models

Best model is selected on the basis of accuracy. The model which provides high accuracy is chosen for predictions. To select the best model we use confusion matrix and ROC Curve.

## Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification problem where output can be two or more classes. Fig. 3.6 illustrates the 4 different combinations of predicted and actual values in the confusion matrix.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 3.6 Confusion Matrix

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

### True Positive

Interpretation: You predicted positive and it's true.

You predicted that a woman is pregnant and she actually is.

### True Negative

Interpretation: You predicted negative and it's true.

You predicted that a man is not pregnant and he actually is not.

### False Positive: (Type 1 Error)

Interpretation: You predicted positive and it's false.

You predicted that a man is pregnant but he actually is not.

### **False Negative: (Type 2 Error)**

Interpretation: You predicted negative and it's false.

You predicted that a woman is not pregnant but she actually is.

Just Remember, We describe predicted values as Positive and Negative and actual values as True and False.

## **ROC CURVE**

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate - yaxis
- False Positive Rate - xaxis

Roc curve uses the average of recall from the classification table to draw the curve. If the curve is more into True positive rate (y-axis) and away from False Positive Rate (x-axis), then it would be more ideal.

AUC stands for "Area under the ROC Curve". That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

From the Fig. 3.7, all the curves are more into true positive rate but XGBoost and Random forest have a higher curve when compared to other algorithms, and it is seen that XGB and Random forest collide at some points, but at some areas Random forest curve is higher than XGB curve.

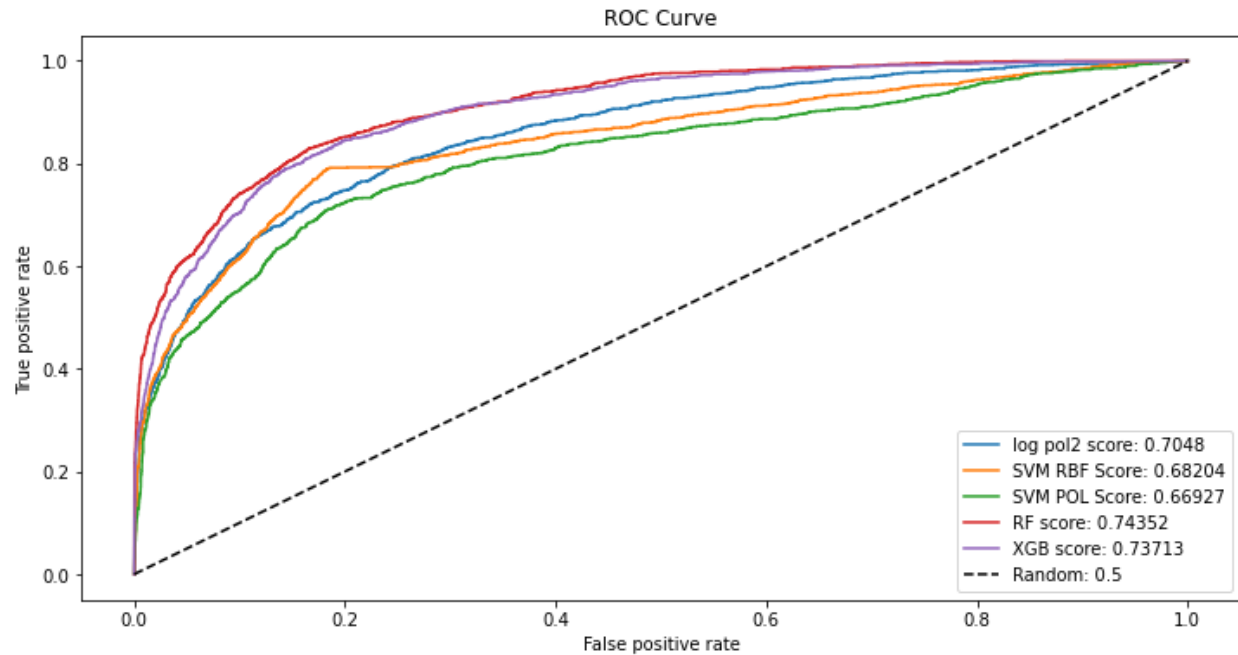


Fig. 3.7 ROC curve for the prediction models

From this we can conclude that the best algorithm suited for our model is random forest classifier.

So, we have used **Random Forest Classifier** for our prediction.

## 4. Web Development

We have used Flask framework and pickle library for web deployment.

### Flask:

Flask is a micro web framework for Python based on Werkzeug, Jinja2 and good intentions. It includes a built-in development server, unit testing support, and is fully Unicode-enabled with RESTful request dispatching and WSGI compliance. Since the web is built on flask framework, deployment is very easy and maintenance is also very less. Flask makes the process of designing a web application simpler. Flask lets us focus on what the users are requesting and what sort response to give back.

### Flask installation

As we are using Python3, we don't have to install virtual environment because it already comes with venv module to create virtual environments. Create a project folder and a subfolder named 'venv' using the command "`py -3 -m venv venv`". Before working on the project, we need to activate the corresponding environment using the command "`venv\Scripts\activate`". Within the activated environment, use the command "`pip install Flask`" to install Flask. After installing flask, create a flask application named "`application.py`"

Python **Pickle** module is used for serializing and de-serializing python object structures. The process which converts any kind of python objects into byte streams is called pickling or serialization. We can convert the byte stream back to python objects by a process called

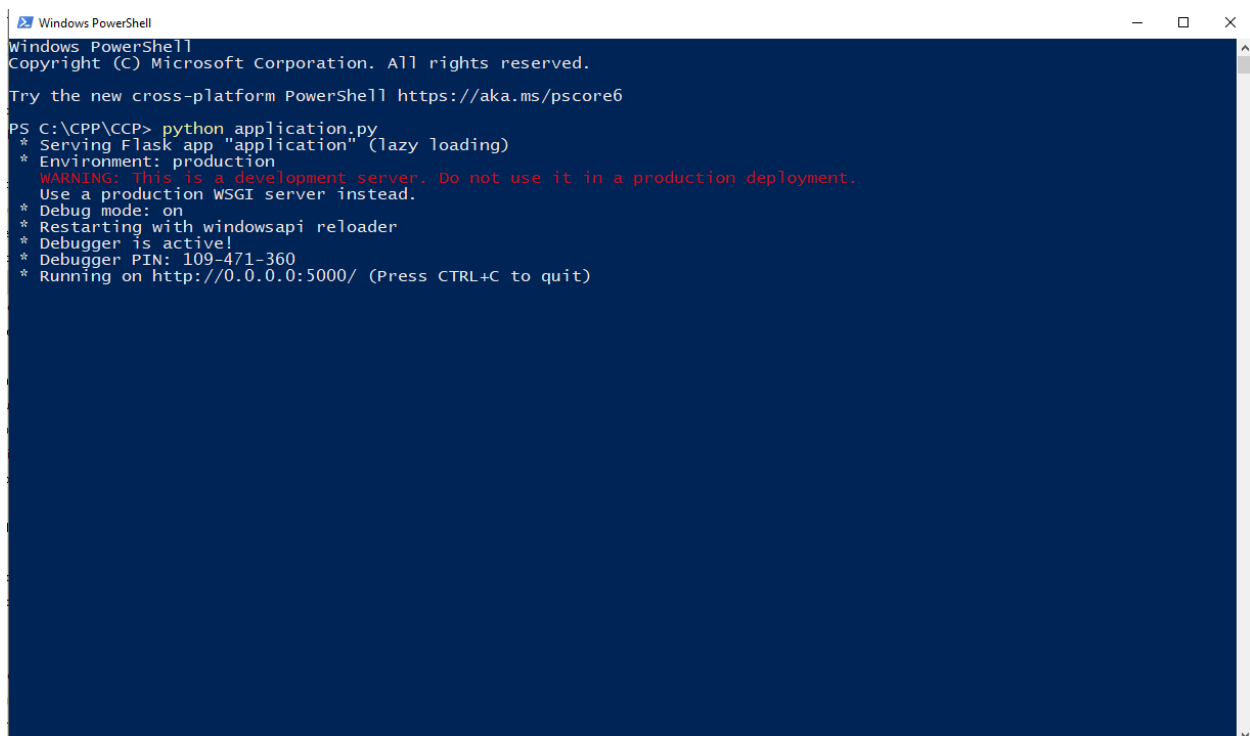


as unpickling. The use of pickling and unpickling is widespread as they allow us to easily transfer data from one system/server to another system/server and then store it in a file or database.

## Html and Templates in flask

Create a new HTML file. We created app.html file.

The Flask Framework looks for HTML files in a folder called templates. We need to create a templates folder and put all our HTML files in that folder. We imported `render_template()` method from the flask framework. `render_template()` looks for a template (HTML file) in the templates folder, then it will render the template for which we asked. We change the return so that now it returns `render_template(app.html)`, this will let us view our HTML file.

A screenshot of a Windows PowerShell terminal window. The window title is "Windows PowerShell". The text inside shows the execution of a Python script. It starts with the PowerShell prompt "PS C:\CPP\CCP> python application.py". The output includes several status messages: "\* Serving Flask app 'application' (lazy loading)", "\* Environment: production", a red warning message "WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.", "\* Debug mode: on", "\* Restarting with windowsapi reloader", "\* Debugger is active!", "\* Debugger PIN: 109-471-360", and finally "\* Running on http://0.0.0.0:5000/ (Press CTRL+C to quit)". The terminal has a dark blue background and white text. The window has standard Windows window controls (minimize, maximize, close) in the top right corner.

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\CPP\CCP> python application.py
* Serving Flask app "application" (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Restarting with windowsapi reloader
* Debugger is active!
* Debugger PIN: 109-471-360
* Running on http://0.0.0.0:5000/ (Press CTRL+C to quit)
```

Fig. 4.1 Screenshot of Windows PowerShell

From fig. 4.1, the important part is where it says Running on <http://127.0.0.1:15000/>.127.0.0.1 means this local computer. The main idea is that 127.0.0.1 and localhost refer to the local computer. Go to that address and you could see the web application. Fig. 4.2, 4.3 and 4.4 illustrates the deployment of Customer Churn Prediction.

The screenshot shows a web browser window with the address bar displaying 'localhost:5000/hello'. The page content is a web application titled 'Customer Churn Prediction'. Below the title is a subtitle 'Predict the probability of Customers who are likely to churn'. The form contains several input fields and radio buttons:

- Gender:** Radio buttons for 'Female' and 'Male'.
- IsActiveMember:** Radio buttons for '0' and '1' (selected).
- HasCreditCard:** Radio buttons for '0' and '1' (selected).
- Age:** A text input field with the placeholder 'Age'.
- Geography:** A dropdown menu with 'Germany' selected.
- Number of Products:** A dropdown menu with '1' selected.
- Balance:** A text input field with the placeholder 'Balance'.
- Estimated Salary:** A text input field with the placeholder 'Estimated Salary'.
- Tenure:** A text input field with the placeholder 'Tenure'.
- CreditScore:** A text input field with the placeholder '300-900'.

At the bottom of the form is a green 'Predict' button.

Fig. 4.2 Screenshot of deployed WebApp

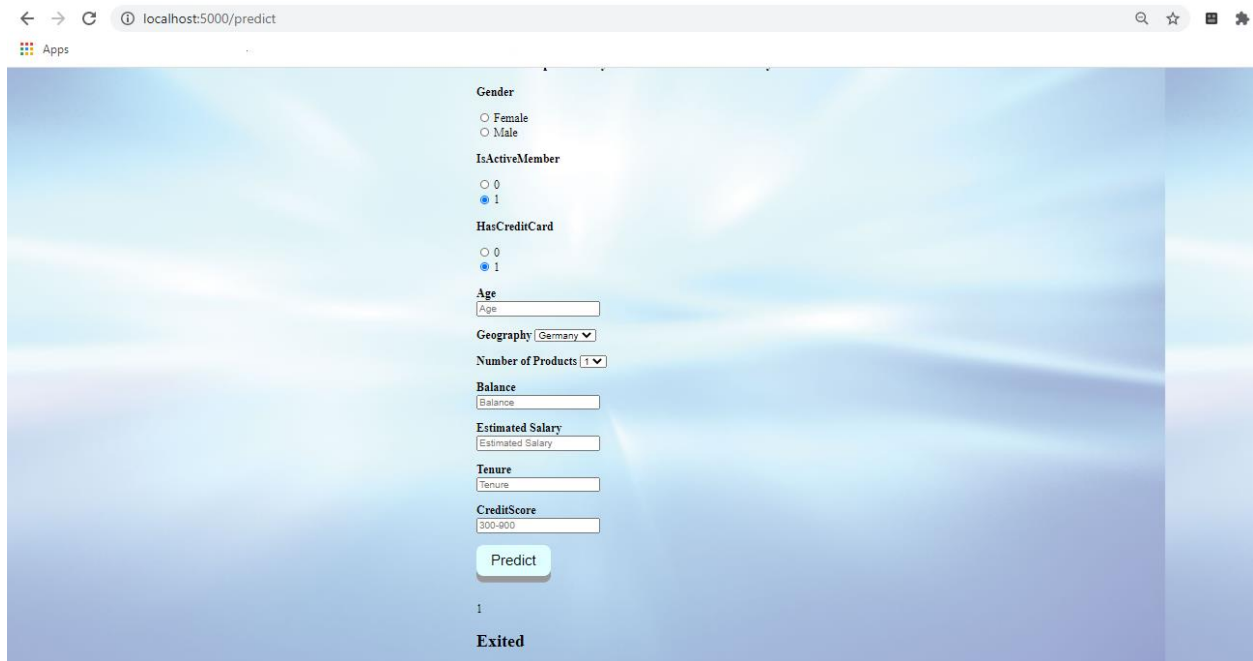


Fig. 4.3 Screenshot of deployed WebApp

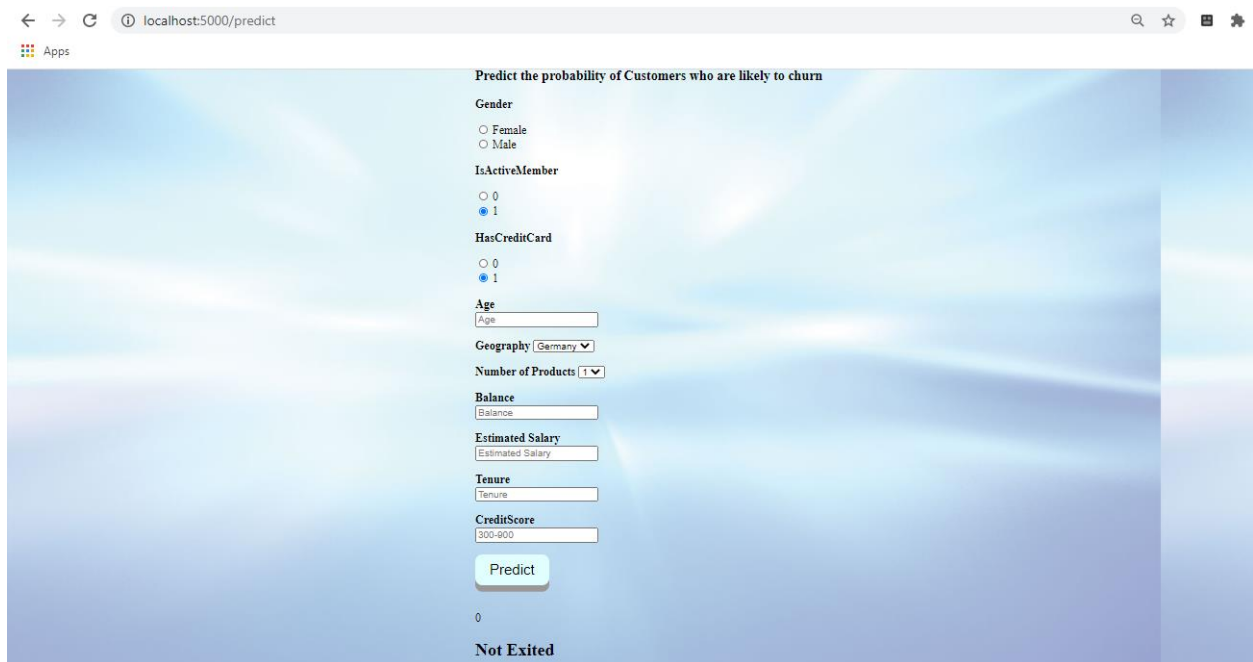


Fig. 4.4 Screenshot of deployed WebApp

## 5. Conclusion

Although all the models yield similar results, the theoretical reasons and theoretical considerations are different. When assessing the statistical learning methods, computational efficiency is important to consider. This is strength of random forest being able to select the strongest independent variables to estimate the final model, without requiring additional information. Hence, the overall best performing model to predict bank customer churn is Random Forest Classifier. This is due its overall performance for all the evaluation measurements and computational efficiency.

Our WebApp will predict whether any customer will be churned or not churned. Once we know about the customers who are on the verge of leaving the association, then we can use various promotional strategies targeted to impress or improve our relation with that customer thus by reducing the bank churn rate.

# References

- [1] [www.kaggle.com](http://www.kaggle.com)
- [2] [www.towardsdatascience.com](http://www.towardsdatascience.com)
- [3] [www.kdnuggets.com](http://www.kdnuggets.com)
- [4] [https://www.youtube.com/watch?v=yieJLP\\_vwbA](https://www.youtube.com/watch?v=yieJLP_vwbA)
- [5] <https://medium.com/@kohlshivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>
- [6] <https://dev.to/sahilrajput/install-flask-and-create-your-first-web-application-2dba>
- [7] <https://medium.com/@kohlshivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>