DATA VISUALIZATION
(CSE3020)

Predicting Customer Churn in Bank Sector

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology
in
Computer Science

By
ABHIJEET TOMER – 18BCE0637
JINESH THAKKER – 18BCE0315
SHAIK HASEEB UR RAHMAN – 18BCE0646

Under the guidance
Of
Prof. Murugan.K

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
(SCOPE)
VIT, Vellore.

June, 2020

DECLARATION

I hereby declare that the Project entitled "***Predicting Customer Churn in Bank Sector***" submitted by me, for the award of the degree of *Bachelor of Technology in CS* to VIT is a record of bonafide work carried out by us under the supervision of Murugan.K.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore
Date: 07-06-2020

ABHIJEET TOMER
JINESH THAKKER
SHAIK HASEEB UR RAHMAN

CERTIFICATE

This is to certify that the project entitled ***"Predicting Customer Churn in Bank Sector***" submitted by

ABHIJEET TOMER – 18BCE0637
JINESH THAKKER – 18BCE0315
SHAIK HASEEB UR RAHMAN – 18BCE0646

VIT, for the award of the degree of *Bachelor of Technology in CS*, is a record of bonafide work carried out by him / her under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The Project fulfils the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore
Date :03-06-2020

Signature of the Guide

INDEX

ABSTRACT

The aim of this project is to introduce a contextual investigation of utilization of one of the data mining techniques, neural network, in knowledge discovery from databases in the banking industry. Churn prediction models are created by academics and specialists to viably oversee and control customer churn so as to hold existing clients. As churn management is an important activity for organizations to hold faithful customers, the capacity to precisely anticipate customer churn is significant. In this project, we tend to use one of the data mining techniques, neural network to predict customer churn in bank. the main focus on customer churn is to determinate the clients who are at risk of going away and analyzing whether or not those customers are value retaining. we want to find out the customers who are probable to churn therefore bank ought to concentrate on those clients, and offer them product according to their needs.

INTRODUCTION

With increased availableness of data, inexpensive storage and processing power, the quantity of data hold on in banking databases is large and perpetually increasing. However, data by itself doesn't provide abundant information. Data processing is employed to find patterns and relationships in data so as to enhance business decision processes

Nowadays, with market soaking and competition being serious, a lot of corporations begin to target Customer Relationship Management (CRM) and Business Intelligence (BI). As existing customer's churning can possible lead to the loss of companies and so decline in profit, churn prediction has been increased vividly within the consumer marketing and management analysis literature over the past few years. Additionally, research suggests that a little modification within the retention rate may result in vital impact on business.

Customer churn may be a concern for many industries; however it's notably acute within the banking sector. It's estimated that the average churn rate for the banking sector is about 15% per month. As churn management is an essential for banks to hold faithful customers, the capacity to precisely anticipate customer churn is critical. Hence, some researches are done on the issue of customer churn, particularly in banking sector. Although there are some papers published on customer churn prediction, few researches are done on the issue of customer churn of banking sector that results in the subsequent issues.

- The cost of attracting new customers is 5 to 6 times over holding on to an existing customer

- Long term customers become less costly to serve, they generate higher profits, and that they may also provide new referrals
- Losing a customer typically results in loss in profit for the bank.

Banks are shifting their attention towards customer churn and are implementing various strategies to prevent customers from getting churned and in order to effectively control customer churn, it's vital to make more effective and accurate client churn prediction model and so this paper proposes a neural network based approach to predict customer churn in bank. Real-world data was used for making a model for customer churn. The main hypothesis was to find out the customers who are likely to churn and so bank should focus on those clients, and offer them products according to their needs.

RELATED CUSTOMER CHURN FIELDS

AUTHORS	FIELD	ALGORITHM USED
Burez and Poel	Pay-Tv company	Logistic regression and Markov chains random forests
Hung et al.	Wireless telecommunication company	Decision tree, Neural network K-means clustering
Nie, G., Wang, G., Zhang, P., Tian, Y., & Shi, Y.	Banking (Credit Cards)	Logistic Regression
Poel and Larivie‘re	European financial services company	Hazard model survival analysis
Shao, J., Xiu, L., & Wenhuang, L.	Commercial Banking	Real AdaBoost, Gentle AdaBoost, & Modest AdaBoost

LITERATURE REVIEW

Model of Customer Churn Prediction on Support Vector Machine

XIA Guo-en, JINWei-dong solved the problem of customer churn in mobile telecommunication industry. They applied support vector machine (SVM) with structural risk minimization on customer churn prediction which improved the prediction abilities of machine learning methods. The dataset used were from UCI database of University of California and a home telecommunication carry. When predicted on first dataset, the accuracy rate was about 0.908 on dataset 1 and 0.5963 on dataset 2.

A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services

Anuj Sharma and Dr. Prabin kumar Panigrahi solved the problem of customer churn in cellular network services. They proposed artificial neural network (ANN) technique to find a best model from stored customer data to predict churn and to prevent the customer's turnover. The dataset used were from UCI database of University of California which deals with cellular service provider's customers and the data pertinent to the voice calls they make. The predicted accuracy for this neural network is 92.35%

Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques

Coussement and Van den Poel applied support vector machines (SVM), Random Forest and Linear Regression in a newspaper subscription churn context in order to construct a churn model. The data used was from a Belgian newspaper publishing company. The accuracy rate achieved was about 84.90% for SVM 87.21% for Random Forest and 84.60% for Logistic Regression.

CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted

Burez and Van den Poel (2006) built a prediction model for European pay-TV company by using Markov chains and a random forest model benchmarked to a basic logistic model. They used the dataset from the warehouse of a pay-TV company. The accuracy rate achieved was about 76.94% for Random Forest , 76.51% for Logistic Regression, 76.93% for Logistic Regression with Markow.

Applying Data Mining to Customer Churn Prediction in an Internet Service Provider

Afaq Alam Khan, Sanjay Jamwal, M.M.Sepehri solved the problem of churn in ISP using various Data mining techniques such as Logistic Regression, Decision tree and Neural Networks. The dataset used was taken from Sepanta Co. which contains the data from 21/3/2005 to 21/5/2006. The accuracy rate achieved was about 87.74% for Decision Tree, 89.01% for Logistic Regression and 89.08 % for Neural Networks.

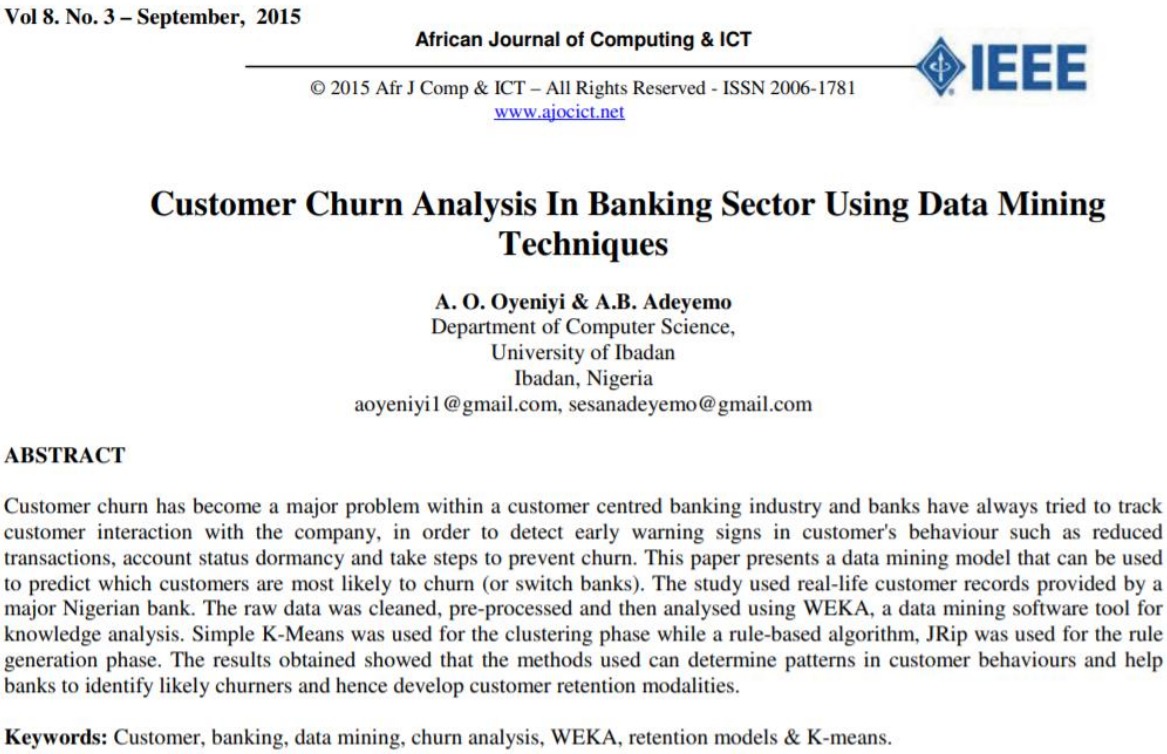
Customer Churn Prediction Based on the Decision Tree in Personal Handy Phone System Service

Luo Bin, Shao Peiji, Liu Juan applied Decision Tree to solve the problem of customer churn in Personal Handy Phone System Service (PHSS). Data of call details of 6000 customers of PHSS during a period of 180 days was used for training the algorithm. The F-measure for test set are 0.95, 0.82, 0.88, respectively.

YEAR	AUTHOR	PROBLEM	METHOD	DATASET	RESULT	LIMITATIONS
2008	XIA Guo-en, JINWei-dong	Customer Churn on mobile Tele communication Industry	SVM with structural risk minimization	The first dataset was taken from UCI where definition of customer churn was a cellular phone customer who does not enjoy all services of telecommunication carry The second dataset was taken from home telecommunication carry where definition of customer churn is that customers with personal access phone removed the phones or cancelled the phone numbers.	The accuracy rate was about 0.908 on dataset 1 and 0.5963 on dataset 2.	Limitations on how to select fitting kernel function and parameter; how to weigh customer samples
2011	Anuj Sharma and Dr. Prabin kumar Panigrahi	Customer Churn in Cellular Network Services	Artificial Neural Networks	The dataset used were from UCI which deals with cellular service provider's customers and the data pertinent to the voice calls they make	The accuracy rate achieved was about 92.35%	
2008	Coussement and Van den Poel	Churn prediction in subscription services	Support vector machines random forests logistic regression	Subscriber database from a Belgian newspaper publishing company	The accuracy rate achieved was about 84.90% for SVM 87.21% for Random Forest and 84.60% for	Though SVM performs better than Logistic Regression but it could not surpass Random Forest

					Logistic Regression .	
2006	Burez and Van den Poel	Churn prediction at pay-TV company	Logistic regression and Markov chains random forests	The dataset was from the warehouse of a pay-TV company.	The accuracy rate achieved was about 76.94% for Random Forest , 76.51% for Logistic Regression , 76.93% for Logistic Regression with Markow.	
2010	Afaq Alam Khan, Sanjay Jamwal, M.M.Sepehri	Churn Prediction in Internet Service Provider	Logistic regression, Decision Tree, Neural Networks	The dataset used was taken from Sepanta Co.	The accuracy rate achieved was about 87.74% for Decision Tree, 89.01% for Logistic Regression and 89.08 % for Neural Networks.	
2011	Luo Bin, Shao Peiji, Liu Juan	Churn Prediction in Handy Phone System Service	Decision Tree	Data of call details of 6000 customers of PHSS during a period of 180 days was used for training the algorithm.	The F-measure for test set are 0.95, 0.82, 0.88, respectively.	

BASE PAPER



Vol 8. No. 3 – September, 2015

African Journal of Computing & ICT

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781
www.ajocict.net

Customer Churn Analysis In Banking Sector Using Data Mining Techniques

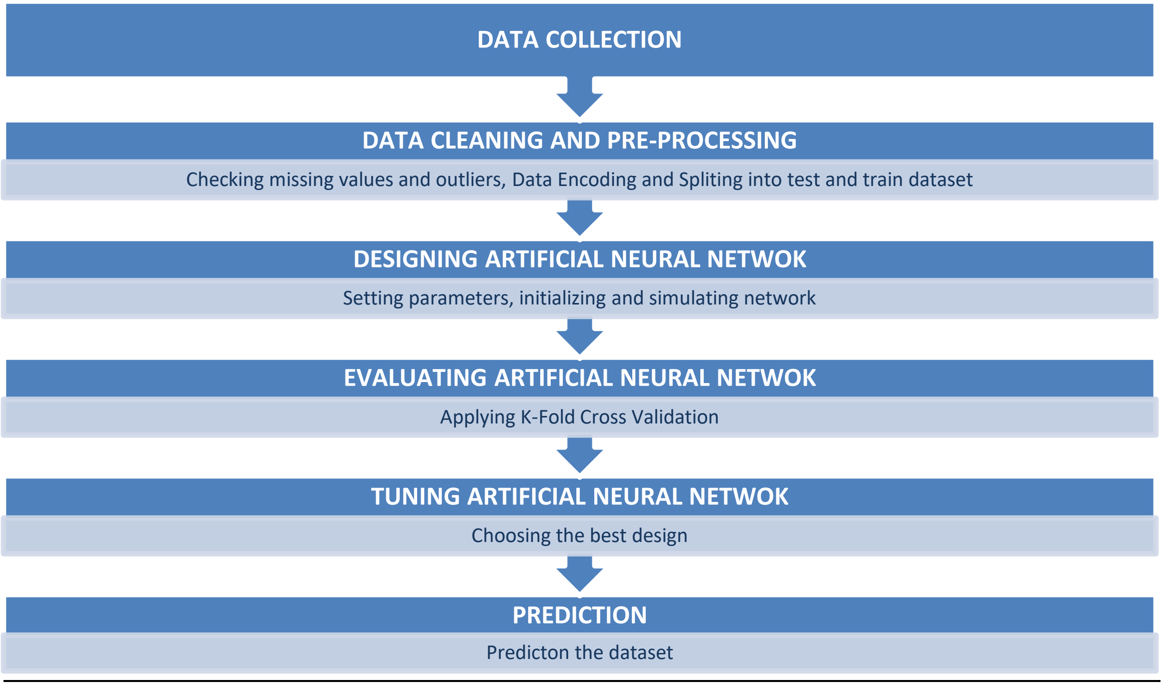
A. O. Oyeniyi & A.B. Adeyemo
Department of Computer Science,
University of Ibadan
Ibadan, Nigeria
aoyeniyi1@gmail.com, sesanadeyemo@gmail.com

ABSTRACT

Customer churn has become a major problem within a customer centred banking industry and banks have always tried to track customer interaction with the company, in order to detect early warning signs in customer's behaviour such as reduced transactions, account status dormancy and take steps to prevent churn. This paper presents a data mining model that can be used to predict which customers are most likely to churn (or switch banks). The study used real-life customer records provided by a major Nigerian bank. The raw data was cleaned, pre-processed and then analysed using WEKA, a data mining software tool for knowledge analysis. Simple K-Means was used for the clustering phase while a rule-based algorithm, JRip was used for the rule generation phase. The results obtained showed that the methods used can determine patterns in customer behaviours and help banks to identify likely churners and hence develop customer retention modalities.

Keywords: Customer, banking, data mining, churn analysis, WEKA, retention models & K-means.

METHODOLOGY



DATA COLLECTION

DATA CLEANING AND PRE-PROCESSING

Checking missing values and outliers, Data Encoding and Spliting into test and train dataset

DESIGNING ARTIFICIAL NEURAL NETWOK

Setting parameters, initializing and simulating network

EVALUATING ARTIFICIAL NEURAL NETWOK

Applying K-Fold Cross Validation

TUNING ARTIFICIAL NEURAL NETWOK

Choosing the best design

PREDICTION

Predicton the dataset

EXPECTED OUTCOME:

The expected outcome is to show that clients who use more bank services are more loyal, so bank should focus on those clients who use less than three products, and offer them products according to their needs.

CODE:

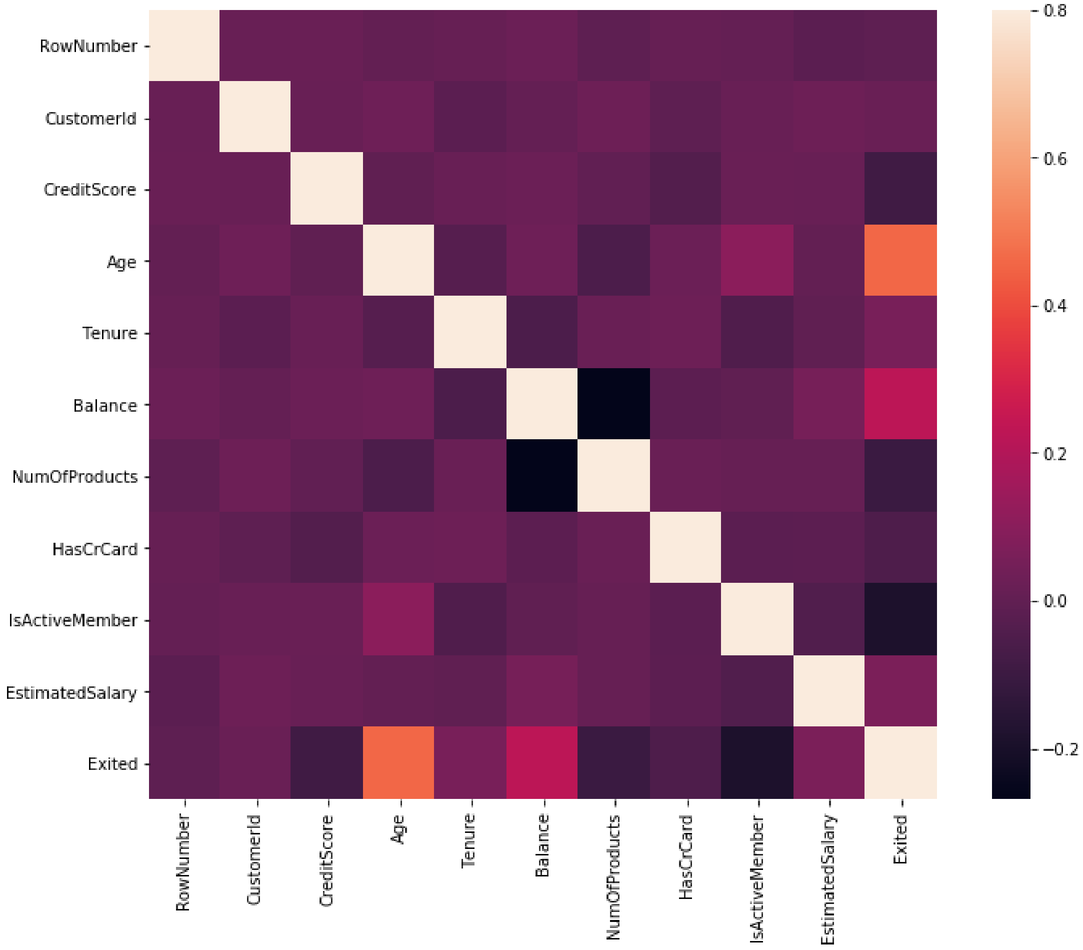
```
# Importing the libraries

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd



# Importing the dataset

dataset = pd.read_csv('Churn_Modelling.csv')

X = dataset.iloc[:, 3:13].values

y = dataset.iloc[:, 13].values


# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

labelencoder_X_1 = LabelEncoder()

X[:, 1] = labelencoder_X_1.fit_transform(X[:, 1])

labelencoder_X_2 = LabelEncoder()

X[:, 2] = labelencoder_X_2.fit_transform(X[:, 2])

onehotencoder = OneHotEncoder(categorical_features = [1])

X = onehotencoder.fit_transform(X).toarray()
```

```
X = X[:, 1:]

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)



# Importing the Keras libraries and packages
import keras
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import Dropout

# Initialising the ANN
classifier = Sequential()

# Adding the input layer and the first hidden layer
classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu', input_dim = 11))
```

```
# classifier.add(Dropout(p = 0.1))


# Adding the second hidden layer

classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu'))

# classifier.add(Dropout(p = 0.1))


# Adding the output layer

classifier.add(Dense(units = 1, kernel_initializer = 'uniform', activation = 'sigmoid'))


# Compiling the ANN

classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])


# Fitting the ANN to the Training set

classifier.fit(X_train, y_train, batch_size = 10, epochs = 100)



# Predicting the Test set results

y_pred = classifier.predict(X_test)

y_pred = (y_pred > 0.5)



# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)

# Evaluating the ANN
```

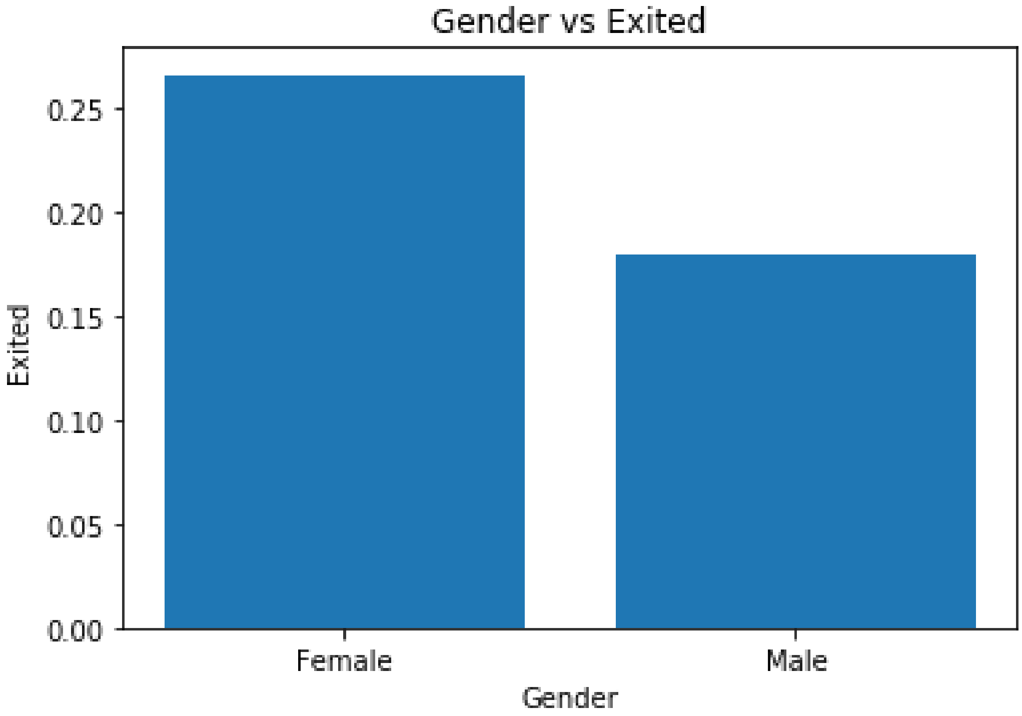
```
from keras.wrappers.scikit_learn import KerasClassifier
from sklearn.model_selection import cross_val_score
from keras.models import Sequential
from keras.layers import Dense
def build_classifier():
    classifier = Sequential()
    classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu', input_dim = 11))
    classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu'))
    classifier.add(Dense(units = 1, kernel_initializer = 'uniform', activation = 'sigmoid'))
    classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
    return classifier
classifier = KerasClassifier(build_fn = build_classifier, batch_size = 10, epochs = 100)
accuracies = cross_val_score(estimator = classifier, X = X_train, y = y_train, cv = 10, n_jobs = -1)
mean = accuracies.mean()
variance = accuracies.std()


# Improving the ANN
# Dropout Regularization to reduce overfitting if needed


# Tuning the ANN
from keras.wrappers.scikit_learn import KerasClassifier
from sklearn.model_selection import GridSearchCV
from keras.models import Sequential
from keras.layers import Dense
def build_classifier(optimizer):
    classifier = Sequential()
    classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu', input_dim = 11))
    classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu'))
    classifier.add(Dense(units = 1, kernel_initializer = 'uniform', activation = 'sigmoid'))
    classifier.compile(optimizer = optimizer, loss = 'binary_crossentropy', metrics = ['accuracy'])
    return classifier
classifier = KerasClassifier(build_fn = build_classifier)
parameters = {'batch_size': [25, 32],
              'epochs': [100, 500],
              'optimizer': ['adam', 'rmsprop']}
grid_search = GridSearchCV(estimator = classifier,
                           param_grid = parameters,
                           scoring = 'accuracy',
                           cv = 10)
grid_search = grid_search.fit(X_train, y_train)
best_parameters = grid_search.best_params_
best_accuracy = grid_search.best_score_
```

OUTPUT :

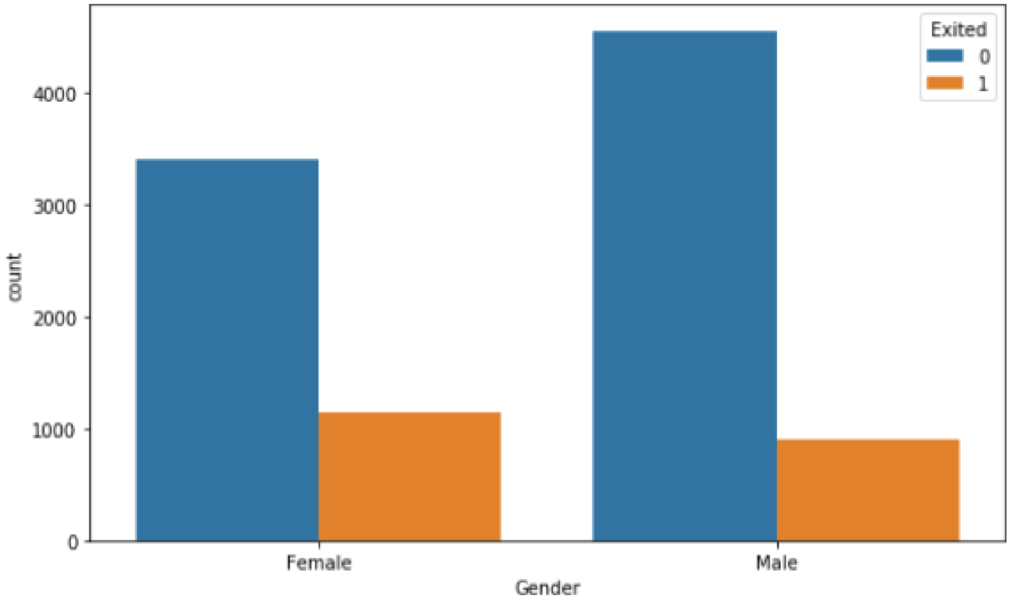


Age and Balance might play an important role for customer churn

Tableau

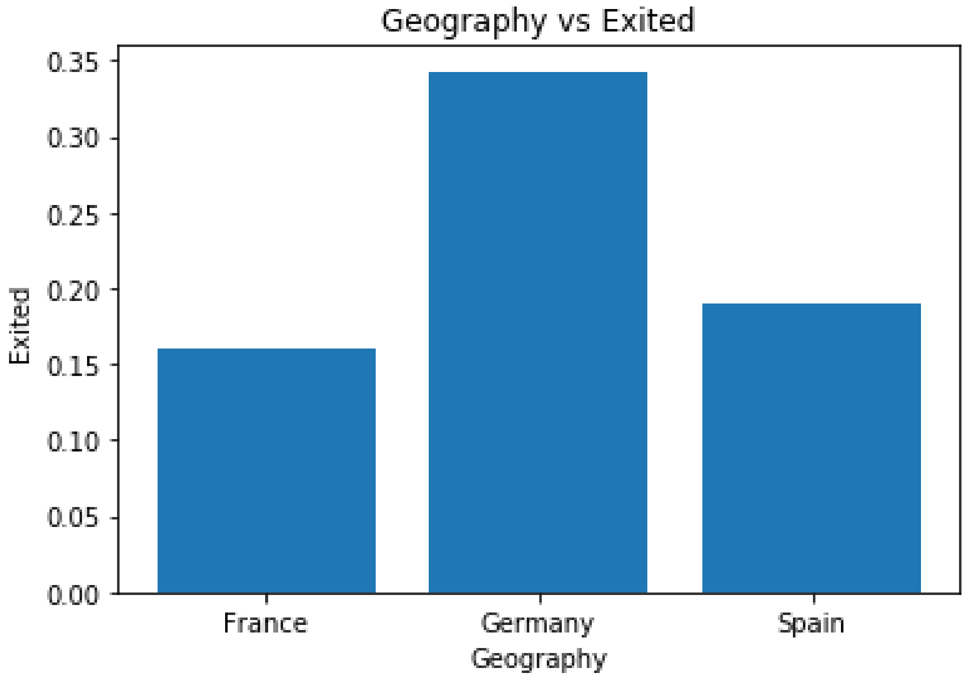


Jupyter

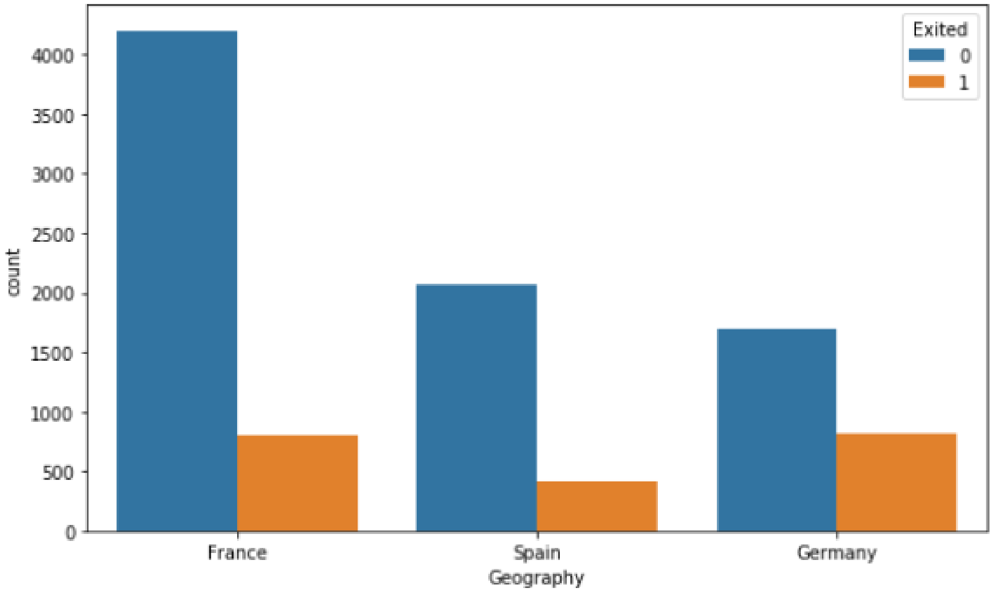


From the output above both from Jupyter and Tableau ,we can conclude that females are more likely to churn.

Tableau

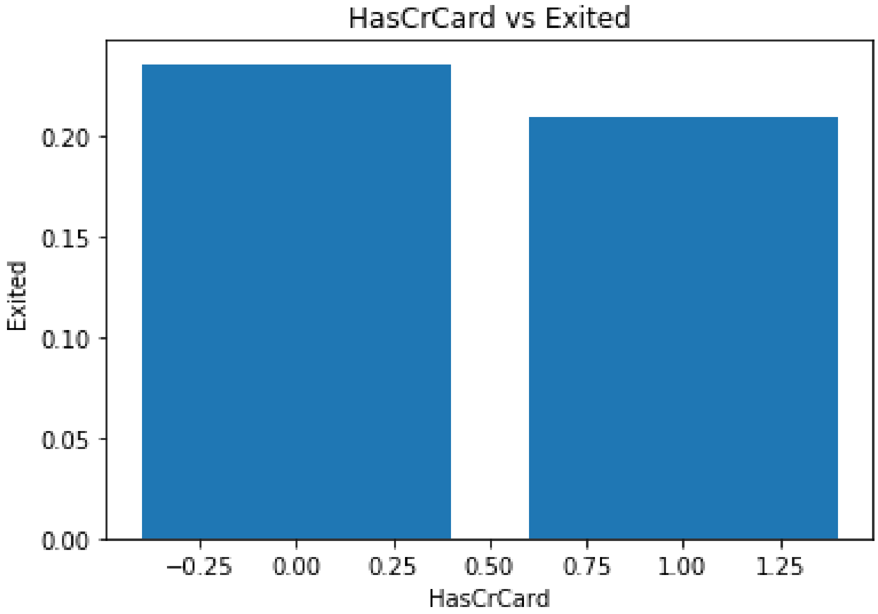


Jupyter

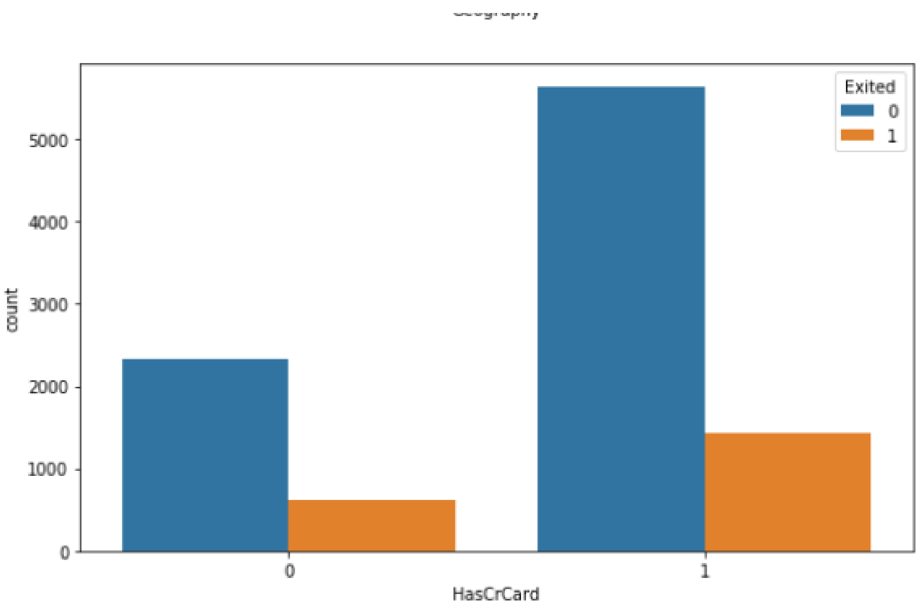


It can be concluded from the graphs above that German people are more likely to churn out of the 3 nationalities

Tableau

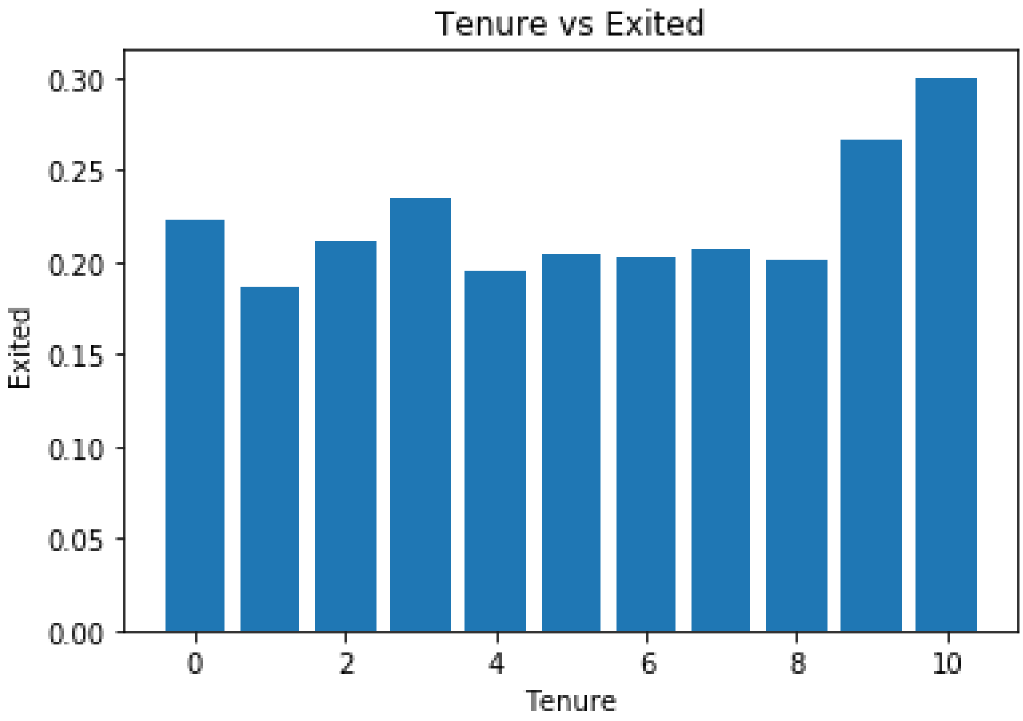


Jupyter

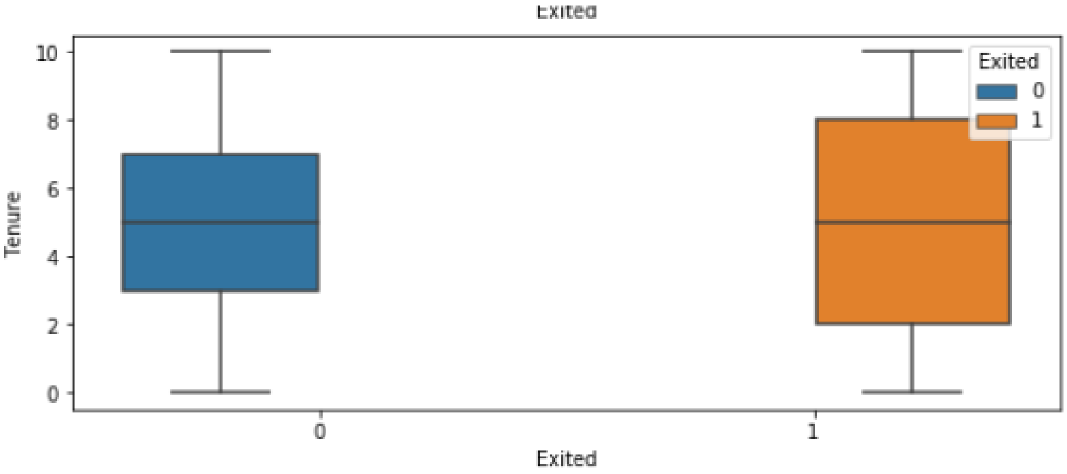


This output suggests the obvious conclusion that people having a credit card are less likely to churn.

Tableau

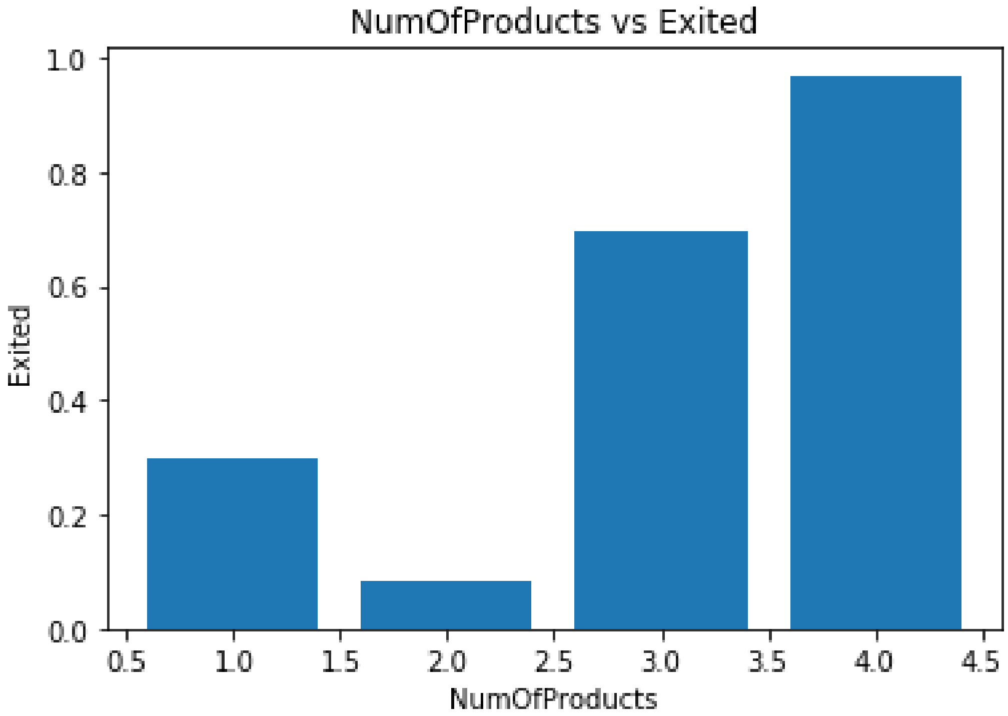


Jupyter

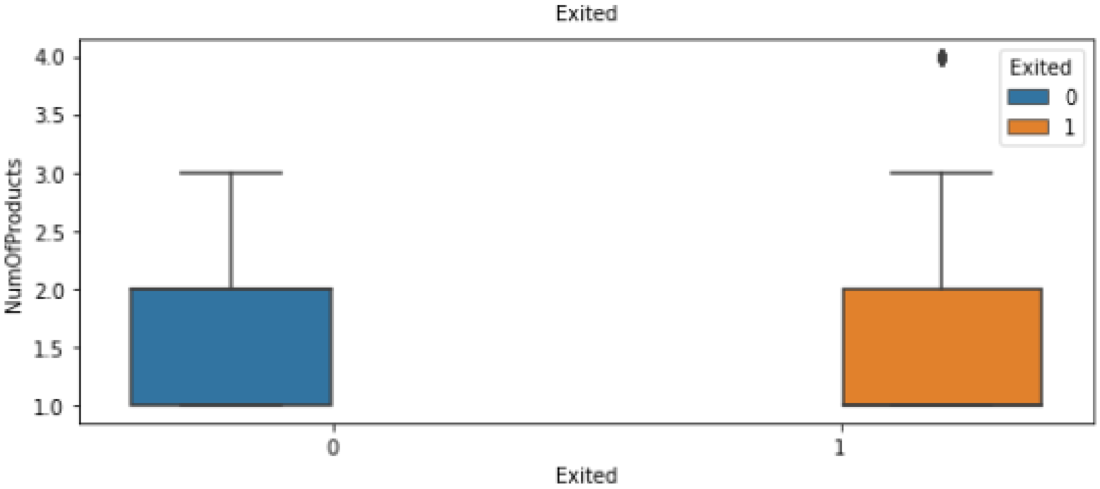


The customer with greater tenure are likely to churn.

Tableau

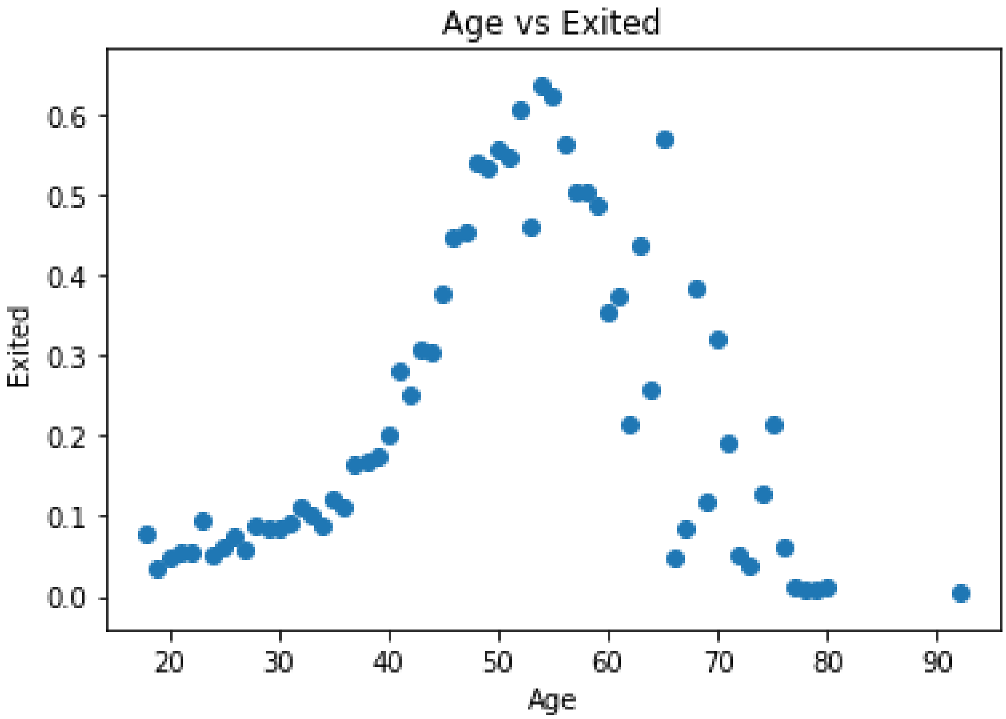


Jupyter

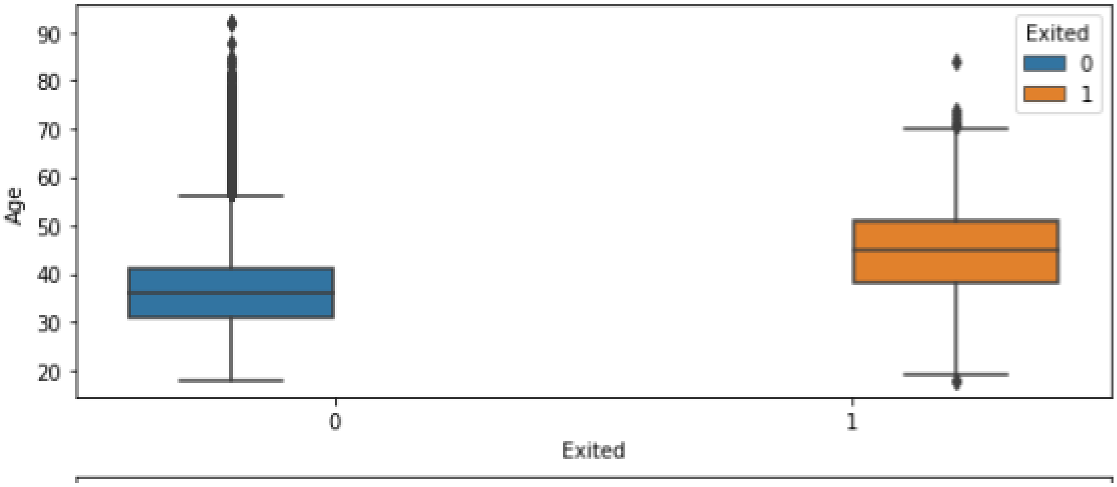


Customer with more number of Products(>3) are likely to churn

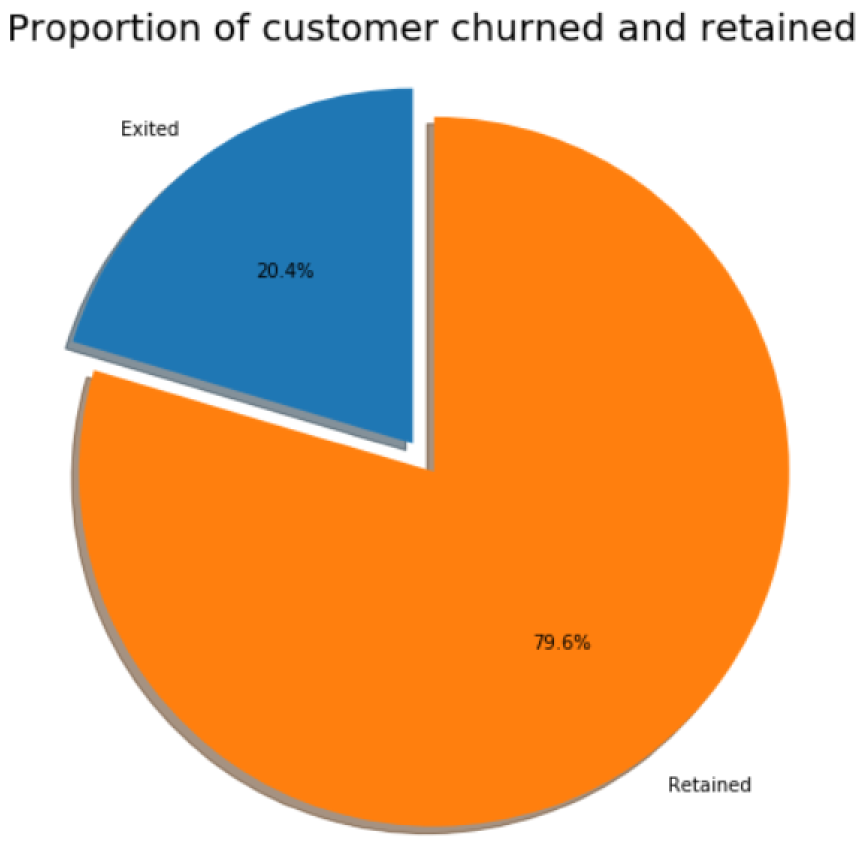
Tableau



Jupyter



The Output above suggests that age between 50-60 are likely to churn.



Pie chart of The data set

CONCLUSION :

We can clearly conclude from the above outputs that German Females Between Age 50-60 having a Credit card are likely to churn.German Banks can use this data to design policies for this very section of customers so that churn factors can be reduced for their own good.

This analysis can also be picked up by the nation French and Spanish as well to help them reduce the churn factor. These results can help the banks help them prevent turnover drops and help them build customer trust over years.

The striking similarity between the actual output and predicted output demonstrates the credibility of the data model. Being quite accurate this data model can be trusted by the banks to calculate the churn factor without worrying about the deviation of the results from the actual values.

REFERENCES :

[1] Bingquan Huang, Mohand Tahar Kechadi, Brian Buckley "Customer churn Prediction in telecommunications", Expert Systems with Applications 39 (2012) 1414–1425.

[2] Adem Karahoca, Dilek Karahoca,"GSM churn management by using fuzzy c-Means clustering and adaptive neuro fuzzy inference system", Expert Systems with Applications 38 (2011) 1814–1822.

[3] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. A. K., & Vanthienen, J. (2003). Benchmarking state of the art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.

[4] Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. European Journal of Operational Research, 138(1), 191–211.

[5] Berry, M. J. A., & Linoff, G. (1999). Data mining techniques: for marketing,

sales, And customer support. Morgan Kaufmann Publishers.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

[6] Hsieh, Nan-Chen., (2004). “An integrated data mining and behavioral scoring model for analyzing bank customers.” Expert Systems with Applications 27: 623-633.

[7] Han, J., Kamber, M., (2003). Data Mining: Concepts and Techniques, Morgan Kaufmann.

[8] Hsieh, Nan-Chen., (2004). “An integrated data mining and behavioral scoring model for analyzing bank customers.” Expert Systems with Applications 27: 623-633.

[9] Au, T., Li, S., and Ma, G., (2003). "Applying and Evaluating to Predict Customer Attrition Using Data Mining Techniques." Journal of Comparative International management 6(1).

[10] Au, W.-H., Chan, K. C. C., and Yao, X., (2003). "A Novel Evolutionary DataMining Algorithm With Application to Churn Prediction." IEEE Transactions on Evolutionary Computation 7(6): 532-545.