Customer Churn Prediction Model

The aim of this project to analyze the bank customer's demographics and financial information which inculdes customer's age, gender. country, credit score, balance and many others to predict whether the customer will leave the bank or not.

About the dataset

The dataset is taken from Kaggle. It contains 10000 rows and 14 columns. The objective of the dataset is to predict whether the customer will leave the bank or not, based on the customer's demographics and financial information included in the dataset.

The dataset has several factors that can influence the customer to leave the bank, which are termed as independent variables. The target variable is the customer's decision to leave the bank, which is termed as dependent variable.

Data Dictionary

Column Name	Description
RowNumber	Row number
CustomerId	Unique identification key for different customers
Surname	Customer's last name
CreditScore	Credit score of the customer
Geography	Country of the customer
Age	Age of the customer
Tenure	Number of years for which the customer has been with the bank
Balance	Bank balance of the customer
NumOfProducts	Number of bank products the customer is utilising
HasCrCard	Binary flag for whether the customer holds a credit card with the bank or not
IsActiveMember	Binary flag for whether the customer is an active member with the bank or not
EstimatedSalary	Estimated salary of the customer in Dollars
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained

In [ ]:

```
#importing the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:
```
#loading the dataset
df = pd.read_csv('churn.csv')
df.head()
```

Out[ ]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure
0	1	15634602	Hargrave	619	France	Female	42	2
1	2	15647311	Hill	608	Spain	Female	41	1
2	3	15619304	Onio	502	France	Female	42	8
3	4	15701354	Boni	699	France	Female	39	1
4	5	15737888	Mitchell	850	Spain	Female	43	2

Data Preprocessing 1

In [ ]:
```
#checking the shape of the dataset
df.shape
```

Out[ ]:
```
(10000, 14)
```

Dropping the unecessary columns - RowNumber, CustomerId, Surname

In [ ]:
```
#drop coulumns
df = df.drop(['RowNumber','CustomerId','Surname'], axis=1)
```

Checking for Null/Missing values

In [ ]:
```
#null values count
df.isnull().sum()
```

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

Checking the data types of the columns

In [ ]:
```
#column data types
df.dtypes
```

Out[ ]:
```
CreditScore          int64
Geography           object
Gender              object
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
Exited               int64
dtype: object
```

Checking for duplicate values

In [ ]:
```
#dulicate values
df.duplicated().sum()
```

Out[ ]: 0

Renaming the column 'Exited' to 'Churn'

In [ ]:
```
#rename column
df.rename(columns={'Exited':'Churn'},  inplace=True)
```

Descriptive Statistics

In [ ]:
```
#descriptive statistics
df.describe()
```

Out[ ]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	Has
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	1000
mean	650.528800	38.921800	5.012800	76485.889288	1.530200	
std	96.653299	10.487806	2.892174	62397.405202	0.581654	
min	350.000000	18.000000	0.000000	0.000000	1.000000	
25%	584.000000	32.000000	3.000000	0.000000	1.000000	
50%	652.000000	37.000000	5.000000	97198.540000	1.000000	
75%	718.000000	44.000000	7.000000	127644.240000	2.000000	
max	850.000000	92.000000	10.000000	250898.090000	4.000000	

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

Out[ ]:

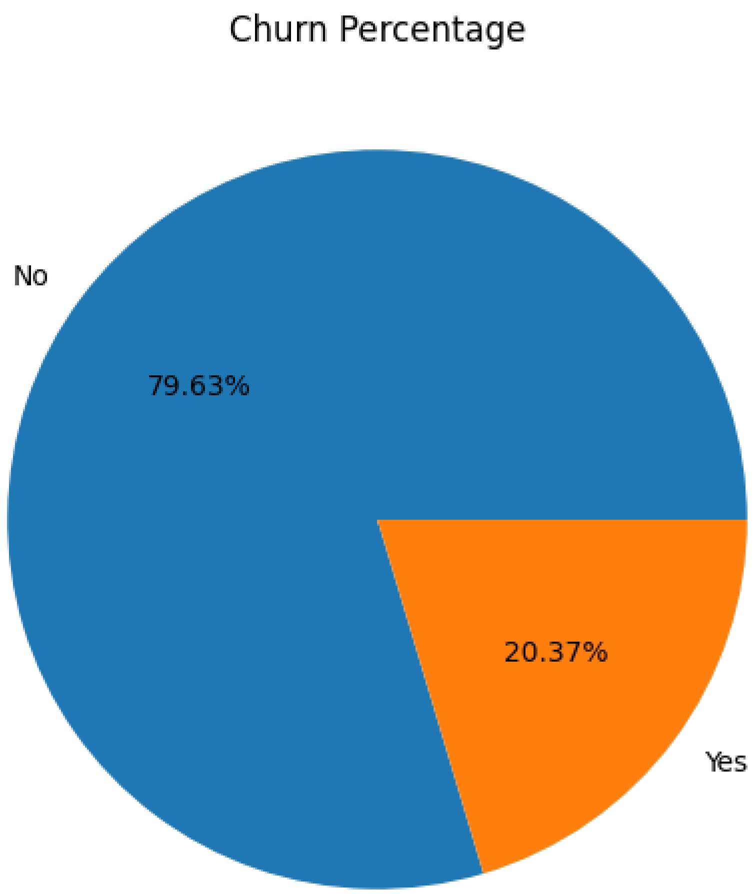
	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCa
0	619	France	Female	42	2	0.00	1	
1	608	Spain	Female	41	1	83807.86	1	
2	502	France	Female	42	8	159660.80	3	
3	699	France	Female	39	1	0.00	2	
4	850	Spain	Female	43	2	125510.82	1	

Explorative Data Analysis

In the exploratory data analysis, I will be looking at the distribution of the data, the coorelation between features and the target variable and the relationship between the features and the target variable. I will start by looking at the distribution of the data, followed by the relationship between the features and the target variable.

Pie Chart for Customer Churn

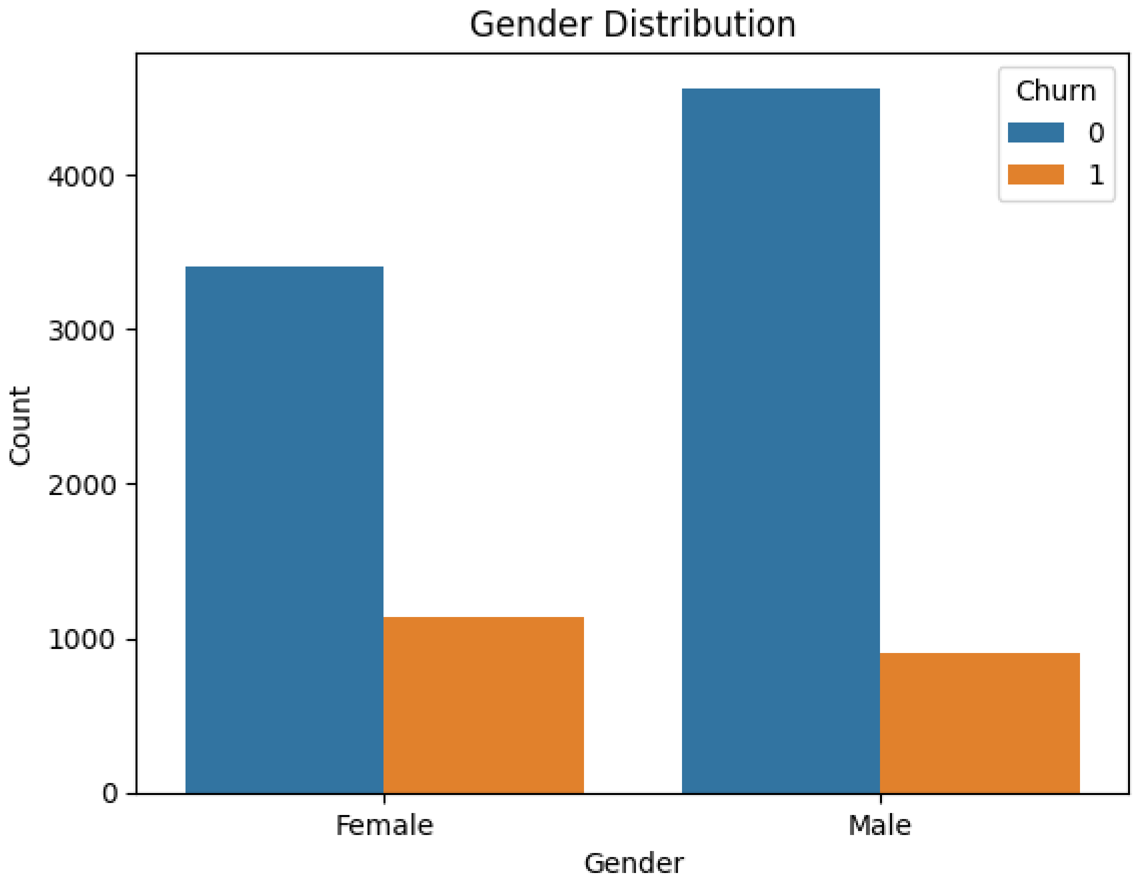
In [ ]:
```
#pie chart
plt.figure(figsize=(10,6))
plt.pie(df['Churn'].value_counts(),labels=['No','Yes'],autopct='%1.2f%%')
plt.title('Churn Percentage')
plt.show()
```



The pie chart clearly visulaizes the customer churn in the dataset. The majority of the customers in the dataset continue to use the serivces of the bank with only 20.4% of the customers churning.

Gender

In [ ]:
```
#gender and customer churn
sns.countplot(x = 'Gender', data = df, hue = 'Churn')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

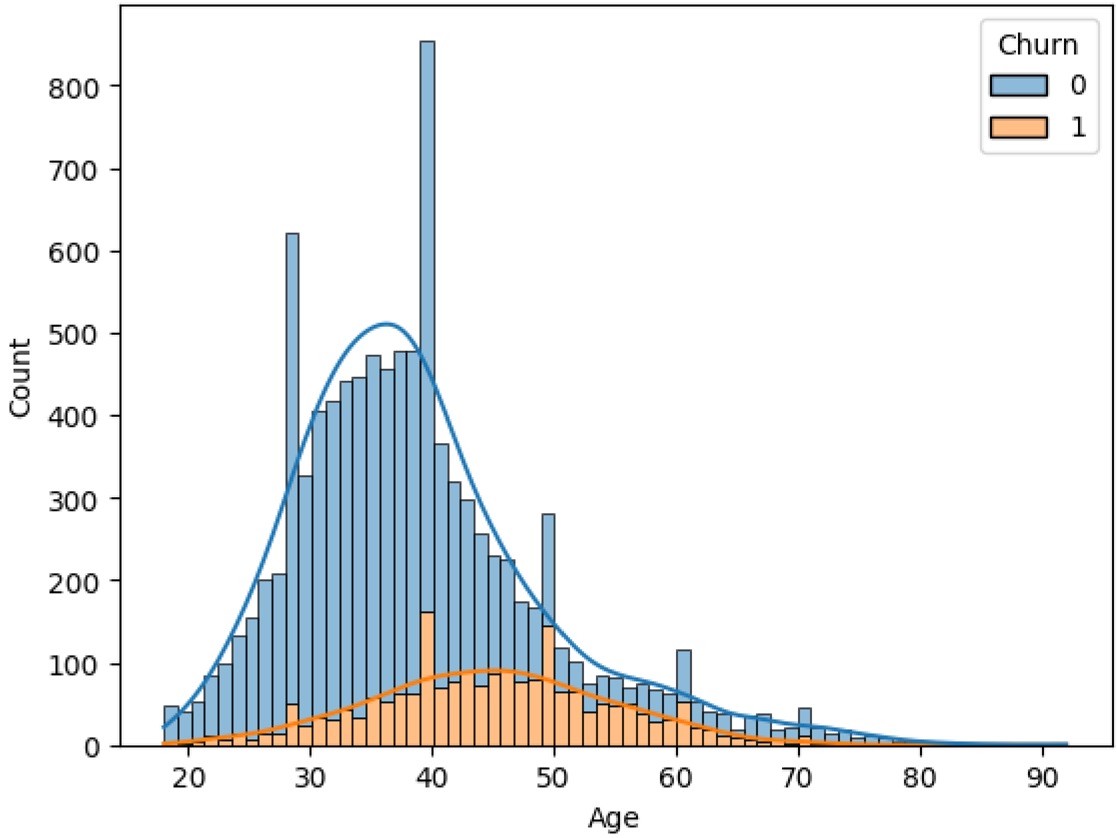


As shown in the graph, majority of the customers are male. But upon looking at the customer churn, we can see that females have more tendency to churn as compared to males. However there is not much difference between the churn count of the two genders so we cannot have a hypothesis regarding the customer churn based on the gender of the customer.

Age Distribution

In [ ]:
```
#histogram for age distribution
sns.histplot(data=df, x="Age", hue="Churn", multiple="stack",kde=True)
```

Out[ ]:
```
<Axes: xlabel='Age', ylabel='Count'>
```

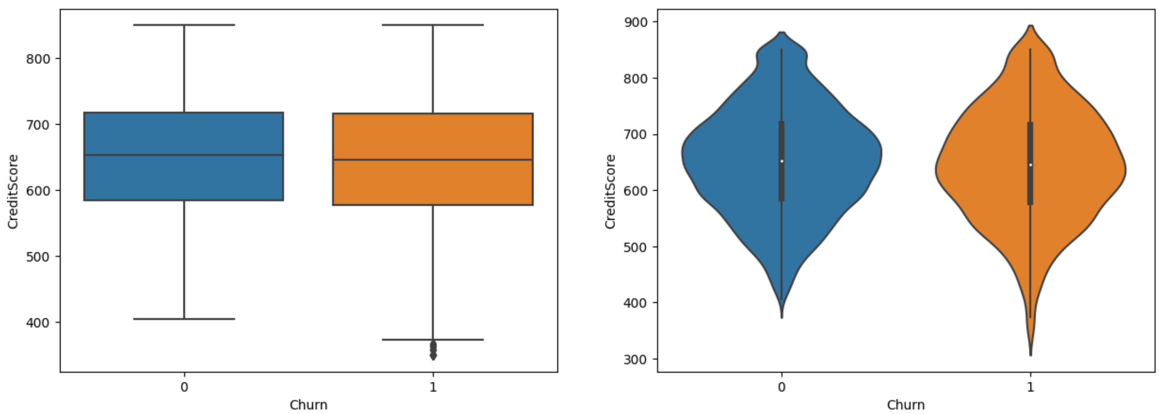


This histtogram visualizes the age distribution and the churn count of the customers. The majority of the customers are from age group 30-40 years old. However the customer churn count is highest for the customersof age 40 and 50. In addition to that customers from age group 20-25 years old count for the lowest churn count. Therefore, age plays a significant role in customer churn, where late adults are more likely to churn as compared to young adults with minimal churn count.

Credit Score

In [ ]:
```
fig, ax = plt.subplots(1,2,figsize=(15, 5))
sns.boxplot(x="Churn", y="CreditScore", data=df, ax=ax[0])
sns.violinplot(x="Churn", y="CreditScore", data=df, ax=ax[1])
```

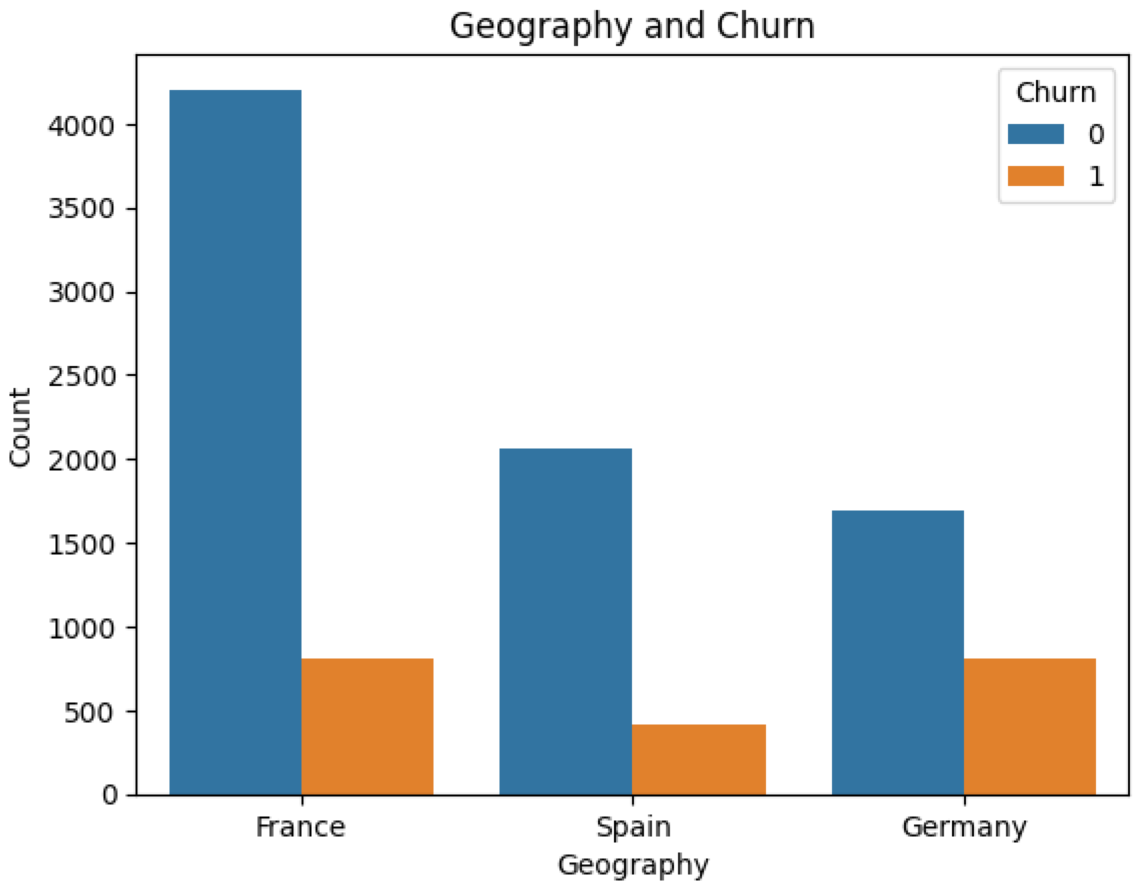
Out[ ]: <Axes: xlabel='Churn', ylabel='CreditScore'>



The boxplot and violinplot shows the distribution of curstomer's credit score along with their churn. In the boxplot, the median of both the churn and non churn customers are almost same. In addition to that, the shape of violinplot is also similar for both the churn and non churn customers. However some churn customers have low credit score, but on the whole, the credit score is not a good indicator of churn.

Customer location

In [ ]:
```
sns.countplot(x = 'Geography', hue = 'Churn', data = df)
plt.title('Geography and Churn')
plt.xlabel('Geography')
plt.ylabel('Count')
plt.show()
```

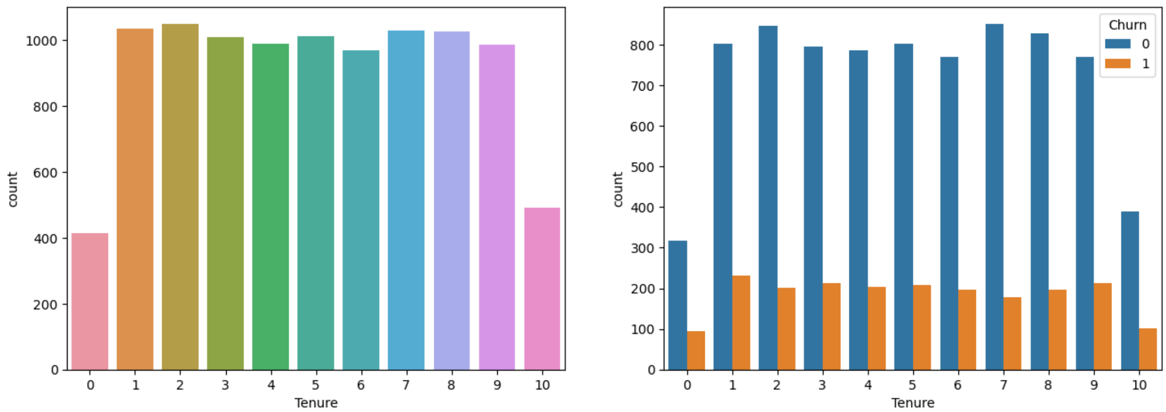


This graphs shows the number of customers from the their repective countries aling with their churn count. Majority of the customers are from France, followed by Spain and Germany. However in contrast to that Germany has the highest number of customer curn followed by France and Spain. From this we can infer that German customers are more likely to churn than the customers from other countries.

Tenure

In [ ]:
```
fig,ax = plt.subplots(1,2,figsize=(15,5))
sns.countplot(x='Tenure',  data=df,ax=ax[0])
sns.countplot(x='Tenure',  hue='Churn',  data=df,ax=ax[1])
```

Out[ ]: <Axes: xlabel='Tenure', ylabel='count'>



Tensure refers to the time (in years) that a customer has been a client of the bank. Majority of the customers in the dataset have a tenure between 1-9 years, having equal distribution among them. There are very few customers with a tenure of less than 1 years or more than 9 years. Looking at the churn of these customers based on their tenure, it can be observed that customers with tenure 1-9 years have higher churn count with maximum in customers with 1 year tenure followed those with 9 year tenure. However customers more than 9 years on tenure counts for the least churn. This is because the customers with higher tenure are more loyal to the bank and less likely to churn.

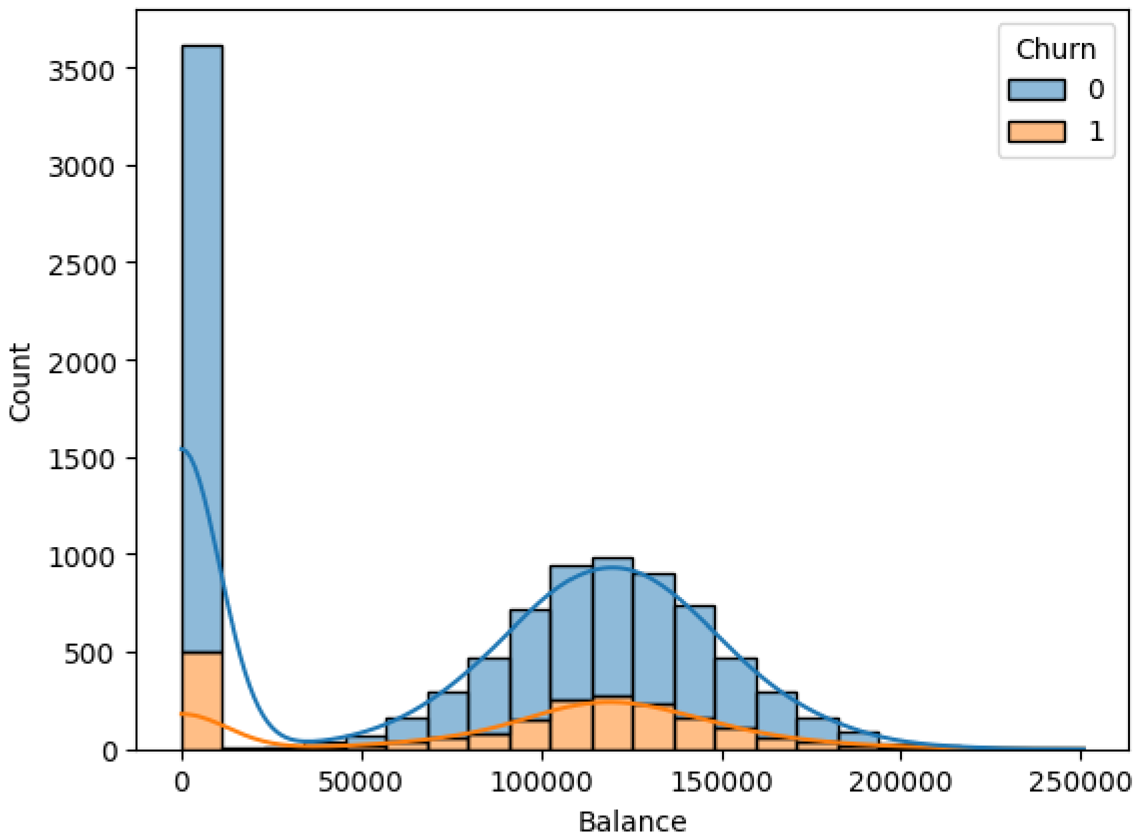
Bank Balance

In [ ]:

```
sns.histplot(data=df, x="Balance", hue="Churn", multiple="stack",kde=True)
```

Out[ ]:

```
<Axes: xlabel='Balance', ylabel='Count'>
```

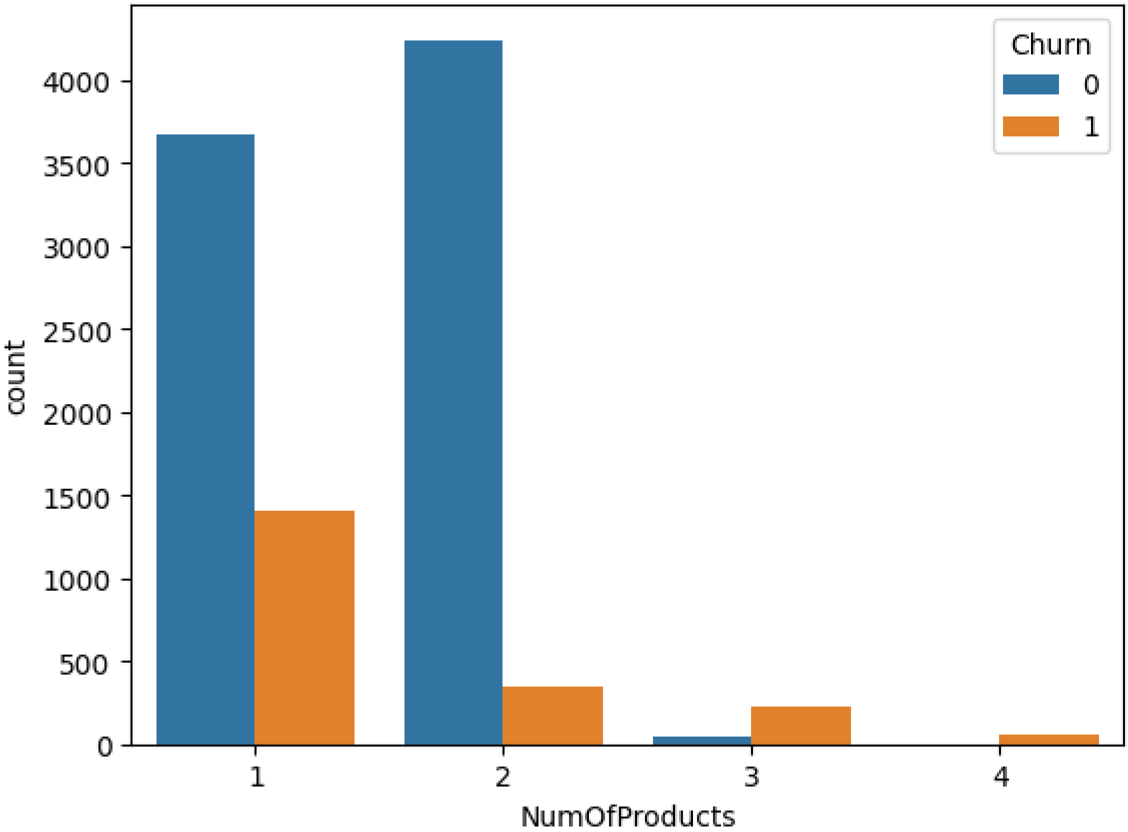


A huge number of customers have zero bank balance which also resulted in them leaving the bank. However, customer having bank balance between 100000 to 150000 are more likely to leave the bank after the customers with zero bank balance.

Number of products purchased

In [ ]:
```
sns.countplot(x='NumOfProducts', hue='Churn', data=df)
```

Out[ ]: <Axes: xlabel='NumOfProducts', ylabel='count'>

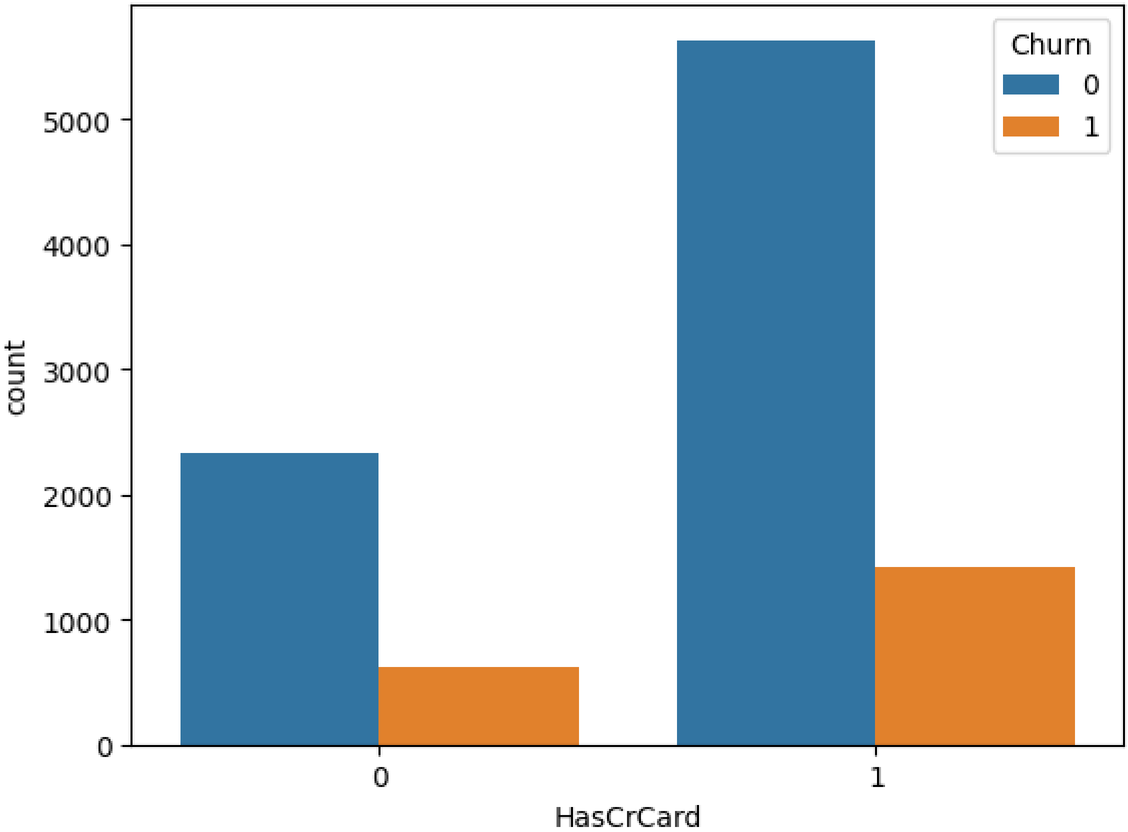


In the dataset, we have customers in four categories according to the number of products purchased. The customers with purchase or 1 or 2 products are highest in number and have low churn count in comparison to the non churn customers in the category. However, in the category where customers have purchased 3 or 4 products the number of leaving customers is much higher than the non leaving customers. Therefore, the number of product purchased is a good indicator of customer churn.

Customers with/without credit card

In [ ]:
```
sns.countplot(x=df['HasCrCard'],hue=df['Churn'])
```

Out[ ]: <Axes: xlabel='HasCrCard', ylabel='count'>

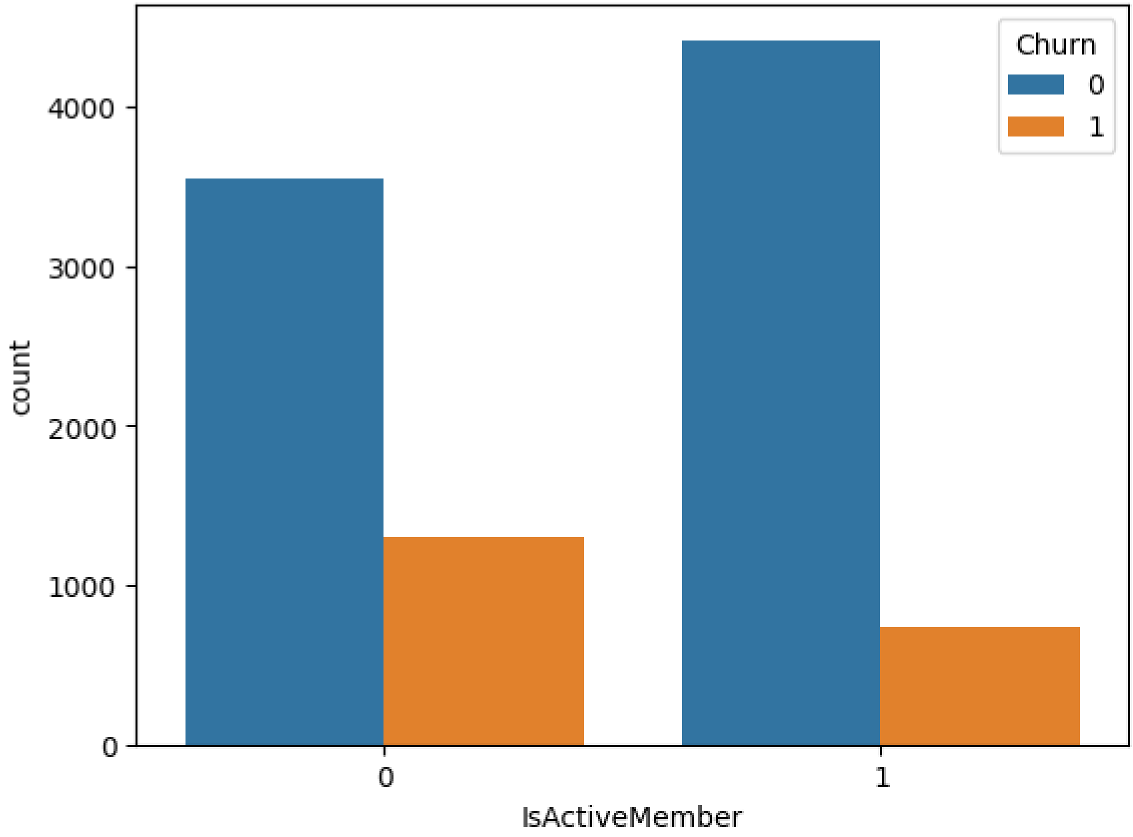


Majoity of the customers have credit cars i.e. nealy 70% of the customers have credit cards leaving 30% of the customers who do not have credit cards. Moreover, the number of customers leaving the bank are more whom have a credit card.

Active Members

In [ ]:
```
sns.countplot(x='IsActiveMember', hue='Churn', data=df)
```

Out[ ]:
```
<Axes: xlabel='IsActiveMember', ylabel='count'>
```

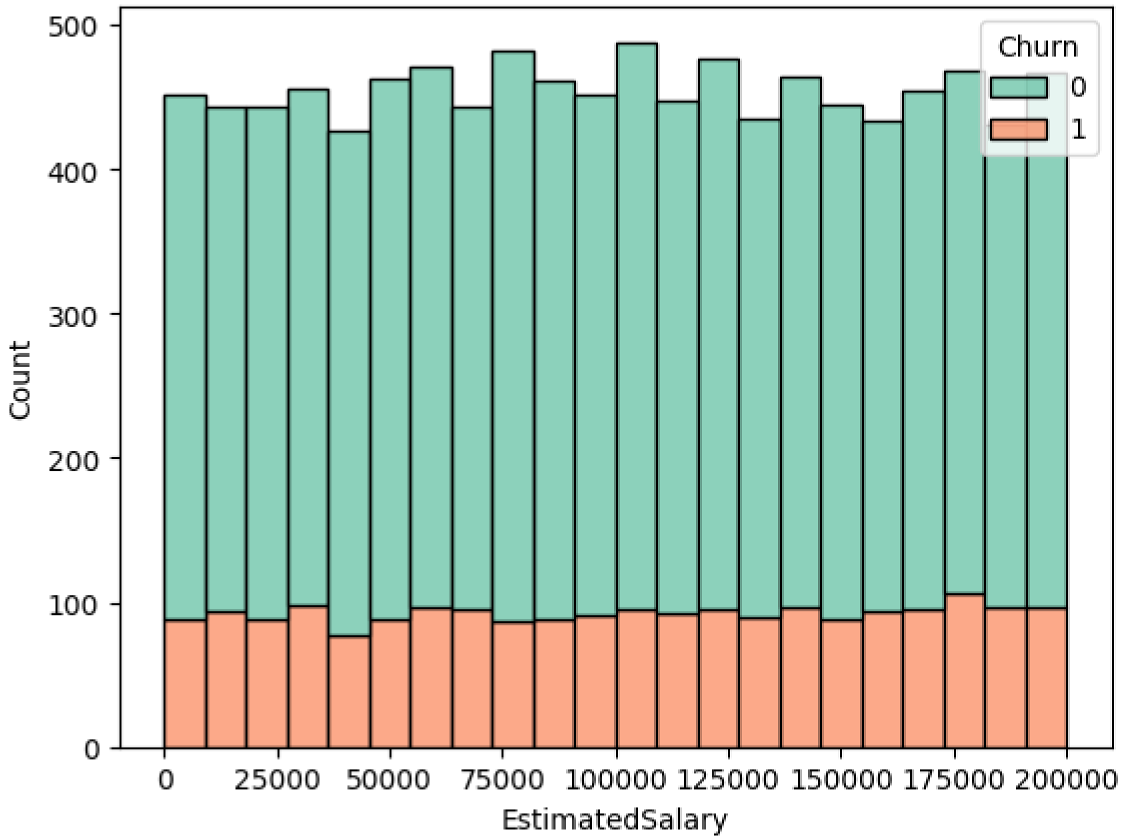


As expected, the churn count is higher for non active members as compared to the active members of the bank. This is because the active members are more satisfied with the services of the bank and hence they are less likely to leave the bank. Therefore, the bank should focus on the non active members and try to improve their services to retain them.

Estimated Salary

In [ ]:
```
sns.histplot(data=df,x='EstimatedSalary',hue='Churn',multiple='stack',palette='S
```

Out[ ]:
```
<Axes: xlabel='EstimatedSalary', ylabel='Count'>
```



This graph shows the distribution of the estimated salary of the customers along with the churn count. On the whole the there is no definite pattern in the salary distribution of the customers who churned and who didn't. Therefore estimated salary is not a good predictor of churn.

Data Preprocessing-2

Label encoding the variables

In [ ]:
```
#label encoding
variables = ['Geography','Gender']
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
for i in variables:
    le.fit(df[i].unique())
    df[i]=le.transform(df[i])
    print(i,df[i].unique())
```

```
Geography [0 2 1]
Gender [0 1]
```

Normalization

In [ ]:
```
#normalize the continuous variables
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['CreditScore','Balance','EstimatedSalary']]  =  scaler.fit_transform(df[['Cred
```

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

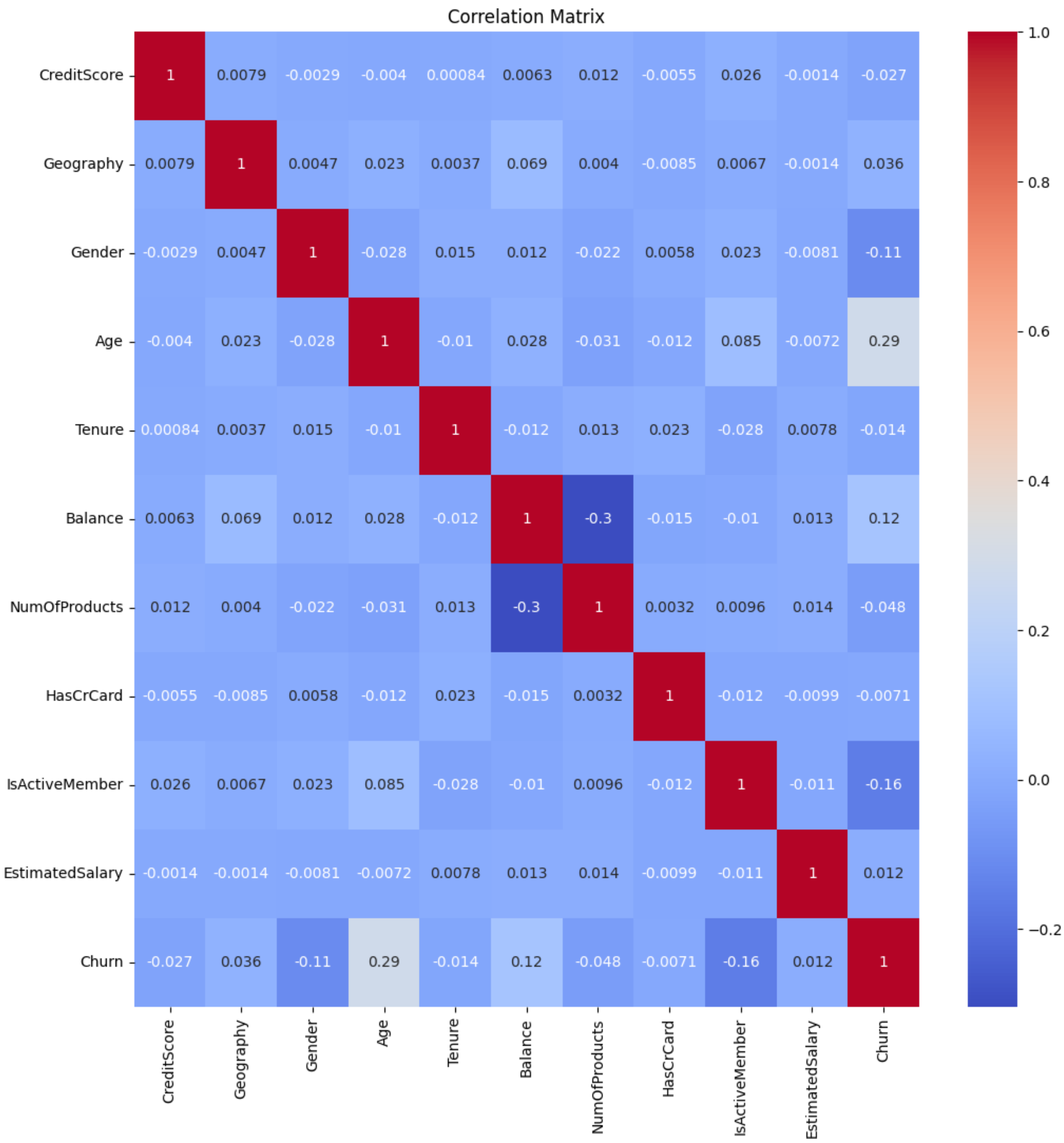
Out[ ]:

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCa
0	-0.326221	0	0	42	2	-1.225848	1	
1	-0.440036	2	0	41	1	0.117350	1	
2	-1.536794	0	0	42	8	1.333053	3	
3	0.501521	0	0	39	1	-1.225848	2	
4	2.063884	2	0	43	2	0.785728	1	

Coorelation Matrix Heatmap

In [ ]:

```
plt.figure(figsize=(12,12))
sns.heatmap(df.corr(),annot=True,cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



There is no significant coorelation among the variables. So, I will proceed to model building.

Train Test Split

In [ ]:
```
#train test split
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(df.drop('Churn',axis=1),df['Churn
```

Churn Prediction

For predicting the churn of customers, depending on the data of the customers, we will use the following models:

- Decision Tree Classifier
- Random Forest Classifier

Decision Tree Classifier

Using GridSearchCV to find the best parameters for the model.

In [ ]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
#creating Decision Tree Classifer object
dtree = DecisionTreeClassifier()

#defining parameter range
param_grid = {
    'max_depth':  [2,4,6,8,10,12,14,16,18,20],
    'min_samples_leaf':  [1,2,3,4,5,6,7,8,9,10],
    'criterion': ['gini',  'entropy'],
    'random_state': [0,42]
    }

#Creating grid search object
grid_dtree = GridSearchCV(dtree, param_grid, cv = 5, scoring = 'roc_auc', n_jobs

#Fitting the grid search object to the training data
grid_dtree.fit(X_train,  y_train)

#Printing the best parameters
print('Best  parameters  found:  ', grid_dtree.best_params_)
```

```
Fitting 5 folds for each of 400 candidates, totalling 2000 fits
Best parameters found:  {'criterion': 'gini', 'max_depth': 6, 'min_samples_leaf':
10, 'random_state': 42}
```

Adding the parameters to the model

In [ ]:
```
dtree  = DecisionTreeClassifier(criterion='gini',  max_depth=6,  random_state=42,  m
dtree
```

Out[ ]: ▾ DecisionTreeClassifier

```
DecisionTreeClassifier(max_depth=6, min_samples_leaf=10, random_state=4
2)
```

In [ ]:
```
#training the model
dtree.fit(X_train,y_train)
#training accuracy
dtree.score(X_train,y_train)
```

Out[ ]: 0.8581428571428571

Predicting Customer Churn from Test set

In [ ]:
```
dtree_pred = dtree.predict(X_test)
```

Random Forest Classifier

In [ ]:
```
from sklearn.ensemble import RandomForestClassifier
#creating Random Forest Classifer object
rfc = RandomForestClassifier()

#defining parameter range
param_grid = {
    'max_depth': [2,4,6,8,10],
    'min_samples_leaf': [2,4,6,8,10],
    'criterion': ['gini', 'entropy'],
    'random_state': [0,42]
    }

#Creating grid search object
grid_rfc = GridSearchCV(rfc, param_grid, cv = 5, scoring = 'roc_auc', n_jobs = -

#Fitting the grid search object to the training data
grid_rfc.fit(X_train, y_train)

#Printing the best parameters
print('Best parameters found: ', grid_rfc.best_params_)
```

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best parameters found:  {'criterion': 'entropy', 'max_depth': 10, 'min_samples_le
af': 8, 'random_state': 0}
```

Adding the parameters to the model

In [ ]:
```
rfc = RandomForestClassifier(min_samples_leaf=8, max_depth=10, random_state=0, c
rfc
```

Out[ ]: ▾ RandomForestClassifier

```
RandomForestClassifier(criterion='entropy', max_depth=10, min_samples_l
eaf=8,
                       random_state=0)
```

In [ ]:
```
#training the model
rfc.fit(X_train, y_train)
#model accuracy
rfc.score(X_train, y_train)
```

Out[ ]:
```
0.8767142857142857
```

Predicting the customer churn from Test set

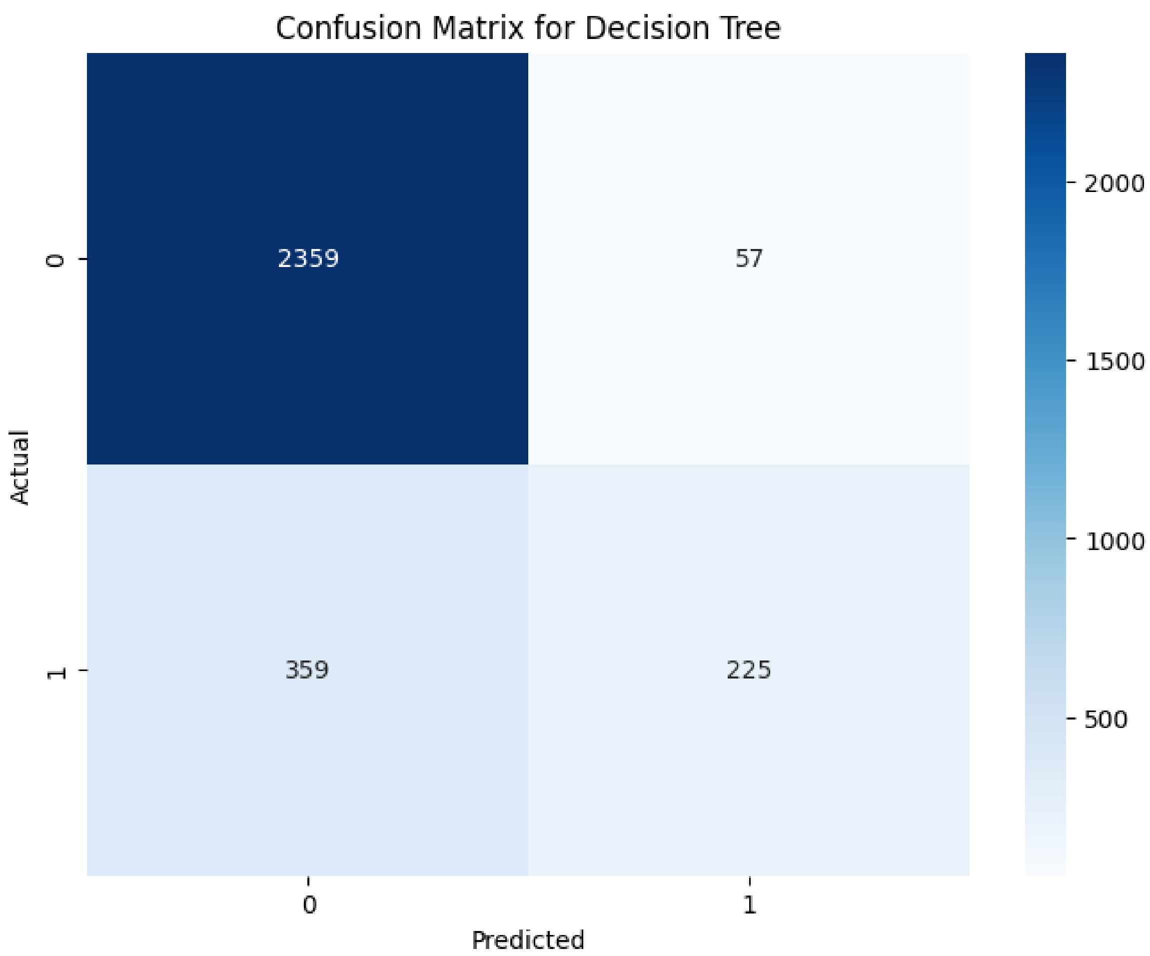
In [ ]:
```
rfc_pred = rfc.predict(X_test)
```

Model Evalution

Decision Tree Classifier

Confusion Matrix Heatmap

In [ ]:
```
#confusion matrix heatmap
from sklearn.metrics import confusion_matrix
plt.figure(figsize=(8,6))
sns.heatmap(confusion_matrix(y_test,dtree_pred),annot=True,fmt='d',cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix for Decision Tree')
plt.show()
```

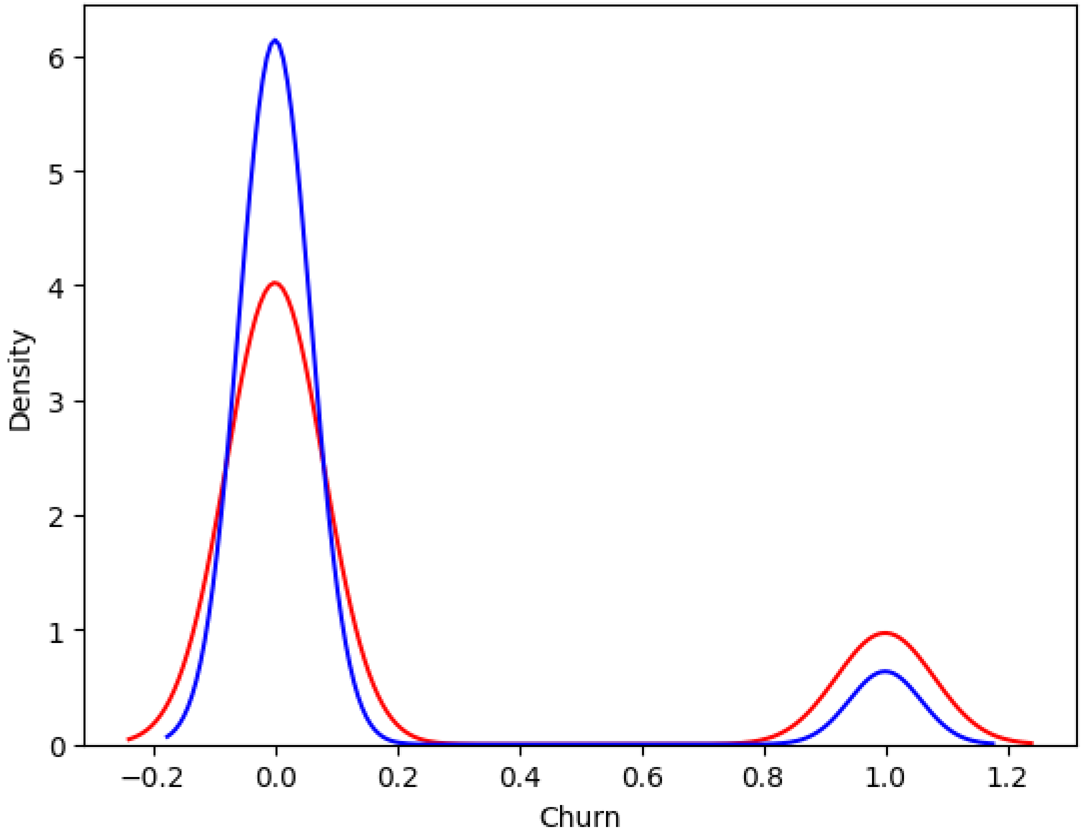


The True Positive shows the count of correctly classified data points whereas the False Positive elements are those that are misclassified by the model. The higher the True Positive values of the confusion matrix the better, indicating many correct predictions.

Distribution Plot

In [ ]:
```
ax = sns.distplot(y_test, hist=False, color="r", label="Actual Value")
sns.distplot(dtree_pred, hist=False, color="b", label="Fitted Values" , ax=ax)
```

Out[ ]:
```
<Axes: xlabel='Churn', ylabel='Density'>
```



The more overlapping of two colors, the more accurate the model is.

Classification Report

In [ ]:
```
from sklearn.metrics import classification_report
print(classification_report(y_test,  dtree_pred))
```

```
              precision    recall  f1-score   support

           0       0.87      0.98      0.92      2416
           1       0.80      0.39      0.52       584

    accuracy                           0.86      3000
   macro avg       0.83      0.68      0.72      3000
weighted avg       0.85      0.86      0.84      3000
```

In [ ]:
```
from sklearn.metrics import accuracy_score, mean_absolute_error, r2_score
print("Accuracy Score: ", accuracy_score(y_test, dtree_pred))
```

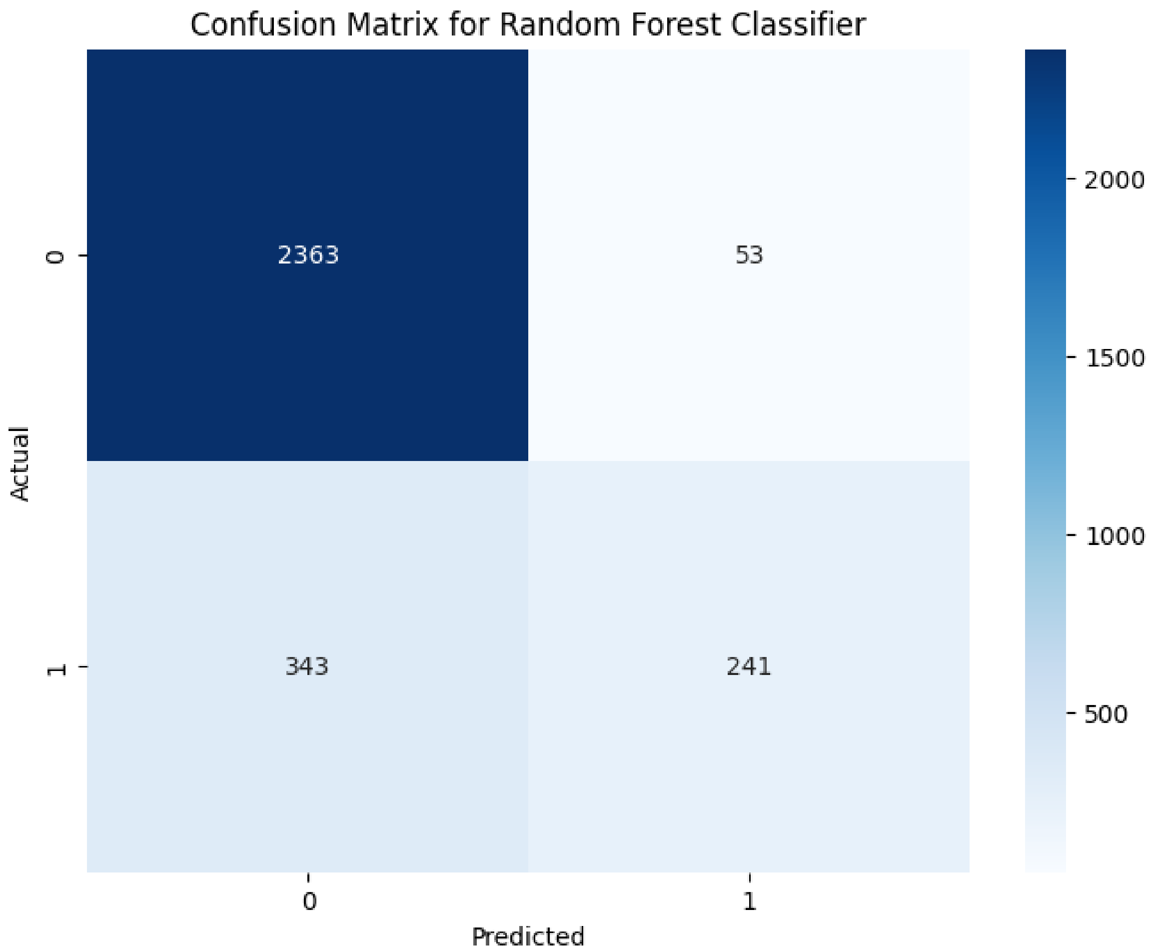
```
print("Mean Absolute Error: ", mean_absolute_error(y_test, dtree_pred))
print("R2 Score: ", r2_score(y_test, dtree_pred))
```

```
Accuracy Score:  0.8613333333333333
Mean Absolute Error:  0.13866666666666666
R2 Score:  0.11548580241313633
```

Random Forest Classifier

Confusion Matrix Heatmap

In [ ]:
```
plt.figure(figsize=(8,6))
sns.heatmap(confusion_matrix(y_test,rfc_pred),annot=True,fmt='d',cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix for Random Forest Classifier')
plt.show()
```

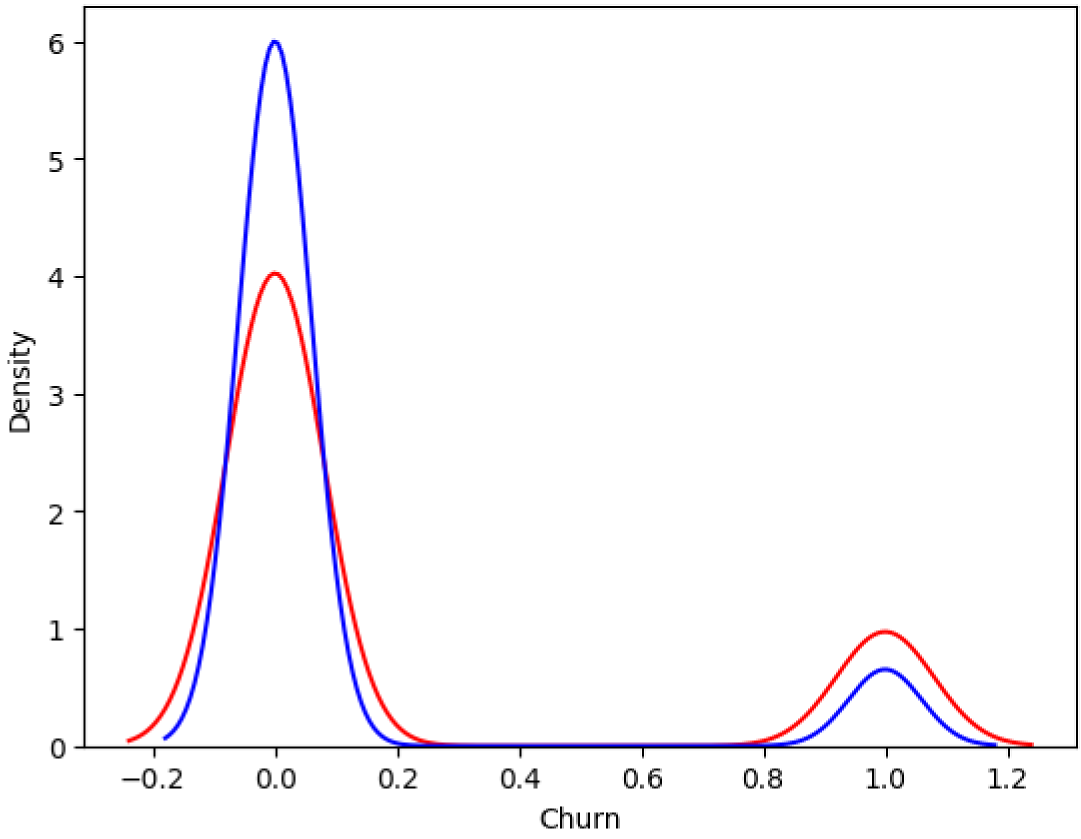


The True Positive shows the count of correctly classified data points whereas the False Positive elements are those that are misclassified by the model. The higher the True Positive values of the confusion matrix the better, indicating many correct predictions.

Distribution Plot

In [ ]:
```
ax = sns.distplot(y_test, hist=False, color="r", label="Actual Value")
sns.distplot(rfc_pred, hist=False, color="b", label="Fitted Values" , ax=ax)
```

Out[ ]: <Axes: xlabel='Churn', ylabel='Density'>



Classification Report

In [ ]:
```python
from sklearn.metrics import classification_report
print(classification_report(y_test, rfc_pred))
```

```
              precision    recall  f1-score   support

           0       0.87      0.98      0.92      2416
           1       0.82      0.41      0.55       584

    accuracy                           0.87      3000
   macro avg       0.85      0.70      0.74      3000
weighted avg       0.86      0.87      0.85      3000
```

In [ ]:
```python
print("Accuracy Score: ", accuracy_score(y_test, rfc_pred))
print("Mean Absolute Error: ", mean_absolute_error(y_test, rfc_pred))
print("R2 Score: ", r2_score(y_test, rfc_pred))
```

```
Accuracy Score:  0.868
Mean Absolute Error:  0.132
R2 Score:  0.15801052345096633
```

Conclusion

From the exploratory data analysis, I have concluded that the churn count of the customersdepends upon the following factors:

1. Age
2. Geography

3. Tenure
4. Balance
5. Number of Products
6. Has Credit Card
7. Is Active Member

Coming to the classification models, I have used the following models:

1. Decision Tree Classifier
2. Random Forest Classifier

Both the models were hyperparameter tuned using GridSearchCV. Both the models have nearly equal accuracy score. But, the Random Forest Classifier has a better accuracy and precision score than the Decision Tree Classifier.