Main objective of this analysis

Churn rate, in its broadest sense, is a measure of the number of individuals or items moving out of a collective group over a specific period. It is one of two primary factors that determine the steady-state level of customers a business will support. For this project I am concerned with churn rate of a bank. The bank wants to know given some details about its customer whether the customer will churn or not in 6 months time.

Dataset

The dataset contains information of a bank regarding their customers. The dataset contain various features of customer and also one target variable which stated whether the customer stayed with the bank or left.

The dataset is quit big which contains 10000 elements and 14 rows. The dataset included various information about the customer

- Name
- customer Id
- credit score
- location of the customer
- gender
- age
- how many products the customer had with the bank
- Whether the customer had credit card or not
- is the customer an active member
- how much is the salary of the customer.
- How long the customer has been part of the bank

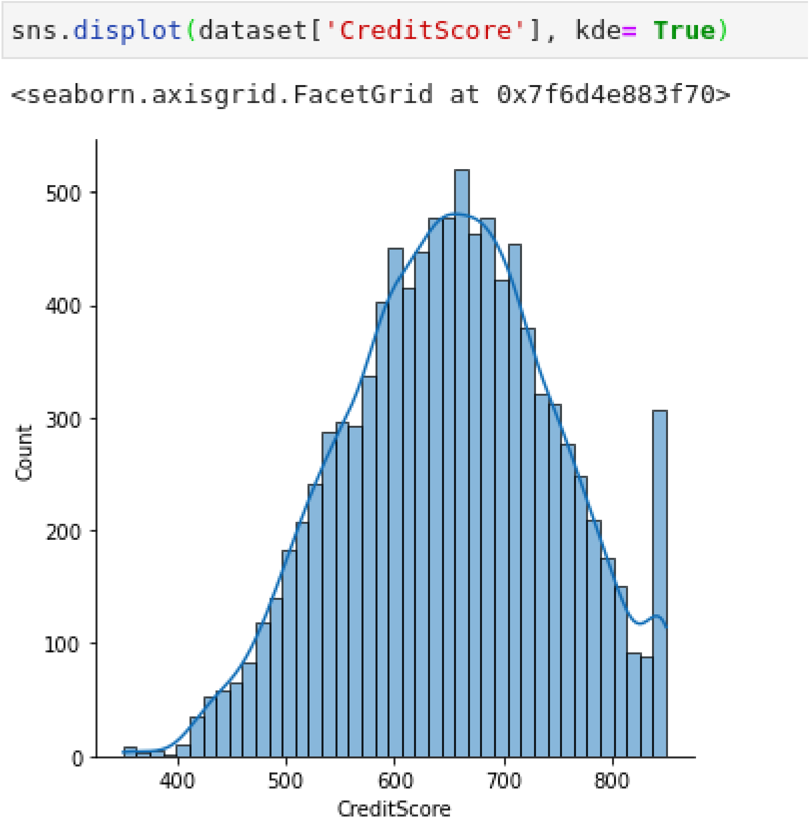
3 Analysis Method to be used

This is a supervised learning problem as the target variable is known to us. Hence, for this assignment I will use various Deep learning methods specially Artificial Neural Network with 3 variations in number of layers, number of nodes in each layer and also changing the activation function. I will find out which one gives the best result in the test set.

This problem is not a CNN or RNN problem. ANN will be sufficient to find the underlying pattern in the data. As the requirement of the project, I have trained 3 variations of the ANN as described above.

Exploratory Data Analysis

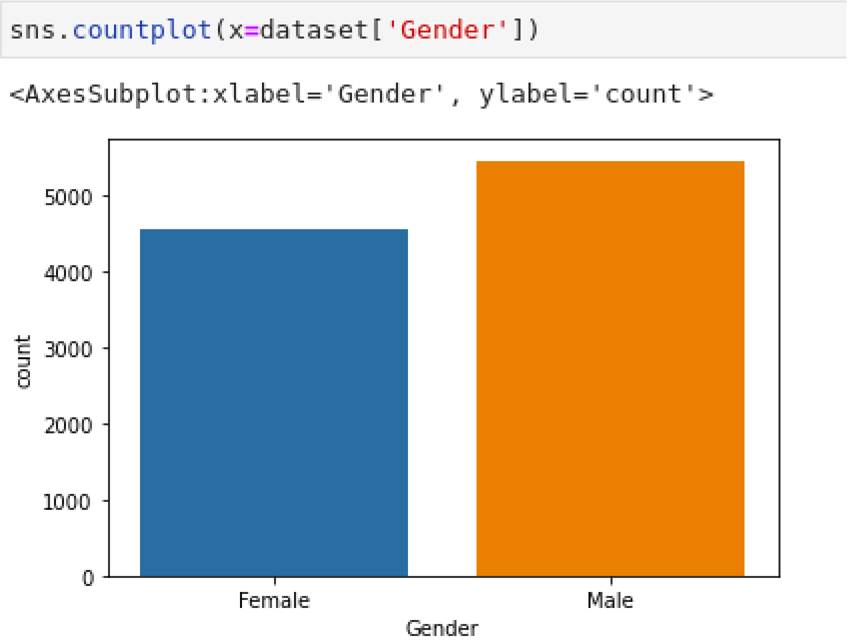
1. Credit Score of the customer to some extent follows a normal distribution. Majority of them lie in the region of 600 - 750



```
sns.displot(dataset['CreditScore'], kde= True)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f6d4e883f70>
```

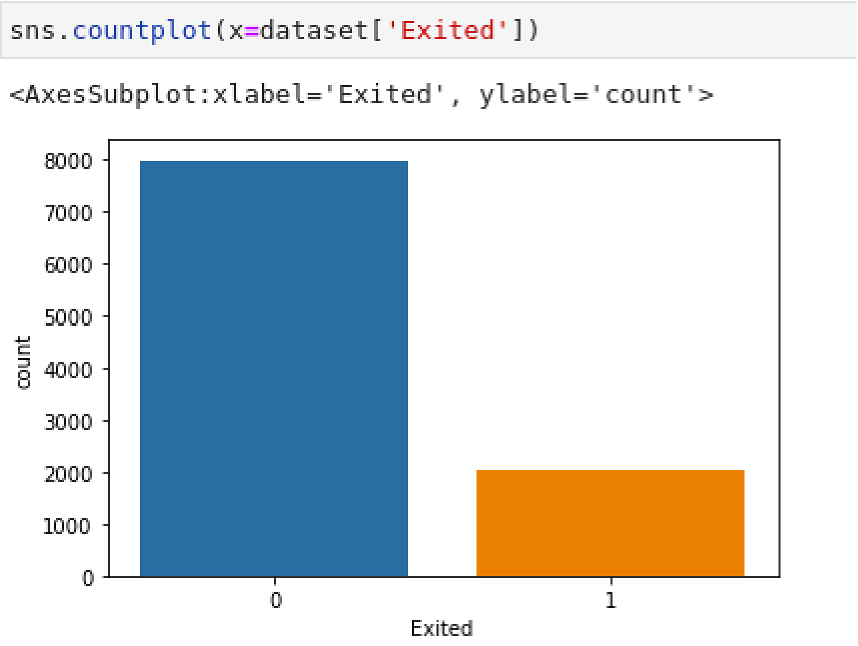
2. There are more male customers than female



```
sns.countplot(x=dataset['Gender'])
```

```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```

3. There are more people who stayed in the bank rather than leaving.



```
sns.countplot(x=dataset['Exited'])
```

```
<AxesSubplot:xlabel='Exited', ylabel='count'>
```

Data Cleaning

There is 13 predictor variables and one target variable – Exited. Two of the predictor variables are qualitative and hence needs to be encoded into dummy variables. Country and Gender variables has been converted into categorical and for Country 3 different columns were introduced and for gender 2 of them were introduced. The good thing about this dataset is that there is no null values that are present and hence did not require any kind of impute to be performed. Once the dummy variables are created, our dataset is split into two parts one for training set and other for test set. I have split the dataset into 75 : 25 ratio.

Model Training

As discussed above, I have trained 3 variations ANN for this project.

1. I have trained one ANN where we have 16 nodes in the input layer, one output node which predicts whether the customer stayed or not and one hidden layer having 6 nodes. The activation function used for the ANN was relu in the hidden layer and sigmoid in the output layer.

2. For the second model I have increased the number of hidden layers to 3 but kept each hidden layers with 6 nodes. I did not change the activation function for the hidden layers.

3. The last ANN model that I have trained have the similar architecture as that of the first one but this time I have changed the activation function for the hidden layer to sigmoid function. I have

Best Model out of 3 models that are trained

With accuracy being the evaluation matrix and also keeping in mind that the model not overfit the dataset i.e. have low bias and low variance, I have chosen the first model as the best one out of the 3 that I have trained. The ANN model had 16 nodes in the input layer, one output node which predicts whether the customer stayed or not and one hidden layer having 6 nodes. The activation function used for the ANN was relu in the hidden layer and sigmoid in the output layer.

Suggestions for Improvements:

I could have many regularisation techniques to improve the accuracy of the model. For examples, Adam optimiser, stochastic gradient descent, ridge regularisation. This techniques make the model better and better. Also since, the bank is operational in 3 countries, we could get more observations about customers. There can also be a variable which indicated whether the customer has given any rating about the service of the bank or not, that would be one variable which would be directly dependent on whether the customer stayed or not.