Decreasing the Customer Churn is a key goal for any business. Predicting Customer Churn (also known as Customer Attrition) represents an additional potential revenue source for any business. Customer Churn impacts the cost to the business. Higher Customer Churn leads to loss in revenue and the additional marketing costs involved with replacing those customers with new ones.

In this challenge, as a data scientist of a bank, you are asked to analyze the past data and predict whether the customer will churn or not in the next 6 months. This would help the bank to have the right engagement with customers at the right time.

Objective

Our objective is to build a machine learning model to predict whether the customer will churn or not in the next six months.

Churn is defined in business terms as 'when a client cancels a subscription to a service they have been using.' In this case members who are not satisfied with bank services or due to their financial problems are likely to churn.

There are any factors influence the reasons for a customer to Churn. It may be the fact that there's a new competitor in the market offering better prices or maybe the service they are getting has not been up to the mark, or customers financial problems so on and so forth.

My approach in this case study is to find out the factors affecting churn by doing EDA on each feature comparing it to target and take the relevant features and build a DecisionTreeClassification model for prediction.

Data-preprocessing / Feature Engineering -

1)Dusing data analysis we found that there was **no missing value in data**

2) we have 5 object type elements - **'ID' is unique and irrelevant for analysis** so we are not going to use it for analysis.

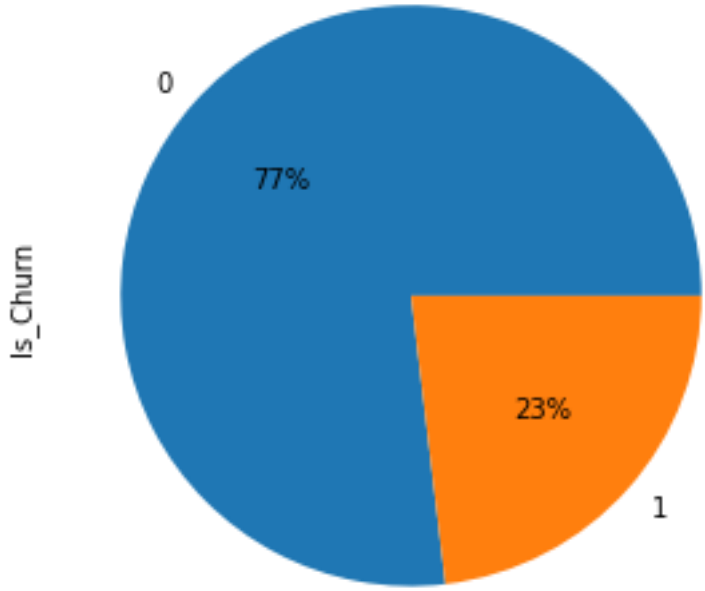
3)Categorical values like Gender Product_Holdings, Credit_Category, Income effects the churn rate - which we are going to analyze.

- **we can encode this categorical data to numerical values for model building**

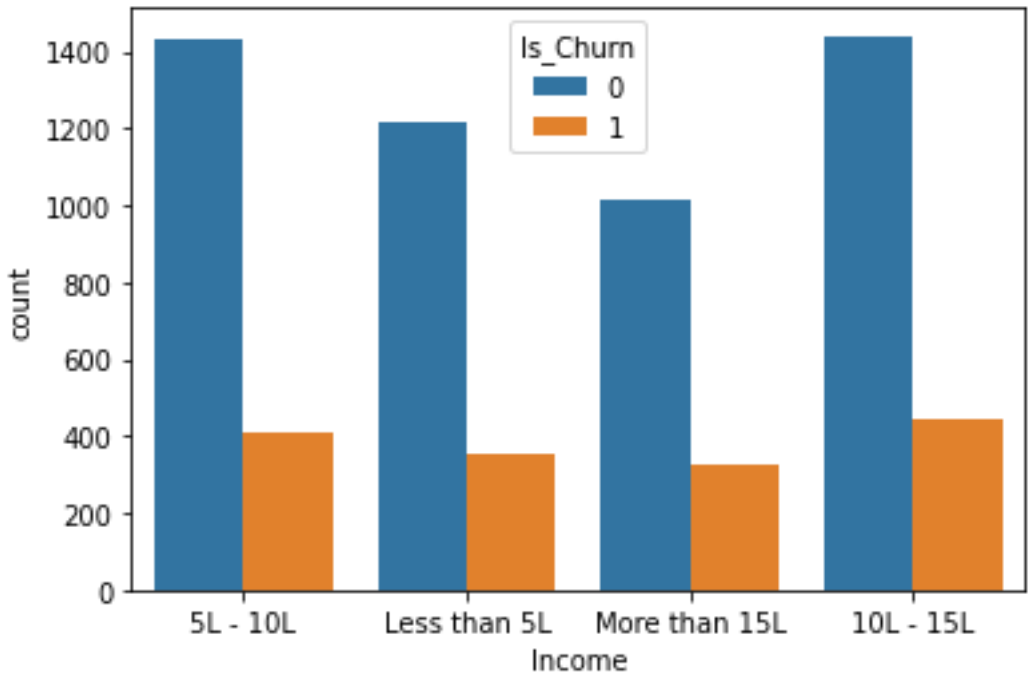
4) the Is_Churn column is int64 making my target binary feature

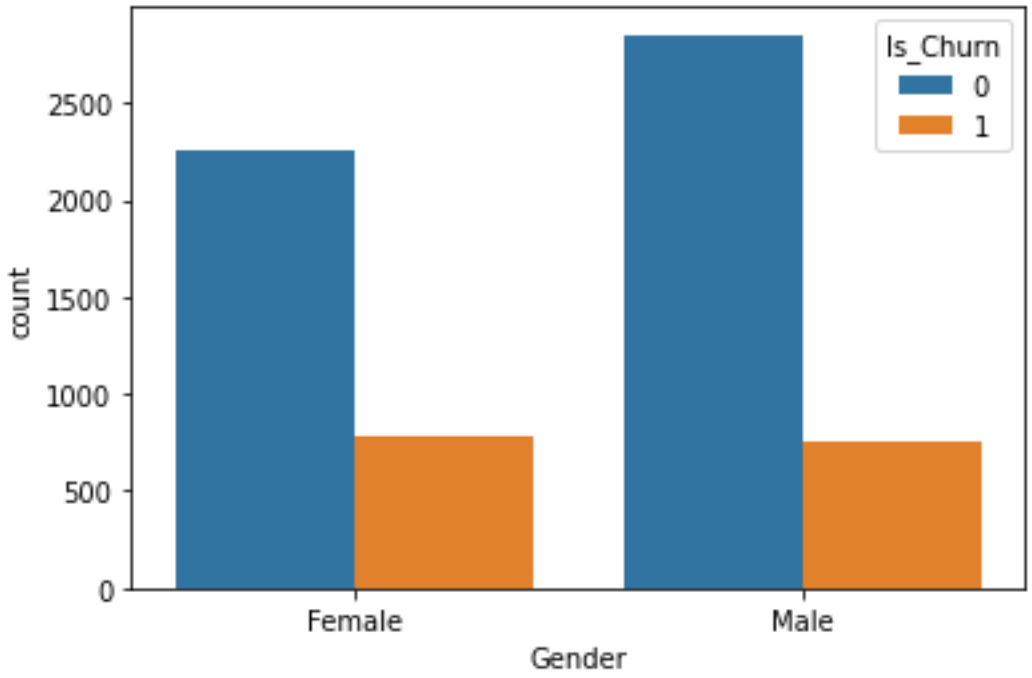
It shows that distribution that almost 23% of our total training data members are churned (1) due to some factors lets see the analysis to find the reason.

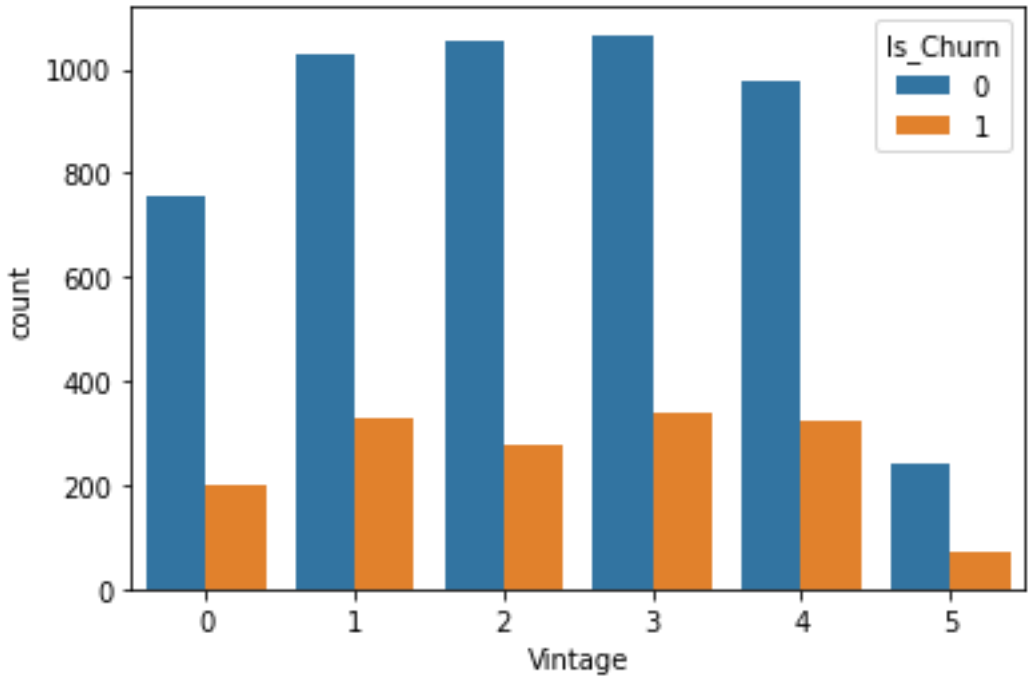
From this we can say that data is **imbalanced and need to up-sampling before creating model**

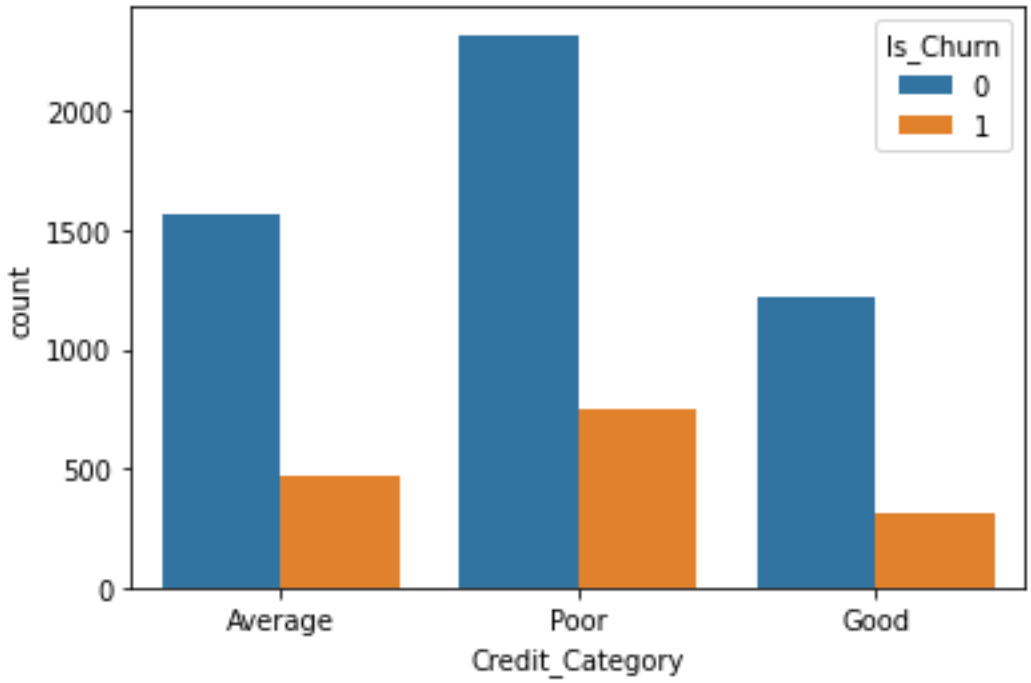


Univariate Analysis -




Is_Churn
0
1
2500
2000
1500
1000
500
0
count
Female
Male
Gender

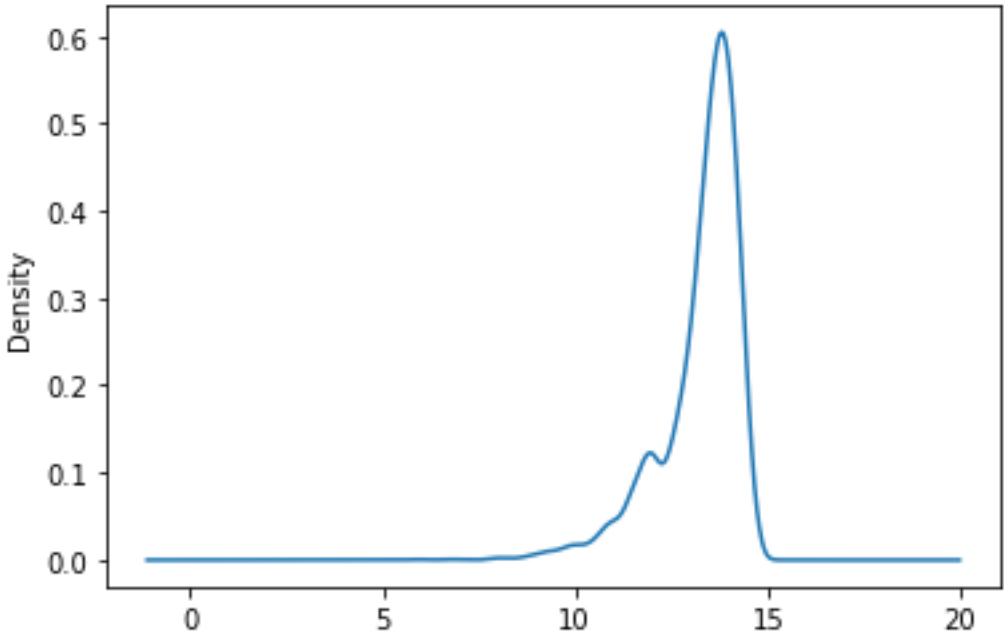

Is_Churn
0
1
1000
800
600
400
200
0
count
0
1
2
3
4
5
Vintage



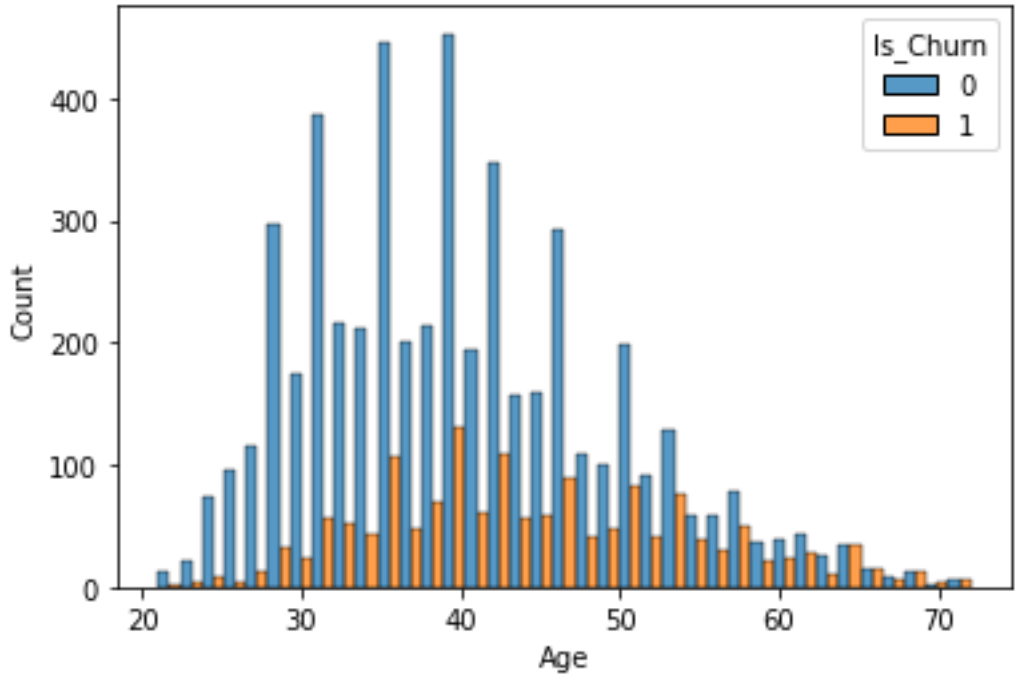
Both male and female are churning are same rate, People with income between 5-10L and between 10-15L have slightly more chances to churn similarly are the bars are giving similar relationship with churn

Data Preparation We need to make sure that the data is in the right form to be used for prediction. Machine Learning models do not work well with categorical inputs. So, we convert the categorical variables in our data set to numerical values by using **one-hot encoding and up-sample our data.**

Also the plot for 'Balance' after **log transform** showing highly right skewed data.



Even if we use this features or not for analysis it is not affecting much our results. But income is an important feature to determine if the customer is financially stable so we will take log/10 for our analysis.



This is showing how age 35-50 people are most likely to churn.

Conclusion –

For predicting the Churn value after 6 months we need time series data which is not into our dataset but DecisionTreeClassification gives us predict_proba() function to predict the probabilities for churn for upcoming future, So that we can take action to retain the customer.