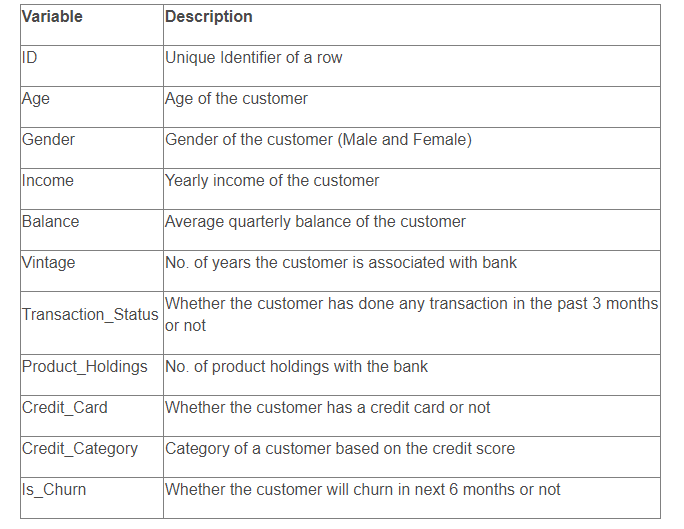
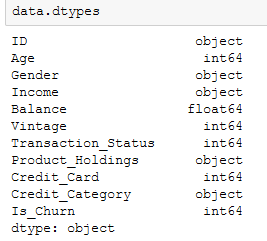
**Data:**

* Data is having 10 independent columns one dependent column
* Shape of data is (6650,11)
* Business description/information of columns/data is

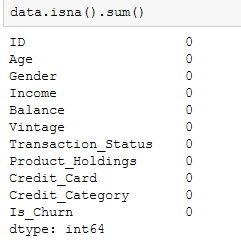


**Exploratory Data Analysis:**

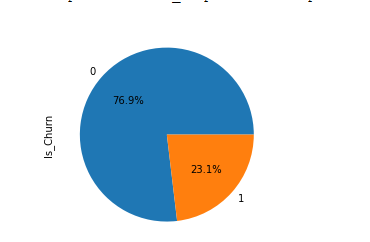
* Types of column in the data is



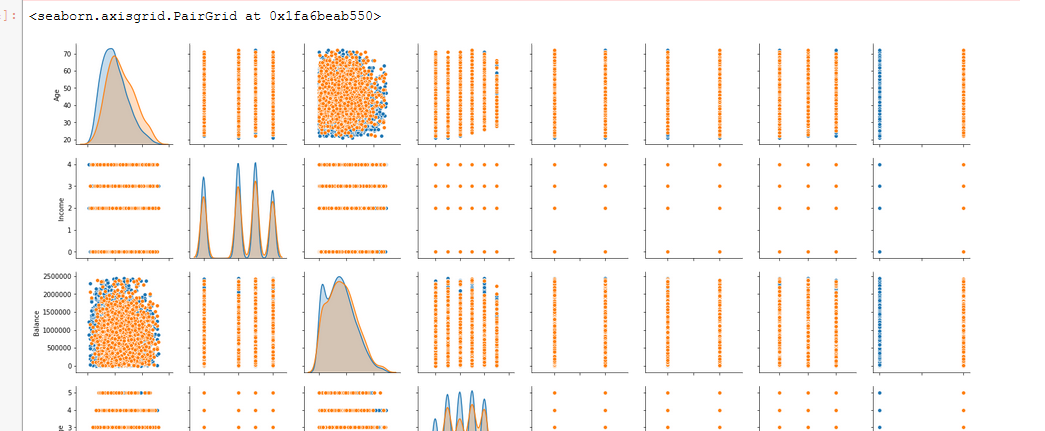
* There are no nulls in the data



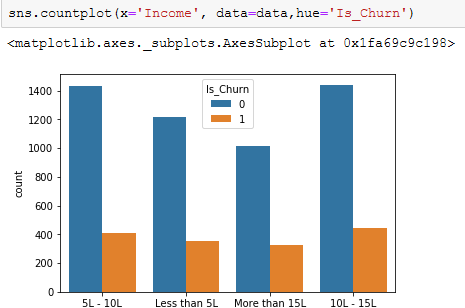
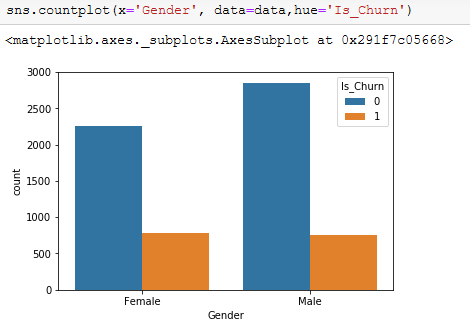
* Percentage of different classes in the data, and we can say that data is not very imbalanced.

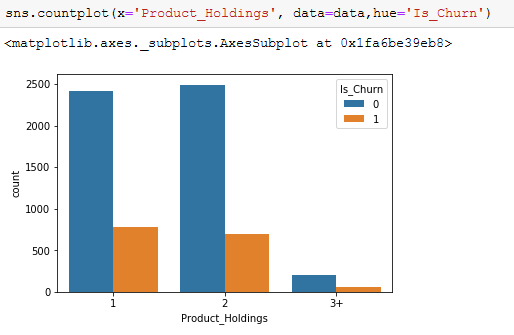
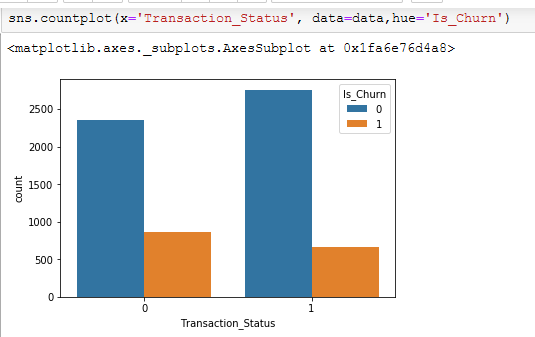


* By using pairplot understood that the distribution of data in different column with respective to columns is almost similar making it difficult to get differentiate factor to predict between classes

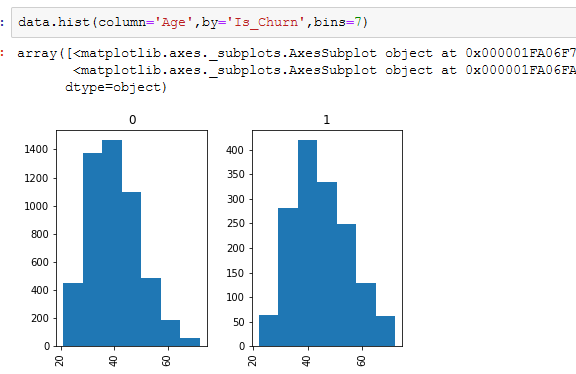
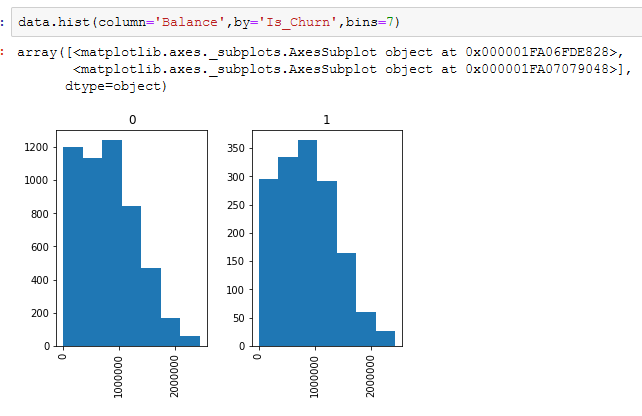


* Tried to understand if any of the string columns has any differentiating factor for predicting churn. There is no specific factor/type/bin to differentiate between classes.

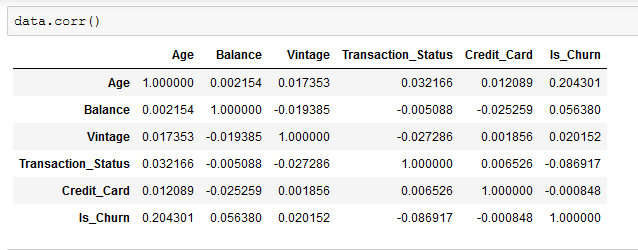
  

* Tried to understand can we divide numerical columns to bins to differentiate between classes but distribution is same for both classes.

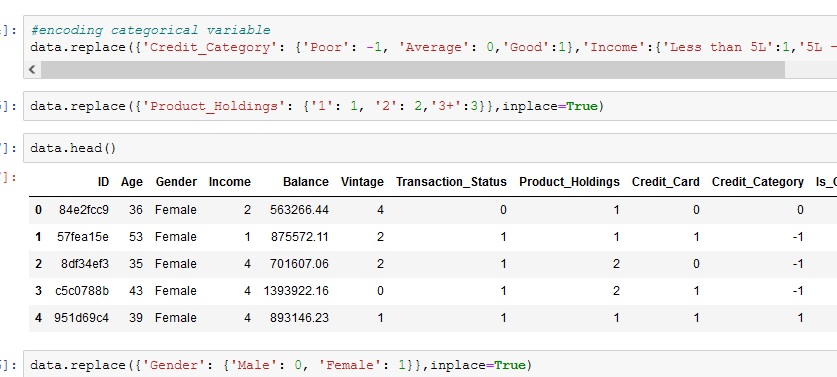
 

* There is no multi collinearity and no correlation between dependent and independent columns



**Data Pre-Processing:**

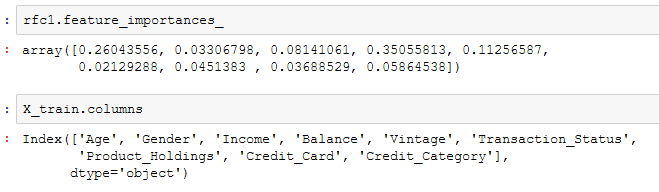
* Replaced string values in column instead of label encoding as we know the information has order example in income less than 5L>5L-10L



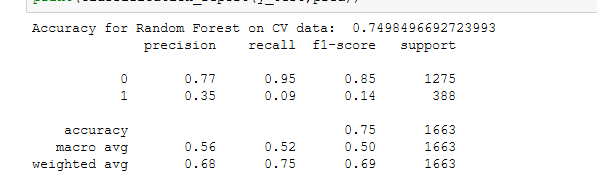
**Modelling Approaches before Feature Engineering:**

**1.Random Forest with GridSearchCV:**

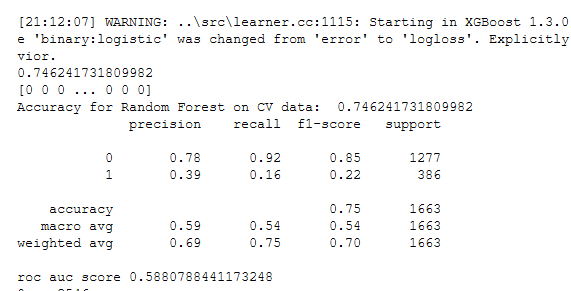
* Tried Random Forest Model using grid with different parameters, and final model parameters are max\_features='auto', n\_estimators= 5000,criterion='gini',max\_depth=40
* Feature importance’s were



* Performance of the model is



* Tried Xgboost and LGBM as well which gave similar performance



* Tried oversampling with SMOTE but performance didn’t improve

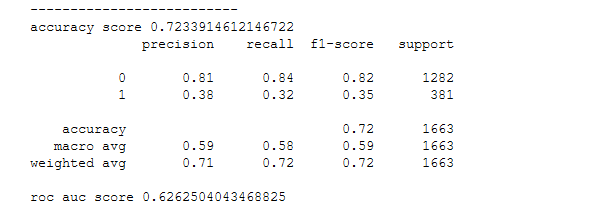
**Feature Engineering:**

* Tried to create new features using important features in above Random Forest Model to improve performance
* New Features created are

1. Balance to Income Ratio (Balance/Income)
2. Balance for 6 months (Balance\*2 as we are predicting churn for 6 months)
3. Networth in bank (Vintage\*Balance)
4. Start age of customer at bank (Age-Vintage)
5. Tran\_cred (Transaction\_status + Credit\_Card to check these features together)
6. Age to Balance Ratio (Balance/Age)

**Modelling Approaches after Feature Engineering:**

* After creating new features used same above models
* Also checked with different thresholds for improving f1 score and fixed threshold 0.65(< 0.65 is 1 else 0)



**Additional Approaches:**

* Also tried PCA as it combines different features to create features, it increased performance on train/test data(given data split into train and test) but on predicting data performance is less than above(submitting in solution checker)
* Also tried Grid search approach to select best parameters which are



