Credit Card Lead Prediction

The goal of this project is to predict wether customer is interested for credit card or not

Background

- In Banking Credit Card is a service provide to customers that they can use money from bank and they can pay back in due time by full payout or in form of installments.
- This service is used by many customers for they requirments like online shopping,buying goods,paying bills etc,.

Background

- Credit Card service give bankers a gud revenue in form of bill payments. And late due fee payment.
- But some customers want to use this service based on there requirements and some not.
- So banks want to check the customers whether they accept their offer or not and approach them with that feedback.
- They require a model which can predict that customer is interested in their product or not.

Outcome

- Our target of this project to build a model which predict customer interest (Yes/No).
- This is binary classification problem.
- The evaluation metric is **roc_auc** score which uses base model probabilities of class1 to predict class outcomes it comes handy for binary classification.
- We need to get high roc_auc score so that our prediction accuracy can be increases.

Data Collection

- We get data from Analytics Vidhya as part of **JOB-A-THON - May 2021.**
- Webpage: https://datahack.analyticsvidhya.com/contest/job-a-thon-2/
- Data has two Datasets:
  - ➔ Train_data 245725 records
  - ➔ Test_data 105312 records (30%)

- we have 70% private test data is there we need to get gud score on 30% public test data as well as 70% private test data.
- Here the data given is from Happy Customer Bank as per challenge.

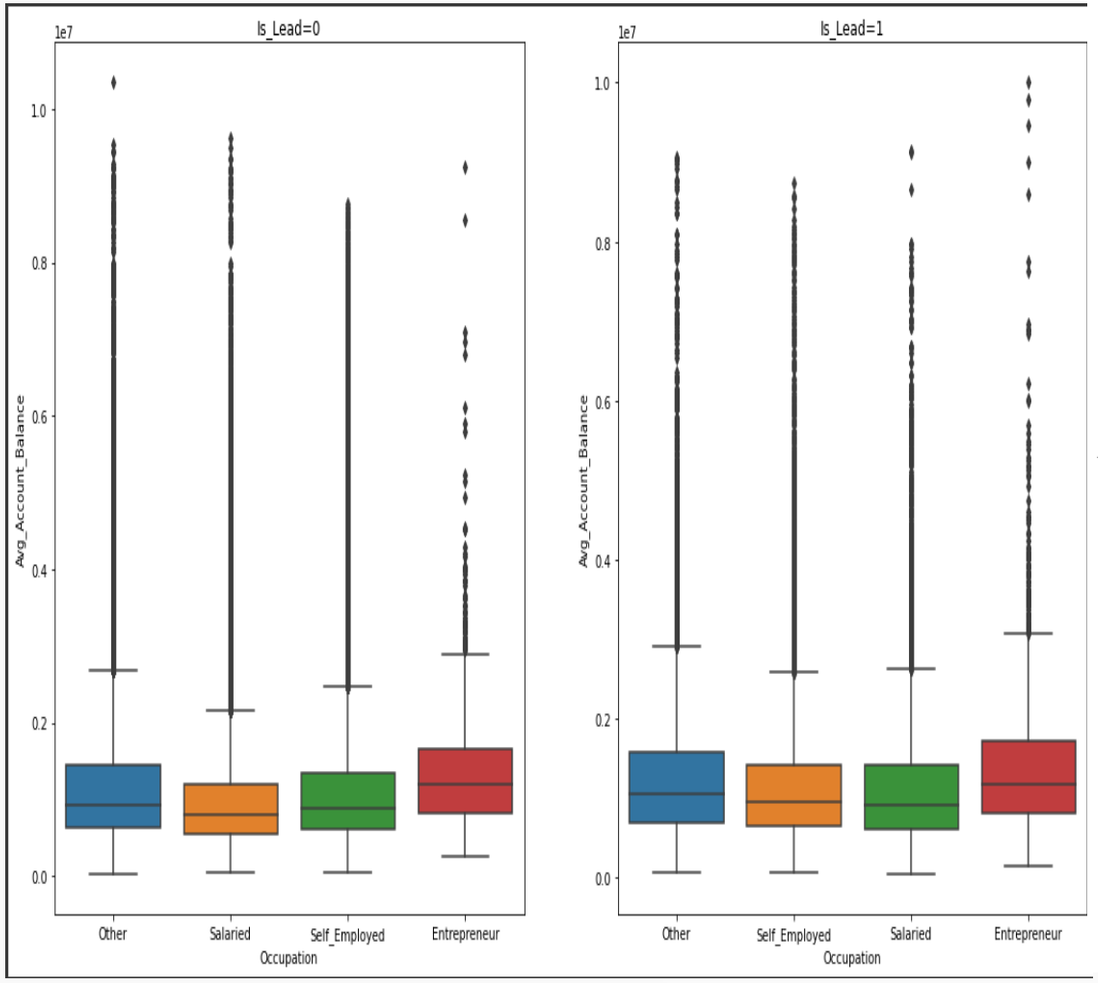
Data Inspection:

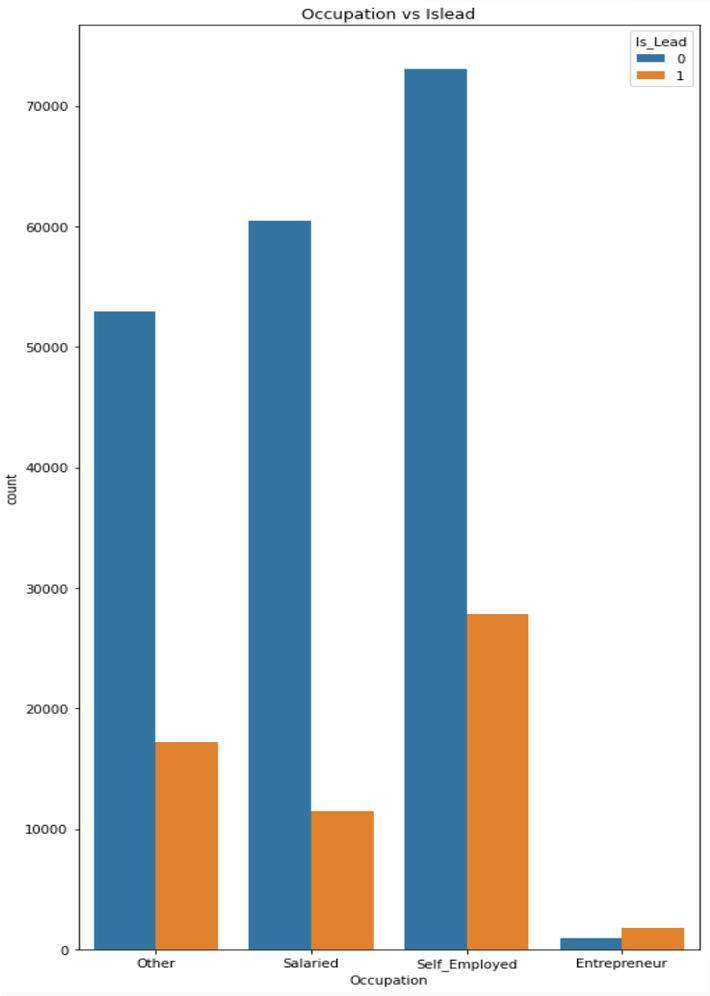
Train Data

Variable	Definition
ID	Unique Identifier for a row
Gender	Gender of the Customer
Age	Age of the Customer (in Years)
Region_Code	Code of the Region for the customers
Occupation	Occupation Type for the customer
Channel_Code	Acquisition Channel Code for the Customer (Encoded)
Vintage	Vintage for the Customer (In Months)
Credit_Product	If the Customer has any active credit product (Home loan, Personal loan, Credit Card etc.)
Avg_Account_Balance	Average Account Balance for the Customer in last 12 Months
Is_Active	If the Customer is Active in last 3 Months
Is_Lead(Target)	If the Customer is interested for the Credit Card 0 : Customer is not interested 1 : Customer is interested

Data visualisation

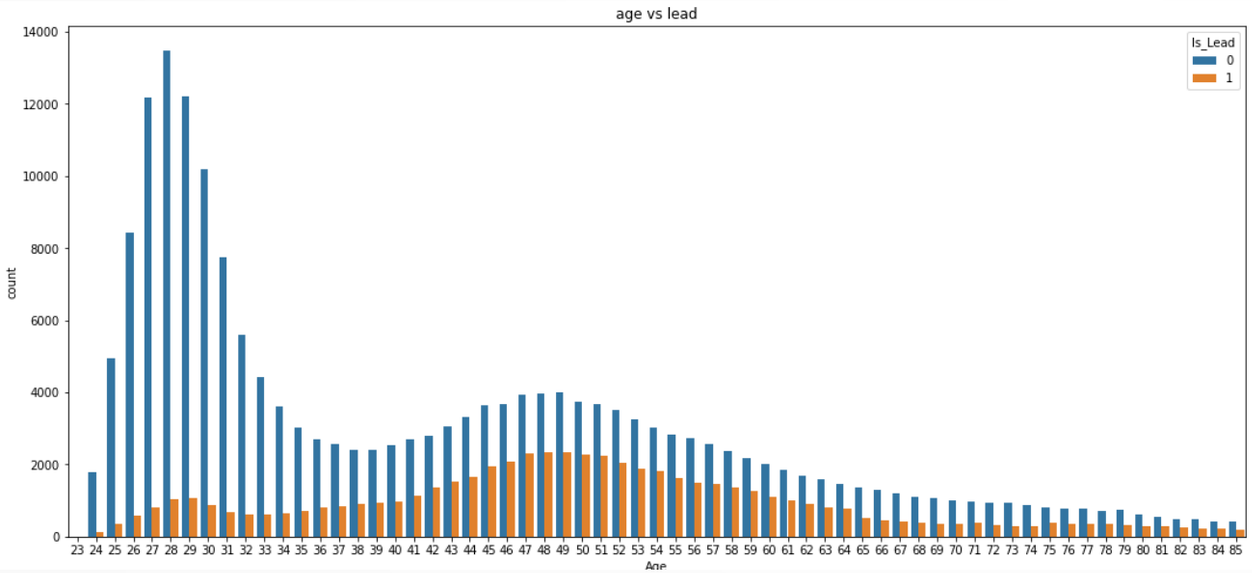
Occupation vs avg Account_Balance and
occupation vs Is_Lead(interested in cred_card)

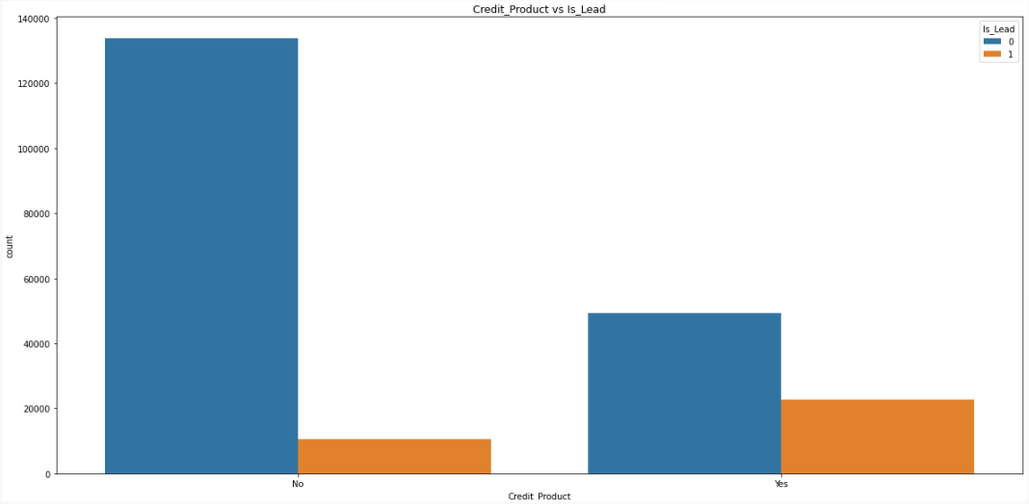




1)From the graph we observe the enterpreneur salary is high than remaining.
2)from occupation vs Is_Lead graph enpreneurs are interested in credit products more and salaried person are less interested
3) the Self_employed are more in numbers and than others

Age vs lead and Credit_product vs Is_Lead





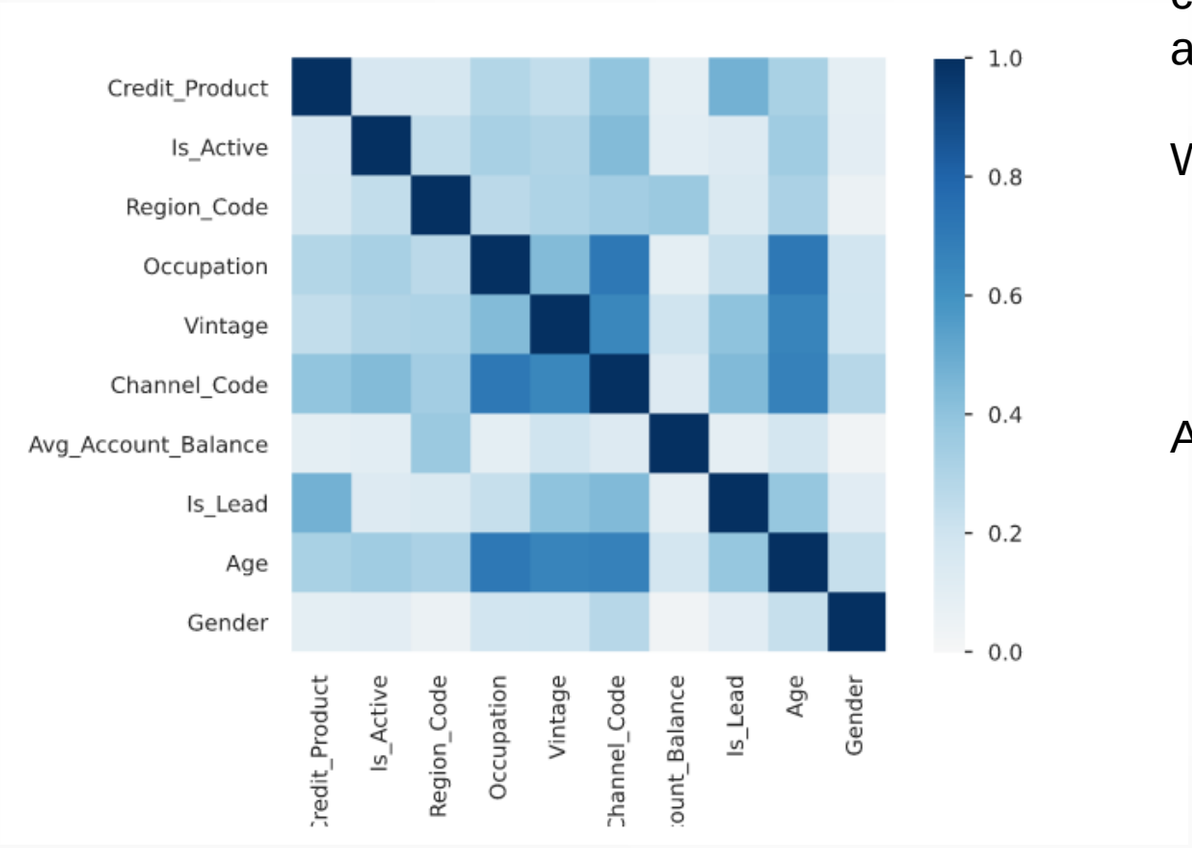
The chances of customer interest in taking credit card is high for age group between 45-55

While the younger age group between 24-32 are not much interested in taking credit product

This age features help in solving the problem

From credit_product vs Is_lead
We observe that the customers who are not using any credit products previous are not likely to take credit cards and who are using they are interested to take credit cards

Correlations



Here are the correlations between features categorical and numerical using phik correlation algorithm

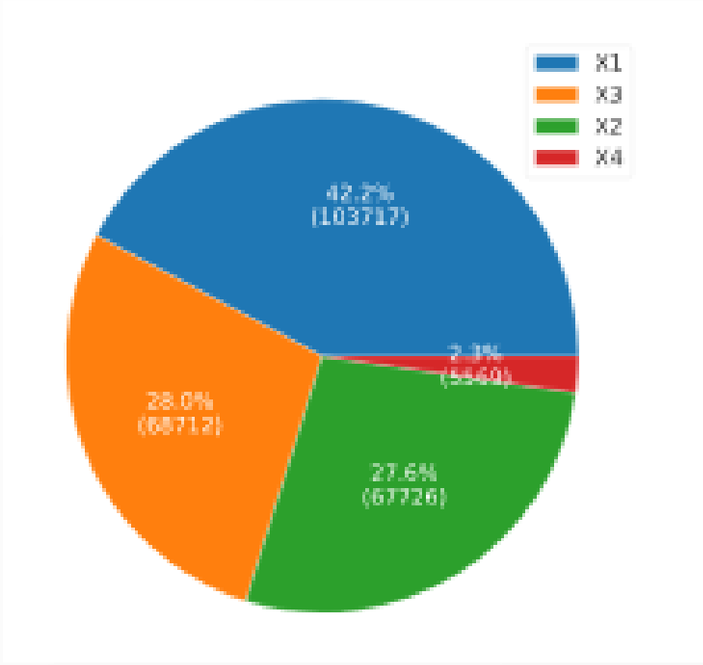
We observe that:

- Channel_Code & Vintage
- Occupation & Channel_Code
- Age & Occupation
- Channel_Code & Age

Are some what corellated.

Channel_Code

Value	Count	Frequency (%)
X1	103718	42.2%
X3	68712	28.0%
X2	67726	27.6%
X4	5569	2.3%



We observe that Channel_Code X1 is used most to contact customer
Second is X3
X4 is used less

Checking Nulls:

- We observe about 11% records have nulls values in Credit_product feature .
- We dont know wether they use credit product before or not so i consider those null values as third category as neutral(NA) and train the model.
- And in stacking the removed the third column neutral so that [No,Yes] values [0,0] denotes neutral(null).
- And tried with filling nulls with ‘No’(Mode) but it doesnt help much for me it doesnt increase score. So considering above process is best suited for model.

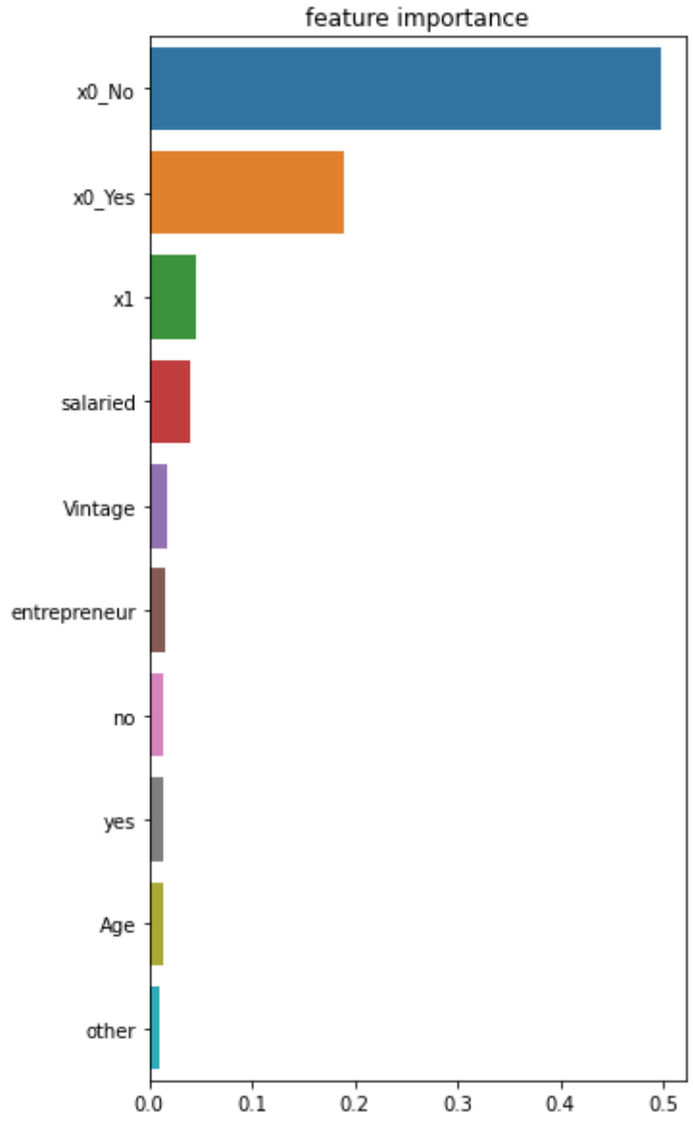
Feature_Engineering

- The categorical Features are simply one hot encoding and counter vectorizer for some features(OHE).
- The numerical Features are used as it is XGBoost algorithm is robust to numerical values.
- -> I got 52 features from above featurization methods.
- I tried Vintage as categorical feature and apply OHE but the perfomance is less on test data so i used as numerical data and it perform better than categorical featurization bcz in unseen data there might be a value which is not present in Train data so that category will loss and the model will underfit on test data.
- Avg_Account_Balance is the feature which contains high high magnitude values but DecisionTree is robust to it magnitude so i havent apply featurization.
- Age is a numerical values but it has significance on taking Credit product. So I use exact values without scaling them.

Modelling

- I use GridSearchCV for hyper parameter tunning and kept the scoring metric as auc_roc that is our validation metric for this problem and got n_estimators=1000,col_sam_bytree=1.0,learning_rate=0.01 for LightGBM,XGBoost.
- With 52 features i run XGBoostClassifier with parameters n_estimators=1330,learning_rate=0.02,max_depth=6,col_sam_bytree=0.5
- I got auc_roc score on train=0.899,cv=0.871,test=0.8732(after submitting)
- I tried with RandomForesh with n_estimators=10,100 but i doesnt perform well on test data it give test score of 0.869 and 0.871. and it take much highest time among the these three models
- I tried LightGBM with n_estimators=1000,max_depth=12 but it give score of 0.864 for me on test.
- I used DecisionTree models because it robust to numerical values and it robust to variance in data.
- So i finalise the XGBoost Model which give highest score on the train data.
- I plotted the confusion matrix on overall train_data and check the TN,FN,FP,TP.
- And saved the probabilities to submissions.csv in Is_Lead column

Feature_Importance



- We observe that credit product with NO have high importance to take credit card
- Next is credit product with No
- Followed by channel X1,occupation,Vintage,Is_Active,Age

Summary

- 30% of customers who are active in the bank are likely to take credit card products. Hence, these customers will be potential to bank to offer credit card.
- The customers who are salaried needs to be targeted first over other categories .
- The bank can prefer “X1” channel to promote credit card products to their customers.

Thanks for the Data provided by Analytics Vidhya