APPROACH DOCUMENT - CREDIT LEAD PREDICTION DATASET

Project Introduction - Credit Card Lead Prediction

Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products, among other offerings.

The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc.

In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

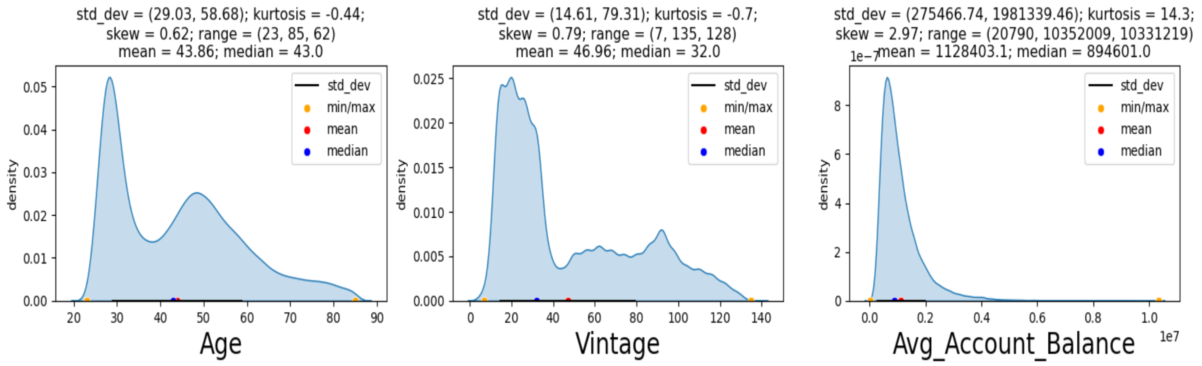
- Customer details (gender, age, region etc.)
- Details of his/her relationship with the bank (Channel_Code,Vintage, 'Avg_Asset_Value etc.)

1. Exploratory Data Analysis

1.Importing Libraries	Numpy, Pandas, Matplotlib, Seaborn
2. Variable Identification	Categorical : [Gender, , Region_Code, Occupation, Channel_Code, Credit_Product, , Is_Active, Is_Lead] Numerical : [Age, Vintage, Avg_Accoun_Balance]
3. Shape of dataframe	245725 rows, 11 cols
4. Type casting features	

2. Univariate Analysis on all numerical features

Feature	Min	Max	Mean
Age	23	85	43
Vintage	7	135	46
Avg_Account_Balance	20790	10351009	1128403



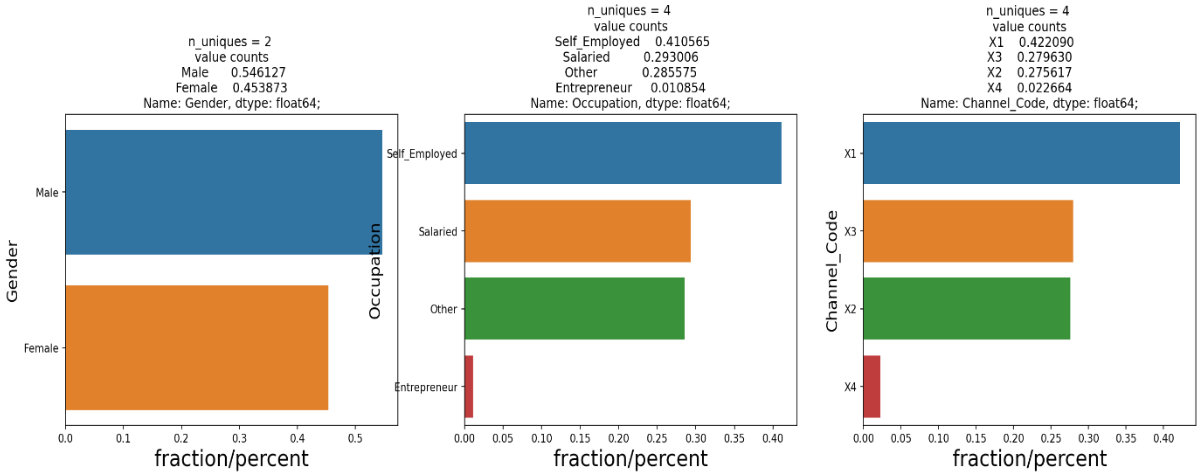
a. Majority of the population is in range of 20-40 years old and we can see a small spike near 50 years of age

b. Majority of customers are don't have very old relationships with the bank.

c. Also, the Average account balance of the population is skewed towards the lower side

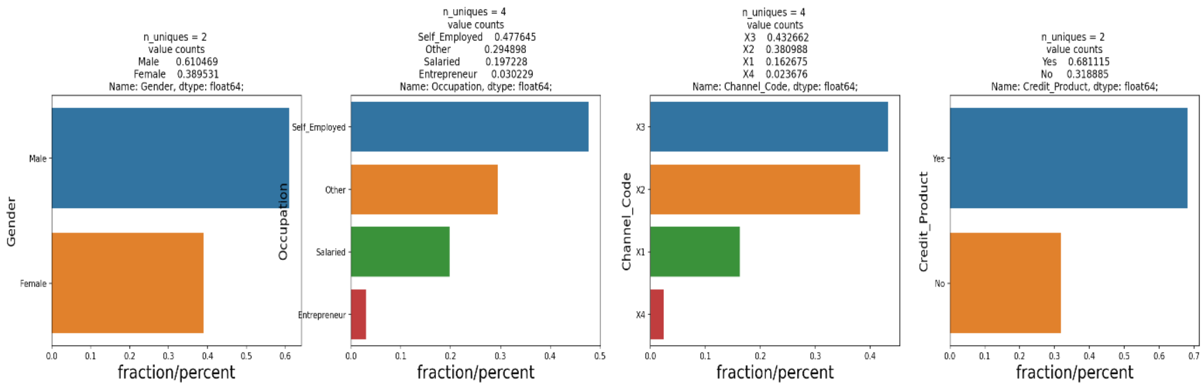
d. All the numerical features are not distributed normally and are positively skewed. We will see later for any outliers.

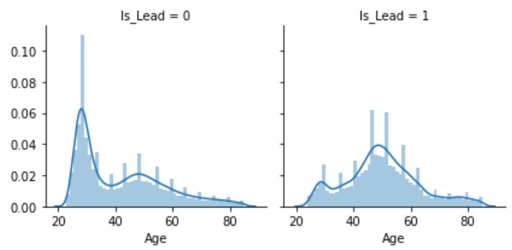
Univariate Analysis on all Categorical Features

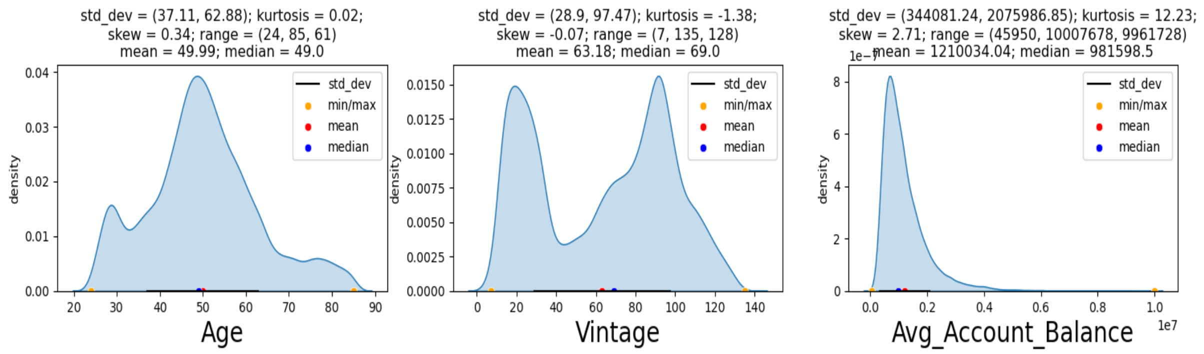


1. 55 % of the overall population are Male and 45% are Females.
2. 41 % of the population are Self Employed followed by 30% Salaried.
3. 42% of the population are from X1 Channel code
4. Only 33% of the population have a Credit product and only 38% customers were active in the last 3 months.

EDA on the customers who are interested in Credit Card (Filter out the data where 'Is_Lead' == 1)



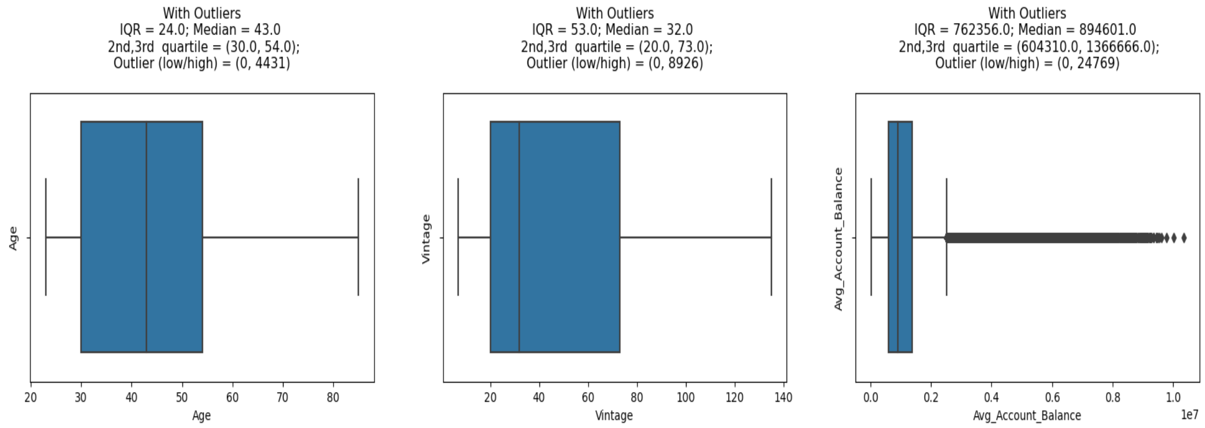




Among the customers who are interested in the Credit Card :

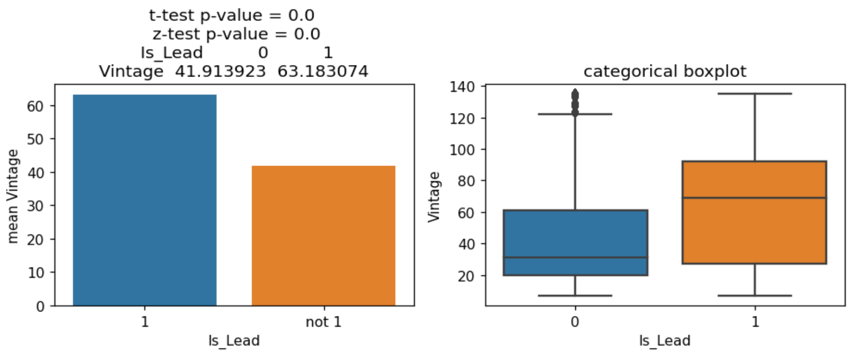
1. 61% are Male population
2. Approx. 47 % accounts for of people who are Self-employed. Only 3% are entrepreneurs
3. 43% of the customers have channel code as X3 followed by 38% from X2.
4. 68% have a credit product while 31% don't.
5. 18% of the customers come from the Region RG268
6. Customers in age group 20-40 are not interested in the credit card, while customers in the age group 40-60 are interested.

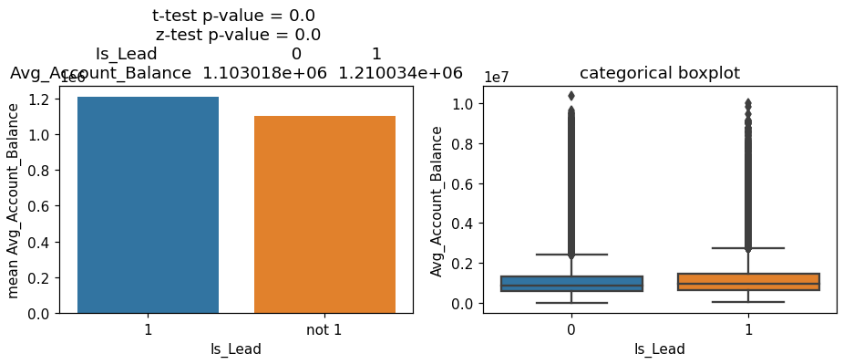
Visualizing Outliers

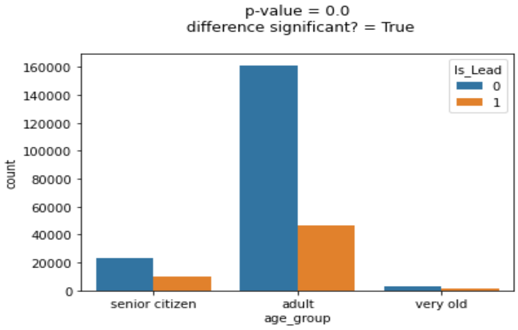


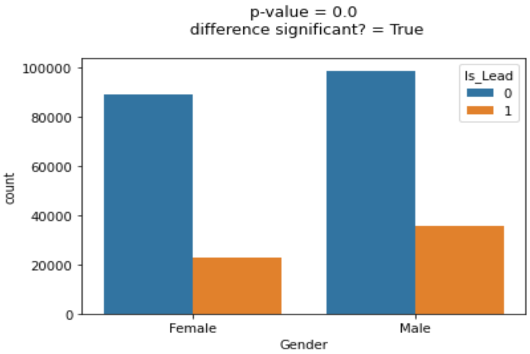
1. Avg Account balance has a high range. It has outliers on the on the higher side. We can log transform / scale the variable to normalize it.

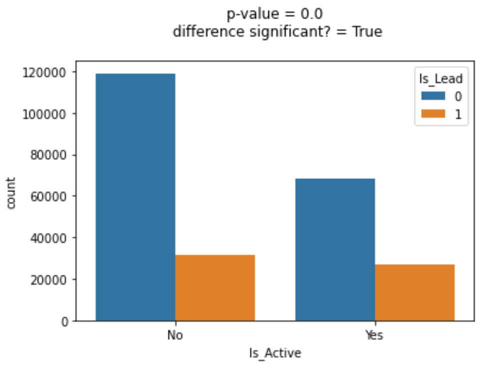
Bivariate Analysis

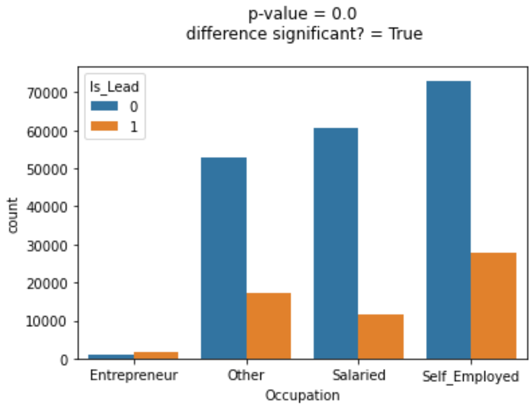












1. Majority of Vintage customer are likely to be interested in Credit card
2. Customers with higher average account balance are likely to be interested in Credit Card.
3. Majority of the customers who are not interested in Credit Cards are adults.
4. Among the customers who are not interested in Credit Card, majority of them ae Females.
5. Majority of the customers are not interested in Credit Card are not active.
6. Among the people who are Self Employed 72% are not very likely to be interested in Credit Card. However, if we take a look among Entrepreneurs, 66% are interested in Credit Card.

Identifying Missing Values

Only 'Credit_Product' column has missing values. We can later replace them by 'Unknown'. Data can be biased if we fill the Nan values by mode of the column.

Model Building Experiment 1

1. Imported the train and test data. Missing values/ Outliers' imputations are performed parallelly on train and test data.
2. Dropped the 'ID' column as it has high cardinality.

3. Segregated the numerical and categorical features
4. Imputed the 'Nan' values in Credit_Product by 'Unknown'.
5. Segregated the train data into train and validation set to work on.
6. Categorical features are encoded using OrdinalEncoder ().
7. Scaled the numerical features using StandardScaler ().
8. Different models are applied, and results are compared to see who gives the highest auc score.

Model	auc_score on train data	auc_score on validation data
KNN	0.756	0.715
Logistic Regression	0.574	0.737
Decision Tree Classifier	0.756	0.754

Model Building Experiment 2

1. Imported the train and test data. Segregated the train data into train and validation set to work on. Missing values/ Outliers' imputations are performed parallelly on train / validation and test data.
2. Dropped the 'ID' column as it has high cardinality.
3. Removed duplicates if any.
4. Segregated the numerical and categorical features
5. Imputed the 'Nan' values in Credit_Product by 'Unknown'.
6. Used Log Transform on 'Average_acount_balance' to so it has a normal distribution.
7. Used LabelEncoder() to transform all the categorical variables at one go.
8. Check if multicollinearity exists.
9. Split the data in train and validation data
10. Scale the training set using MinMaxScaler()

Model	auc_score on train data	auc_score on validation data
KNN	0.802	0.754
Logistic Regression	0.566	0.754
Decision Tree Classifier	0.756	0.754
Logitboost	0.864	0.862
Xgboost	0.893	0.871
Xgbclassifier	0.885	0.873
LGBM Classifier	0.881	0.873

Cross Validation Results

	CV1	CV2	CV3	CV4	CV5	CV Mean	CV Std Dev
LGBM	64.730080	64.813206	65.238489	65.035036	64.915611	64.946484	0.199158
KNN	61.632400	62.053123	62.160857	62.068307	62.159289	62.014795	0.219529
CART	54.726493	55.146966	55.102905	55.297721	54.808222	55.016461	0.240342
LR	28.433451	28.860131	29.821450	29.120300	28.182957	28.883658	0.638294

** Final model selected is LGBM based on roc score and cross validation results **

Snapshot of output is attached below

	ID	Is_Lead
0	VBENBARO	0.042984
1	CCMEWNKY	0.874960
2	VK3KGA9M	0.069017
3	TT8RPZVC	0.023126
4	SHQZEYTZ	0.022801