JOB-A-THON- May 2021

Solution by Ujjwal Sharma

Email – ujjwalsharma191297@gmail.com

Ph - +91 8957581819/8299106018

Problem Statement -

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.

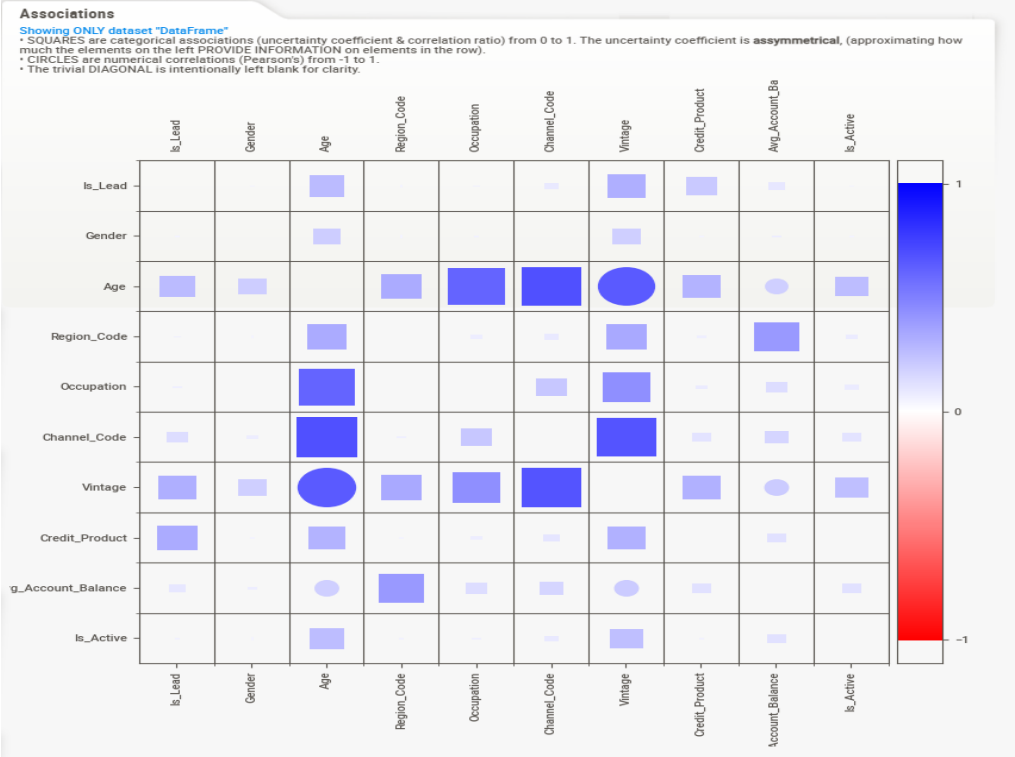
Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

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

EDA *

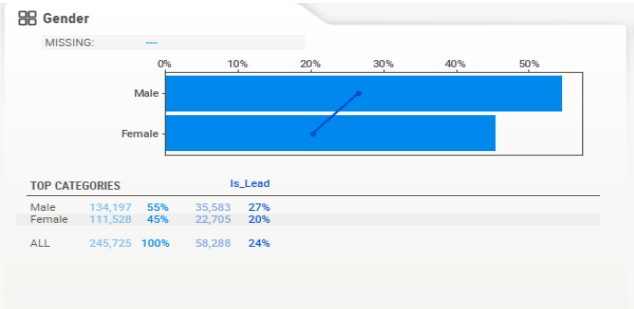
*detailed reports & plots can be found in “eda_report_train.html” and “comparison_report.html” files.

Overall associativity among all sensors:



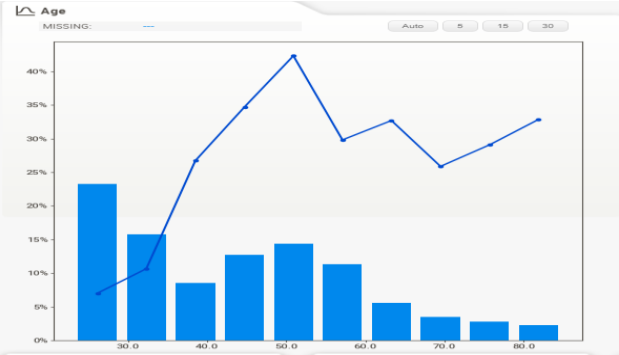
Categorical features:

1. Gender:
    a. Male is taking more loan than Female.

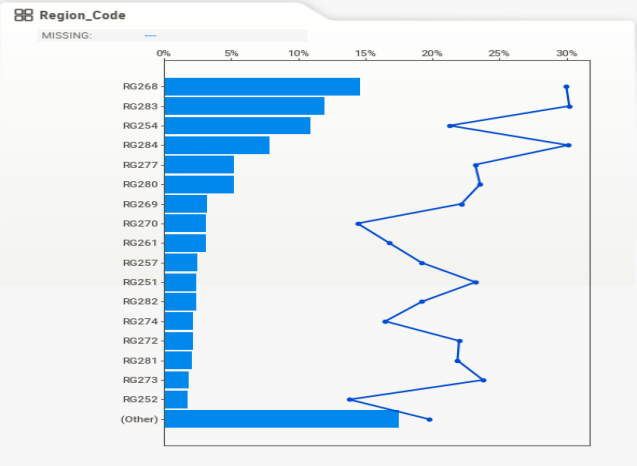


2. **Age:**

a. In the second mode target=1 is increasing dramatically.

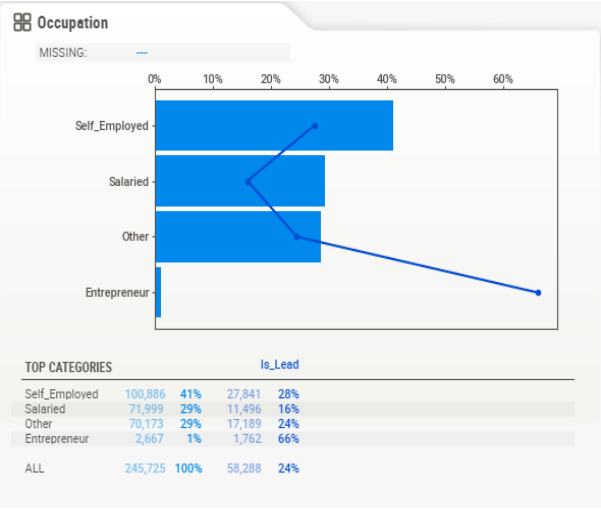


3. Region_Code:

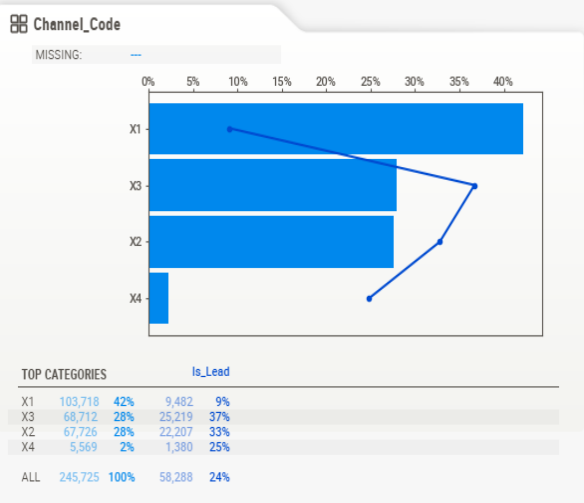


4. Occupation:

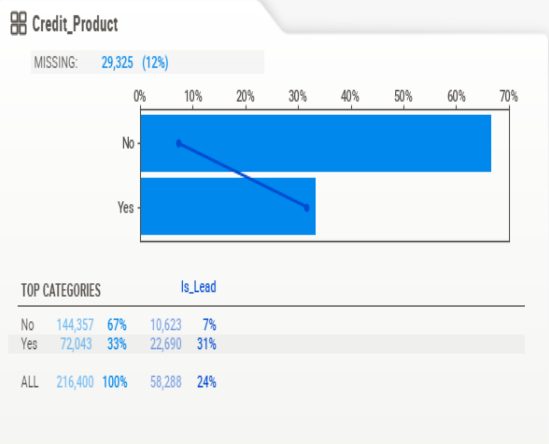
a. Entrepreneurs tends to take mode loan so it is an important feature.



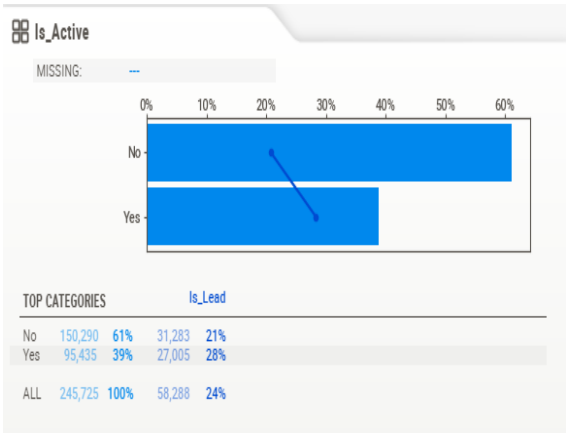
5. **Channel_Code:**
    a. Channel_code category “X1” has lowest target=1 and “X3” has highest.



6. **Credit_Product:**
    a. This is an important feature but contains 12% missing value so we need to impute it carefully.

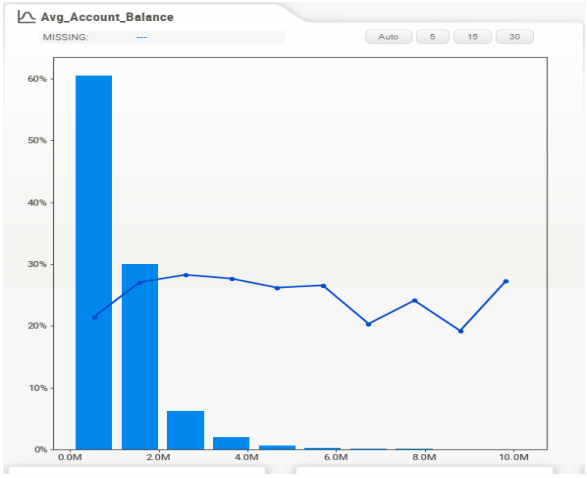


7. **Is_Active:**
    a. **“Yes”** Category in this feature increases the chances of target=1.

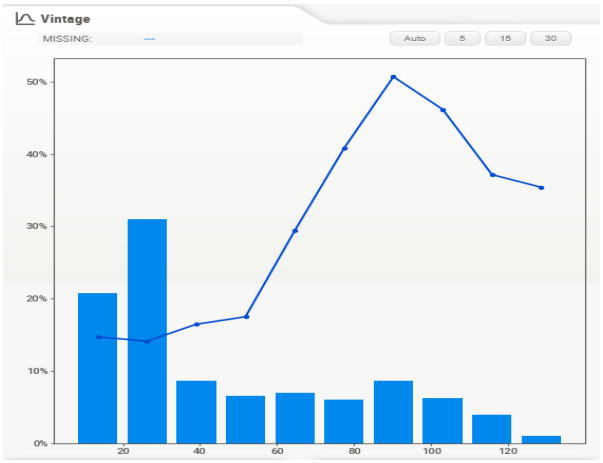


Continuous features:

1. **Avg_Account_Balance:**
    a. We can use bin in this to create an important feature.



2. **Vintage:**
    a. This is an important feature on the second mode there is a spike in target feature.



Important Features for target are -

1. Avg_Account_Balance
2. Vintage
3. Credit_Product
4. Occupation
5. Age

Missing Value imputation:

Feature "Credit_Product" contains 16% missing values. I have used following 4 techniques to impute missing values-

1. Treat missing value as third category
2. Filling it by mode
3. Impute missing value with KNN
4. Predict missing value by building Model for it.

Out of these 3, first one worked better in this data.

Feature Engineering and Selection

1. Dividing "Avg_Account_Balance" in 3 category based on the 50%, 75% data distribution and created another categorical feature “Richness”.
2. Created combined feature of Richness and Occupation.
3. Created combined feature of Age and Occupation.
4. Created combined feature of Credit_Product and Occupation.

Dealing with imbalance dataset

To deal with imbalance dataset I have used SVMSMOTE to oversample the training dataset.

Modeling

1. I have tried different models and StratifiedKFold for selecting the models.
2. I have use roc_auc_score for evaluation and comparison of models.
3. I have ensembled following three models for better predictions-
    a. LGBMClassifier
    b. XGBClassifier
    c. CatBoostClassifier

With these techniques and ensembled model I have managed to get 0.002 less roc_auc_score from the first ranker. My roc_auc_score is **0.87194 on public leaderboard.**