# **JOB-A-THON - May 2021**

## Credit Card Lead Prediction

Analysis and Solution by Aditya Raj

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.)

Solution:

The current solution is based on assumptions on EDA(graphs and other details included in Jupyter notebook .ipynb file) and ascertaining the data type as well as feature engineering.

1.Data collection : All data used for training is provided by the competition admins, no external data is used for training the model.

2. Data Pre-processing : A total of 10 features are provided.

EDA study was carried out based on pandas\_profiling on training set. Attached is av\_credit\_card\_lead\_eda.html for reference.

Categorical columns

0 ID

1 Gender

3 Region\_Code

4 Occupation

5 Channel\_Code

7 Credit\_Product

9 Is\_Active

Numerical Columns

2 Age

6 Vintage

8 Avg\_Account\_Balance

Creating a basic heat map of the training set we find that though other features are nearly correlated with ‘Is\_Lead’ but their removal depends on how they fit while modelling the data.

Chart

Description automatically generated

Vintage highly correlated with Age and Channel Code

Occupation high correlation with Channel code and Age

Age high correlation with Vintage, Channel code and Age

When observing all the the categorical columns, some data was found to be NaN for Credit\_Product

Chart

Description automatically generated

We fill Credit\_Product with ‘Unknown’ as there are missing values in test set too on which we need to predict.

Now when looking at numerical columns we observe that data corresponding to Vintage, Age and Avg\_Account\_Balance is skewed.

Also, Avg\_Account\_Balance has few outliers,

A screenshot of a computer

Description automatically generated with low confidence

* I tried both approaches of removing(values >0.999 and values<.001 quantile)as well as keeping these outliers. Though keeping them provided better AUC score.

For age, I tried binning it into 4 categories of young, middleage, aged, old as age\_category.

We observed the ls\_Lead (0/1) value distribution was differing. Hence, the model can learn the distribution in bin category and provide accurate results.

I also tried binning Vintage also in years feature but that did not help the model accuracy.

All the categorical values were then label encoded into numerical values.

I also, tried the approach of one-hot encoding the categorical values but that did not improve the model. Also, Tree-models struggle if there are a large number of levels/features, regardless of how much data we have. Hence, keeping less features and label-encoded values.

Since, we had few skewed features in dataset we scale the training set and test set using StandardScalar.

As, this is a class imbalance problem, the data distribution is uneven between ‘Is\_Lead’: (0/1)

Approach used:

1. Oversampling :Impute synthetic data via SMOTE, so make Is\_Lead-0 and Is\_Lead :1 value to equal up. More synthetic data points for 1 was added.
2. Undersampling: Train with random selection of equal number of Is\_Lead:0 to Is\_Lead :1. So reduce higher volume of Is\_Lead :0
3. Continue with maximum training so that model understands the Is\_Lead:1 data points with more confidence

The step 1 & 2 resulted in lower AUC score compared to step 3. Hence final model was trained without synthetic of equal valued for Is\_Lead:1 and Is\_Lead:0

3. Model Training : For model training, I have tried various classifiers such as Decision Tree, Random Forest, k-means, Logistic Regression, SVC , XGBoost, SGDClassifier with GridSeacrch and LightGBM with GridSearch. I have not tried ANN due to past experience of simple interpretable models performing better on such tasks.

Among tried models, LightGBM gave best AUC Score and no overfitting. [.0.872743928714212]

* I tried to train with maximum examples as possible to avoid overfitting and underfitting. Hence, distribution set of .9 for training and .1 for test was used.

I did try out with distribution set of.7 and .3 but observed the AUC score get better with more training on provided data. Since the created model has same AUC score as on public dataset. I don’t think this approach has overfitted on data yet.

* The model I have used is LightGbm with hyperparameter tuning using GridSearchCV for 4 cross validations on the dataset.
* I tried ensemble modelling by majority voting ensemble using different hyperparameter value models for LightGBM. But, I am assuming the properly tuned model should be performing better on private set.
* I have not tried other ensemble combination for LightGBM with other models like Random forest, XGBoost etc,

Model results:

Overall accuracy of Light GBM model: 0.8609449395678184

Graphical user interface, application, table

Description automatically generated

* I have also tried with few top features as feature\_fraction=0.7 in model params

Chart, line chart

Description automatically generated

AUC score: 0.8695915099762787

Chart, treemap chart

Description automatically generated

4.Solution submission: The best performing model is used to predict\_proba for probability values of both classes for ‘Is\_Lead’. The vales corresponding to 1 is used for prediction.