Credit Card Lead Prediction

Hello!

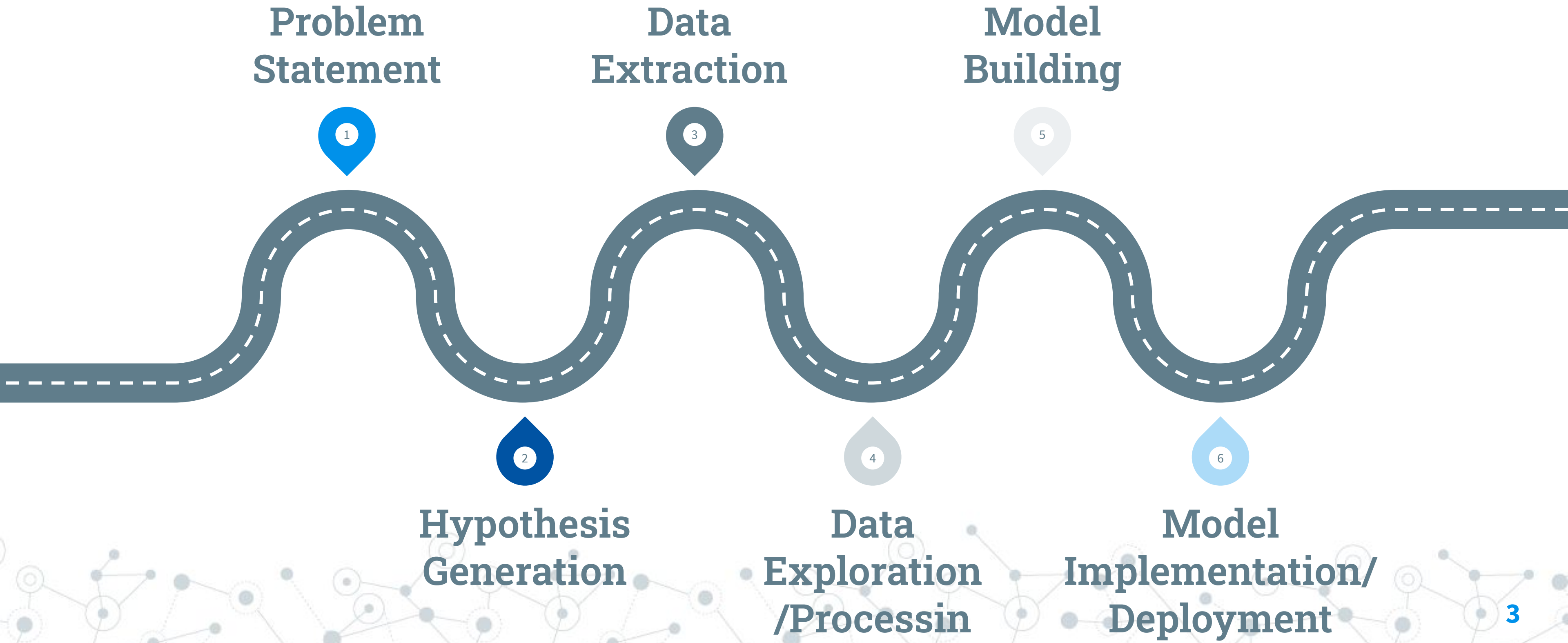
I am Brooklin Santosh A G S

A budding Data Scientist

You can find me at:

https://www.linkedin.com/in/brooklinsantosh/

Data Science Life Cycle



1. Problem Statement

Predict the lead to cross sell Happy Customer Banks credit card to its existing customers

Detailed Problem Definition

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 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.)

2. Hypothesis Generation

Few of the hypothesis:

- Occupation affects the Credit card lead
- Monthly income affects the Credit card lead
- Credit score affects the Credit card lead
- Age affects the Credit card lead
- Previous credit product affects the Credit card lead
- Mode of communication is significant due to Credit card lead
- Old customers have more intended towards the Credit card lead

3. Data Collection

Data collection contd.

- Data is provided by the Happy Customer Bank
- No external data is allowed
- We have very limited customer demographic as well as relationship with bank details
- We have three different dataset, train, test and sample submission
- Lets see the data dictionary of train dataset in the next slide.

Data dictionary of the given train dataset

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

4. Data Exploration / Transformation

Train vs Test

Train

Shape - (245725, 11)

Missing values - Credit_Product

% of Missing data in Credit_Product column - ~12%

Test

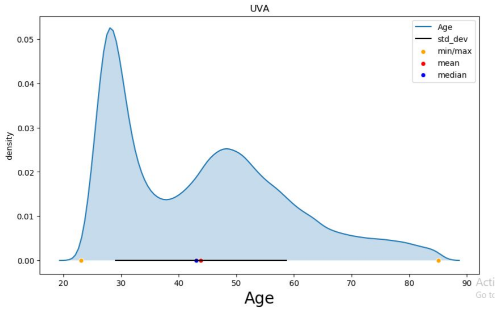
Shape - (105312, 10)

Missing values - Credit_Product

% of Missing data in Credit_Product column - ~12%

Data Exploration - Numerical

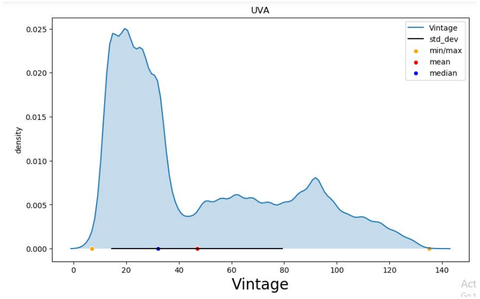
Age



It looks like bimodal distribution.

It is positively skewed, platykurtic distribution

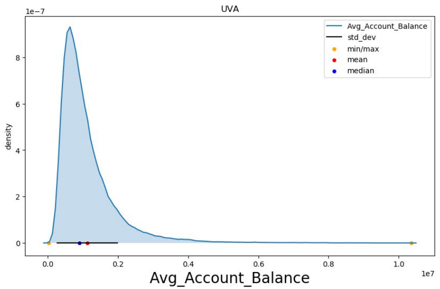
Vintage



This is not a normal distribution.

It is positively skewed, platykurtic distribution

Avg_Account_Balance



Positively skewed, leptokurtic distribution

Has outliers on the right tail.

Data Exploration - Categorical

- This is an imbalanced dataset
- Explored all the categorical columns in the chek_data.ipynb file.
- None of the datasets (train/test) have a missing category in either of it, say a category present in train and not present in test, vice-versa
- Channel_Code is encoded and Region_Code has more categories.

Data Exploration contd.

	Age	Vintage	Avg_Account_Balance	Is_Lead
Age	1.000000	0.477790	0.121379	0.210291
Vintage	0.477790	1.000000	0.134562	0.208096
Avg_Account_Balance	0.121379	0.134562	1.000000	0.063248
Is_Lead	0.210291	0.208096	0.063248	1.000000

Kendall's Tau correlation matrix between the numerical columns.

Why Kendall?

Kendall's Tau is robust, normality of the data is not required, monotonic relationship

Exploration contd.

Salaried	X1	0	0.913346
		1	0.086654
	X2	1	0.741125
		0	0.258875
	X3	1	0.658516
		0	0.341484
	X4	0	0.634584
		1	0.365416

The conversion rate of X2 and X3 channel codes for the salaried person is very high as highlighted in the image.

Exploration contd.

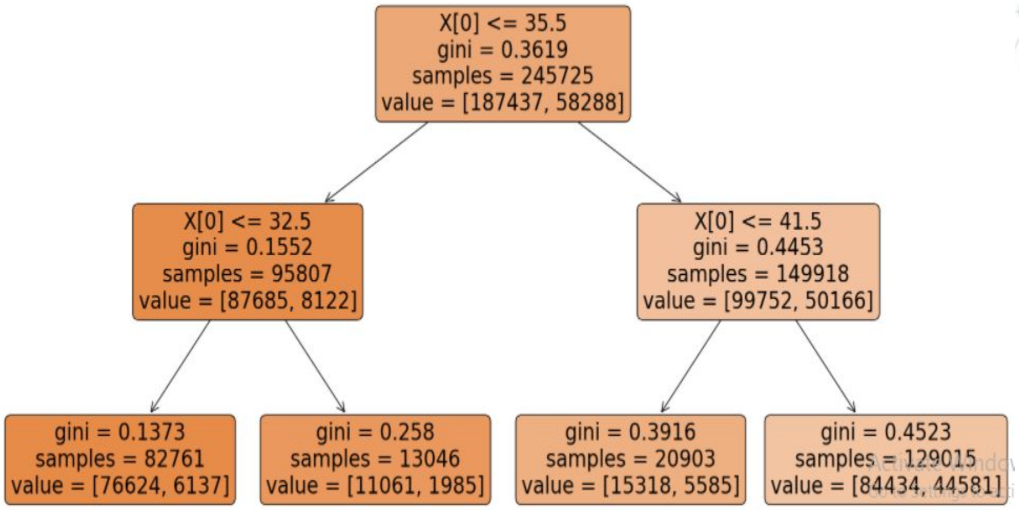
Credit_Product	
No	0.073588
None	0.851662
Yes	0.314951

The rows which have missing values in Credit_Product column is having very high conversion rate. We can create this a new features.

Similarly we have reviewed all the hypothesis we have created, to find new insights.

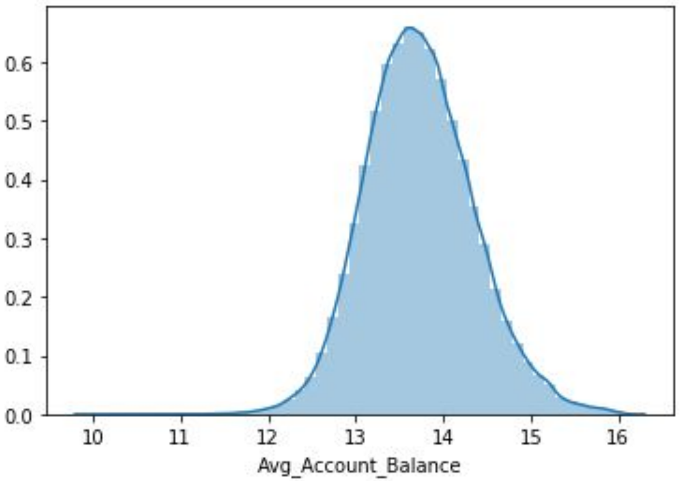
Data Transformation

In order to tackle the bimodal distribution of Age column, we have converted into category by binning with the use of Decision Tree to find the bins



Done the same for Vintage column too.

Outliers treatment



Created a new feature for the Avg_Account_Balance column to tackle the outliers by taking log.

Tried different transformations and selected log over others.

Feature Engineering

- Label encoded all the categorical columns using LabelEncoder.
- Region_Code has many categories, to tackle that I have encoded all the categorical features by frequency and mean encoding and added them as new features
- By checking the different hypothesis and figured out lot of information regarding different groups and created the features mentioned in the next slide

Feature Engineering contd.

- Credit_Product_Missing
- Salaried_X2
- Salaried_X3
- Salaried_Vintage_Bin3
- Active_Age_Bin1
- Occupation_Avg_Account_Balance>Median
- Active_Occupation
- etc

5. Model Building

Cross Validation

- Selected StratifiedKFold since the target class is imbalanced
- Decided with 5 folds cross validation.
- Created a kfold column in the train dataset and kept the fold number so that I can do like holdout validation for 5 iterations
- Predicting for test data in all 5 folds, then blend the predictions so that it will be more generalised.

Model Building

Tried all the below models and selected lgbm, xgb and catboost classifiers since they performed well.

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Adaptive Boosting
- LightBGM
- XGBoost
- CatBoost

Hyper parameter tuning

- Using Optuna tuned the hyper parameters for the selected 3 models
- Trained the model in cross validated manner and predicted for the validation data and test data, considering each cross validation iteration as individual model
- Got 5 test predictions,blended them and got a good public score for all the 3 selected models

Stacking

Model stacking is an efficient ensemble method

Decided to stack the models.

6. Model Implementation

Structure of the project

```
.
├── input
│   ├── train.csv
│   └── test.csv
├── src
│   ├── create_folds.py
│   ├── train.py
│   ├── inference.py
│   ├── models.py
│   ├── config.py
│   └── model_dispatcher.py
├── models
│   ├── model_rf.bin
│   └── model_et.bin
├── notebooks
│   ├── exploration.ipynb
│   └── check_data.ipynb
├── README.md
└── LICENSE
```

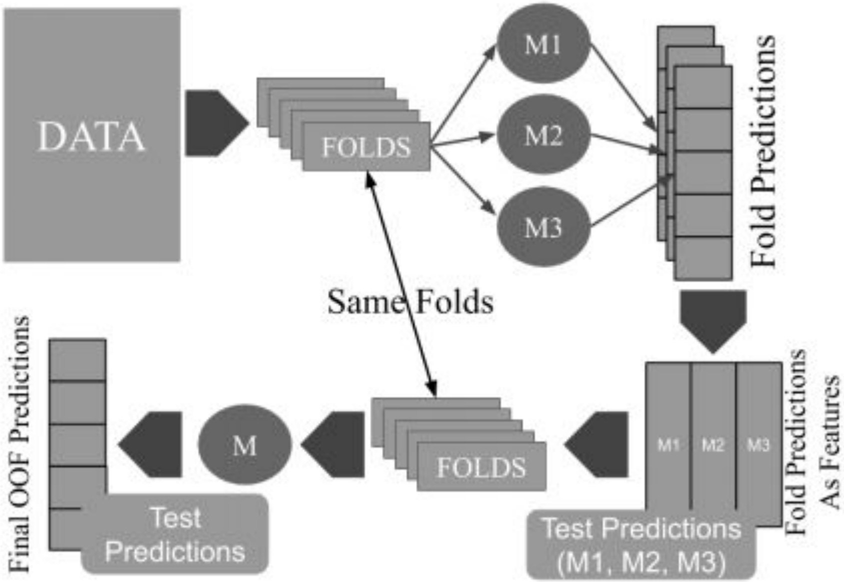
This is the folder structure, I am going to use. The idea is to create a clean reproducible code so that it will be easy to understand and also can be reused.

Everything will be tracked in GIT.

Model Architecture

This is the architecture I am going to use.

Same folds will be used throughout this lifecycle.



Model Implementation

- I have used the Optuna tuned LGBMClassifer as model1
- I have used the Optuna tuned CatBoostClassifier as model2
- I have used the Optuna tuned XGBClassifier as model3
- Instead of meta classifier, I have again blended the model predictions

Scope of Improvement

I haven’t try the below due to time constraint,

- Any kind of feature selection.
- Hypothesis testing to know more about the features
- Neural Networks
- AutoML
- Multiple layer stacking

Thanks!

You can find me at:

https://www.linkedin.com/in/brooklinsantosh/

https://github.com/brooklinsantosh

https://www.kaggle.com/brookie210

brooklinsantosh@gmail.com