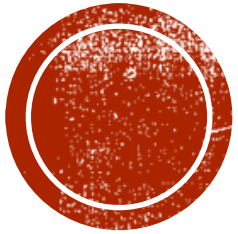
HEALTH INSURANCE LEAD PREDICTION

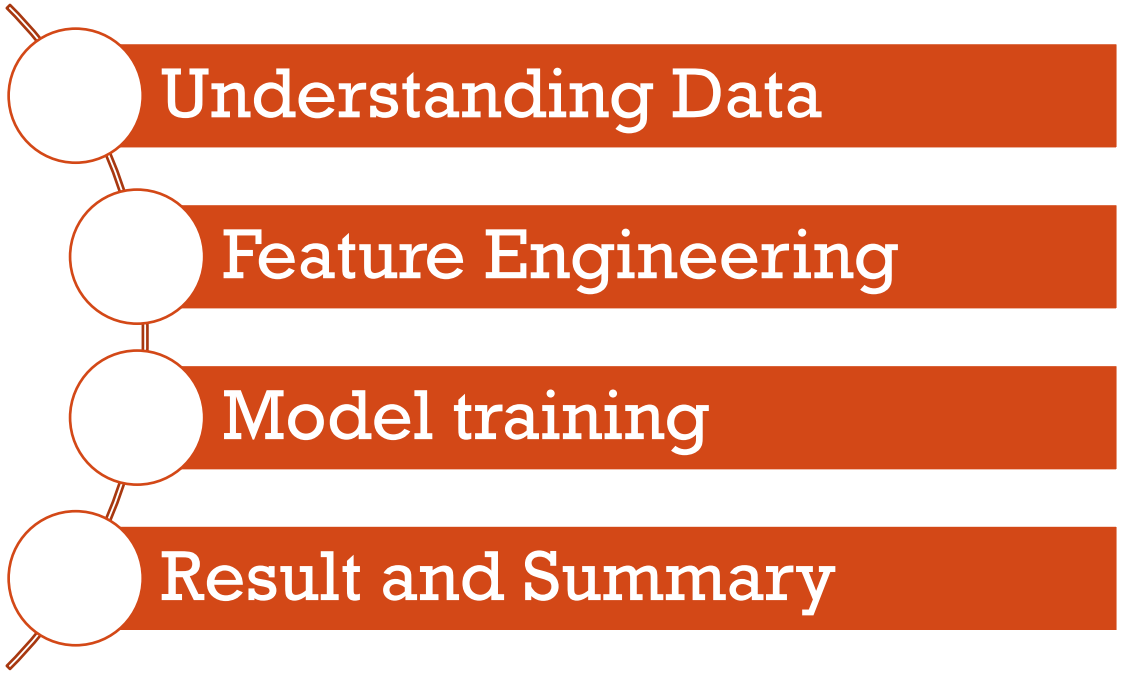
Abhishek Singh Rathore
Senior Data Scientist



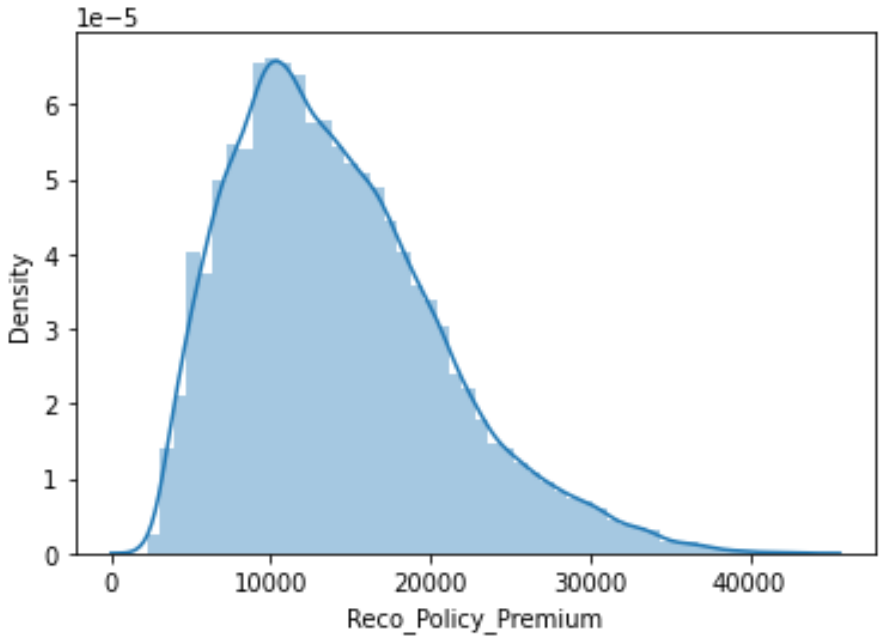
PROJECT GOAL

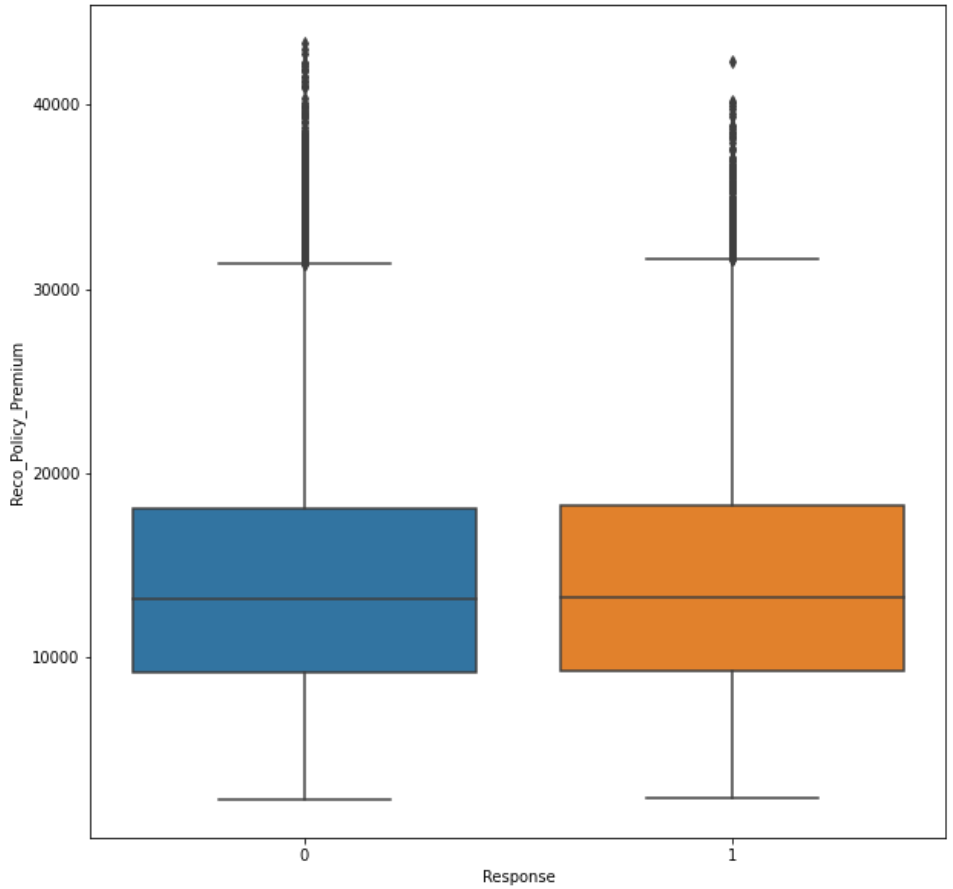
to predict whether the person will be interested in their proposed Health plan/policy given the information about the customer

APPROACH


Understanding Data
Feature Engineering
Model training
Result and Summary

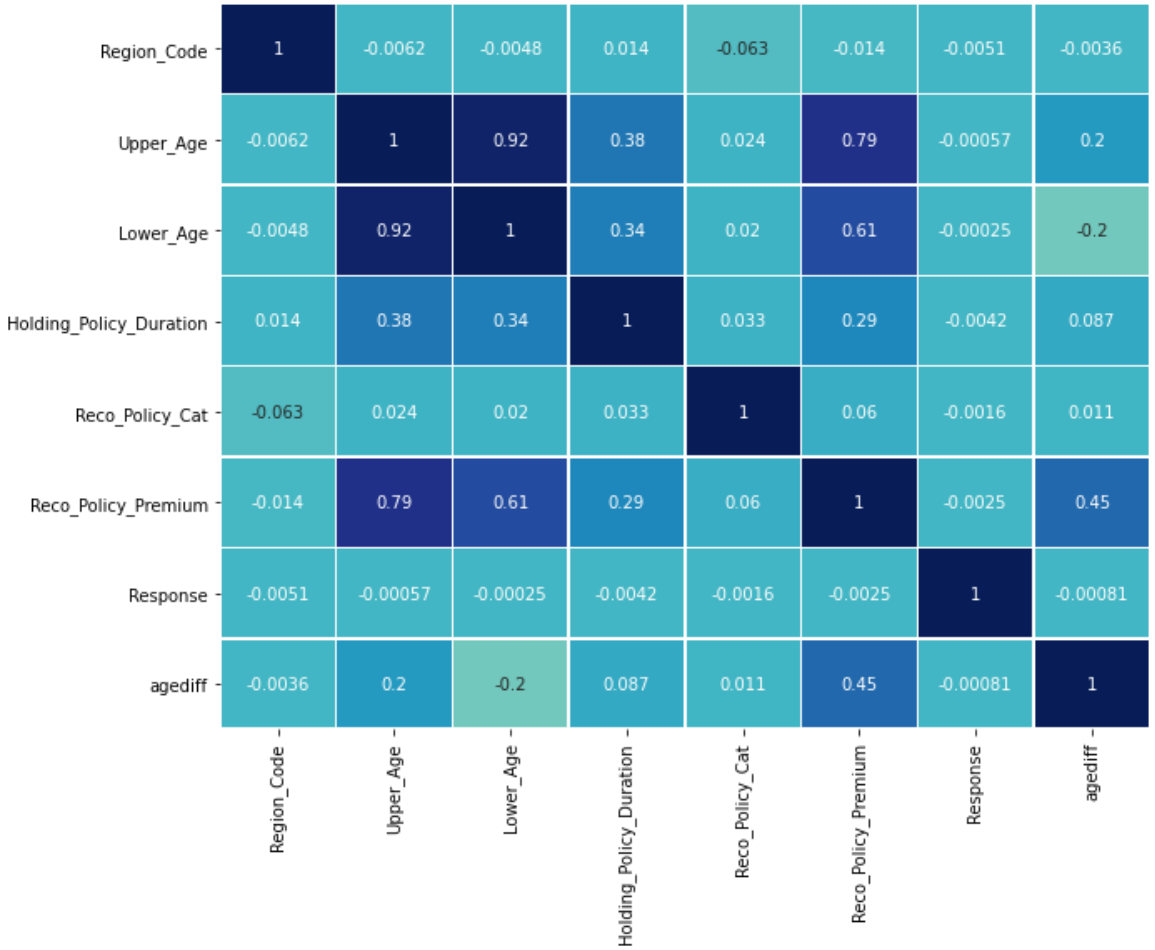
Understanding Data





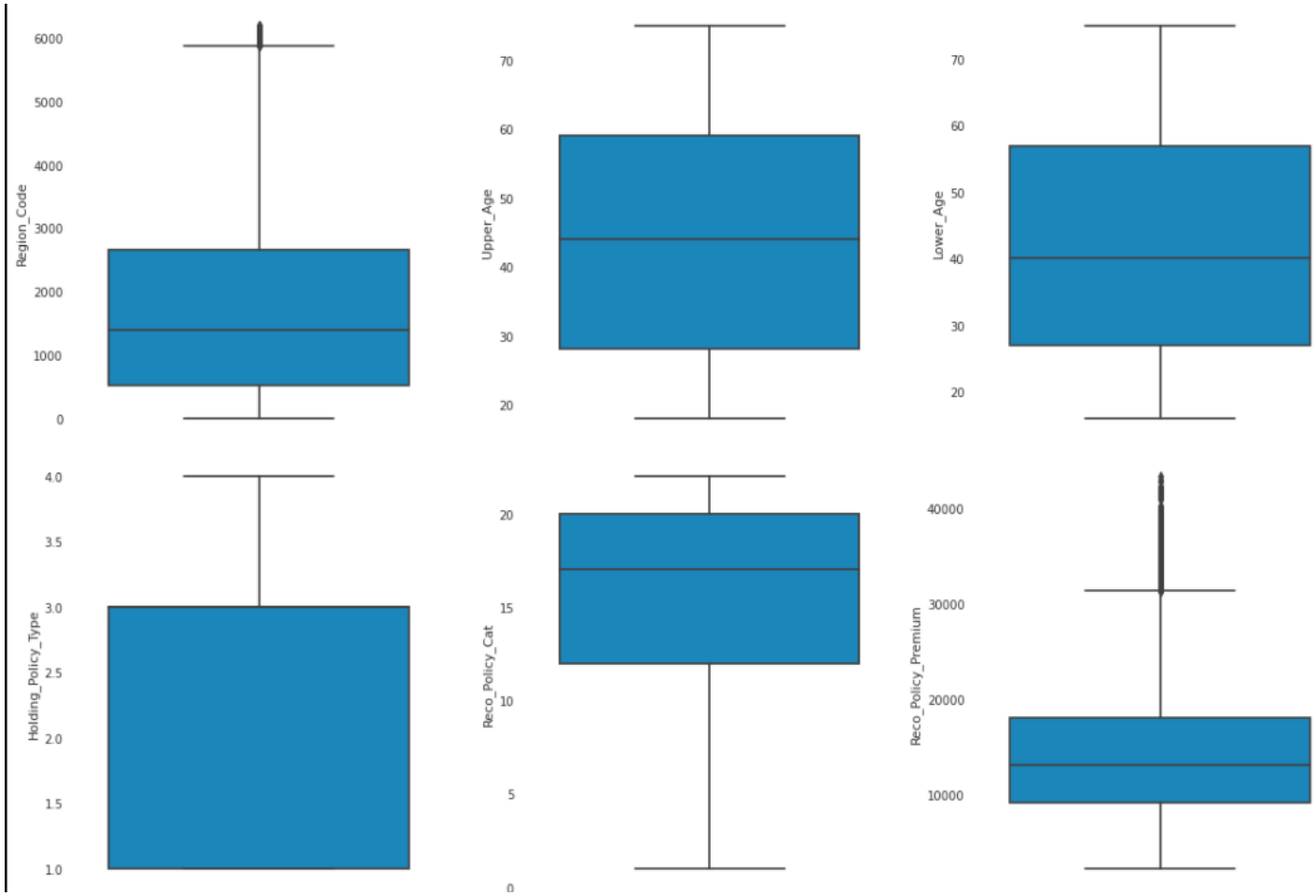
Median value of Recommended policy premium is pretty much same across interested and not interested customers

Understanding Data



Premium is correlated to Lower and upper age, which themselves are correlated

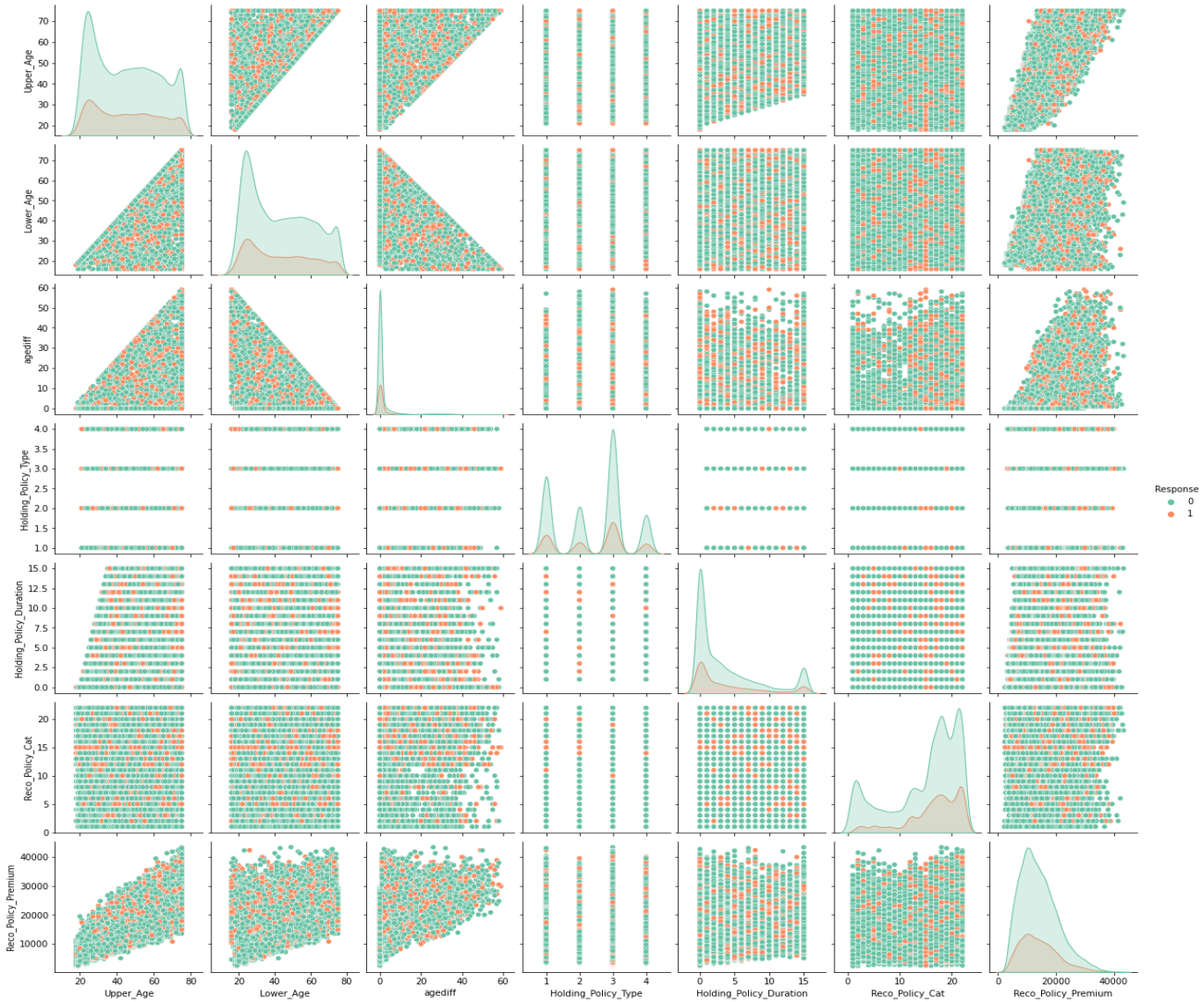
Understanding Data



Some outliers are evident

Understanding Data

Pairplot to understand relations and patterns between variables taking Response into account



Understanding Data

Key points

- 222 Region codes are new in test data
- Frequency encoding would be a good approach
- Recommended policy premium is correlated to the age of the customer(s)
- If data related to holding policy is missing then that customer could be taken as a new customer
- Smart group level imputation for Health Indicator would be a better approach

Feature Engineering

Key points

- New feature for the age difference (Upper age – Lower)
- Encoding 14+ in Holding Policy Duration as 15 and them imputing 0 for nulls
- Logic behind imputing 0 is to treat those cases as new customers
- Fill ‘nc’ (new customer) Holding Policy Type for missing values
- Using category level mode to impute the Health Indicator for the available category levels and global mode of training data in remaining cases
- New feature using Holding Policy Duration
- Category and label encoding
- New column to flag out any new customer

Model Training

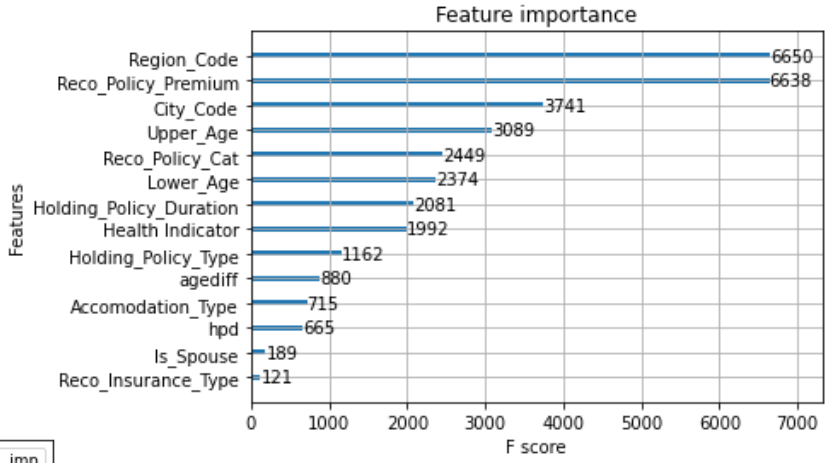
Salient points

- Using LightGbm, Xgboost and catboost algorithms
- Giving categorical feature index in catboost (to make the algorithm understand which columns are categorical)
- Hyper parameter tuning
- Cross validation and using class weight is a good approach

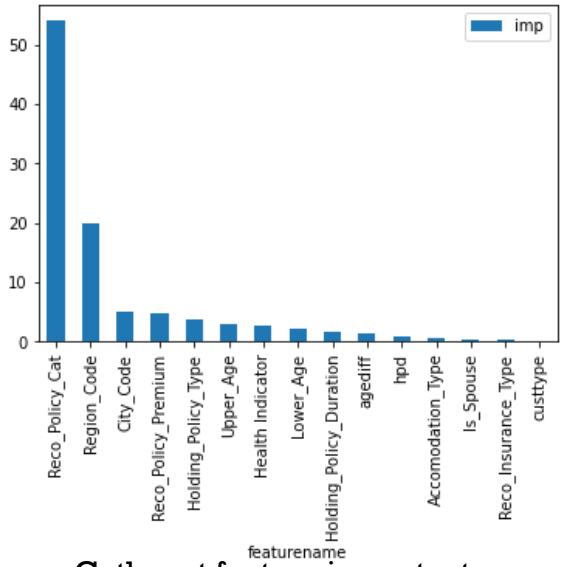
Result and Summary

Final model: Using Catboost
Validaton Roc_Auc : 0.80

Test Roc_Auc : 0.80 (Submission score)



Xgboost feature important



Catboost feature important

Summary:

- Region code is an important feature
- Recommended Policy category and Policy premium are also important features driving the decision of the customer

Thanks