Name : Sachinkumar Uttamrao Ranveer
Designation : Data Engineer (Nykaa 15 July to 21 August)
Education : Mtech Machine Learning(IIIT Allahabad) [2017 to 2019]

Val AUC = 0.67

Problem Statement :

The finance companies give different kinds of services to its customers including insurance policies, loans, EMI services. In order to expand the business the company targets certain customers who may subscribe to one of the services provided by the companies. To whom a company should target in order to expand the business can be answered using machine learning. The historic data of the customer can be used to train the ML model and later the same can be used to effectively make decisions regarding recommending particular policies to the customer and whether to target the particular customer or not. Here we are going to answer whether the given customer will subscribe to recommended service or not provided his historic data in the form of city he lives in, age, already a customer, heath index etc.

Approach :

First I have done lots of exploratory data analysis to come up with some insights from the data, after doing analysis I managed to get 4 new features.

a) Couple Response Code :

Some things which occur rarely can give lots of information about the other events or the event we are interested in. Keeping this in mind I found that the couples were occurring rarely in the dataset so I added couple_response_code that is how likely it is for the couples to respond positively.

b) Big City :

The people living in the city are normally educated. They tend to have knowledge about what offers, online schemes different companies are offering. They are also more focused on investing money on different schemes whereas people living in small towns are not much aware of schemes so if the customer lives in a big city there is a good chance that he will be interested in different schemes proposed by the companies. So, to know whether the city is a big city or not I am counting the number of regions that fall in that city. If the number of regions that fall in the city are greater than some threshold I will consider that as a big city.

c) Prob Estimate :

```
Old Guy + Already hold policy for long duration (Old Customer) + Rich
meaning owns the house
```

Using the intuition I have come up with this new feature. It is common that old people are prone to health related issues, the second one is some customers who are taking services from the last few years can be considered as trustworthy and the person who is rich can subscribe to as many schemes as he wants. I used this knowledge to get a prob_estimate of how likely the customer is going to respond positively to the given scheme.

d) City Response :

People living in some city tend to be more active in some dedicated field. Our aim is to find out whether the people living in some city are active in health related policies. To know it I have used a response feature given in the dataset. If for some city people the probability of people showing the interest in recommended policy is reasonably higher than other cities we can think of that city as active in the 'Health' related policies.

As we have both categorical and continuous features I tried different encoding techniques for categorical features. I tried one hot encoding , frequency encoding, hash encoding specially to deal with large cardinality categorical features, response encoding. The numerical features were taken care of by standard scaler (x - mean/sd).

After all the preprocessing and cleaning of the data I mainely trained 4 models namely Adaboost, Xgboost, Catboost, Logistic Regression. I got 0.68 val auc on Catboost but it was overfitting so chose to drop it. For logistic regression and Adaboost performance was moderate and not much impressive. I spent lots of time on training and fine tuning the Xgboost model. Sometimes it was overfitting , sometimes underfitting. I tried different parameters available to make it work properly.

During the training my model was only predicting only labels as 0 that is for most of the time it was predicting labels as class 0. I realised that I am facing a class imbalance problem as there are more data points belonging to class zero. To deal with class imbalance problems I tried classical methods such as oversampling, undersampling and SMOTE. I also combined undersampling with SMOTE, but it did not increase my score beyond 0.67.