Anadolu Hayat Emeklilik Datathon

(Ökkeş Donbaloğlu - Muhammed Ali Karslı)

Overview

We have the data of customers of Anadolu Hayat Emeklilik insurance company. The goal is to predict which customer is likely to buy a product of the given seven product:

- HU06
- HU07
- HU11
- HU12
- HU13
- HU15
- HU19
- UA (not bought)

Train data has 852.719 rows and 96 columns. Some of the important columns are label, gender, occupation, income, age, segment, number of child and the other insurances that are already bought. Label column shows us what customer bought. We will try to predict label column in test data.

The difficult part of the problem

When we try to implement random forest classifier model, it has very good accuracy. But that's not a good thing. Before explaining why, let's look at the counts of the products:

Label: HU06 - Row count: 3.178

Label: HU07 - Row count: 3.728

Label: HU11 - Row count: 439

Label: HU12 - Row count: 676

Label: HU14 - Row count: 10.754

Label: HU15 - Row count: 415

Label: HU19 - Row count: 759

Label: UA - Row count: 832.770

It's clear that there is an imbalanced data. %97.6 of data is UA (not bought). That's why we need to look other metrics such as weighted-f1 score, precisions:

	precision	recall	f1-score	support
HU06	0.01	0.24	0.03	635
HU07	0.02	0.88	0.03	746
HU11	0.00	0.76	0.01	88
HU12	0.00	0.39	0.00	135
HU14	0.05	0.34	0.08	2151
HU15	0.01	0.59	0.02	83
HU19	0.01	0.09	0.01	152
UA	1.00	0.32	0.48	166554
accuracy			0.32	170544
macro avg	0.14	0.45	0.08	170544
weighted avg	0.98	0.32	0.47	170544

Classification report looks like this after training the model. We used shuffle=True when splitting the data for train and test. We gave the parameters as below:

(n_estimators=250, max_features= 3, max_depth = 5, min_samples_leaf = 2, min_samples_split = 2, random_state=42)

We have also tried to give different class_weights. We calculated class_weights with this formula:

```
total observations / (number of classes * observations in class)
```

After all, the model is still not be able to predict the products that has less counts.

The Way to Solution

As seen before, this data is very imbalanced. It requires to use other methods that are specialized in these types of data. SLIM (sparse linear methods) is one of the topics.

Contact

Ökkeş Donbaloğlu: okkesdonbaloglu25@gmail.com