Feature Engineering and Supervised Learning

Yu Qiu

Feature Engineering

Table 1: Transaction data

	user_id	signup_time	purchase_time	purchase_value	device_id	source	browser	sex	age	ip_address	class
0	22058	2015-02-24 22:55:49	2015-04-18 02:47:11	34	QVPSPJUOCKZAR	SEO	Chrome	M	39	7.327584e+08	0
1	333320	2015-06-07 20:39:50	2015-06-08 01:38:54	16	EOGFQPIZPYXFZ	Ads	Chrome	F	53	3.503114e+08	0
2	150084	2015-04-28 21:13:25	2015-05-04 13:54:50	44	ATGTXKYKUDUQN	SEO	Safari	M	41	3.840542e+09	0
3	221365	2015-07-21 07:09:52	2015-09-09 18:40:53	39	NAUITBZFJKHWW	Ads	Safari	M	45	4.155831e+08	0
4	159135	2015-05-21 06:03:03	2015-07-09 08:05:14	42	ALEYXFXINSXLZ	Ads	Chrome	M	18	2.809315e+09	0

Feature Engineering

Table 2: IP address look-up table

	lower_bound_ip_address	upper_bound_ip_address	country
0	16777216.0	16777471	Australia
1	16777472.0	16777727	China
2	16777728.0	16778239	China
3	16778240.0	16779263	Australia
4	16779264.0	16781311	China

Feature Engineering

Three main problems of the features in this dataset that we need to handle:

- How use the time related raw data in the model
- How to do encoding of the categorical features
- How to handle the imbalanced label data

Feature Engineering

Operations on time related raw data

- Combine the sign up time and first transaction time to get the interval between sign up and first transaction
- Convert the time of sign up to the days of a year and the seconds of a day
- Convert the time of transaction to the days of a year and the seconds of a day

Feature Engineering

Encoding of categorical features:

- For the categorical features which only have several unique variables, we can do the one-hot encoding to convert it to numerical variables
- For the categorical features which have many different variables, such as the devices_shared, we can count the frequency and then encode them by frequency

Feature Engineering

Methods to handle imbalanced data

- SMOTE (Synthetic Minority Over-sampling Technique)
- Import weight to balance the dataset during training phase

Supervised Learning

I only pick two models, logistic regression and random forest, at this phase of our project.

During the training phase , I apply five-fold cross validation to choose the hyperparameters for these two models according to different metrics, such as f1 score and recall.

The final result shows that the performance of random forest is way better than logistic regression.

Supervised Learning

I also use the feature importance generated by random forest to make some predictions and recommendations. This table shows the feature importance.

	importance
interval_after_signup	0.417489
purchase_days_of_year	0.125334
purchase_seconds_of_day	0.081799
signup_seconds_of_day	0.078478
n_dev_shared	0.072574
signup_days_of_year	0.054064
purchase_value	0.043650
age	0.039094
n_country_shared	0.024125
n_ip_shared	0.017022
sex	0.007894
browser_Chrome	0.006778
source_SEO	0.006321
browser_FireFox	0.005802
source_Ads	0.005355
source_Direct	0.005003
browser_Safari	0.004798
browser_IE	0.003601
browser_Opera	0.000819

Conclusion and Recommendation

I look into the relation between some features and the class.

Firstly, I check the relation of the n_dev_shared and the class. It is obvious that the ratio of fraud is higher as more accounts shared one device.

class	0	1
n_dev_shared		
0.0	104966	461
0.2	4403	371
0.4	152	172
0.6	37	87
0.8	13	32
1.0	1	5

Conclusion and Recommendation

Secondly, I looked into the relation between interval after sign up and the class.

What surprised me is that more than half of frauds happened only in one second after the account is signed up. I think this means that more than half of frauds are caused by bot.

	interval_after_signup
class	
0	5194911.0
1	1.0

Conclusion and Recommendation

Finally, I apply my model to get the probability of a transaction is a fraud and then convert this probability to a score. Based on the scores, I gave these recommendations:

- green: 1 - 3 pass
- grey: 4 - 7 need manual investigation
- red: 8 - 9 decline