Predicting Ads Deal Probability

Danijel Kopčinović, IT Market

The Problem

When selling used goods online, a combination of tiny, nuanced details in a product description (e.g. well taken photos, nicely formatted copy etc.) can make a big difference in drumming up interest.

Avito, Russia's largest classified advertisements website, is deeply familiar with this problem. Sellers on their platform sometimes feel frustrated with both too little demand (indicating something is wrong with the product or the product listing) or too much demand (indicating a hot item with a good description was underpriced).

Our task is to predict demand for an online advertisement based on its full description (title, description, images, etc.), its context (geographically where it was posted, similar ads already posted) and historical demand for similar ads in similar contexts. With this information, Avito can inform sellers on how to best optimize their listing and provide some indication of how much interest they should realistically expect to receive.

The task was originally published on Kaggle: https://www.kaggle.com/c/avito-demand-prediction

The Data

The data we work with consists of:

- item_id - Ad id.
- user_id - User id.
- region - Ad region.
- city - Ad city.
- parent_category_name - Top level ad category as classified by Avito's ad model.
- category_name - Fine grain ad category as classified by Avito's ad model.
- param_1 - Optional parameter from Avito's ad model.
- param_2 - Optional parameter from Avito's ad model.
- param_3 - Optional parameter from Avito's ad model.
- title - Ad title.
- description - Ad description.
- price - Ad price.
- item_seq_number - Ad sequential number for user.
- activation_date- Date ad was placed.
- user_type - User type.
- image - Id code of image. Ties to a jpg file in train_jpg. Not every ad has an image.
- image_top_1 - Avito's classification code for the image.
- **deal_probability** - The target variable. This is the likelihood that an ad actually sold something. It's not possible to verify every transaction with certainty, so this column's value can be any float from zero to one.

Finding Predictors for the Target Variable

After initial data processing, our main task is to find good (highly correlated) predictors for the deal_probability variable. The initial set of predictors are the initial data variables but we also include their "interactions"/combinations to cover as much variability in the target variable as possible.

Correlating all these predictors with deal_probability gives us the following table:

	Predictor	Correlation
1	parent_category_name_Лсчные_вещс	0.3249185
2	category_name_Одесда__обувь__аксессуадјы	0.2147931
3	parent_category_name_Услугс__Х__description_length	0.2108996
4	category_name_Пдједлосенсе_услуг__Х__description_length	0.2108996
5	parent_category_name_Услугс	0.2105639
6	category_name_Пдједлосенсе_услуг	0.2105639
7	parent_category_name_Услугс__Х__category_name_Пдједлосенсе_услуг	0.2105639
8	category_name_Детская_одесда_с_обувь	0.1978882
9	param_1_Сенская_одесда	0.1965910
10	parent_category_name_Услугс__Х__user_type_Private	0.1920129
11	category_name_Пдједлосенсе_услуг__Х__user_type_Private	0.1920129
12	item_seq_number__Х__parent_category_name_Услугс	0.1895534
13	item_seq_number__Х__category_name_Пдједлосенсе_услуг	0.1895534
14	image_top_1__Х__parent_category_name_Услугс	0.1870032
15	image_top_1__Х__category_name_Пдједлосенсе_услуг	0.1870032
16	price__Х__parent_category_name_Услугс	0.1700630
17	price__Х__category_name_Пдједлосенсе_услуг	0.1700630
18	item_seq_number	0.1690906
19	parent_category_name_Недвсссмость	0.1646846
20	parent_category_name_Недвсссмость__Х__description_length	0.1621378
21	image_top_1	0.1606214
22	param_1_Для_девочек	0.1479474
23	price__Х__parent_category_name_Недвсссмость	0.1470145
24	parent_category_name_Tdjансподјт	0.1467983
25	parent_category_name_Tdjансподјт__Х__description_length	0.1462346
26	image_top_1__Х__parent_category_name_Tdjансподјт	0.1441984
27	item_seq_number__Х__parent_category_name_Недвсссмость	0.1435174
28	price__Х__parent_category_name_Tdjансподјт	0.1424918
29	image_top_1__Х__description_length	0.1424293
30	param_1_С_пдјобегом	0.1418551
31	parent_category_name_Tdjансподјт__Х__param_1_С_пдјобегом	0.1418551
32	category_name_Автомобслс__Х__param_1_С_пдјобегом	0.1418551
33	param_1_С_пдјобегом__Х__description_length	0.1416224
34	parent_category_name_Tdjансподјт__Х__user_type_Private	0.1404075
35	image_top_1__Х__param_1_С_пдјобегом	0.1397038
36	category_name_Автомобслс	0.1395902
37	parent_category_name_Tdjансподјт__Х__category_name_Автомобслс	0.1395902
38	price__Х__description_length	0.1392518
39	category_name_Автомобслс__Х__description_length	0.1392317
40	price__Х__param_1_С_пдјобегом	0.1387086
41	image_top_1__Х__category_name_Автомобслс	0.1375089
42	price	0.1368594
43	price__Х__category_name_Автомобслс	0.1363071
44	category_name_Квадјтсдјы	0.1342821
45	parent_category_name_Недвсссмость__Х__category_name_Квадјтсдјы	0.1342821
46	category_name_Квадјтсдјы__Х__description_length	0.1331503
47	category_name_Автомобслс__Х__user_type_Private	0.1323884
48	param_1_С_пдјобегом__Х__user_type_Private	0.1323248
49	parent_category_name_Недвсссмость__Х__user_type_Private	0.1320766
50	param_1_Дјемонт__стдјостельство__Х__description_length	0.1295317

To better understand the meaning of the predictors, here is a brief translation of some words from Russian to English:

```
* Лсчные вещс : Laser products
* одесда обувь аксессуадjы : clothes accessories shoes
* Услуг : Services
* Пdjедлосенсе услуг : Paid services
* Детская одесда с обувь : Children's clothes with shoes
* Сенская одесда : Women's clothes
* Недвсссмость : Undecided
* Для девочек : For girls
* Tdjансподjт : Transport
* С пdjобегом : With mileage
* Автомобслс : Automotive
* Djемонтстdjостельство : Demonstration
* Сдам : Rent
* Квадjтсdjы : Quads
* Пdjодам : Sell
* Tdjансподjт пеdjевозкс : Transit train
```

The "___X___" sequence of characters means the interaction/combination of two variables.

We can see that the top (mostly correlated) predictor for the deal_probability is the parent category of "Laser products". This means that if the parent category is "Laser products", then the deal probability increases. Coefficient 0.3249185 is the Spearman coefficient. Its magnitude indicates solid positive correlation between the two variables. Similarly, we can see that:

- the deal_probability increases for the category "clothes accessories shoes"
- the deal_probability increases with the ad text description length for the parent category "Services"
- the deal_probability increases for "Paid services" within the "Services" category etc.

Let's have a look at some predictors negatively correlated with the deal_probability:

	Predictor	Correlation
4198	parent_category_name_Лсчные_вещс__X__user_type_Private	-0.2575794
4199	parent_category_name_Лсчные_вещс__X__description_length	-0.2625350
4200	image_top_1__X__parent_category_name_Лсчные_вещс	-0.2800965
4201	price__X__parent_category_name_Лсчные_вещс	-0.2906887
4202	item_seq_number__X__parent_category_name_Лсчные_вещс	-0.3131590

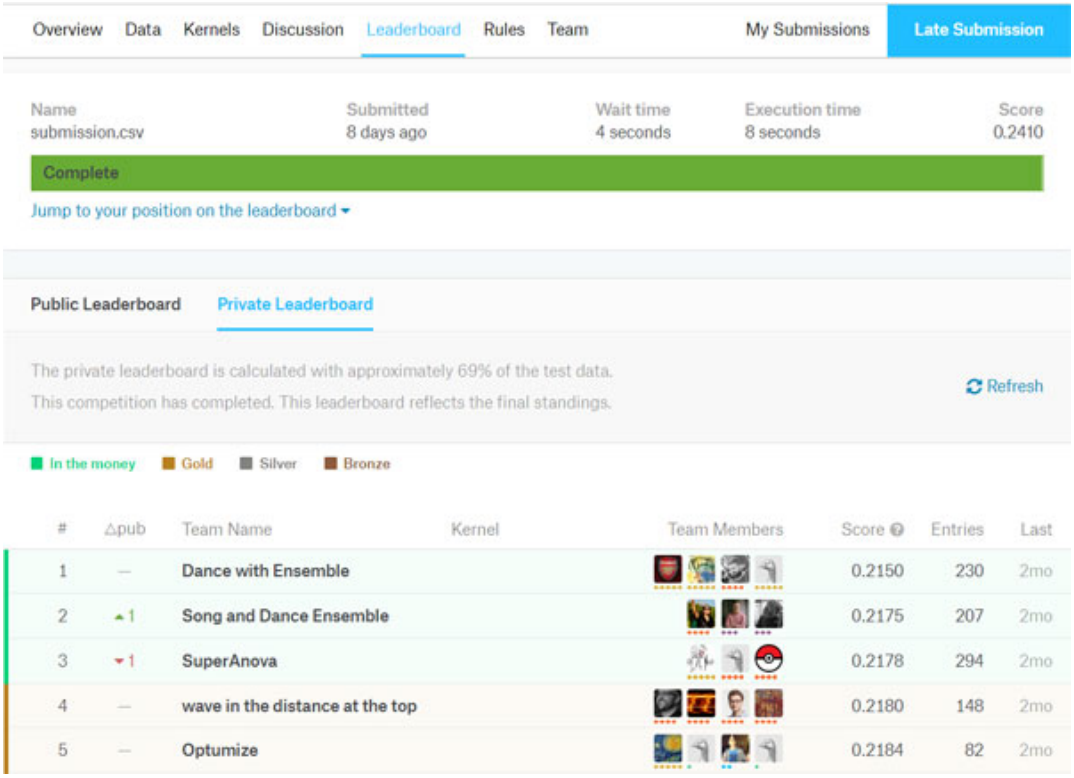
We can see that the deal_probability decreases in the category "Laser products" for "Private" users, for greater "description length" (this is a bit surprising, but it's what the data suggests), for greater price and for greater item_seq_number. Greater item_seq_number indicates greater number of ads by the same seller, so the conclusion would be that if a seller within "Laser products" publishes many ads, he/she actually decreases the probability of the deal/sale. The image_top_1 is an internal Avito image categorization so we cannot describe it verbally, but is also decreases the deal_probability for the "Laser products".

Sometimes it's not easy to interpret the interactions and combinations of predictors, but if they do provide an important impact on the target variable (indicated by the correlation coefficient) then they should be included in the model to increase its quality.

Regression Models

Once we've set up the most important predictors, we build the regression models for the deal_probability. We use 3 different methods: linear regression, xgboost and neural network. All the methods gave us roughly 20% R-squared value meaning that they managed to explain 20% of the deal_probability variations. This is not too good and it indicates that there are other variables influencing the deal_probability variable. Some of those would probably be the quality of the ad images that we didn't take into account.

All in all, we predicted the target variable for the test set in the challenge and scored a solid result of 0.2410 RMSE (root mean squared error). The top results were around 0.21 RMSE.



Conclusion

Data science provides many different tools and approaches to analyze data and make educated business decisions.

Are you analyzing your data? ;)

For further information please contact **danijel.kopcinovic@itmarket.hr**.