

Avito Demand Prediction Challenge, Kaggle

2nd place solution

Valentina Biryukova, Moscow

Sergei Fironov, Saint-Petersburg

Savva Kolbachev, Minsk

July 7, 2018

Table of contents

1. Introduction

2. Data

3. Features

4. Validation

5. Models

6. Stacking

7. Other technical details

8. 1st place solution

9. Conclusion

1	—	Dance with Ensemble
2	▲ 1	Song and Dance Ensemble
3	▼ 1	SuperAnova
4	—	wave in the distance at the top
5	—	Optumize
6	—	Dmitry Larko
7	▲ 7	Light in June
8	▲ 2	CortexLabs

Introduction

Formulation of the problem

- Avito is the Russian largest classified advertisements website;
- The aim of the competition was to predict demand for a specific good.

Объявления Магазины Для бизнеса Помощь [Вход и регистрация](#)

Avito Авто Недвижимость Работа Услуги ещё... [Подать объявление](#)


Любая категория Поиск по объявлениям Москва Станция метро Найти

☐ только в названиях ☐ только с фото


Все объявления в Москве 7 792 989

Личные вещи 3 000 464 Для дома и дачи 627 264 Услуги 105 521 Для бизнеса 56 824
Транспорт 2 389 583 Бытовая электроника 543 779 Работа 99 066
Хобби и отдых 792 306 Недвижимость 107 062 Животные 71 120

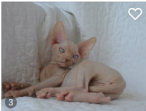
Кошки



Котёнок канадского сфинкса
10 000 р.
м. Площадь революции
Сегодня 11:13




Канадский сфинкс. Мальчик
Цена не указана
м. Охотный ряд
Сегодня 12:32




Канадский сфинкс
компаньон
15 000 р.
м. Измайловская
Сегодня 11:20


Вы смотрели



Palit GeForce GTX 780 3Gb
11 000 Р
Москва, м. Октябрьская



Мониторы 20-29" есть о...
3 500 Р
Москва, м. Фрунзенская



Канадский сфинкс
Цена не указана
Москва, м. Комсомольская

What you need to succeed

2nd place solution took:

- 64 GB + 64 GB + 32 GB RAM;
- 1 TB data storage;
- Two NVIDIA GTX 1080Ti;
- 500\$;
- 3-4 hours everyday work for 3 people for 2 months;
- ~nonstop servers work.

Data

- Train: 1.5KK lines, test: 0.5KK;
- 5 GB of similar data for unsupervised learning,
~70 GB of pics and data for ads duration in zip format;
- Data types: ids, numerical data, categorical data, dates, counters, texts, geo data, pics, results of other models;
- Target in $[0, 1]$ - probability of being sold (or something else);
- Evaluation: RMSE.

Data example

	item_id	user_id	region	city	parent_category_name	category_name	param_1	param_2	param_3
0	b912c3c6a6ad	e00f8ff2eaf9	Свердловская область	Екатеринбург	Личные вещи	Товары для детей и игрушки	Постельные принадлежности	NaN	NaN
1	2dac0150717d	39aeb48f0017	Самарская область	Самара	Для дома и дачи	Мебель и интерьер	Другое	NaN	NaN
2	ba83aefab5dc	91e2f88dd6e3	Ростовская область	Ростов-на-Дону	Бытовая электроника	Аудио и видео	Видео, DVD и Blu-ray плееры	NaN	NaN

	title	description	price	item_seq_number	activation_date	user_type	image	image_top_1	deal_probability
	Кокоби(кокон для сна)	Кокон для сна малыша,польз...	400.0	2	2017-03-28	Private	d10c7e016e03247a3bf2d13348...	1008.0	0.12789
	Стойка для Одежды	Стойка для одежды, под веш...	3000.0	19	2017-03-26	Private	79c9392cc51a9c81c6eb91eceb...	692.0	0.00000
	Philips bluray	В хорошем состоянии, домаш...	4000.0	9	2017-03-20	Private	b7f250ee3f39e1fedd77c141f2...	3032.0	0.43177

Data example at Avito site

Все объявления в Казани

Транспорт

Вспомогательные и аксессуары

Аксессуары

city

parent_category

category

param_1

This ad has no param_2 and param_3

Назад Следующее →

★

Усилитель для радиостанций сибирского диапазона 200 Вт

title


2 000 ₽

price

№ 742345681, размещено 6 мая в 23:30

👁 638 (+1)

Добавить заметку



image

Показать телефон
8 986 XXX-XX-XX

Написать сообщение

Азат
Компания
На Avito с июня 2014
Завершено 25 объявлений

10 объявлений пользователя

Контактное лицо
Азат
Адрес
Казань, м. Проспект Победы

Avito

×

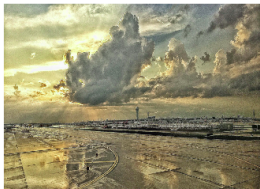
Features

Features

- Texts: SVD over tf-idf and/or countvectorizer, fasttext with different word preparation and interactions, byte pair encoding for nn, doc2vec, other features like sentences length, word overlap, special symbols;
- Categorical data: statistics with average price or duration on 1-3 interactions, robust to overfitting target encoding with all available data;
- Pictures: key points, pic characteristics, vectors from pre-trained models: VGG16, ImageNet, ResNet50, MobileNet;

Features (pictures)

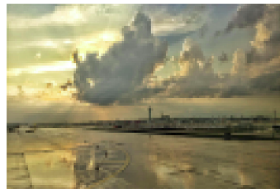
Something about pics:



4.008 +- (1.849)



4.464 +- (1.828)



3.003 +- (2.199)



3.243 +- (1.828)



4.879 +- (1.699)



2.033 +- (1499)

Features (pymorphy)

	param_1	param_2	param_3	title	description	text	descr
0	Постельные принадлежности	NaN	NaN	Кокоби(кокон для сна)	Кокон для сна малыша,пользовались меньше месяц...	постельный принадлежность кокоби кокон сон	кокон сон малыш пользоваться маленький месяц ц...
1	Другое	NaN	NaN	Стойка для Одежды	Стойка для одежды, под вешалки. С бутика.	другой стойка одежда	стойка одежда вешалка бутик
2	Видео, DVD и Blu-ray плееры	NaN	NaN	Philips bluray	В хорошем состоянии, домашний кинотеатр с blu ...	видео плеер	хороший состояние домашний кинотеатр настроить...
3	Автомобильные кресла	NaN	NaN	Автокресло	Продам кресло от0-25кг	автомобильный кресло автокресло	продать кресло

Features (interactions)

```
['ккнсврдлов снсврдлов млшсврдлов плзвлссврдлов мншсврдлов мсцсврдлов цвтсврдлов срьсврдлов',  
 'стйксмр одждсмр вшлксмр бтксмр',  
 'хршмрств сстнрств дмшнйрств кнттрств нстртрств рбттрств смтрств тврств тргрств',  
 'прдмттрстн крслттрстн',  
 'хршмттрстн сстнттрстн',  
 'мрорнбрг очнорнбрг удбнорнбрг',  
 'стлкнжгрд мртнжгрд пркскйнжгрд рйннжгрд цвтнжгрд тмннжгрд фтнжгрд',  
 ''',  
 'кндцнрксндр гдрслтлксндр рлксндр элктрчскксндр стклпдмнкксндр мтрксндр дплнтлнкснд  
 р устнвлнксндр сгнлзцксндр птскксндр оргнлксндр хэнксндр мшнксндр мстнксндр мшнксндр  
 плнстксндр хдксндр тхнчскксндр испрвнксндр дкмнтксндр прдкксндр кзвксндр гнлйксндр  
 ржвйксндр мшнксндр блскксндр ндчтксндр вднксндр фтксндр цнксндр адквтнксндр тргкснд  
 р мнмлнйксндр осмтрксндр кксндркксндр рйнксндр нвгксндр стднксндр улксндр встчнксндр  
 крглквскйксндр',  
 'пнкврнж пйврнж ждтврнж лбщврнж хэйкврнж кмплктврнж идтврнж бтлчкврнж нблшврнж смчкврнж бс  
 врнж пншкврнж слмнврнж рзгврвтврнж кштврнж сдтврнж прдтсврнж рбнкврнж врсврнж хчтврнж игртв  
 рнж']
```

Features (continue)

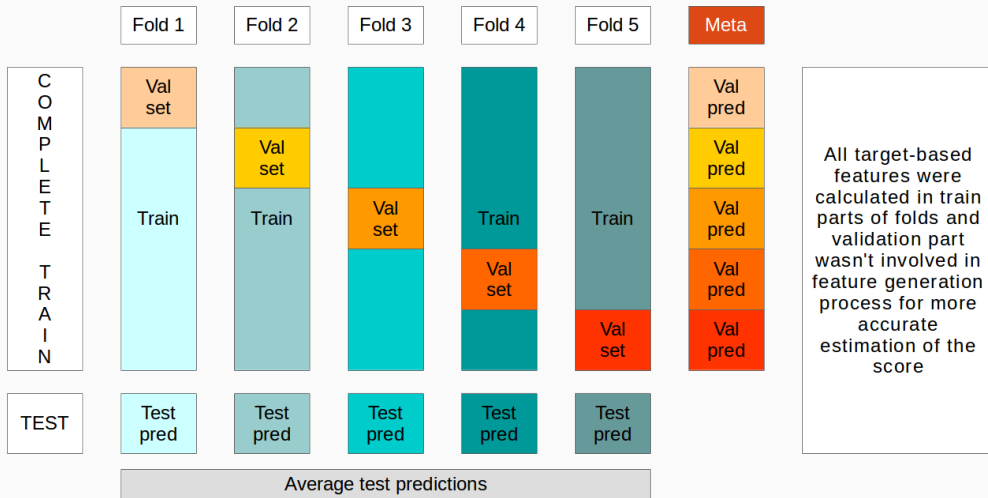
- Numerical data: the most important numerical feature were $\log_{10}(\text{price})$ minus means from statistics, all other numerical features;
- Geo data: information about cities, population, territory, average salaries, number of districts, nearest large cities, and their categorical statistics;
- Dates: they were used only in time series features (that wasn't good enough);
- Predictions from other unsupervised models: clustering of text vectors, pics feature that was given by Avito.

Validation

Validation

- Very sensitive to overfitting folds-in-folds validation without taking DateTime into account;
- All features that use target were calculated into folds where the validation set was treated as test set;
- Then test predictions from all folds were averaged;
- Predictions from validation set were used as meta for stacking;
- Our validation strategy had very good correlation with leaderboard;
- Different models had different gap with lb, but their direction was identical.

Validation scheme

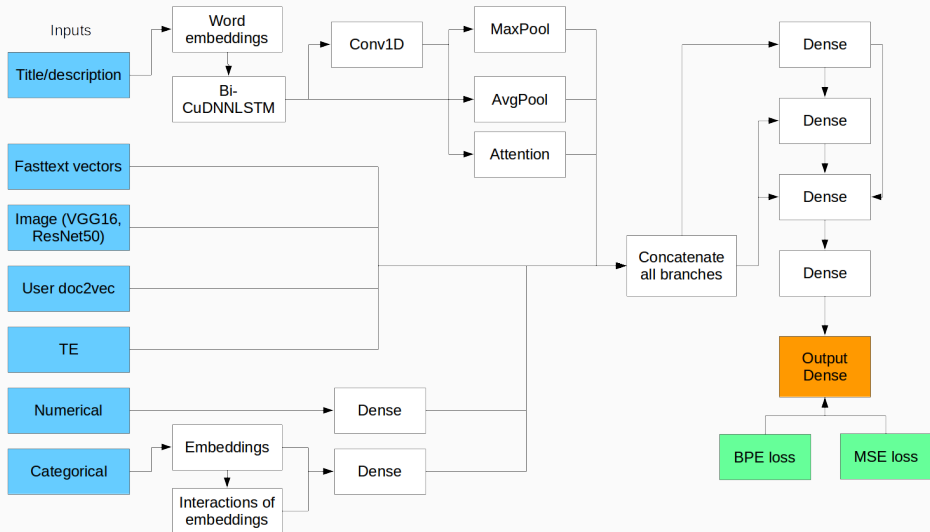


Models

Neural Networks

- Neural Networks: Our best single model (0.2163 on public) is a neural network with different branches:
 - FM like style for categorical features with embeddings;
 - Numerical features;
 - Concatenated fasttext vectors;
 - Concatenated image vectors;
 - BiLSTM for words and BiLSTM for characters with concatenated max, avg poolings with attention;
 - Target encoded features for categorical features and their second and third order interactions;
 - Users 2 vectors features;
- Some details: cyclic LR, Nadam optimiser, plenty of BNs, big dropouts;
- Most of the branches have a dense layer before concatenating them.

Neural Networks scheme



Other models

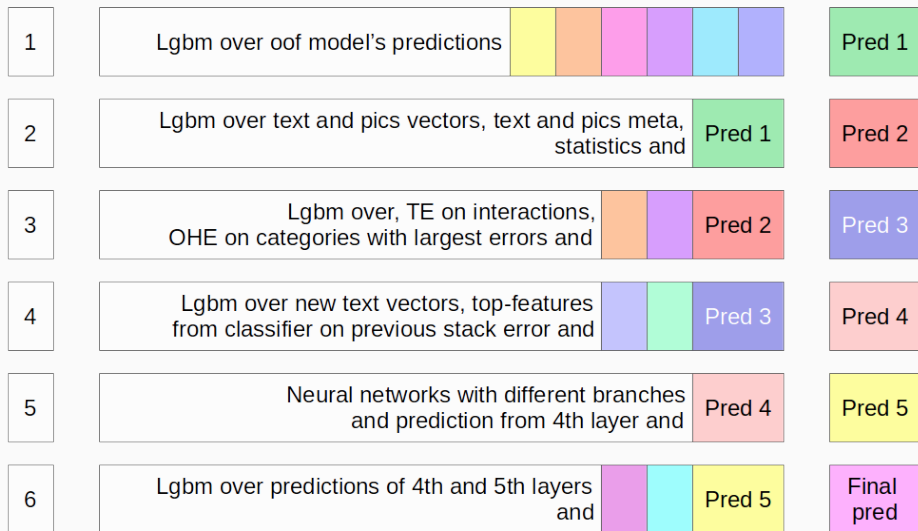
- LightGBM: Surprisingly, fasttext vectors were helpful for lightgbm single model as well (0.2185 on public). SVD over TFIDF transformation was a nice feature for our models. Numerical features, statistics, $\log_{10}(\text{price})$ minus means from statistics, and target encoded categories made the rest of the work;
- Bunch of weak models: FM_FTRL, Ridge, CatBoost, XGBoost.

Stacking

6 layers stack

- First: not deep lgbm over out of folds predictions;
- Second: lgbm with first layer predictions, text and pics vectors, text and pics meta, statistics;
- Third: prediction from 2nd layer, new predictions, TE on interactions, OHE on categories with largest errors;
- Forth: prediction from 3rd layer, new predictions, new text vectors, top-features from classifier on previous stack error;
- Fifth: neural networks with different branches and prediction from 4th layer;
- Sixth: predictions of 4th and 5th stack layers and new model predictions.

Stack scheme



Other technical details

- Unusual libraries: tsfresh for time series, pymorphy for Russian language processing;
- Best ideas: robust TE, full accurate statistics, fasttext in different ways, neural networks with branches of all available data;
- Kaggle discussions were not useful to us at that time;
- All teammates did their roles during competition: one made neural networks, another one on features and plain lgbms, third one on features, stacking and coordinating team.

1st place solution

Important points

- After four participants merged into single team, they got their four good strong approaches combined;
- Their stack was already better than ours, so diversity of models is very important;
- One good feature that could have made us winners: difference between $\log_{10}(\text{price})$ and prediction of $\log_{10}(\text{price})$ based on other features;
- This feature looks like «real market value of a product» and it's difference with set price should determine buyers choice.

Conclusion

If you have any questions fill free to connect with us in ODS chat:

- Ask Sergei (@sergeif) about stacking and nn models;
- Ask Savva (@thinline72) about nn models and features;
- Ask Valya (@mytykritik) about lgbm models and features.