CLASSIFIED

Classified is Jordi Beernink, Roel Bouman, Jeffrey Luppes, Gerdriaan Mulder and Thijs Werrij

Sberbank Russian Housing Market Competition

https://www.kaggle.com/c/sberbank-russian-housing-market

The competition

Sberbank heavily relies on models to predict the value of property

Goal: Predict prices of realty in Moscow area

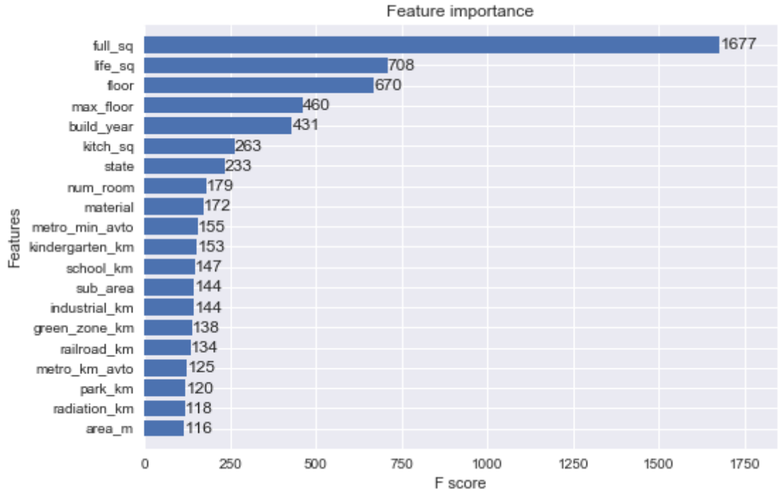
Data set consists of:

- train.csv, data from Russian property market
- macro.csv, data on Russia's macroeconomy and financial sector

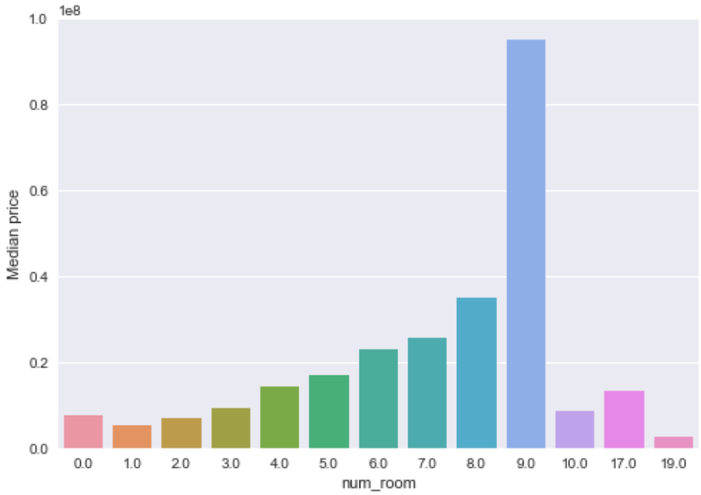
Data exploration

First two weeks we explored the data and discussed findings:

- 30471 entries with 291 features
- Exploratory work: Python / Apache Spark / Leaflet.js
- Good feel for the data, early identification of outliers/missing data



Most important features



Median price vs. number of rooms

Pre-processing

- Training data was supplied as numerical and categorical data
- To test methods, we transformed the data using one-hot encoding for categories, but preserving numerical data
- To deal with missing data, imputing data has been tried (for example KNN and SVD-based methods), but were ultimately too costly to develop internally
- Simpler methods like using mean, modus, median offered no improvement over not replacing NaN values
- Not replacing NaN values means that some numerical data can not be used. Categorical treats NaN data as a separate category

Evaluation

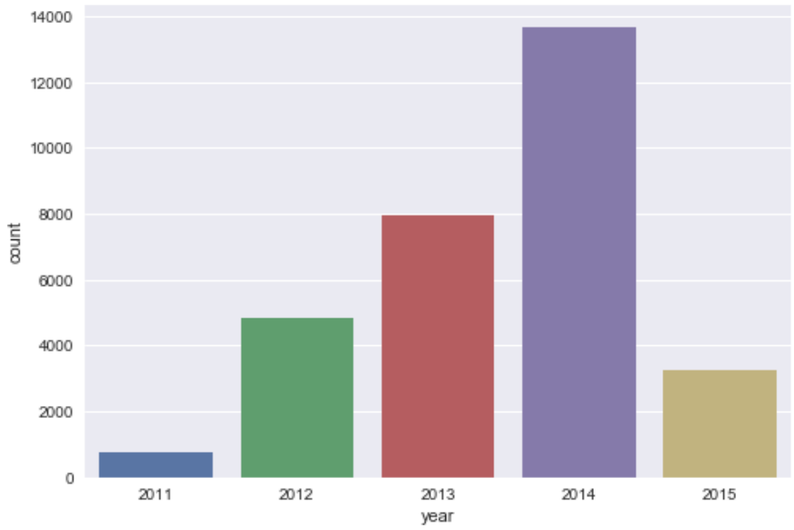
Root Mean Squared Logarithmic Error (RMSLE) $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log \hat{y}_i)^2}$

Split the train data between data before 2015 and in 2015, for internal validation

Little correlation with validation used by Kaggle submissions

Possibly fraud in data, or discrepancies in the time-price correction

Result: usefulness of validation greatly decreased



Testing different methods

Tested several methods:

- Linear Regression
- Random Forests
- AdaBoost
- k-nearest neighbors (KNN)
- Stochastic gradient descent (SGD)
- Deep Learning (Keras)
- XGBoost (Extreme Gradient Boosting)

Methods are easy to use, readily available, and proven to work decently

XGBoost outperformed all other methods, which later became even more notable…

XGBoost ensemble

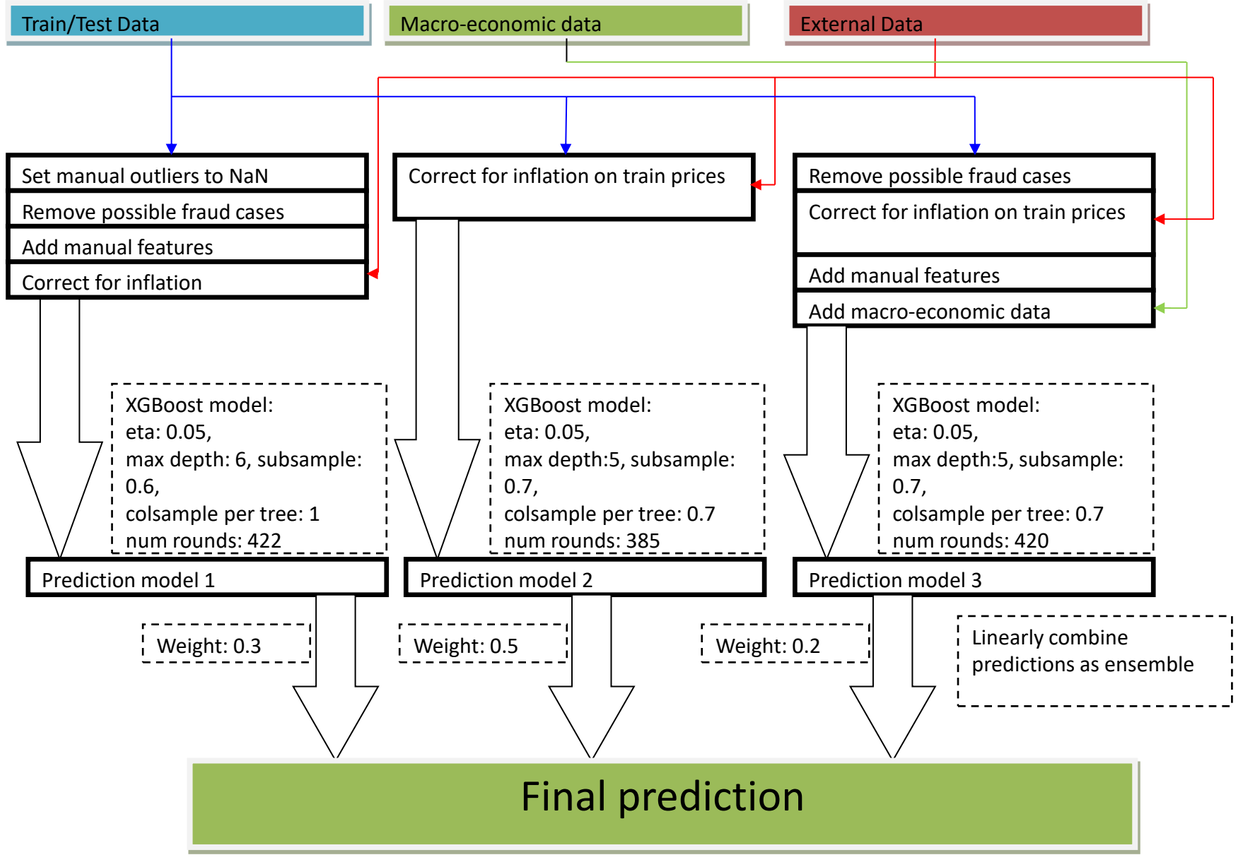
Kernel published by Andy Harless

Used 3 models from others.

Widely used in the competition

We improved it and outperformed the original significantly.

Original now somewhere around rank 600.



Results

Model	Extra info	RMSLE	Kaggle Rank
XGBoost	Ensemble with three models	0.31038	134
XGBoost	Ensemble with three models	0.31062	420
XGBoost	Single model	0.32575	1856
Gradient Boosting Regressor	trained on 2015	0.41384	2767
DNN	10-layer	0.467445	2870
Linear Regression	trained on 2015	0.49689	2897
Decision Tree	trained on 2015	0.5846	3020
SGD Regressor	Naive	0.5956	3021
XGBoost	Baseline Model	0.67333	3034
Random Forest	trained on 2015	0.75239	3040
KNN Regressor	trained on 2015	0.93122	3050
Random Forest	Naive	6.12138	3072

Best obtained score from each implementation is shown, if submitted.

Conclusion

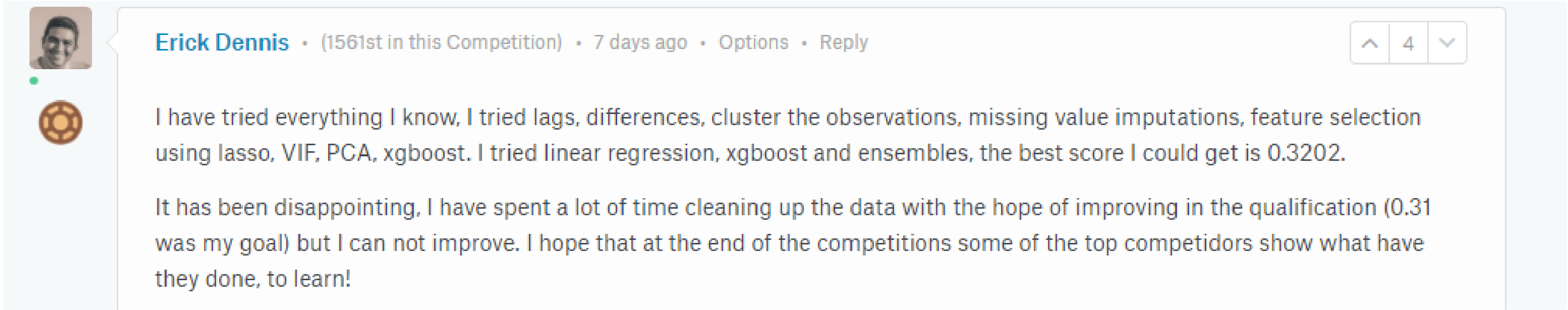
- Top scores very close together
- Ensembles, Ensembles, Ensembles..
- XGBoost outperformed other models
- Pre-processing helped tremendously
- Compensating for inflation

Conclusion

- Felt “wrong” to take advantage of code posted in public
- But at the same time it accelerated our progress, and it was standard practice
- We were working as a community on the same ideas!

Words on Sberbank's competition

- Blatant fraud and tax evasion in data set (5% of data?!)
- Wrong data (laziness) or errors
- Data based on developers/investors/owner's head office location instead of property itself.
- Kaggle: new scores with new data set, old scores not recalc'd until 4 days ago
- Fluctuating leaderboard

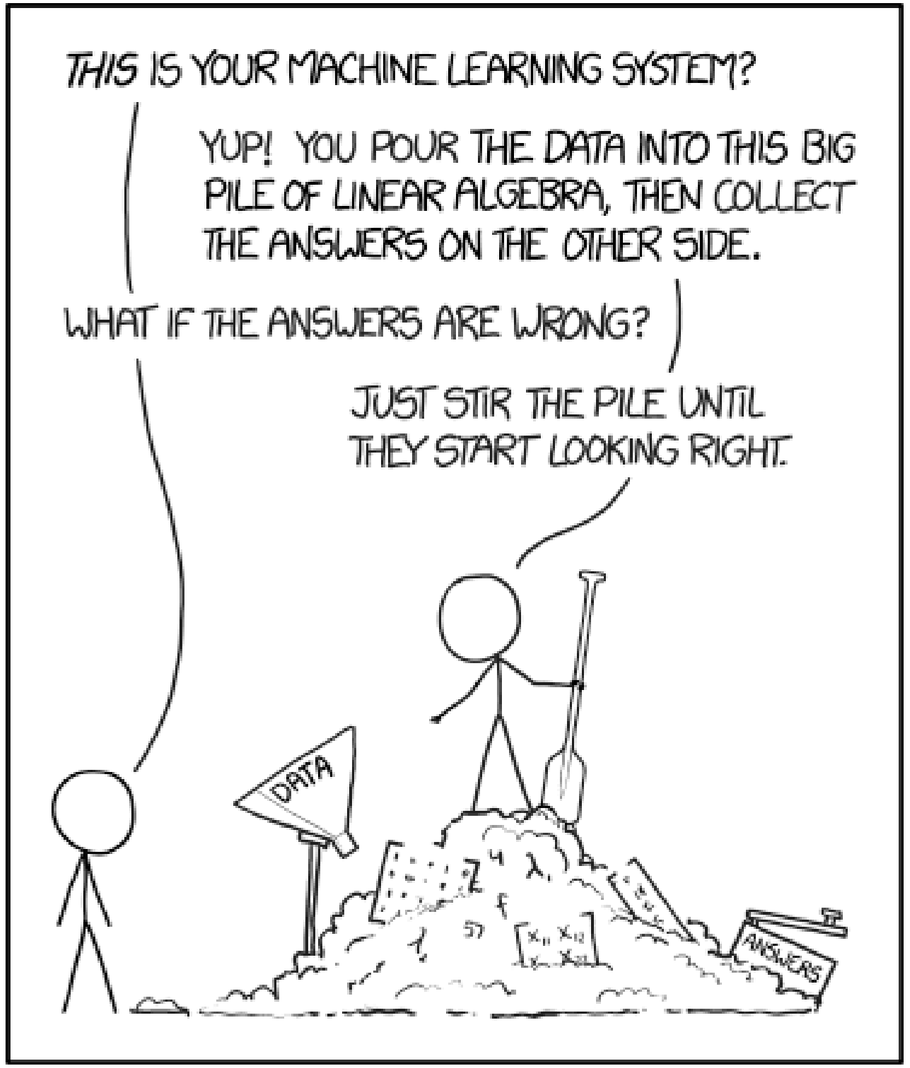


Erick Dennis • (1561st in this Competition) • 7 days ago • Options • Reply

4

I have tried everything I know, I tried lags, differences, cluster the observations, missing value imputations, feature selection using lasso, VIF, PCA, xgboost. I tried linear regression, xgboost and ensembles, the best score I could get is 0.3202.

It has been disappointing, I have spent a lot of time cleaning up the data with the hope of improving in the qualification (0.31 was my goal) but I can not improve. I hope that at the end of the competitions some of the top competidors show what have they done, to learn!



https://xkcd.com/1838/

Reflection

Ensembling, git, putting in more and more hours, letting people do their thing and report back

Yet, someone posts something which blows up. Do we follow suit?

Forming a team (restricting submissions) was premature

Using Kernels to code more interesting because of privacy (kernels set to non-public now)

Team CLASSIFIED

Sberbank Russian Housing Market Competition

XGBoost ensemble

Current rank 130-ish, silver (top 4-5 percent)

Thank you for your attention!