Machine Learning Engineer Nanodegree

Capstone Proposal

Dmitry Chudinovskikh

August 24th, 2017

Domain Background

Housing market in Russia is subject to various fluctuations that depend on a plethora of factors – from macroeconomic shifts to the changes in the local real estate landscape. Sberbank, the largest Russian bank, hopes to find the dependencies between the real estate prices in Moscow and various factors that might be affecting them to make future analysis of the housing market more precise.

My personal motivation for investigating this particular problem is that Moscow is the city where I lived for over 20 years and thereby have at least some knowledge about its housing market. I suppose that the said domain knowledge might be helpful in filtering out false conclusions.

Problem Statement

Given the dataset with over 30000 records and 292 features, I need to create an algorithm that would classify the records by the features that affect the target variable (price) the most. Once that is done, a regression algorithm needs to be applied to find dependencies between the macroeconomic situation in the country and the target variable. However, I shall pursue other ways to approach this issue as well, with the aim of the project being to predict the sale price of each property.

The training data is from August 2011 to June 2015, and the test set is from July 2015 to May 2016. The dataset also includes information about overall conditions in Russia's economy and finance sector.

Datasets and Inputs

All the datasets have been downloaded from the Kaggle competition page (https://www.kaggle.com/c/sberbank-russian-housing-market/data) with the description taken directly from the said page as well:

- train.csv, test.csv: information about individual transactions. The rows are indexed by the "id" field, which refers to individual transactions (particular properties might appear more than once, in separate transactions). These files also include supplementary information about the local area of each property.
  https://www.kaggle.com/c/sberbank-russian-housing-market/download/train.csv.zip
  https://www.kaggle.com/c/sberbank-russian-housing-market/download/test.csv.zip

- macro.csv: data on Russia's macroeconomy and financial sector (could be joined to the train and test sets on the "timestamp" column)
  https://www.kaggle.com/c/sberbank-russian-housing-market/download/macro.csv.zip

- sample_submission.csv: an example submission file in the correct format

https://www.kaggle.com/c/sberbank-russian-housing-market/download/sample_submission.csv.zip

- data_dictionary.txt: explanations of the fields available in the other data files
https://www.kaggle.com/c/sberbank-russian-housing-market/download/data_dictionary.txt

Solution Statement

After exploring the data and eliminating missing records, I shall reduce the dimensionality of the data by applying a PCA algorithm to the data (I will also try to apply an ICA algorithm because I expect it to be more accurate). Once that is complete, my initial guess would be to apply a Naïve Bayes algorithm for the classification purposes due to the size of the dataset. However, as it was mentioned above, I shall try other approaches as well to come up with the best classification results.

After that is done, a regression algorithm will be implemented to find dependencies between the target variable and the data on Russia's macroeconomy and financial sector.

Benchmark Model

I shall use DanilaSavenkov's model as a benchmark model (https://github.com/Danila89/sberbank_kaggle). It has the Root Mean Squared Logarithmic Error of 0.31180 and I will try to match that number and possibly achieve a better result.

Evaluation Metrics

The results will be evaluated on the Root Mean Squared Logarithmic Error between their predicted prices and the actual data. The target variable, called price_doc in the training set, is the sale price of each property.

Project Design

Data exploration and cleaning will be implemented first, then dimensionality reduction algorithm will be applied - in particular either a Principal component analysis or Independent Component Analysis algorithm. After that, a classification algorithm will be applied to the data (Naïve Bayes will be an initial attempt which then will be compared with other approaches such as Random Forest and Gradient Boosting).

Then, a regression algorithm will be implemented to find dependencies between the target variable and the data on Russia's macroeconomy and financial sector.

Some domain background description as well as the comments to the code will be added to the Jupyter notebook to make the analysis more intuitive.