Machine Learning Engineer Nanodegree

Capstone Project

P6: Sberbank Russian Housing Market

Report

I. Definition

Project Overview

Regression analysis is a form of math predictive modeling which investigates the relationship between variables. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about these factors and their predictions?

The main factor that we're trying to understand or predict is a target (a dependent variable). The features (independent variables) are the factors we suppose to have an impact on the dependent variable. Using this set of variables, we generate a function that maps inputs to outputs. The training process continues until the model achieves the desired level of accuracy.

The project investigates **supervised learning** as a part of regression analysis that uses a known (training) dataset to make predictions. This dataset includes input data and response values. The supervised learning algorithms seek to build models which make predictions of the response values for a new dataset. A test dataset is used to validate the model.

Housing costs are a sphere in the real economy for applying supervised learning. They demand a significant investment from both consumers and developers. And when it comes to planning a budget—whether personal or corporate—the last thing anyone needs is uncertainty about one of their budgets expenses. Sberbank, Russia's oldest and largest bank, helps their customers by making predictions about reality prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building.

Although the housing market is relatively stable in Russia, the country's volatile economy makes forecasting prices as a function of apartment characteristics a unique challenge. Complex interactions between housing features such as a number of bedrooms and location are enough to make pricing predictions complicated. Adding an unstable economy to the mix means Sberbank and their customers need more than simple regression models in their arsenal.

The project solutions are applied to the real housing costs and consist of two main parts:

1. preparation of data for analysis (selection of variables, deletion of records containing too many empty values, digital encoding categorical variables, etc.);
2. application of a set of machine learning algorithms in regression analysis in order to identify the most effective of them.

The project was built on the basis of the competition offered on the site https://www.kaggle.com.

Here popular Python resources (numpy, pandas, matplotlib, scikit-learn, keras, etc.) for building the regression models are applied.

The most valuable side of this project is the investigation of real data and the attempt to approximate the predictions on them to the threshold of 0.7-0.8 for the coefficient of determination.

Proble Statement

Sberbank is challenging programmers to develop algorithms which use a broad spectrum of features to predict real prices. Algorithm applications rely on a rich dataset that includes housing data and macroeconomic patterns. An accurate forecasting model will allow Sberbank to provide more certainty to their customers in an uncertain economy.

Metrics

The wide spectrum of metrics for regression was chosen and documented.

1. Explained variance regression score.

If $\hat{y}$ is the estimated target output, y the corresponding (correct) target output, and Var is Variance, the square of the standard deviation, then the explained variance is estimated as follow:

$$\texttt{explained\_variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

2. Coefficient of determination.

If $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value, then the score R^2 estimated over n_{samples} is defined as

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2} \quad \text{where} \quad \bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i.$$

3. Mean squared error.

If $\hat{y}_i$ is the predicted value of the i-th sample, and y_i is the corresponding true value, then the mean squared error (MSE) estimated over n_{samples} is defined as

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

4. Mean absolute error.

If $\hat{y}_i$ is the predicted value of the i-th sample, and y_i is the corresponding true value, then the mean absolute error (MAE) estimated over n_{samples} is defined as

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \left| y_i - \hat{y}_i \right|.$$

5. Median absolute error.

II. Analysis

Data Exploration

The datas the for the investigation is a large number of economic indicators for pricing and prices themselves (train.csv and test.csv). Macroeconomic variables are collected in a separate file for transaction dates (macro.csv). In addition, the detailed description of variables is provided (data_dictionary.txt).

For practical reasons, I have not analyzed all the data and have chosen the following independent variables:

1. the dollar rate, which traditionally affects the Russian real estate market;
2. the distance in km from the Kremlin (the closer to the center of the city, the more expensive);
3. indicators characterizing the availability of urban infrastructure nearby (schools, medical and sports centers, supermarkets, etc.) ;
4. indicators of a particular living space (number of rooms, floor, etc.);
5. proximity to transport nodes (for example, to the metro);
6. indicators of population density and employment in the region of housing accommodation.

All these economic indicators have a strong influence on price formation and can be used as a basic set for regression analysis. Examples of numerical variables: the distance to the metro, the distance to the school, the dollar rate at the transaction moment, the area of the living space. Examples of categorical variables: neighborhoods, the nearest metro station, the number of rooms.

The goal of the project is to predict the price of housing using the chosen set of numerical and categorical variables. The predicted target is not discrete, for the training set all the values of this dependent variable are given, and therefore it is necessary to apply the regression algorithms of supervised learning.

The data preprocessing confirmed the assumption: these variables are in a sufficiently strong relationship with the target variable. They are used as the basis for building different types of models in several forms: only numerical variables, numeric and categorical variables transformed into numeric or binary code.

Exploratory Visualization

To realize the project it was necessary to use a lot of visualization tools at all stages: data tables, distributions of quantities, correlation maps, the graphical comparison of predictions and real values, representation of the feature importance for specific algorithms, operation processes of neural networks, etc.

Algorithms and Techniques

To compare the prediction quality, I chose the most effective (for financial indicators) regression ensemble algorithms and different types of neural networks: multilayer perceptrons, convolutional and recurrent neural networks.

1. ScikitLearn: Gradient Boosting Regressor, Bagging Regressor, MLP Regressor.
2. Keras: multi-layer perceptrons (MLP), convolutional neural networks (CNN), recurrent neural networks (RNN).
3. Numpy, Pandas, ScikitLearn: data preprocessing.
4. Matplotlib, Seaborn: data visualization.

In addition, I was wondering what the highest accuracy rate will be achieved by each of the presented algorithms and whether the predicted trends of price change for all used types of techniques will coincide.

Benchmark

The benchmark regressor among investigated models is the Gradient Boosting algorithm. It has the best level of all the evaluation metrics. We should notice that the Bagging algorithm results are really close to Gradient Boosting.

The CNN model for 44 numeric and categorical features demonstrates the best predictions among neural networks.

III. Methodology

Data Preprocessing

Data processing consisted of the following important steps:

1. deleting rows with a lot of missing data,
2. filling a small amount of missing data by linear interpolation,
3. the addition of a macroeconomic indicator,
4. transforming categorical variables into discrete numerical and binary encoded features,
5. checking the coding and eliminating the differences between the category variables in the training and test sets.

Implementation

Two ensemble Scikit-Learn algorithms (Gradient Boosting and Bagging), Scikit-Learn Multi-Layer Perceptron Regressor, three types of Neural Networks (Keras) were applied to three sets of the features (numeric, numeric and categorical, numeric and encoded categorical).

Such a wide range of algorithms allowed to determine the approximate level of achievable accuracy of test predictions for this data set: 70-72%. Identifying the most effective algorithms in the sphere of real financial indicators is also an important task for machine learning in general.

A detailed technical report on the algorithm parameters and the architecture of each particular model is presented in the Jupyter notebook format.

Refinement

Improvements in performance were achieved by optimizing the parameters of the algorithms or developing the structure of the neural networks. As a result, in many models, the indicator "Coefficient of Determination" (for example) changed from the beginning level 66-67 % to the final level 70-72 % on the test data.

IV. Results

Model Evaluation and Validation

All the methods of measurement listed in the section "Metrics" were used to evaluate the accuracy and efficiency of algorithms.

The best indicators:

1. Ensemble Algorithms.

```
<_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_>
Numeric Features; Gradient Boosting Regressor
<_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_>
EV score. Train:  0.86189746402
EV score. Test:  0.720761771784
----------
R2 score. Train:  0.86189746402
R2 score. Test:  0.720678730918
----------
MSE score. Train:  0.251150449123
MSE score. Test:  0.558895012634
----------
MAE score. Train:  0.31458911313
MAE score. Test:  0.400739497884
----------
MdAE score. Train:  0.174402117839
MdAE score. Test:  0.19910250734
```

2. Neural Networks.

```
<_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_>
Numeric and Categorical Features; CNN Model
<_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_><_>
EV score. Train:  0.750772050101
EV score. Test:  0.703572108338
----------
R2 score. Train:  0.750702571788
R2 score. Test:  0.703490345929
----------
MSE score. Train:  0.453367207317
MSE score. Test:  0.593287318946
----------
MAE score. Train:  0.400844047897
MAE score. Test:  0.432183848017
----------
MdAE score. Train:  0.204027751396
MdAE score. Test:  0.208331465191
```

Justification

Participation of the project in the competition allowed to improve the constructed models, compare indicators with other competitors and evaluate the results.

V. Conclusion

Free-Form Visualization

As a final visualization, I chose the image of the predictions of all models on a single graph. As can be seen from the illustrations, the predictions are very close to each other and determine the overall price dynamics quite clearly.

Reflection

The prediction of financial values is quite complex due to the strong dependence of the indicators on each other, the influence of the time factor and uncertainty. To achieve greater approximation to real data is one of the closest and achievable tasks of machine learning.

It was interesting for me to work with the project precisely because of the large range of variables in real data and the possibility to advance the understanding this field of activity.

Improvement

There are many possible ways to improve the modeling: studying of other sets of variables, combining of existing algorithms in ensembles, developing the architecture of built neural networks in the project, applying the existing neural networks with a complex structure from the external sources, etc.

VI. Bibliography

1. Amy Gallo. A Refresher on Regression Analysis. Harvard Business Review, 2015.
2. Model evaluation: quantifying the quality of predictions (http://scikit-learn.org/stable/modules/model_evaluation.html)
3. Keras: The Python Deep Learning library (https://keras.io/).