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 70-80 percentages.

Problem 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;
2. coefficient of determination;
3. mean squared error;
4. mean absolute error;
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.

As expected, 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

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.

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

.

Refinement

.

IV. Results

Model Evaluation and Validation

.

Justification

.

V. Conclusion

Free-Form Visualization

.

Reflection

.

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/).