# Machine Learning Engineer Nanodegree

## Capstone Project

# P6: Sberbank Russian Housing Market

## Capstone Proposal

### Domain Background

**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.

### 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.

### Datasets and Inputs

The basis 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.

## Solution Statement

The project solutions 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.

## Benchmark Models

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

1. Scikit Learn: Gradient Boosting Regressor, Bagging Regressor, MLP Regressor;
2. Keras: multi-layer perceptrons (MLP), convolutional neural networks (CNN), recurrent neural networks (RNN).

## Evaluation Metrics

The wide spectrum of metrics for regression are planned to use and document:

1. explained variance regression score;

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

## Project Design

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 were 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.

## 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/).