Capstone Project Proposal

Elias Laura Wiegandt

Sberbank Russian Housing Market Kaggle competition

May 15, 2017

1 Domain Background

For many individuals, buying and selling housing will be the biggest financial decision they make during their lifetime. A bad decision, such as buying to dearly or selling too cheap, can mean financial ruin. A model for housing prices can be used as a benchmark to avoid buying overpriced or selling underpriced. This can mitigate the risk the individual buyer and seller faces during a housing trade. Machine Learning has been applied to Housing Prices before, see e.g. [2], where Ridge Regression and Support Vector Machines (SVMs) are applied.
Not that when a house is overpriced the seller wins and the buyer loses. Disregarding payment to estate agent or bank, buying and selling housing is essentially a zero sum game. It is assumed that there is decreasing marginal utility of purchasing power (the gain from "winning" the housing trade). This is known as "Gossen's law" and generally accepted in Economics, see [1]. This causes both the buyer and seller to wish to decrease the likelihood of over- or underpricing of the housing, as their downside is bigger than their upside. This is the raison d'tre for constructing a model that predicts housing prices.

2 Problem Statement

My proposed capstone project is to build a model for Russian housing prices, as part of this Kaggle competition: Sberbank Russian Housing. This is a regression problem, where the inputs are features of the housing, its neighborhood and the macro economic environment. Further descriptions of the data can be seen in section 3. The output of the project will be a model that predicts housing prices given the dwelling's feature, its neighborhood and the macro economic environment.

3 Data

Sberbank has made data on houses and macro economic data available: Sberbank Data. The data is also included in the uploaded zip folder. The data consists of three headline categories of data: dwelling specific, neighborhood specific and macro economic.
The housing data set describes 14 features of the dwellings at the time they were sold. One of

these 14 features is the target variable, the price the dwelling was sold at. The data includes variables for the number of rooms, when a carport were built and other variables specific to the dwelling itself. The neighborhood specific data includes 101 variables describing the neighborhood, e.g. how many cafs there are within a specific radius of the dwelling and distance to the nearest kindergarten. The macro economic dataset includes 288 macro economic variables such as GDP growth and inflation measures.
In total, this makes for 1 label and 402 features. In the uploaded zip folder, a `data_dict.txt` with all variables listed is included.
Sberbank has included both a training dataset for training and validation, and a set of test-features that can be used to test the performance of the final model. To do this, a csv datafile with predictions of the prices for the dwellings in the test set have to be uploaded to the Kaggle website, which then returns the Root Mean Squared Logarithmic Error (RMSLE) for the predictions. This will be used as the evaluation metric for this project. The RMSLE is given as:

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(p_i+1\right)-\log\left(a_i+1\right)\right)^2} \tag{1}$$

Where ϵ is the RMSLE, n is the total number of observations, p_i is the prediction, a_i is the actual value for observation i and $\log(x)$ is the natural logarithm of x.
Though Sberbank has already divided the dataset into a test and training set, I expect to use the training dataset for all computation, splitting this into test and training set. This is simply to achieve a higher speed during the project, as uploading predictions to Kaggle is a time consuming process.

4 Benchmark Model

As part of the project on Boston housing prices in the Machine Learning Engineer Nanodegree, a decision tree regressor was fitted to data. The thought process of this project will serve as a model for my capstone project; first getting to grabs with the data, then applying a machine learning algorithm and training it. I expect to supply more visualizations of findings throughout the capstone project, but the basic skeleton of the analysis will be the same.
My success criterion is to build a model than can beat an out-of-the-box Random Forrest Regressor.

5 Solution Statement

I will build a machine learning algorithm for predicting house prices, mainly focusing on ensemble learning models such as gradient boosting, Adaboost and random forest. I will also try using a Ridge Regression, LASSO or Elastic Net for comparison, as these models can be interpreted more easily than an ensemble learning model. At the time of writing, the best contributions have an RMSLE of 0.31. I do not expect to be able to match this, but my goal is to achieve an RMSLE lower than 0.5.

Before I can apply any of these machine learning algorithms, a thorough exploratory data analysis (EDA) will be required. I expect to include relevant parts of the EDA in my final project submission. The Kaggle competition includes several resources I expect to take advantage of; other participants have already published some EDAs or other material on the competition. There is also a lively forum. If I uncover any hitherto unknown findings during my EDA, these findings will be uploaded to the forum in an iPython Notebook.
As part of the EDA, data preprocessing will also be applied. This will be outlier analysis, feature scaling, normalization, feature selection and dimension reduction with e.g. PCA.

6 Evaluation Metrics

As noted earlier, I will use RMSLE as described in equation 1 to evaluate my model. I will be comparing my model against a benchmark Random Forrest Regression Model.

7 Project Design

The work flow could look either like this:

1. EDA
   (a) Analyze the 14 dwelling specific variables
      i. Plot distribution of all variables
      ii. Analyze outliers, remove if necessary
      iii. One-hot encode all class variables
      iv. Analyze scatter plots of all variables against housing prices
      v. Analyze correlation matrix

   (a) Analyze the 101 neighborhood specific variables
      i. Look at all variables' description
      ii. As there are too many variables to plot meaningfully, implement automatic outlier removal process
      iii. One-hot encode all class variables
      iv. Join data with the target variable, housing prices
      v. Construct gradient boosting model that predicts housing prices from all neighborhood specific variables
      vi. Run feature selection on the model.
      vii. Analyze results, choose most important neighborhood features
   (b) Analyze the 289 macro economic variables
      i. Look at all variables' description
      ii. As there are too many variables to plot meaningfully, implement automatic outlier removal process

iii. One-hot encode all class variables
iv. Join data with the target variable, housing prices
v. Construct gradient boosting model that predicts housing prices from all macro economic variables
vi. Run feature selection on the model.
vii. Analyze results, choose most important macro economic features

(c) Combine datasets

i. Join dwelling specific data with macro economic and neighborhood data
ii. Plot histograms of all joined economic and neighborhood features
iii. Analyze histograms, remove outliers and transform data if necessary
iv. Analyze scatter plots of economic and neighborhood featuers against target variable, housing price

2. Building Machine Learning Algorithm

(a) Build pipeline that takes in the joined dataset
(b) Implement feature selection algorithm in the pipeline
(c) Set up different sklearn models to be tried in the pipeline, both for the feature selection and the model itself. This could e.g. be

i. Gradient Boosting
ii. K-Nearest-Neighbors
iii. Adaboost
iv. Random Forrest (will be used as benchmark)
v. Ridge Regression
vi. Elastic Net
vii. LASSO
viii. OLS

(d) Evaluate performance of the models
(e) Choose best performing model, as measured by RMSLE
(f) Tune best performing model with cross validation
(g) Analyze feature importance from best performing model

3. Test performance on Sberbank test set

References

[1] Laurence S. Moss. The Laws of Human Relations and the Rules of Human Action Derived Therefrom. By Hermann Heinrich Gossen. Translated by Rudolph C. Blitz with an introductory essay by Nicholas Georgescu-Roegen. Cambridge: M.I.T. Press, 1983. Pp. 460. USD 47.50. *The Journal of Economic History*, 44(04):1130–1132, 1984.

[2] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.