CSC440 Project Proposal
Predicting Of House Price With Its Parameters

Andrii Osipa

October 2016

1 Problem Statement

House price prediction is actual problem as it is always a task to evaluate home before selling in a proper way, so it will be sold for its value and will not stay for very long time "on sale" because its value was set too high. House price is variable that depends on many other parameters in a very unclear way. Therefore we have a problem to find dependence of a price from house parameters, select ones which affect the price and drop those which does not and still maintain the interpretability of the price got in result. There is a big amount of possible attributes to describe a house that may affect house's price.

This Kaggle contest offers to invent price prediction algorithm using rather big number of attributes and probably invent new attributes.

Difficulties that come in the project are feature selection and invention and also dealing with noisy data that is presented in the dataset.

Other task is to produce a model that will give estimations of a good level, so the evaluation will give score which is not much lower from top of the Kaggle's leaderboard on this problem.

2 Data

Dataset provided by Dean De Cock - Ames Housing Dataset. It contains list of different 79 attributes for each home. These variables explain almost every property of home including its type, age, lot size, type of street and district. Data is from real homes located in Ames, Iowa. There are number of continuous, discrete variables and ordinal and nominal categorical variables. Total amount of records in dataset is 1460 in training and 1459 in test part.

3 Algorithm choice

I am planning to try random forest and regression in this task. On my opinion, regression will give better result but random forest may give very simple model

and still have god performance. I think of trying different types of regressions and algorithms for regression learning process. Actual problem is how each attribute affects real price and are all attributes needed, therefore regressions must be built for different sets of attributes. Moreover, I think that there must be invented new attributes from given ones, that have significant affect on the price and still can be well interpreted. Feature invention also will require some specialist interference so new variables make logical sense.

4 Performance evaluation

Result will be evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

5 References

This problem was observed in different works with different datasets. Previous works contained less variables that explain homes and used different regression techniques to find the regression.

1. **Kaggle. House Prices: Advanced Regression Techniques**
   `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`
2. **Multivariate Regression Modeling for Home Value Estimates with Evaluation using Maximum Information Coefficient**
   `http://www.acisinternational.org/Springer/SamplePaper.pdf`
3. **Machine Learning for a London Housing Price Prediction Mobile Application**
   `http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf`
4. **Predicting Home Prices from Various Factors**
   `http://sites.stat.psu.edu/~sesa/stat460/Project/greg.pdf`
5. **House Price Prediction**
   `http://terpconnect.umd.edu/~lzhong/INST737/milestone2_presentation.pdf`