Proposal

March 27, 2019

Erin Witmer
CSC 440

1 Dataset Choice

The proposal is to analyze the dataset: "House Prices: Advanced Regression Techniques" available via Kaggle. This dataset aligns with my educational background and previous work experience. My undergraduate degree is in finance and economics. I am a Chartered Financial Analyst and have over a decade of work experience in the investment industry. My previous work includes asset valuation and pricing.

Real estate, specifically residential real estate is an asset that is unique in many ways. Residential real estate is an illiquid asset with high transaction costs. Unlike, for example, shares of common stock in Apple (APPL), every residential dwelling is a unique asset. Transactions often occur between unsophisticated parties with highly asymmetric information. There is a large degree of emotion and economically irrational behavior involved each transaction and the resulting sale price. These include psychological phenomena such as anchoring to a purchase price, and factors independent of market forces motivating buy/sell decisions such as relocation or new baby. As such, I am intrigued by the challenge of modeling this highly inefficient market.

2 Team composition

I (Erin Witmer) will be working solo on the final project. I am a part-time student pursuing a M.S. in Computer Science.

3 Goal of the Analysis

Utilize supervised learning techniques to develop a model for predicting the final sales price of homes in the test data set. Performance will be measured by Root Mean Squared Logarithmic Error.

4 Planned technical approach

4.1 Data visualization

The first step will be to understand the data better through basic visualization.

4.2 Data preprocessing

A cursory review of the data reveals some missing data values, frequent use of NaN indicating features not relevant to the property and inconsistent coding of features. Data preprocessing will be required to clean the dataset.

4.3 Model selection

The baseline model used will be Naive Bayes. More complex modeling will be compared to this as a baseline. Naive Bayes will require additional preprocessing to bin the prediction categories.

A preliminary analysis suggests the most significant challenge with this dataset will be working with the high dimensionality of the data. A preliminary review of relevant literature suggests a model to explore is the Lasso Model. In essence, it is a linear regression model that balances the trade-off between fit of the data and feature proliferation and aids in dimensionality reduction through feature selection. Adaptive Lasso is an enhancement to Lasso which I also plan to explore. In addition to regression models, I plan to compare the results of regression models with those of classification models and explore relevant literature on optimal algorithms for classification in high dimensional space.

In addition to these techniques, I also plan to explore the affect of principal component analysis on results, as a different technique for reducing dimensionality.

From an economic perspective, it seems as though first clustering followed by running these regression/classification analyses may also improve the efficacy of the model. That is because the buyers of homes in the 50-100K range may prioritize different features than buyers of homes in the 500-750K range. So first clustering the data, perhaps with the EM method, will be explored as an extension to the stand alone regression/categorization analyses.

5 Role of team members

I will be performing all of the steps outlined above as well as delivering the final in class presentation.