Latika Bhurani, Siddhesh Chavan, Srishti Jhanwar, John Maxwell, Erik Osland, Rashmi Arora

Abstract

Buying a house is often the most expensive purchase in a person's lifetime. Having an accurate valuation before purchasing can greatly benefit buyers. Companies such as Zillow and Trulia attract users specifically by offering them accurate estimates of housing prices. We aim to survey various machine learning techniques for modeling housing prices.

The Data

We will be working with a sample of housing sales of individual residential properties in Ames, Iowa from 2006 to 2010. There 1,460 training cases and 1,459 test cases with no labels. This data was gathered during a unique time in the US housing market. In 2006 the US started to enter a recession and foreclosure rates increased. In October 2007, the U.S. Secretary of the Treasury called the bursting housing bubble "the most significant risk to our economy". [1] These factors will undoubtable affect the prices of houses in this dataset as well as which houses were able to sell. We will keep this in mind while analyzing the data.

The dataset is primarily comprised of features describing the houses. It contains no information about the buyers. Since there are few variables describing the surrounding area we'd like to explore adding census data if time permits. This dataset is anonymized to protect people's personal information so there is no address or block level identification only a feature labeling 25 distinct neighborhoods. If we are able to add census data we will need to tag census blocks to each neighborhood. It is possible that a neighborhood could be comprised of multiple census blocks with very different geodemographics or a census block could be split between adjacent neighborhoods. In these cases we will need to aggregate and split census data to generate geodemographics at the neighborhood level.

Tasks

Each model will be performing a regression to predict the sales prices of houses in the dataset. We will start our process with exploratory data analysis (EDA) to better understand the data we're working with. Following this we will engineer features derived from the current dataset and if time permits add census data. We will train and fine tune parameters for each model in the survey. The results from these models will be compared on the basis of RMSE (root-mean-square error) as well as what subgroups they misclassify to paint a more thorough picture of their strengths and weaknesses. Finally, we will visualize our findings in a way that succinctly communicates the factors that contribute to home prices.

Techniques

We'll be surveying 4 broad techniques on this dataset. Within each technique we may test different implementations, feature transformations, and tuning parameters to get the best comparison.

- Linear Regression - This is a good baseline for comparing against other techniques due to its simplicity and tendency not to overfit. We may also try similar techniques such as quantile regression.
- Gaussian Process - A probabilistic prediction is valuable because we may not want to provide predictions if we're uncertain of it. Giving no estimate may be more desirable than a highly inaccurate guess in a production system.
- Tree Based Techniques - Ensembles of trees are powerful because they can easily handle both numeric and categorical variables, can expose interactions between variables, and can be more interpretable than some techniques. We may look at random forest, RIPPER (Repeated incremental pruning to produce error reduction), and various boosting techniques as part of this.
- Neural networks - Neural networks are attractive because of their ability to engineer their own features

Tools

We will be working with tools in the python ecosystem for data manipulation, model training, and visualization. Numpy and Pandas are adept at data manipulation and cleaning. Scikit-learn will be used to train linear regression, gaussian process, and tree based models while TensorFlow will be used for neural networks. Matplotlib and seaborn will provide tools for EDA and post-model visualizations.

Acknowledgments

The Ames Housing dataset [2] was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

[1] http://www.amstat.org/publications/jse/v19n3/decock.pdf
[2] https://web.archive.org/web/20100918181527/http://afp.google.com/article/ALeqM5hWSjWmGJ4YXTh3PM5kOC7csTT48g