CSE 847: Machine Learning—Project Proposal

An Exploration and Implementation of Automated Valuation Models to Learn and Predict the Value of Real Estate

Mick Langford langfo37@msu.edu
Mike Lingg linggmic@egr.msu.edu
Jordi Lucero luceroj2@msu.edu

February 17, 2017

1 Problem Description

Automated Valuation Models (AVM) have become increasingly popular as the real estate market has embraced the World Wide Web as a source of accurate, up to the minute data.[3] Banks have also shown great interest in using AVMs to help mitigate fraud by human appraisal.[4] Our goal is to explore various machine learning techniques to implement an AVM and predict the true value of a house based on features commonly found on real estate listings. Our data will be drawn from the Nashville, TN housing market, using a dataset posted on Kaggle[1].

We will begin by exploring linear regression models that take into account physical attributes of each house and location. Further work will be performed exploring nonlinear models, such as deep learning with neural networks and decision trees, which can be compared and contrasted. Additional work may be performed to explore missing feature estimation.

2 Related Work

An obvious and popular example is Zillow's proprietary Zestimate®. Zillow uses a closed source AVM that takes into account special features of the home, location, and market conditions. Zillow admits to using features such as physical attributes, tax assessments, and prior transactions. Zillow claims to have data on 110 million homes and estimates on approximately 100 million homes.[5]

Relevant papers include the doctoral dissertation of Lowrance which explores and compares various linear models on housing data for the Los Angeles County.[6] Park and Bae explore machine learning algorithms such as C4.5, RIPPER, Naive Bayesian, and AdaBoost.[7] Bin performed a study that estimates a hedonic price function using a semiparametric regression.[8] This may be particularly useful for real estate listings that are incomplete or for data that is entered erroneously. Bourassa et al. consider the spatial dependence of house prices, which is intuitively an important factor.[9][10] Kauko et al. research neural network models to help investigate segmentation in the housing market of Helsinki, Finland.[11] Azadeh et al. present an algorithm based on fuzzy linear regression and a fuzzy cognitive map to handle uncertainty in the housing market and improve the analysis of housing price fluctuations.[12] Fan et al. introduce a decision tree approach for modeling and predicting house prices.[13]

3 Project Data

For this project we are working with multiple data sources pulled from real housing sales data.

Nashville Housing Data The Nashville housing data set is a list of home sales in the Nashville, Tennessee area, provided by Kaggle[1]. This data set includes 29 fields of data for 56635 entries. However, nearly half of the entries have gaps in information, which will have to be accounted for. We further augmented this data set by using an geocoding service provided by the United State Census Bureau to add the zip code, latitude, and longitude for entries where a match could be found.

King County Housing Data The King County housing data set is a list of home sales in the King County, Washington area, provided by Kaggle[2]. This data set includes 20 fields of data for 21614 entries, with none of the entries missing any data.

Advanced Regression Techniques Data The

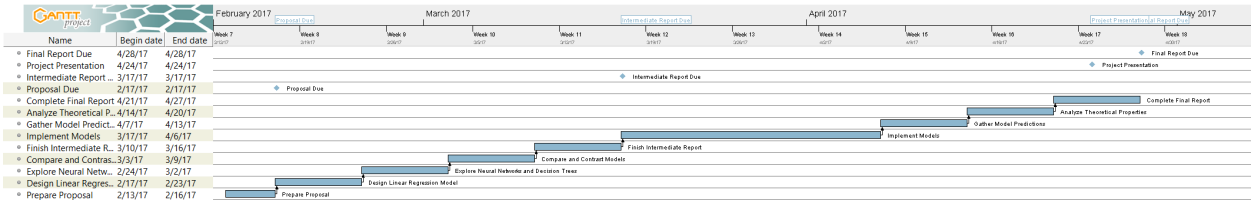

Figure 1: Gantt Chart showing our project milestones and expected completion dates.

Advanced Regression Techniques data is a list of home sales, provided by Kaggle. This data includes 79 features of housing data for 1460 homes. The data has no gaps, except for some N/A data.

Grand Rapids Data The Grand Rapids data is a list of home sales in the Grand Rapids area, provided by the real estate listing service Redfin. This data includes 16 fields for 9646 entries. The data set also has gaps in information for about half of the entries.

The different data sets provide a variety of input to test machine learning techniques against. Data with more features should provide more accurate results. One challenge will be to determine the best way to manage entries with missing information. Two methods are being considered. The first method is to simply substitute missing values with their associated mean values. Another method is to consider the values that are present for each entry, find the nearest neighboring entry, by way of Euclidean distance, and substitute the missing values with that neighbor's values.

For pre-processing the data, we will be normalizing the data set, by dividing each feature value by the difference between that feature's maximum and minimum values. Another problem being considered is how to properly manage fields containing categorical values. Our approach will be to treat a field with n categories as n binary features, indicating whether that entry is of the associated category or not.

4 Project Progress

Figure 1 shows our initial project milestones and timeline to completing them.

4.1 Linear Regression

Initial work has been done on developing working linear regression models. These models include a closed, form ridge regression model, a standard linear descent model and a stoicastic linear descent model. All three models perform roughly the same, with a mean squared error is better than the results of MATLAB's built in fitlm function. One interesting behavior noted is that the highest accuracy is achieved at lower house values, where the highest density of data points is sitting. Additional work is planned to try to eliminate the few large error points, perhaps by seeing if this is caused by a bad feature.

4.2 Logistic Regression

Two working logistic regression models have been created. Both models separate the houses into categories by value. The categories start at the lowest value house and put all homes in the same $ 5k range into a category. The first model creates a set of model weights for each category and places a house in the category producing the highest logistic regression result. The second model creates a a set of model weights that categorize each house price as being higher or lower than the current category, then for each category the house price is higher for, the estimated value is increased by 5k.

4.3 Decision Tree

4.4 Neural Network

Our approach to this problem with a neural network is to treat it as a classification problem. The range of sale prices for the entire data set will be partitioned into n classes, where each class represents an equal number of entries. The network will then be fitted to a training data set, and finally predict each entry in the test data set by fitting it into one of the n classes.

Our initial approach is to use a simple feed-forward network with fully-connected layers. Each data point will have d features extracted from the data. Each feature in the input will have its value send as input to each neuron in the first layer of the network, where each neuron has its own associated

weight and activation function. Each neuron's output from the first layer will be sent as input to each neuron in the second layer, which will also have its own weight values and activation function. Finally, the output from each neuron in the second layer will be sent as input to each neuron in the final layer, again having its own weights and activation function. The final layer will have k outputs, representing which class the data point falls into.

Our current method is to use a rectified linear function as the activation function for each layer, with k neurons in the third layer, $k * 2$ neurons in the second layer, and $k * 4$ neurons in the third layer.

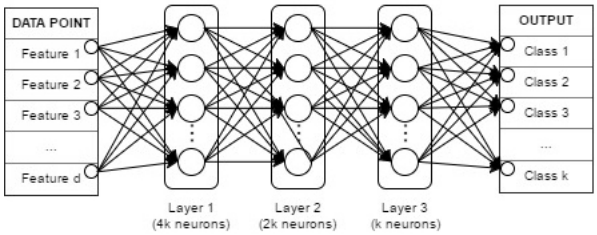

Figure 2: Architecture of our network.

The input data is randomly shuffled and split into a training set and testing set, where the testing data set is 10% of the size of the training set. The model is trained for 50 epochs, using an Adaptive Gradient Descent (Adagrad) algorithm to fit and optimize weights throughout the model, with an initial learning rate $\eta = 0.1$. Techniques are being experimented with to decay the learning rate with relation to the result of the loss function for each epoch.

Displayed in Table 1 and Figures 3 and 4 are the results of running the data sets through the current prototype network. Two different trials were run on each data set, with the first trial learning and predicting against 5 classes of target values, and the second trail against 10 classes of target values. Again each class represents an equally sized partition of the data's target values. A prediction of target class indicates that the given data point's sale price will fall within the boundaries that define that class's partition.

	5 classes		10 classes	
	Train	Test	Train	Test
Nashville	59.02%	59.01%	38.03%	37.58%
King County	52.20%	52.15%	31.36%	30.08%
Grand Rapids	56.45%	52.26%	33.39%	32.85%

Table 1: Observed performance from neural network.

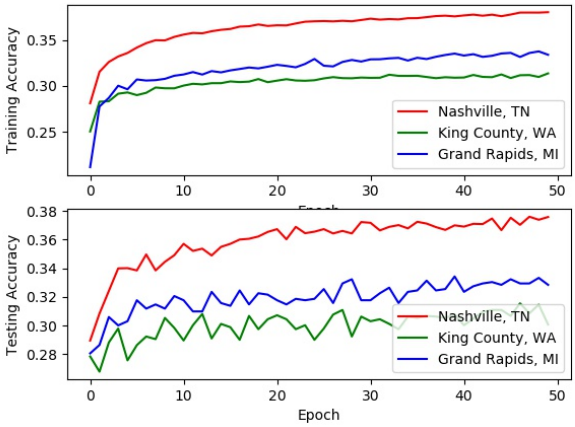

Figure 3: Results for classifying into 5 classes.

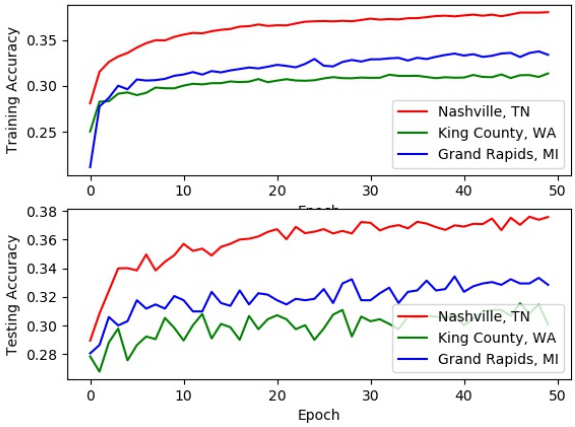

Figure 4: Results for classifying into 10 classes.

As it can be seen from these results, the network performs better when classifying data points into a smaller number of targets, which makes perfect sense. Our goal from this point forward is to attempt to improve accuracy by investigating further improvements on preprocessing the input data, experimenting with different hyper-parameters for the network, and experimenting with varying network architectures. When accuracy improves, the goal with then be to increase the number of target classes in order to provide a more valuable estimation of the sales price of a house.

4.5 Additional Work

In addition to the existing models having been developed, we are considering creating a Support Vector Machine model to categorize house prices by features. ?? Model compare and contrast?

References

[1] *Nashville Housing Data: Home value data for the booming Nashville Market* Retrieved from `https://www.kaggle.com/tmthyjames/nashville-housing-data/`

[2] *House Sales in King County, USA: Predict house price using regression* Retrieved from `https://www.kaggle.com/harlfoxem/housesalesprediction`

[3] *Data-driven property valuations: the real deal?* Retrieved from `http://blog.kaggle.com/2010/06/21/data-inc-are-avms-soothsayers-or-the-real-deal/`

[4] Schroeder, Steve. *Fighting Fraud: A combination of collateral assessment and AVMs can maximize mortgage-fraud management* Retrieved from `http://www.scotsmanguide.com/Residential/Articles/2005/10/Fighting-Fraud/`

[5] *What is a Zestimate? Zillow's Home Value Forecast.* Retrieved from `http://www.zillow.com/zestimate/`

[6] Lowrance, R. E. (2015). *Predicting the Market Value of Single-Family Residential Real Estate* (Doctoral Dissertation). New York University. Retrieved from `http://gradworks.umi.com/36/85/3685886.html`

[7] Park, B., & Bae, J. K. (2015). *Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data.* Expert Systems with Applications, 42(6), 2928-2934. doi:10.1016/j.eswa.2014.11.040

[8] Bin, O. (2004). *A prediction comparison of housing sales prices by parametric versus semi-parametric regressions.* Journal of Housing Economics, 13(1), 68-84. doi:10.1016/j.jhe.2004.01.001

[9] Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). *Predicting House Prices with Spatial Dependence: Impacts of Alternative Submarket Definitions.* SSRN Electronic Journal. doi:10.2139/ssrn.1090147

[10] Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). *Spatial Dependence, Housing Submarkets, and House Prices.* SSRN Electronic Journal. doi:10.2139/ssrn.771867

[11] Kauko, T., Hooimeijer, P., & Hakfoort, J. (2002). *Capturing Housing Market Segmentation: An Alternative Approach based on Neural Network Modelling.* Housing Studies, 17(6), 875-894. doi:10.1080/02673030215999

[12] Azadeh, A., Ziaei, B., & Moghaddam, M. (2012). *A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations.* Expert Systems with Applications, 39(1), 298-315. doi:10.1016/j.eswa.2011.07.020

[13] Fan, G., Ong, S. E., & Koh, H. C. (2006). *Determinants of House Price: A Decision Tree Approach.* Urban Studies, 43(12), 2301-2315. doi:10.1080/00420980600990928