Team 110 Project Proposal: House Price Prediction in Los Angeles

Team 110: Chen Zhang, Kai Ni, Shaojuan Liao, Shengchen Liu, Zheng Kuang, Jinjun Liu
Georgia Institute of Technology
Atlanta, USA
{czhang613,knimr3,sliao33,sliu651,zkuang30,jliu788}@gatech.edu

ABSTRACT

An iterative map is designed to help house-buyers buy a perfect house in Los Angeles

KEYWORDS

house price, machine learning, D3, data mining

ACM Reference Format:
Team 110: Chen Zhang, Kai Ni, Shaojuan Liao, Shengchen Liu, Zheng Kuang, Jinjun Liu. 2019. Team 110 Project Proposal: House Price Prediction in Los Angeles. In *Project proposal*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

House purchase is a big decision in most people's life. A good housing price prediction model that can integrate multiple factors is required for both house buyers and sellers when making an important financial decision [Banerjee and Dutta 2017].

In this project, we aim at developing an accurate house price prediction model in Los Angeles area with integration of multiple community/environmental data and local economic indicators. We will use various machine learning algorithms starting with basic regression techniques and move forward to advanced machine learning algorithms such as ensemble learning and deep learning to help improve the prediction. The results will be presented in the form of visually interactive map.

Project Proposal, 2019 Spring,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 EXISTING SOLUTIONS

2.1 Real estate websites

Currently real estate websites such as Zillow, Craigslist, Redfin, and Trulia have provided detailed physical features and historical transaction of a property. However, there is still absence of information, such as the neighborhood quality, the school information, crime rate, etc. Although they have been working on house price prediction methods extensively, the accuracy is not good enough. Zillow claims their housing price prediction algorithm, 'Zestimate', only estimates about 50 % of houses within the 5 % of their selling prices [Zillow 2014]. For Trulia, only 48.2 % of houses have Trulia-estimated prices to be within the 5 % range of their actual sold prices [Trulia. 2014].

2.2 Academia

Research groups also work on prediction models. These models are important precedences, although most of the models only consider the physical features.

2.2.1 Data preprocessing. Data preprocessing would have a significant impact on performance in this

project. Kotsiantis [Kotsiantis et al. 2006] reviewed multiple aspects of data preprocessing, including data cleaning, normalization, transformation, feature extraction and selection. Dozens of methods were discussed to handle missing values, outliers, redundant and interdependent features, and imbalanced sample representation.

Dealing with textual attributes is necessary for analysis of neighborhood quality. Conventional algorithms like support vector machine [Ahmed and Moustafa 2016] and random forest [Liaw et al. 2002] works well in regression and classification based on textual attributes of the houses [Khamis and Kamarudin 2014; Ng and Deisenroth 2015; Park and Bae 2015]. The concepts and methods will certainly provide guidance in our analysis.

2.2.2 Comparison of machine learning methods. Deep and shallow neural networks are distinguished by the depth of their credit assignment paths, which are chains of possibly learnable, causal links between actions and effects [Schmidhuber 2015]. Deep Learning in Neural Networks can be generally classified into three types: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. In terms of house price analysis, Hedonic house price equation is widely used nowadays for analyzing house prices, especially for spatial autocorrelation in transaction prices [Basu and Thibodeau 1998]. Lu [Lu et al. 2014] proposed a modified model called "grey relational analysis" for predicting house prices in Taiwan market. This method calculates the weighted synthesis of the top ten matching instances through various weighting strategies. This method outperformed the instance-based approach such as KNN.

Modern research leverages the development of multilayer deep neural networks by combining the textual and visual attributes and increase the accuracy of prediction [Ahmed and Moustafa 2016] in terms of MSE and R-squared. By utilizing this approach on large scale of image datasets, Simonyan et al. [Simonyan and Zisserman 2014] shows that the prediction accuracy will improve as we increase the depth of the convolutional neural network. Sirignano et al.[Sirignano et al. 2016] predicts the mortgage delinquency transition rate using deep neural networks with the optimal as 5 hidden layers and each with 140-200 nodes. Such model is good at modeling non-linear effects. Limsombunchao [Limsombunchai 2004]compares the traditional econometric method of hedonic price model and the artificial neural network (ANN) model. ANN yields better prediction especially when there is a structural change, but has 'black box' problem. Bourassa et al. [Bourassa et al. 2010] considers the spatial effect using different models such as neighbors residuals as second stage regression, geostatistical model and trend surface model. Empirical results conclude that a geostatistical model with disaggregated submarket variables performs best.

Support vector machine (SVM) has also been shown as an efficient method. Mu et al. [Mu et al. 2014]compared three methods: SVM, least squares support vector machine (LSSVM) and partial least squares (PLS) on the housing data of Boston. SVM outperformed the others in terms of accuracy and running time. Phan et al. [Phan 2018]extended the analysis on housing data of Melbourne City by comparing 5 methods including linear and polynomial regressions, regression tree, neural network and SVM. Similarly SVM performed better than the others. He also demonstrated that data preprocessing, including removing missing data and outliers, transforming data such as 'log' and PCA, and reducing data by stepwise and boosting techniques, significantly influenced prediction accuracy.

3 PROJECT FEATURES

This project will ideally provide an accurate housing price prediction model based on multiple factors of the property, and an interactive visualization product could provide a user-friendly platform. The following attributes in Table1 will be considered (not limited to) :

Table 1: Project Features

Category	Features
Physical	Area, No. of beds/baths, Style, Open space, Transaction history, Tax rate
School-related	Average score, Test score, Ethnicity, Teacher/student ratio, Low-income ratio
Environmental	Incident reported, Drainage, Infestation
Community-related	Crime rate, Household income, Highest education, Ethnic structure, Proximity to grocery/retail, Proximity to public transport, Proximity to health care

4 PROJECT PLAN

The general plan of this project can be generalized into four steps: data acquisition, feature engineering, house price prediction and data visualization.

4.1 Data Acquisition

Data of this project comes from three resources:

- Kaggle Zillow Home Value Prediction:
  This data set provides a full list of real estate properties in LA area. Transactions are given as ground truth label for the prediction. The data is big which includes about 6 million housing transaction records with more than 50 variables such as detailed house features, location, and historic transaction and so on.
- Craigslist
  A web crawling program will be implemented to scrape information from Craigslist using available API.
- Local information in the neighborhood
  Information in the neighborhood are available through city-data.com [Advameg 2014].

4.2 Feature Engineering

Feature engineering is the crucial step. The performance of the model heavily depends on the quality of the data set. Some commonly used processing techniques covers filling missing values, converting numerical features to categorical features, normalization and transformation, etc.

4.3 House Price Prediction

Various machine learning models will be investigated for predicting house prices. Regression will be used as baseline models. More advanced models such as random forest and neural networks will be implemented.

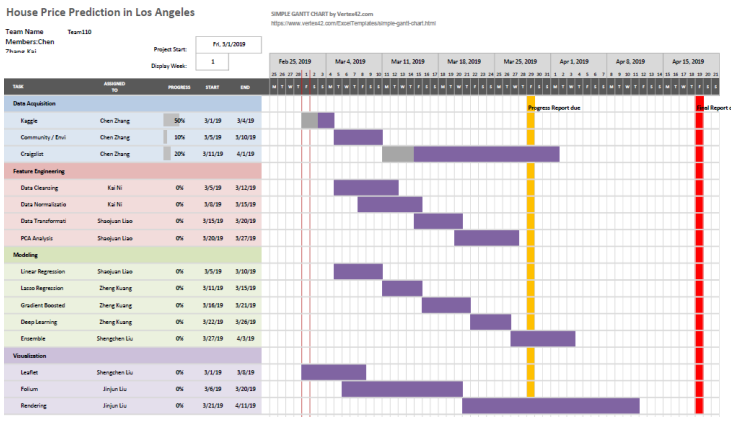
4.4 Data Visualization

An interactive map will be implemented using map libraries such as Leaflet and Folium. The map will contain features such as filtering, sorting, search and zooming.

5 PLAN OF ACTIVITIES

The progress will be measured as two phases for examination: midterm and final.The detailed plan of activities is in the following Gantt Chart 1:

Figure 1: Plan of activities in Gantt chart



6 EVALUATION METRICS

Data will be randomly splitted: 80 % for training, and the rest for validation. Mean Absolute Error **(MAE)** will be used to evaluate the performance. We will compare the predicted price with the true price from the evaluation data set. The log error is defined as:

$$logerror = \log(Estimate) - \log(SalePrice) \quad (1)$$

Mean Absolute Error (MAE) is defined as :

$$MAE = \sum_{i=1}^{n} \mid y_i - x_i \mid /n \quad (2)$$

7 RISKS AND PAYOFFS

Data scraped from web needs parsing and cleansing. Many records have missing values so the quality of the data set may be affected. Also, deep learning models may result in overfitting or underfitting. So the performance may not be as significant as expected.

8 EXPENSES AND COSTS

All the data sets are free to obtain. the expenses and costs will be contributed to the use of computation resources in AWS for model training. Amazon EC2 P2 Instances provides NVIDIA Tesla K80 GPU which has 12GB of GPU memory. The price rate is $0.9/hr according to the description on AWS website [Amazon 2014] . We will apply for Student Credits to utilize the computing instances with overall expense around $300.

REFERENCES

Inc. Advameg. 2014. City Data for Los Angeles. (March 2014). Retrieved March 3, 2019 from http://www.city-data.com/city/Los-Angeles-California.html

Eman Ahmed and Mohamed Moustafa. 2016. House price estimation from visual and textual features. *arXiv preprint arXiv:1609.08399* (2016).

Amazon. 2014. Amazon Webservice. (March 2014). Retrieved March 3, 2019 from https://aws.amazon.com/ec2/instance-types/p2/

Debanjan Banerjee and Suchibrota Dutta. 2017. Predicting the housing price direction using machine learning techniques. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, 2998–3000.

Sabyasachi Basu and Thomas G Thibodeau. 1998. Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics* 17, 1 (1998), 61–85.

Steven Bourassa, Eva Cantoni, and Martin Hoesli. 2010. Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of Real Estate Research* 32, 2 (2010), 139–159.

Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin. 2014. Comparative study on estimate house price using statistical and neural network model. *International Journal of Scientific & Technology Research* 3, 12 (2014), 126–131.

SB Kotsiantis, Dimitris Kanellopoulos, and PE Pintelas. 2006. Data preprocessing for supervised leaning. *International Journal of Computer Science* 1, 2 (2006), 111–117.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

Visit Limsombunchai. 2004. House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference*. 25–26.

Binbin Lu, Martin Charlton, Paul Harris, and A Stewart Fotheringham. 2014. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science* 28, 4 (2014), 660–681.

Jingyi Mu, Fang Wu, and Aihua Zhang. 2014. Housing value forecasting based on machine learning methods. In *Abstract and Applied Analysis*, Vol. 2014. Hindawi.

Aaron Ng and Marc Deisenroth. 2015. Machine learning for a London housing price prediction mobile application. *Final Project, Department of Computing, Imperial College London* (2015).

Byeonghwa Park and Jae Kwon Bae. 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications* 42, 6 (2015), 2928–2934.

The Danh Phan. 2018. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. In *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*. IEEE, 35–42.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. 2016. Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470* (2016).

Trulia. 2014. Trulia Estimates. (March 2014). Retrieved March 3, 2019 from https://www.trulia.com/info/trulia-estimates/

Inc. Zillow. 2014. Zestimate. (March 2014). Retrieved March 3, 2019 from https://www.zillow.com/zestimate/