Team 110 Final Report: 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 interative map is designed to help house-buyers buy a perfect house in Los Angeles

KEYWORDS

house price, machine learning, Leaflet.js, data mining

ACM Reference Format:
Team 110: Chen Zhang, Kai Ni, Shaojuan Liao, Shengchen Liu, Zheng Kuang, Jinjun Liu. 2019. Team 110 Final Report: House Price Prediction in Los Angeles. In *Final Report*. ACM, New York, NY, USA, 10 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 successfully developed an accurate house price prediction model in Los Angeles area with the integration of multiple community/environmental data and local economic indicators. We used 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. We found that the prediction improved upon incorporation of neighborhood data. The results are presented in the form of a visually interactive map.

Final Report, 2019 Spring,

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

2 SURVEY

2.1 Real estate websites

Currently real estate websites have provided detailed physical features and historical transaction of a property. However, there is still absence of information, including neighborhood quality, school information, crime rate, etc. Also the accuracy isn't good enough. Zillow's housing price prediction algorithm 'Zestimate', only estimates about 50 % of houses within the 5 % of their selling prices [Zillow 2014].

2.2 Academia

Research groups also work on prediction models. They are important precedences, although most of them only consider 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.

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]. Hedonic house price equation is used for analyzing house prices, especially for spatial auto-correlation 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.

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] showed that the prediction accuracy will improve as we increase the depth of the convolutional neural network. Sirignano et al.[Sirignano et al. 2016] predicted 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]compared 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 a 'black box' problem. Bourassa et al. [Bourassa et al. 2010] considered 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]compared 5 methods including linear and polynomial regressions, regression tree, neural network and SVM. SVM outperformed the others. Additionally data preprocessing, including removing missing data and outliers, transforming and reducing data, significantly influenced prediction accuracy.

3 EXPERIMENTS

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

3.1 Data Acquisition

As mentioned in our proposal, we need physical, community-related, school-related and environmental features of a property for price prediction.

3.1.1 Physical features of properties. The physical features are provided by Kaggle in Zillow competition [Kaggle 2014]. One dataset contains physical features of all properties at LA, Ventura and Orange county with 58 columns such as number of bedrooms, bathrooms, total area, built year, etc. The other dataset contains all properties that have transactions between Oct 2016 and Apr 2017.

However, the data do not provide the detailed address of a property, so we decided to use reverse geocoding API which takes coordinates (latitude and longitude) and return addresses and zip code [Google 2019]. 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.

One thing we noticed from the original Kaggle dataset is that, the zipcode provided is not accurate. Sometimes the zipcode is not existing. Again, we decided to use reverse geocoding generated addresses and zipcode. These zipcodes are important since our future data merge with neighborhood data is mostly relying on zipcode.

3.1.2 Local information in the neighborhood. One important consideration for house buyers is the quality of assigned schools for a property. Houses within good school district become more attractive and likely to be sold in higher price. School-related features were harvested from school digger API [schooldigger 2019]. We first collected all school data by searching schools in zip codes. In total 2186 public, charter and private schools are identified. Based on the school list, detail information of each school such as the school type, school level, rating (provided by school digger website), enrollment, student/teacher ratio, free lunch ratio and ethnic structure.

Community-related features will certainly impact the quality of a neighborhood. We did data crawling on bestplaces.com [bestplaces 2019], a website that provides neighborhood information by zip code. All the following features have been successfully extracted using html parsing through beautifulsoup: crime index, population, average age, ethnic structure, average household income, commute time. Note that the crime index is a numeric value provide by bestplaces after incorporation of violent crime and property crime data released from police department. Proximity to health care resources is also a crucial feature of a property. The hospital data was directly downloaded and filtered from medicare.gov [medicare 2019].

We tried to look into environmental damages reported in certain area. However, limited data source can be found. Also, available environmental damages are reported by county rather than zip code. This will be less interesting to us, since we expect the zipcode-level environmental data for prediction. Thus, we decided not to incorporate the environmental features.

All data collection is done by python. The codes are available on GitHub [github 2019].

3.2 Feature Engineering

Data pre-processing and feature engineering are critical in machine learning in terms of: reduce dimensions and strengthen relationships between features and target attributes.

The goal is to remove irrelevant, redundant, noisy and unreliable data. In this project, we focus on (1) removing features containing excessive null values and outliers; removing redundant/interdependent features; categorizing features which by themselves are meaningless.

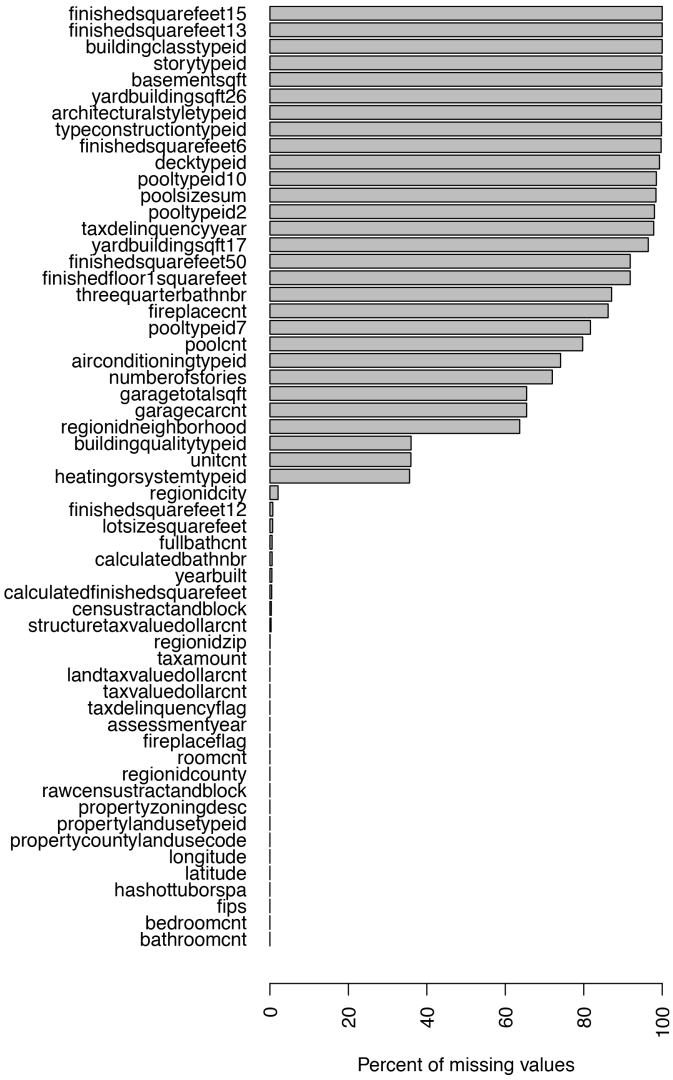
3.2.1 Remove irrelevant data. During the data mining process, we noticed that building types such as apartments, factory, and warehouse would dramatically obscure the relationship between predictive variables and target features. Therefore in this project, we focused on single houses by selecting property type id = 261/279. This reduced data size by 18 %.

A few features were very useful for visualization but less meaningful for machine learning. These include latitude, longitude, assessmentyear and we removed these features.

3.2.2 Remove features containing excessive deviating values. By counting the ratio of entities that contain missing values for each feature, we identified 26 out of 57 features where more than 50 % of entities are missing as shown in Figure 1. Removing these features significantly reduced the dimension of our dataset. Three features, buildingqualitytypeid, unitcnt and heatingorsystemtypeid have 1/3 missing values. Buildingqualitytypeid and heatingorsystemtypeid has 5 and 4 major classes each so that we categorized these two features and included missing values as another class. Unitcnt has

only one dominant class = "1" so we removed this feature too. The other features typically have much less than 1 % missing values. We simply removed the entities which contained these missing values.

Figure 1: Percent of Missing Value



Propertylandusetypeid only has one value 261, so we removed this feature.

3.2.3 Remove redundant / interdependent features. We found pairs of features that were highly redundant or interdependent. We only maintained one feature shown by in the column 'Feature 1' in Table1

3.2.4 Other features. Propertycountylandusecode, Propertyzoningdesc, Rawcensustractandblock, censustractandblock and Regionidneighborhood highly depended on zipcode. Therefore we only kept zipcode (Regionidzip).

3.3 House Price Prediction Model Testing

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

3.3.1 Evaluation. To evaluate the performance of the algorithm, we firstly used a small sub-dataset which has 60,000 instances to do the training and testing. We divided the dataset into 80 % as training and 20 % as testing.

Two ways of evaluation are Mean Absolute Error **(MAE)** and house value. We compared 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 :

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

Meanwhile we also evaluated our performance by house value. Although the house value data were not provided in the original dataset, we harvested the house value through zillow API with scrapped addresses.

3.3.2 Decision Tree Algorithm. We implemented a decision tree algorithm to predict the house prices. As described above, after data cleaning and dimension reduction, we only kept 22 attributes out of 57 original attributes from Kaggle dataset. However, the Kaggle dataset only provides the log error of the house prices, which is defined as the logarithmic difference between the actual transaction price and the prices predicted by Zillow Zestimate. Thus,

Table 1: Feature Correlations

Correlation	Feature1	Feature2	Feature3	Feature4
1	bathroomcnt	calculatedbathnbr	fullbathcnt	
1	calculatedfinishedsquarefeet	finishedsquarefeet12		
1	regionidzip	regionidcounty	regionidcounty	fips
0.95	taxvaluedollarcnt	structuretaxvaluedollarcnt	landtaxvaluedollarcnt	taxamount

we used the Zillow API [Zillow 2014] to get the last sold price after reverse geocoding. The zillow API also returns the last sold date. We extracted the year from sold date as one of the attributes of the algorithm.

The sklearn package in Python is used to implement the decision tree algorithm. 'Entropy' was chosen as the criterion. All the attributes were normalized before using. To avoid overfitting problem, we set the maximum depth of the tree according to the Figure 2.

With the max depth of decision tree increases, the mean absolute error firstly decreases and then increases. It seems that the Max Depth = 8 is the best choice.

We choose Max Depth = 8 and then investigated the probability density function (PDF) of the logerror shown as Figure 3. We can see that the distribution concentrated at logerror = 0. However, the logerror is still a bit large. This is a preliminary testing of machine learning models prior to incorporation of neighborhood dataset. We believed that the logerror could be reduced upon considering neighborhood data.

Figure 2: MAE at different max depth of decision tree

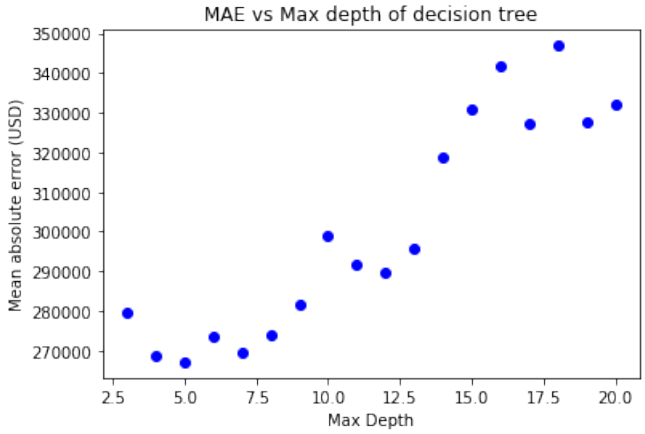
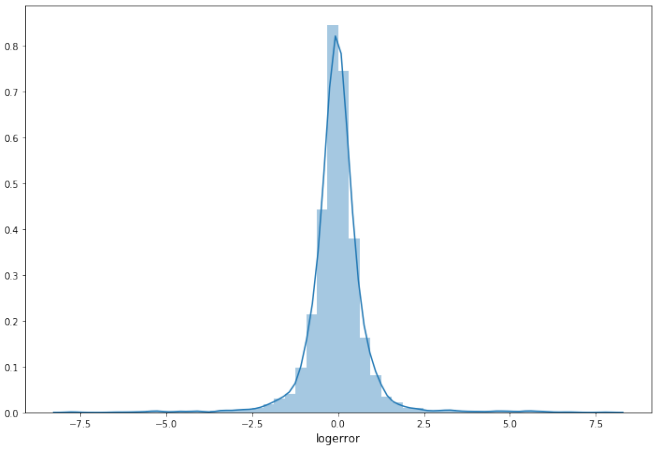


Figure 3: Logerror distribution at Max Depth = 8



3.4 House Price Prediction using neighborhood information

As aforementioned, we hypothesized that house price not only depends on intrinsic characteristics such as year built, square feet, number of bedrooms and bathrooms, but also depends on the environment where it is located. Zipcode largely determines the neighborhood but it is meaningless by itself.

To test this hypothesis, we compared the prediction performance using the original data, which contains mainly intrinsic features, and the expanded

data with neighborhood information included, which consists of traffic, crime, population, school and hospital. Six regression models were used here, linear regression, ridge regression, lasso regression, support vector regression, random forest, and a naive ensemble method which integrates the predictions by the first five methods. Two kernels were used for support vector regression, "linear" and "rbf". Except "linear regression", all other models were optimized with gridsearch and cross validation. Two metrics were used, mean squared error of predicted house price v.s. true house price, and score which is the coefficient of determination R^2 of the prediction. The logerror distribution and house value distribution is shown below in Figure 4 and Figure 5

Figure 4: Log Error Distribution

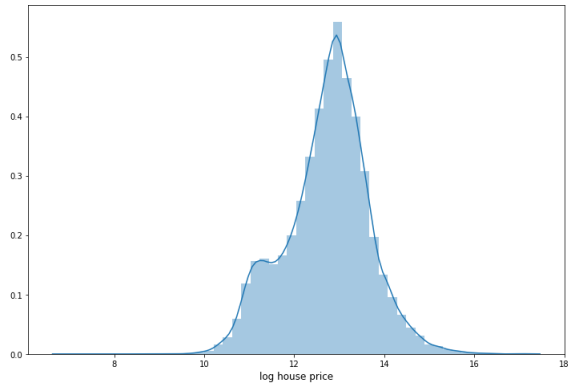
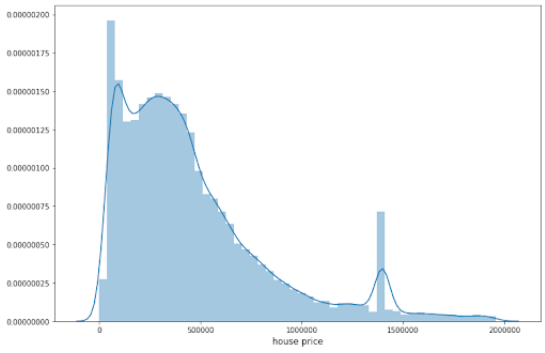
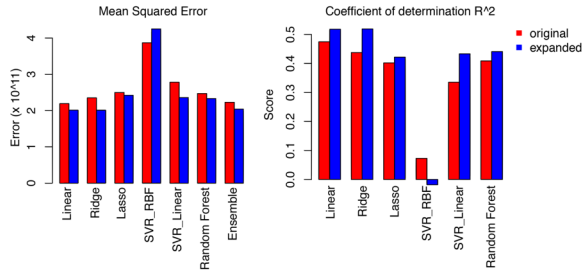


Figure 5: House Value Distribution



The comparison of calculated mean squared error and coefficient determination of R^2 of Linear, Rigde, Lasso, SVR_RBF, SVR_linear and Random Forest is shown in Figure 6. Linear regression and ridge regression performed the best given their lowest errors and highest scores. The performance of Lasso regression, Ensemble and SVR with linear kernel followed. SVR with RBF kernel performed the worst. By comparing the two SVR models, we can clearly see that the radial basis function kernel, RBF, is not very suitable for our dataset while the linear kernel works much better. More interestingly, for all these linear-based algorithms, the expanded dataset worked better than the original dataset given the lower errors and higher scores. This provided a strong evidence to support our primary hypothesis in this paper that including environmental information will improve house price prediction. In addition, Linear regression worked better than Ridge regression when training on the house intrinsic features. However when we expanded with environmental data, Ridge outperformed Linear regression.

Figure 6: Comparison of calculated mean squared error(left) and coefficient determination of R^2(right)



We next analyzed the importance of features for prediction, see the Figures 7 and 8. Within the intrinsic features from the original data, square feet, bathroom and bedroom counts, house quality, lot size, year built are the most important features. When we incorporated neighborhood information, some of them became the top ten important features, including population composition, household income and student composition. We believe these make a lot of sense in house price prediction.

It is in agreement with observations that the ethnic structure, school quality and economic level within a neighborhood can certainly impact the house price.

Figure 7: Important features using original data

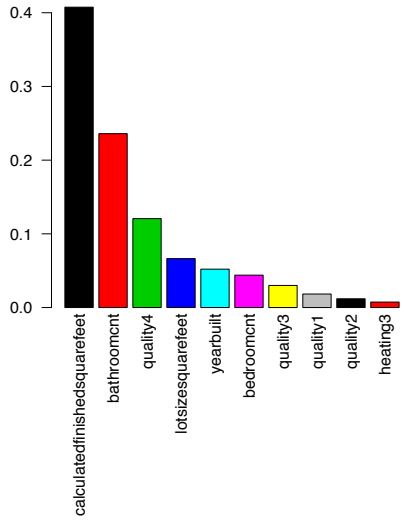
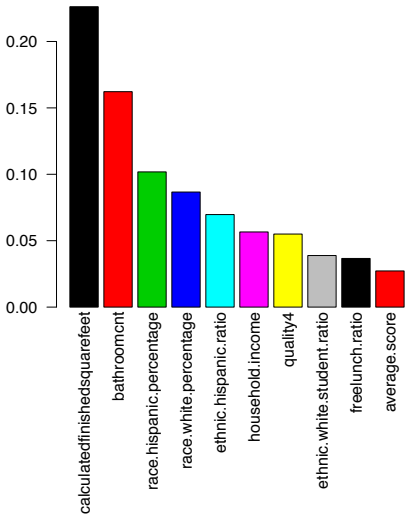


Figure 8: Important features using expanded data



3.5 Data Visualization

An interactive map is implemented using Leaflet, an open-source JavaScript library for mobile-friendly interactive maps. Leaflet is designed with implicity, performance and usability in mind. It works efficiently across all major desktop and mobile platforms [leaflet 2019].

Basic usages of Leaflet has been investigated by group members to build an interactive map. In this project we achieved the following features:

- Reading data from preprocessed files. The loaded file are our house price prediction results including street address, real price, predicted price and 3 key features of a property
- Showing results on map with the area shape highlighted when hovering mouse over
- Adding pops on map layers
- Displaying information of geolocation is displayed when a user click on an object
- Changing zoom levels can display groups of house information interactively in color-coded circles.
- Layer groups and controls. Users can select 'Satellite layer' shown in Figure 9, or 'Streets layer' shown in Figure 10 as background. Optional layers such as 'houses' can be selected to provide the information that users are willing to know. The map website is located at https://shengchen-liu.github.io/portfolio/Map/los_angeles.html.

Figure 9: Satellite layer

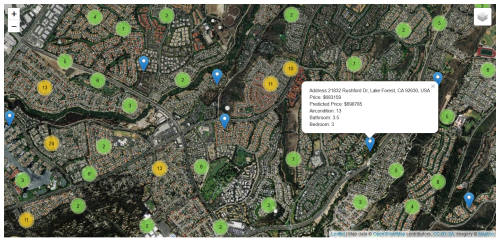
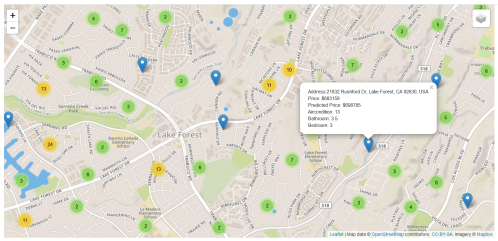


Figure 10: Street layer



Making the map compatible with mobile devices. Using user's geolocation, the map can set its initial view automatically at the center of user's location as shown in Figure 11. After clicking each spot or zoom in manually, an individual house information is shown in Figure 12. Street address, real price, predicted price and 3 key features of a property are displayed upon a mouse click.

Figure 11: Overview of the interactive map

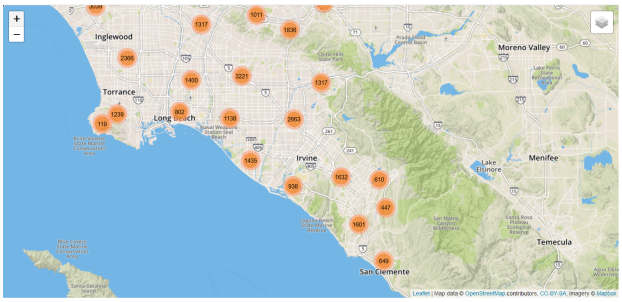
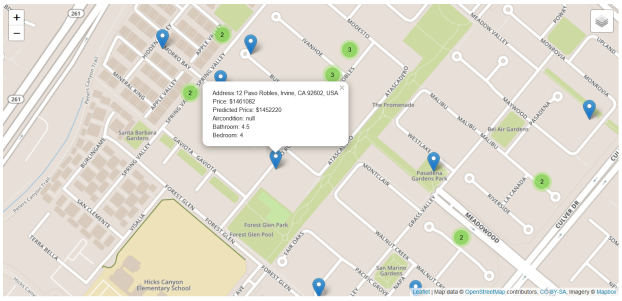


Figure 12: Overview of the interactive map

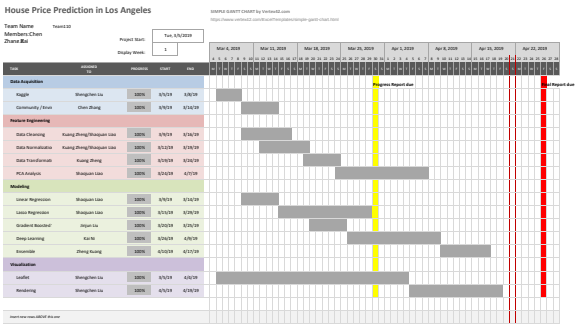


4 PROJECT ACTIVITY AND TEAM WORK

In our group we split the job based on personal experience and expertise. Table 2 briefly describes dedication of each member:

We work as a team in the past 8 weeks on this project. All problems can be solved by group discussion in a timely manner. The detailed timeline is in the following Gantt Chart Figure 13:

Figure 13: Timeline and activities in Gantt chart



5 CONCLUSION

In this project, we developed an accurate house price prediction model with integration of multiple community/school data, and visualized the result on an interactive map. The model of house value prediction was improved by considering neighborhood information. This information can be useful for both house buyer and seller to gather information and estimation of a property prior to bid/transaction. The innovations of our ideas include:

- Scraping information of the local neighborhood and combine them with physical features of properties
- Analyzing feature importance and remove irrelevant data
- Conducting house price prediction using machine learning models.
- Identifying important features for house value prediction
- Improving prediction value model with incorporation of neighborhood data
- Building an interactive map using Leaflet.js to achieve layer controls and mobile compatibility

Table 2: Team Work

Member	GTID	Contribution
Chen Zhang	czhang613	scrape house addresses and locations through reverse geocoding, crawl community and school data using beatifulsoup, organize team meeting and draft report skeleton
Kai Ni	knimr3	Data processing and modeling scraping house price value Initial model testing
Shaojuan Liao	sliao33	Data processing and feature engineering filter redundant and interdependent features, remove missing values/features Initial model testing
Shengchen Liu	sliu651	Initial idea and big picture guidance Kaggle data collection Building up Leaflet interactive map
Zheng Kuang	zkuang30	Data preprocessing, feature engineering and modeling filter redundant and interdependent features, remove missing values/features one-hot encoding, integrate data from different sources, Linear regression, Ridge, Lasso, SVR, Random Forest, Ensemble, Compare original and expanded data, draft Readme
Jinjun Liu	jliu788	Data processing and modeling (decision tree), Data filtering and data cleaning draft poster

REFERENCES

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

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.

bestplaces. 2019. Best Places API. (March 2019). Retrieved March 30, 2019 from https://www.bestplaces.net/

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.

github. 2019. GitHub. (March 2019). Retrieved March 30, 2019 from https://github.com/cse6242-team110-spring2019/CSE6242_Project/tree/master/code/scraper

Google. 2019. Google Maps Platform: Geocoding API. (March 2019). Retrieved March 3, 2019 from https://developers.google.com/maps/documentation/geocoding/start

Kaggle. 2014. Kaggle Competition: Zillow Price. (March 2014). Retrieved March 3, 2019 from https://www.kaggle.com/c/zillow-prize-1

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.

leaflet. 2019. leaflet.js. (March 2019). Retrieved March 30, 2019 from https://leafletjs.com/index.html

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.

medicare. 2019. Medicare. (March 2019). Retrieved March 30, 2019 from https://data.medicare.gov/

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.

schooldigger. 2019. School digger API. (March 2019). Retrieved March 30, 2019 from https://developer.schooldigger.com/?r=schooldiggerexceldialog

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).

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