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

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 2017). Although there have been existing methods for house price prediction, the accuracy isn’t good enough. Besides, most prediction models only adapt physical features of a property, leaving important features unconsidered, such as neighborhood quality, school information, crime rate, etc. 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 to make the prediction. The results will be presented in the form of visually interactive map.

Data

- **Physical properties of houses**: number of bedrooms, bathrooms, total area, built year, etc.
- **School-related**: school type, school level, rating, enrollment, student/teacher ratio, free lunch ratio and ethnic structure
- **Community-related**: crime index, population, average age, ethnic structure, average household income, commute time, hospital data

Physical properties of houses data are collected from Kaggle Zillow competition (~6 million records), school-related data and community-related data are scrapped from website by using API. The above data were merged using ZIP code.

Data pre-processing and feature engineering

- **Remove irrelevant data**
- **Remove features containing excessive deviating values**
- **Remove redundant / interdependent features**

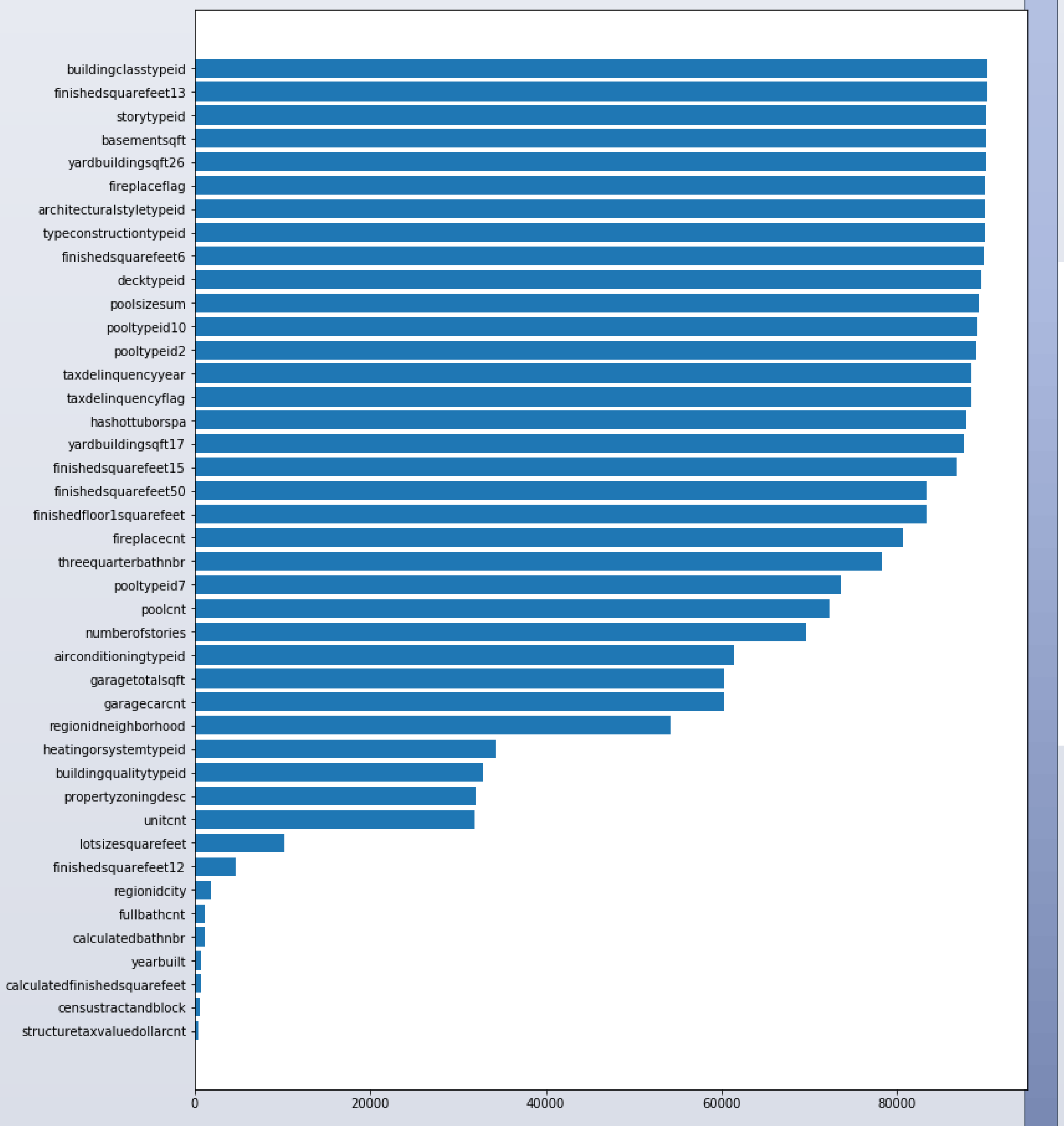

Figure 1. percent of missing value of each attribute

Prediction modeling

We used a small sub-dataset which has 60,000 instances to do the training and testing. Mean Absolute Error (*MAE*) and *logerror* were used to evaluate the results.

$$MAE = \sum_{i=1}^{n} |EstimatePrice_i - SalePrice_i|/n$$

$$logerror = \log(EstimatePrice) - \log(SalePrice)$$

Decision Tree:

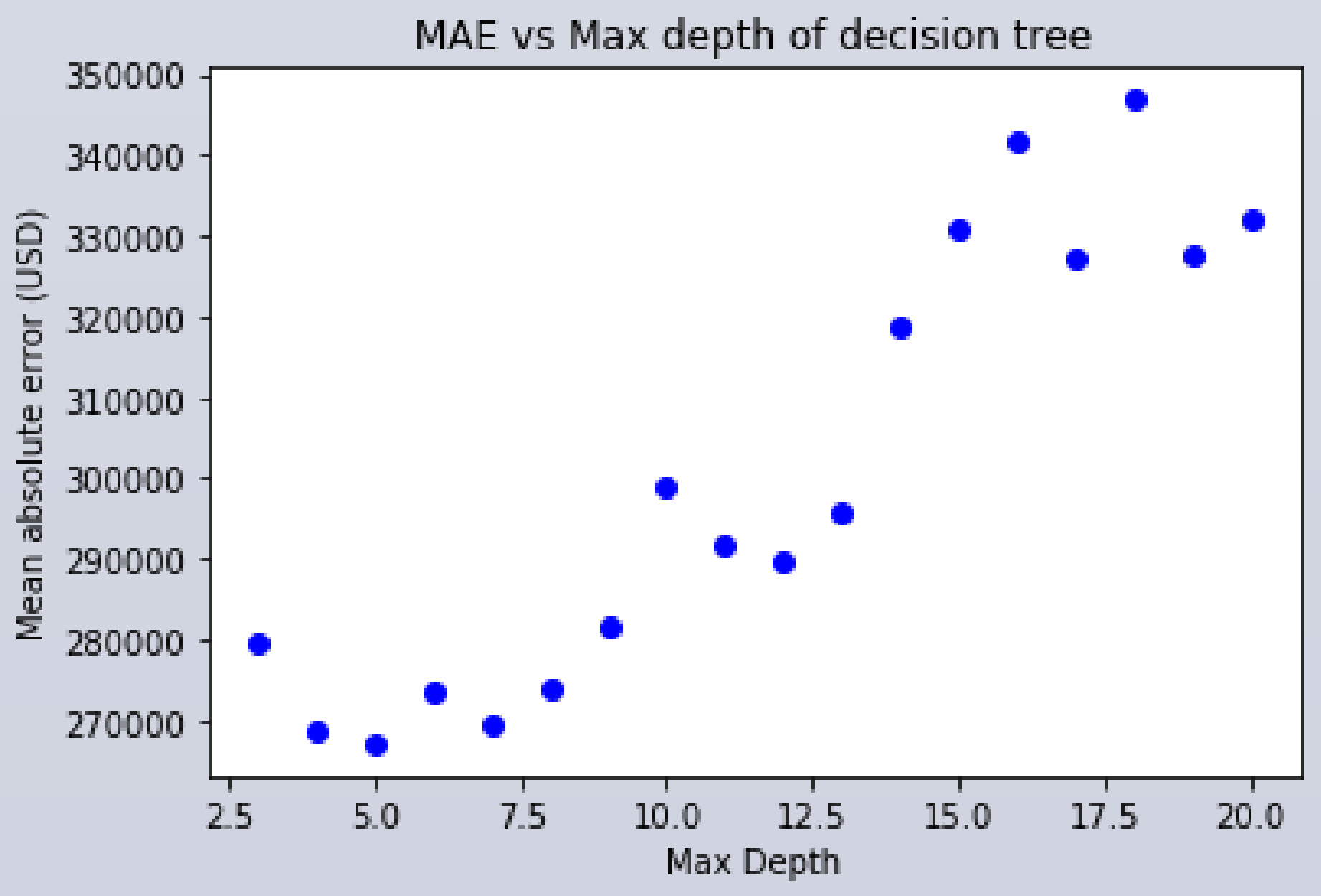

Figure 2. The MAE at different max depth of decision tree

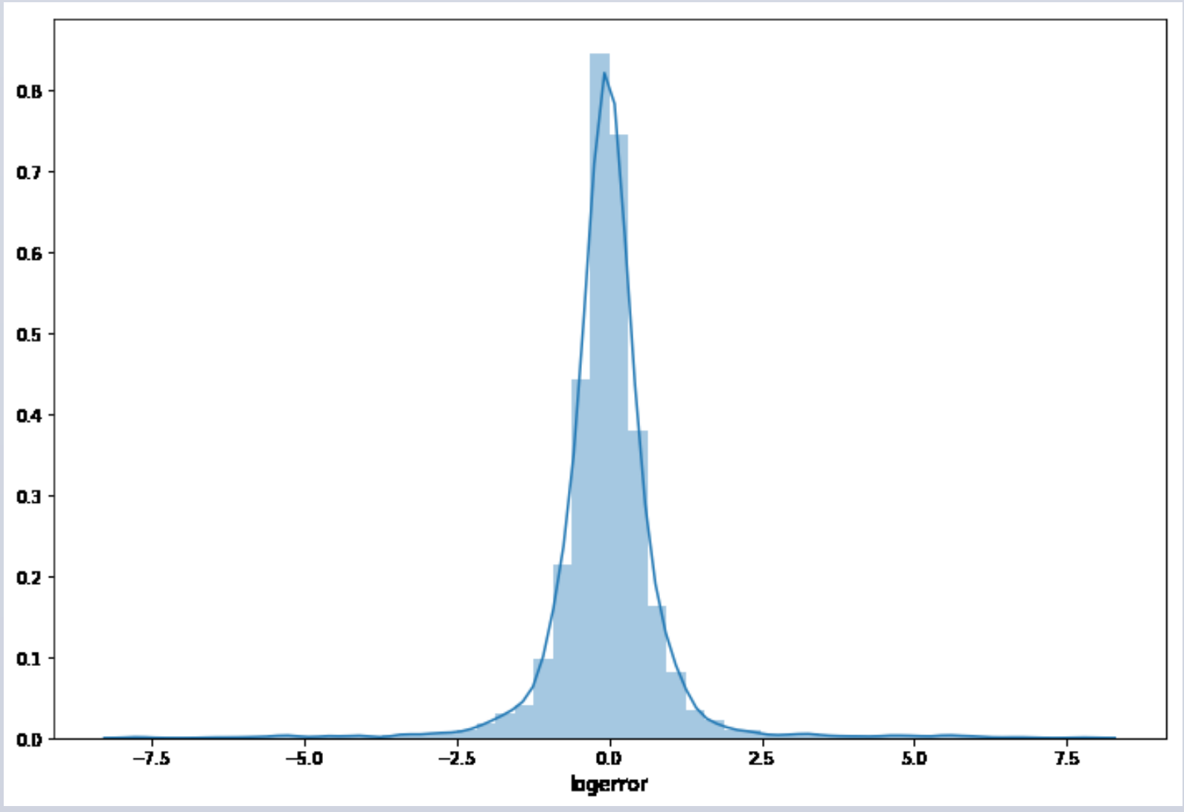

Figure 3. logerror distribution at Max Depth = 8

Prediction using neighborhood information

Six regression models were used here, they are **linear regression, ridge regression, lasso regression, support vector regression, random forest, and a naïve ensemble method**.

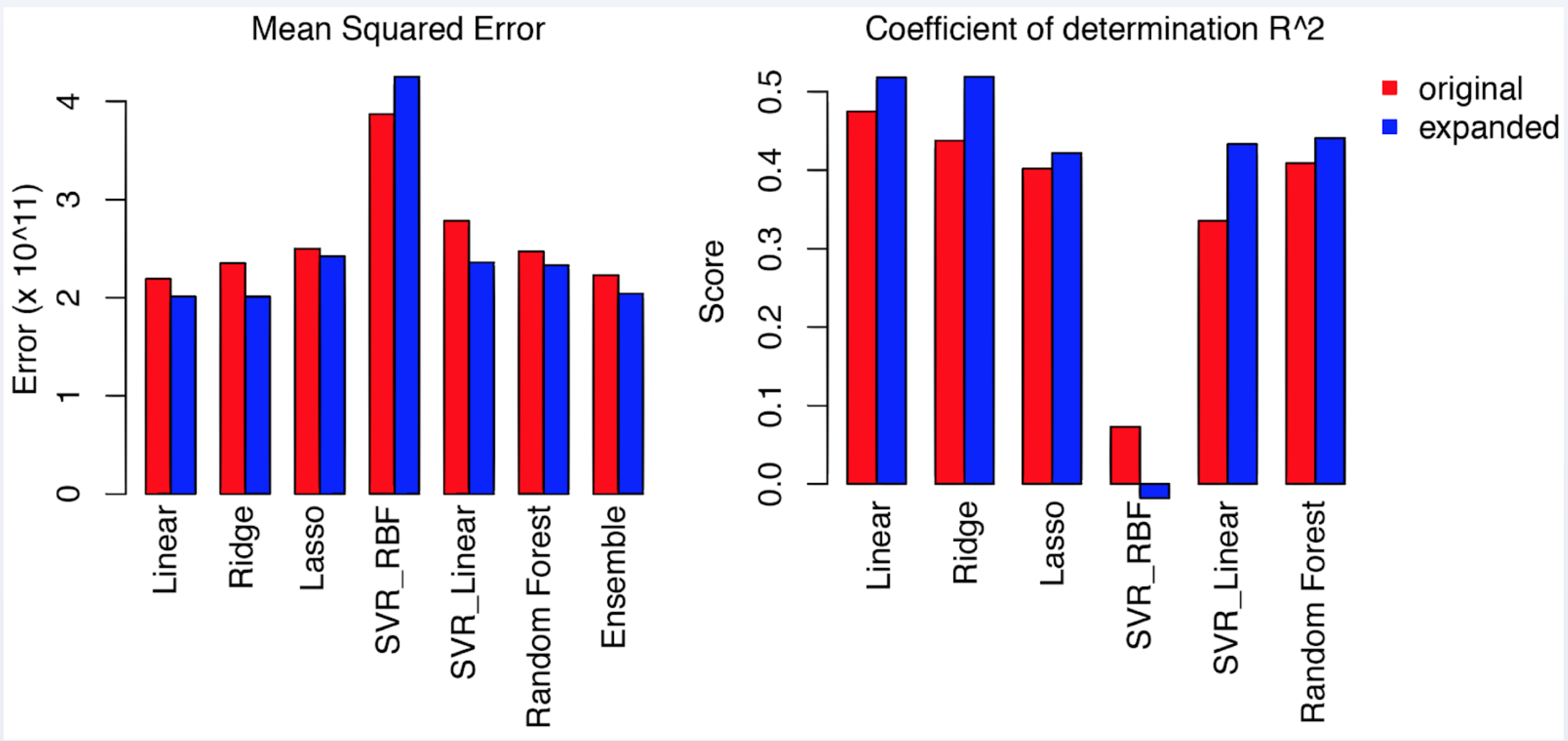

Figure 4. Comparison of calculated mean squared error(left) and coefficient determination of R^2(right)

For all these linear-based algorithms, the **expanded dataset** worked better than the **original dataset**, giving the lower errors and higher scores.

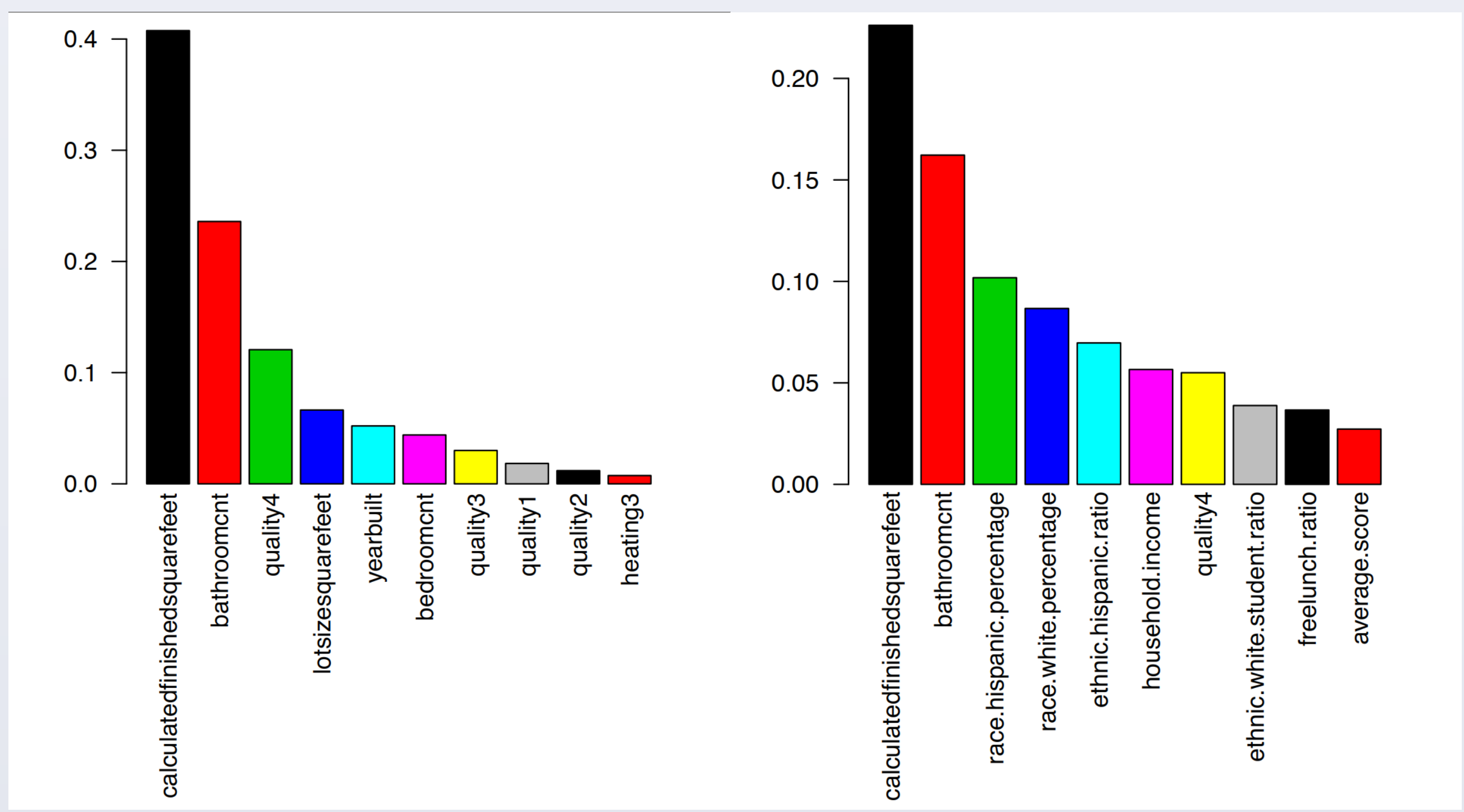

Figure 5. Important features using original data (left) and expanded data (right)

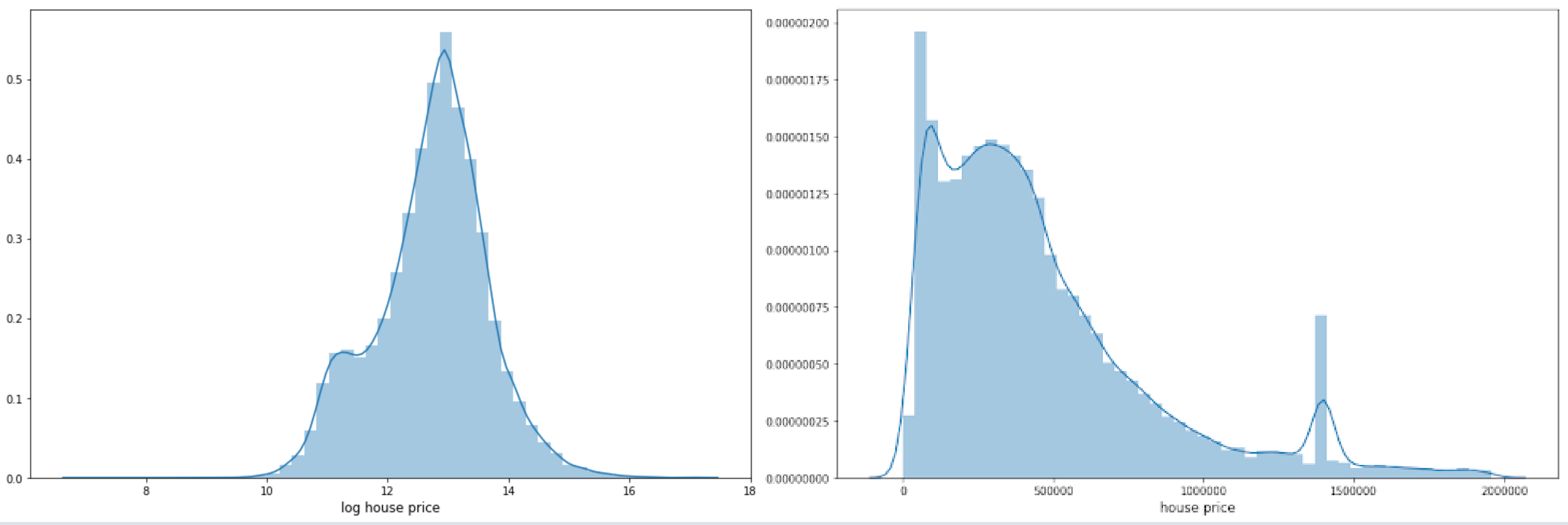

Figure 6. logerror (left) and house price (right) distribution

Map visualization

An interactive map is implemented using **Leaflet**, an open-source JavaScript library for mobile-friendly interactive maps.

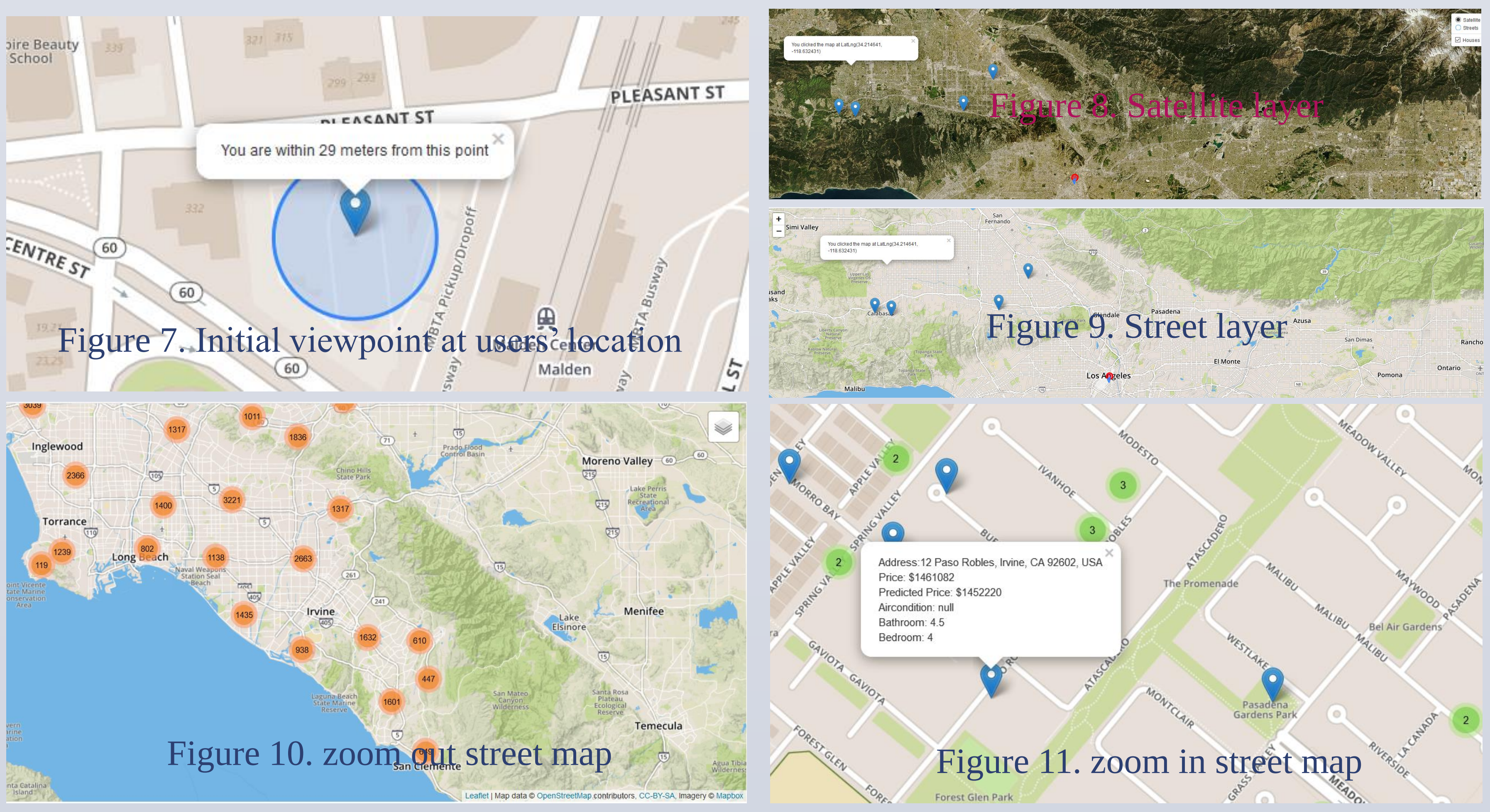
PLEASANT ST
You are within 29 meters from this point
CENTRE ST
60
Malden
Address:12 Paso Robles, Irvine, CA 92602, USA
Price: $1461082
Predicted Price: $1452220
Airconditioin: null
Bathroom: 4.5
Bedroom: 4

Figure 7. Initial viewpoint at users' location

Figure 8. Satellite layer

Figure 9. Street layer

Figure 10. zoom out street map

Figure 11. zoom in street map

Conclusion

We developed an accurate house price prediction model with integration of multiple community and school data, and visualized the result on an interactive map. The innovations of our ideas include:

- Scrape information of the local neighborhood and combine them with physical features of properties;
- Analyze feature importance and remove irrelevant data;
- Conduct house price prediction using machine learning models;
- Build an interactive map using Leaflet.js to achieve layer controls and mobile compatibility.