Evaluating Machine Learning Models for House Price Prediction in San Francisco and Los Angeles

Anandita Juneja

Ajaiy Praveen Thiruchelvam

Lifu ‘Mario’ Ma

1. Executive Summary

The objective of this report is to explore and evaluate the most popular machine learning models for predicting house prices, and analyze their performance on a real-world dataset of house prices in San Francisco and Los Angeles. In this project, the dataset was created through web scraping, API, and secondary sources, and it contains housing price, number of bedrooms, number of bathrooms, city, sqft, neighborhood, population, number of households in the neighborhood, median income, and average housing price per neighborhood.

Eight models were evaluated in this report, including simple linear regression, PCR, LASSO regression, decision tree, trimmed decision tree, random forest, XGBoost, and a hybrid model combining LASSO, Random Forest, and XGBoost. The RMSLE was used as the main metric for evaluation, and after data cleaning, outlier removal, and model evaluation, the hybrid model was found to be the best model for predicting house prices accurately.

By utilizing the predictions generated by the model, policymakers can make more informed decisions about zoning, land use, and development policies. Additionally, businesses and individuals can make more informed decisions related to property investment, development, and sales.

2. Background and Domain Knowledge

Estimating house prices is a complex challenge, but with the growth in data availability and advancements in machine learning algorithms, it has become feasible to develop models that can accurately predict the cost of homes. Several machine learning models have been created and utilized in recent years for predicting house prices, each with its unique strengths and limitations.

This report aims to explore the most popular machine learning models employed for predicting house prices, evaluate their efficacy, and analyze their performance on a real-world dataset of house prices in San Francisco and Los Angeles. The objective of this report is to offer an extensive overview of different machine learning models used for predicting house prices and to assess their suitability for this task.

Accurately predicting house prices can provide valuable insights for city planners and policymakers who are responsible for ensuring an adequate supply of housing to meet the needs of the population. Our model can assist these stakeholders in analyzing the demand for housing and identifying areas where there may be a shortage or surplus of housing. By utilizing the predictions generated by our model, policymakers can make more informed decisions about zoning, land use, and development policies. Specifically, by comparing the house price and income situation in San Francisco and Los Angeles, policymakers can gain a deeper understanding of the housing market and develop targeted policies to address specific issues in each city.

In this project, we utilized our domain knowledge and expertise to build several models to predict housing prices based on a variety of features such as the property's price, address, square footage, neighborhood, and various other relevant characteristics gathered from the Trulia website. In addition, we enriched our dataset by joining it with median income salary data for each neighborhood. This approach allowed us to create more comprehensive and accurate models that take the socioeconomic factors that may affect housing prices in each neighborhood

into account. By leveraging these models, businesses and individuals can make more informed decisions related to property investment, development, and sales.

3. Traditional attempts for House Price Prediction

The House Price Index (HPI) is a commonly used tool for measuring changes in residential housing prices across countries. It is a weighted, repeat sales index that calculates average price changes based on repeat sales or refinancing of the same properties. The data is obtained by reviewing mortgage transactions on single-family properties, whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac since 1975. The HPI is useful for estimating changes in mortgage defaults, prepayments, and housing affordability in specific geographic areas. However, it is not an efficient way to predict the price of a specific house, as it is a rough indicator calculated from all transactions.

Other factors, such as median price income, also play an important role in predicting house prices. Recent studies have explored the potential of machine learning to address this problem, but most have focused on comparing the performance of different models rather than combining them. To tackle this issue, we have developed a project that combines generic house data, such as square footage, number of bedrooms, and price obtained from web scraping the Trulia website, with median income data of the house's neighborhood. This combination allows us to predict the price of a house based on its locality accurately.

4. Analysis

	LASSO	Trimmed Decision Tree	Random Forest	XGBoost	Hybrid Model
RMSLE	0.4579	0.3683	0.3137	0.3163	**0.3051**
RMSE	715790.3834	597218.2383	532336.6114	603489.6343	536439.8642
R-squared	0.3871	0.5733	0.6610	0.5643	0.6558

Fig 1: Model Conclusion

In this report, our group evaluated eight models: simple linear regression, PCR, LASSO regression, decision tree, trimmed decision tree, random forest, XGBoost, and a hybrid model combining LASSO, Random Forest, and XGBoost. We used RMSLE (Root Mean Squared Logarithmic Error) as the main metric for evaluation, which is commonly used in regression problems with a large range of values. Given the wide range of house prices, we selected RMSLE as our metric.

$$\text{RMSLE} = \sqrt{(log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

Fig 2: Formula of RMSLE

4.1 Data Cleaning

Our group created the data file through web scraping, API and some secondary sources. The final dataset used for housing price prediction contains housing price, number of bedrooms, number of bathrooms, city, sqft, neighborhood, population, number of households in the neighborhood, median income, and average housing price per neighborhood.

	price	bed_room	bath_rom	sqft	city	neighborhood	population	number of households	median income	avg_price_per_neighborhood
Data Type	numeric	numeric	numeric	numeric	string	string	numeric	numeric	numeric	numeric
Null Percentage	0.00	0.04	0.06	0.03	0.00	0.00	0.00	0.00	0.00	0.00

Fig 3: null-value description

By observing the null-value description table, we can see that there is small percentage of null values in bed_room, bath_room and sqft. Theoretically, since the house price may be related with the number od bedrooms, bathrooms and sqft, it may not be reasonable to use median values to fill them. As a result, our group decided to drop the null values.

After dropping null values, our group made histograms for each variable. Some variables like bed_room are extremely right-skewed. For data integrity, our group also decided to drop all the outliers.

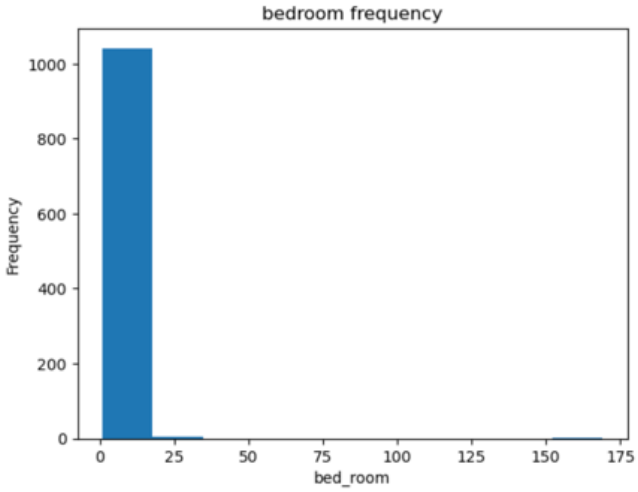

Fig 4: Sample histogram: bed_room

4.2 Simple Linear Regression Model

Our group started with simple linear regression model, and checked heteroscedasticity and multicollinearity for it. For heteroscedasticity, we chose to use BP test. The P-value of BP test is close to 0, which shows severe heteroscedasticity problem. For multicollinearity, we chose to use VIF test. All the VIF test results are over 5, which also implies strong multicollinearity problem.

As a result, it may not be appropriate to directly use simple linear regression model for this dataset. As a result, to fix these problems, our group decided to use principal component regression model.

4.3 Principal Component Regression Model

To apply PCR, our group first fitted principal component analysis to the variables, and draw the scree plot for principal components.

Surprisingly, only one component can explain all the variables well. We fitted the principal component and all the dummy variables to the dataset, with the R-squared score of 0.32.

To check heteroscedasticity problem, we drew the residual plot with the predicted values against residuals.

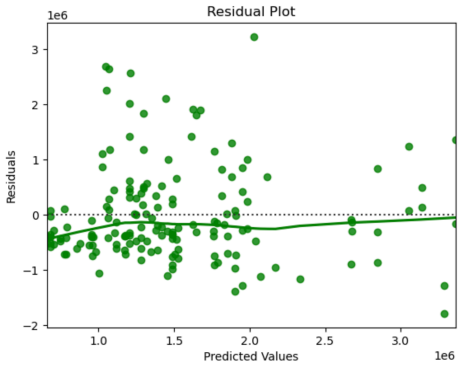

Fig 5: Residual Plot

Based on the residual plot, there seems to be no serious heteroscedasticity problem since the points are scattered almost randomly around the horizontal line at zero.

4.4 CV LASSO Regression Model

Another way to do feature selection is through LASSO regularization. To select the lambda of LASSO, we chose to use cross validation. Based on the result of LASSO regularization, sqft,

population, median income and avg_price_per_neighborhood are selected into the final regression model. The RMSLE of LASSO regression is 0.4579, OOS R-squared is 0.3871.

4.5 Decision Tree Model

Tree models are another kind of powerful tools to make predictions. Compared with linear regression models, tree models have less restrictions for the distribution of the data, and can also catch non-linear relationship. As a result, we first chose to fit the full decision tree model.

Here's the feature importance plot of this tree model:

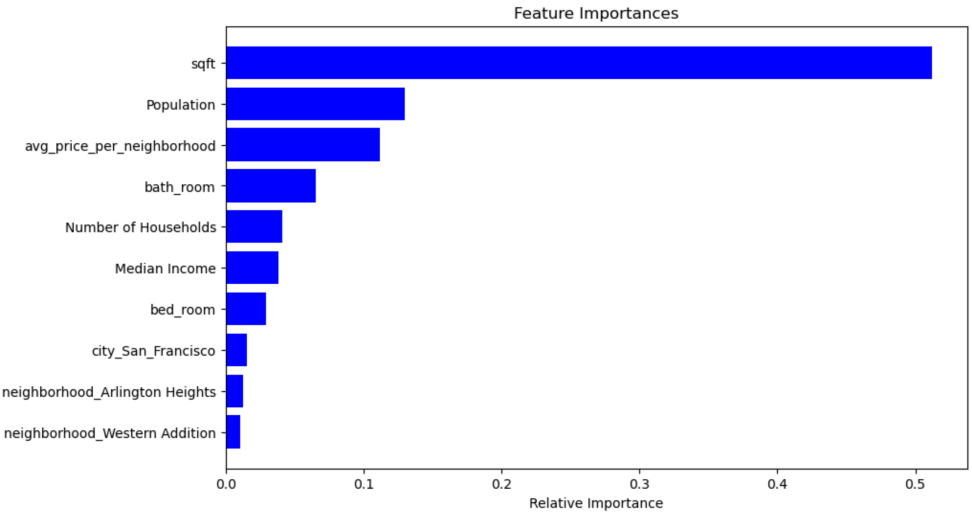

Fig 6: Feature Importance Plot

Based on the importance plot, the sqft of the house and the population, avg_price_per_neighborhood seems to be the most important features that influence the housing price, which are similar with the result of LASSO.

However, the full decision tree model seems to overfit the data a lot, with in-sample RMSLE score of 0.0056, but Out-of-sample RMSLE of 0.3787.

4.6 Trimmed Decision Tree Model

To avoid the overfitting problem, our group decided to trim the full model through cross validation. We achieved this mainly by controlling the min_samples_leaf and max_depth during the training process. In the end, based on our trials, when min_samples_leaf = 20 and max_depth = 10, the model works the best with OOS RMSLE of 0.3683.

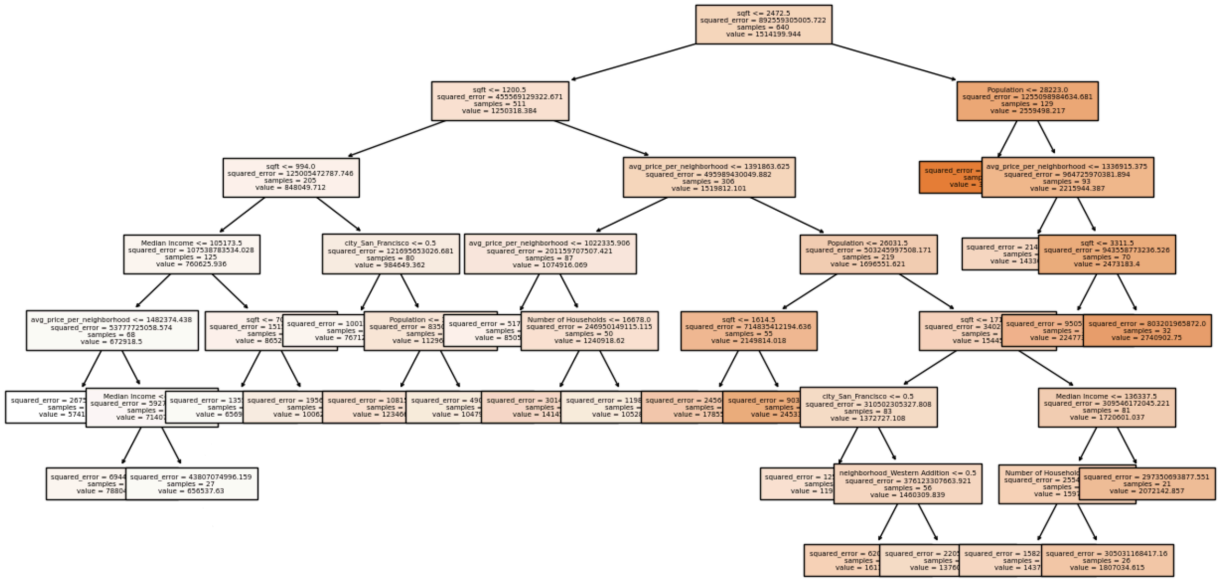

Fig 7: Sample Tree Plot

4.7 Random Forest

Another way to kind of solving the overfitting problem is to use the ensembled results of a batch of models through bootstrapping, and random forest is a method to realize this. Finally, the OOS RMSLE of random forest is 0.3137, which is the smallest by now.

4.8 XGBoost

We can also try to use boosting methods to tackle this problem. Boosting is a combination of connected weak learners to achieve a strong learner. In this case, we tried XGBoost, with OOS RMSLE of 0.3163.

4.9 Hybrid Model

Finally, to achieve the best result, we can also try to combine all the results of our models. We selected the top 4 models by now, which are LASSO, trimmed decision tree model, random forest and XGBoost. Then we calculate the optimized weights to compute the final ensembled prediction, with the 0.25 weight for XGboost, 0.65 for Random Forest and 0.05 for LASSO regression. Now, the OOS RMSLE achieved the best, which is 0.3051, and we selected the hybrid model as our final chosen prediction model.

5. Recommendations and Business Values

5.1 Less population density, Larger the squared feet, Higher the house price

For individual investors and housing agencies, when making real estate investment, they can also take the population density into account, which may be easily neglected. Based on all the models I built, apart from sqft, population is the most important indicator of house price. Also, based on the tree plot of trimmed decision tree and the coefficient of LASSO regression, the less the population density, the higher the house price will be. As a result, population density can also help real estate investors to make informed decisions about buying or selling properties, as well as estimating potential returns on investment.

5.2 Better Urban Planning

For city planners and policymakers, house price predictions can assist them in understanding the demand for housing and identifying areas where there may be a shortage or oversupply of housing. By utilizing our model, they can compare the house price and income situation in SF and LA, and make more informed decisions about zoning, land use, and development policies.

Based on our model output, the house price in SF will be more expensive than LA, but based on the median income, SF may be less affordable than LA, which indicates the city planners should consider building more government subsidized houses.

5.3 Help Make More Appropriate Tax Policies

For government entities, such as tax assessors, can use house price predictions to assess the value of properties for taxation purposes. Traditional property assessments are typically done once every few years, which can result in outdated assessments that do not reflect current market conditions. By using house price predictions, tax assessors can update property valuations more frequently, ensuring that assessments remain accurate and up-to-date.

6. Summary & Conclusion

This report explores and evaluates the most popular machine learning models for predicting house prices and analyzes their performance on a real-world dataset of house prices in San Francisco and Los Angeles. The objective is to provide valuable insights for city planners and policymakers who are responsible for ensuring an adequate supply of housing to meet the needs of the population. The dataset includes features such as price, address, square footage, neighborhood, and socioeconomic factors that may affect housing prices. Eight models were evaluated, and a hybrid model combining LASSO, Random Forest, and XGBoost was found to be the best model for predicting house prices accurately. The predictions generated by the model can assist policymakers in analyzing the demand for housing and identifying areas where there may be a shortage or surplus of housing. On comparing the affordability of houses in SF and LA, we come to a conclusion that LA is more affordable.