House Prices Prediction using Machine Learning

Gang Wu
gangwu@stanford.edu

1 Introduction

Real estate always been a hot topic, especially in the recent years. In this project, I work on a practical problem on this topic: given a set of features (e.g. location, build year, etc.) for a house, how much will it sell? The answer to this question will attract great interest to house buyers and sellers. I will explore various machine learning techniques in this project to help answer this question.

2 Problem Definition

2.1 Input & Output

The input to my problem would be a set of feature values for a house, such as 'year built', 'lot size', 'living room size', 'school district', etc. The output is the predicted sale price.

As a concrete example, below is a simplified set of feature values for a house as an input:

$[' year\_built', ' lot\_size', ' livingroom\_size', ' school\_score'] = [2001, 3000, 800, 9]$

The output would be its sale price in USD: 650000.

2.2 Evaluation Metric

For this work, I use Root-Mean-Squared-Error (RMSE) to evaluate the model [1]:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(y'_t - y_t)}{T}}$$

Here, y'_t denotes the logarithm of predicted house price and y_t denotes the logarithm of real house price. Taking logs will help make sure that errors in predicting expensive houses and cheap houses affect the result equally.

3 Baseline Implementation

3.1 Data

The baseline dataset I used is 'Ames Housing dataset' obtained from Kaggle [2]. This dataset is small (1461 houses each with 81 features on the training set) and out of date. I use it only for initial implementation of different ML algorithms and building up a baseline framework. For the next step of this project, a web crawler will be built to obtain more data from online real estate websites.

3.2 Exploratory Data Analysis

Some exploratory analysis on the 'Ames Housing dataset' is shown in Figure 1. From the distribution of the sale price, we can see most of the houses are in 129K to 214K range. Figure 1 (b) shows the

living room area v.s. sale price, which roughly have a linear relationship. Figure 1 (c) shows the overall quality v.s. sale price. It shows the overall quality is a very strong indicator of the sale price.

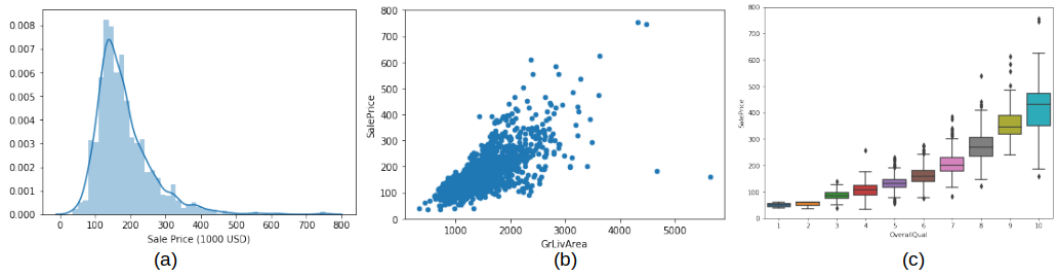

Figure 1: (a) sale price distribution. (b) living room area v.s. sale price (1000 USD). (c) overall quality v.s. sale price (1000 USD).

Top 6 correlated features with sale price is shown in the table below:

Feature Name:	OverallQual	GrLivArea	GarageArea	TotalBsmtSF	1stFlrSF	FullBath
Correlation:	0.81	0.70	0.65	0.61	0.59	0.59

Table 1: Most important features correlating with sales price.

3.3 Preprocessing

The dataset is splited into two parts: 70% on the $training\_set$ for training and 30% on the dev_set for cross validation. I also extracted all the 37 numerical features while ignoring the rest 43 categorical features for a simple baseline implementation. The missing feature values are filled using the median value of the feature. This ends up with the $training\_set$ dimension $(1022, 37)$ and dev_set dimension $(438, 37)$.

3.4 Models & Experimental Results

Figure 2 (a) shows the linear regression results [3] on $training\_set$ and dev_set and Figure 2 (b) shows the results of XGBoost[4]. The corresponding RMSE value is shown in Table 2. From the result, we can see the XGBoost model helps remove the few outliers and also significantly improved the results on the dev_set.

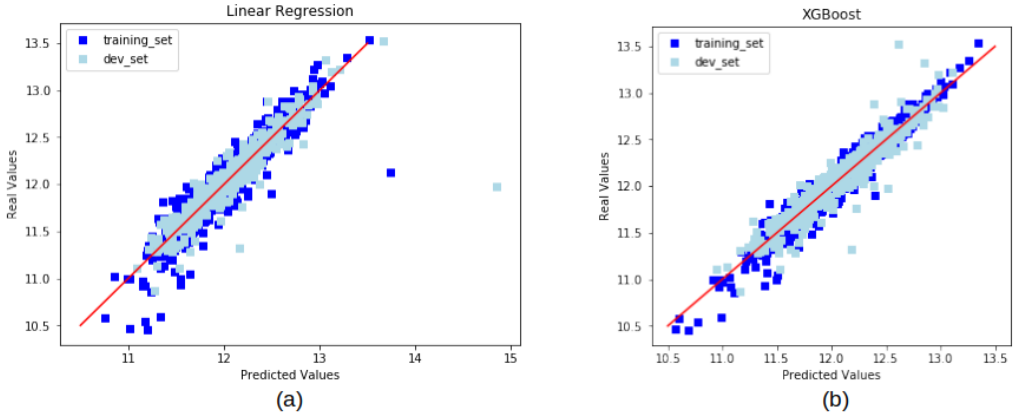

Figure 2: (a) Linear regression. (b) XGBoost.

RMSE	TrainSet	DevSet
LR	0.142	0.184
XGB	0.136	0.148

Table 2: RMSE with linear regression and XGBoost.

3.5 Next Step

The next step of this project is to obtain more data from the real estate websites. Also, more feature engineering work and machine learning models need to be implemented and tuned.

References

[1] https://en.wikipedia.org/wiki/Root-mean-square_deviation.

[2] https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

[3] https://en.wikipedia.org/wiki/Linear_regression.

[4] https://xgboost.readthedocs.io/en/latest/.