Boston Housing Report

# Executive Summary

## Problem Statement

The challenge is to train and compare a linear regression, random forest and XGBoost model. By using either GRID search or RANDOM search PLUS k-fold cross validation to tune model hyperparameters, we will find the most effective model at predicting assessed value for property. Besides, we will especially explore on owner-occupied property and year property was last remodeled to see if these variables influence the assessed value for property as the city of Boston expected.

## Key Findings

* Although year remodeled and owner-occupied are not useful in determining assessed value, as the city of Boston expected, owner-occupied properties and recently remodeled houses do have higher average assessed value. However, houses built in 1990s don’t seem to have higher assessed value.
* Both random forest model and XGBoost model have an RMSE lower than $57854, with random forest of $56668.65 and XGBoost of $53606.67.
* Based on the same recipe, the random forest model performs the best at the hyperparameters of number of trees = 175, min\_n = 5, with an average RMSE of $55483.299 and an average RSQ of 0.864, while the XGBoost model performs the best at 2000 of trees and tree depth of 5.
* The XGBoost model might be underfitted as its learning rate is 0.006 (close to 0).
* The XGBoost model is good at predicting assessed value of regular and average houses that were built after 1920 in Hyde Park and Cambridge, while it performs bad in estimating assessed value for larger and superior houses that were built before 1915 in Jamaica Plain and Dorchester.
* The XGBoost model tends to predict higher assessed value as the estimate errors between predicted value and actual value are mostly negative.

## Metrics

We will evaluate the models using RMSE (Root Mean Square Error) to measure the error of a model in predicting assessed value for property. Formally, RMSE is defined as follows:

## Analysis

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

The left graph above shows the original distribution of assessed value, and we can see the distribution is skewed to right, which might affect our prediction. After transforming into logged assessed value, as shown in the right graph, the assessed value becomes relatively descent normal distribution.

Chart

Description automatically generatedChart, scatter chart

Description automatically generated

These two graphs above are related to the city of Boston’s concerns. The graph on the left side seems to indicate that, overall, owner-occupied does affect a property’s assessed value, while the scatter plot on the right side shows that there is no significant relationship between year remodeled and assessed value. Moreover, since year remodeled has many zero and missing value, we will exclude it in our model.

Chart

Description automatically generatedChart, funnel chart

Description automatically generated

Chart

Description automatically generated

We select our predictors using business sense and variable importance. As shown above, the importance of the same variable is quite different in linear model and random forest model, but we will mainly focus on the importance graph of random forest model and compare with the importance chart of XGBoost model to determine useful predictors.

As we can see above, yr\_remod and own\_occ are not shown in the top 10 variables of random forest and XGBoost importance chart, indicating that they are not good determinators in predicting house assessed value.

Besides, we used forward selection to test the importance of each variable and check on the changes of the evaluation metrics (RMSE, MAE, RSQ), then decided which variables should be used in the model. (Therefore, we excluded variables with low importance such as kitchen style, number of kitchens, zip code…)

A picture containing chart

Description automatically generatedChart, histogram

Description automatically generated

Background pattern

Description automatically generatedBackground pattern

Description automatically generated

Graphs above explain the city of Boston’s concerns towards owner occupied, built year and remodeled year. As we can see, owner-occupied houses do have a higher average assessed value as the city of Boston expected, and homes have been recently remodeled tend to have higher assessed value (also higher than the average total assessed value of $448563.6387). In terms of houses built in the 1990s, their assessed value is not significantly higher than houses built in other decades, but this might be influenced by some outliers such as extremely expensive old houses.

# Methodology

1. Data partitioning
   * Split the data into 70/30 train/test split using random sampling
2. Data preprocessing
   * Formula
     1. av\_total ~ land\_sf + living\_area + yr\_built + num\_floors + r\_ovrall\_cnd + r\_int\_cnd + r\_fplace + median\_income + r\_ext\_cnd + r\_ext\_fin + r\_bldg\_styl + r\_full\_bth + r\_half\_bth + r\_bth\_style + r\_kitch\_style + population + pop\_density
   * Numeric Predictor Pre-Processing
     1. Replaced missing numeric variables with median
     2. Added a new variable called home age using year built
   * Categorical Predictor Pre-Processing
     1. Replaced missing categorical variables with “unknown”
     2. Dummy encoded categories with 1s and 0s
     3. Removed variables that are highly sparse and unbalanced
     4. Assigned a previously unseen factor level to a new value
3. Model specification
   * Linear Regression
     1. Mixture = 1, penalty = 0.01
   * Random Forest
     1. A random forest model creates a collection of decision trees and use these trees to benchmark each other and chooses the most common prediction as the final result.
   * XGBoost
     1. Extreme Gradient Boosting is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset’s features.

## Data Dictionary

|  |  |  |
| --- | --- | --- |
| Name | Description | Keep |
| PID | Unique 10-digit parcel number | Ignore |
| ZIPCODE | Zip code of parcel | Ignore |
| OWN\_OCC | One-character code indicating if owner receives residential exemption as an owner-occupied property | Ignore |
| AV\_TOTAL | Assessed value for property i.e. what you are predicting | Keep |
| LAND\_SF | Parcel’s land area in square feet (legal area) | Keep |
| YR\_BUILT | Year property was built | Converted |
| YR\_REMOD | Year property was last remodeled | Ignore |
| HOME\_AGE | Calculated by 2022 minus year built | Keep |
| LIVING\_AREA | Living area square footage of the property | Keep |
| NUM\_FLOORS | # of levels in the structure located on the parcel | Keep |
| STRUCTURE\_CLASS | Structural classification of commercial building | Ignore |
| R\_BLDG\_STYL | Residential building style | Keep |
| R\_ROOF\_TYP | Structure roof type | Ignore |
| R\_EXT\_FIN | Structure exterior finish | Keep |
| R\_TOTAL\_RMS | Total number of rooms in the structure | Ignore |
| R\_BDRMS | Total number of bedrooms in the structure | Ignore |
| R\_FULL\_BTH | Total number of full baths in the structure | Keep |
| R\_HALF\_BTH | Total number of half baths in the structure | Keep |
| R\_BTH\_STYLE | Residential bath style | Keep |
| R\_KITCH | Total number of kitchens in the structure | Ignore |
| R\_KITCH\_STYLE | Residential kitchen style | Keep |
| R\_HEAT\_TYP | Structure heat type | Ignore |
| R\_AC | Indicates if the structure has air conditioning (A/C) | Ignore |
| R\_FPLACE | Total number of fireplaces in the structure | Keep |
| R\_EXT\_CND | Residential exterior condition | Keep |
| R\_OVRALL\_CND | Residential overall condition | Keep |
| R\_INT\_CND | Residential interior condition | Keep |
| R\_INT\_FIN | Residential interior finish | Ignore |
| R\_VIEW | Residential view | Ignore |
| POPULATION | Population of people in the zip code | Keep |
| POP\_DENSITY | People per square mile | Keep |
| MEDIAN\_INCOME | Median income of the residence of that zip code | Keep |
| CITY\_STATE | City name and state | Ignore |

# Model Performance & Evaluation

## Model Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Partition | RMSE | RSQ | MAE |
| Linear Regression | Train | 72123.89 | 0.7596399 | 52839.00 |
| Test | 74834.56 | 0.7487202 | 53153.28 |
|  |  |  |  |  |
| Random Forest | Train | 30801.30 | 0.9612302 | 22580.74 |
| Test | 56668.65 | 0.8612859 | 40018.87 |
|  |  |  |  |  |
| XGBoost | Train | 40346.18 | 0.9251259 | 30450.33 |
| Test | 53606.67 | 0.8711012 | 38330.51 |

RMSE (Root Mean Squared Error) measures the standard deviation of residuals. It represents the square root of the average of the squared difference between the original and predicted values in the dataset. (The lower, the better)

MAE (Mean Absolute Error) measures the average of the residuals in the dataset. It represents the average of the absolute difference between the actual and predicted values in the dataset. (The lower, the better)

RSQ (R-squared) is a scale-free score which represents the proportion of the variance in the dependent variable which is explained by the model. (The higher, the better)

Overall, XGBoost model performs the best with the lowest RMSE and MAE and the highest RSQ within the test dataset. A lower RMSE and MAE means that the model performs better as the differences between actual and predicted values are small. A higher RSQ means that the model could explain larger proportion of the variance in the target variable.

In this case, a lower RMSE and MAE indicates that using the XGBoost model, we are making less error in predicting assessed value, and a higher RSQ demonstrates that our model could explain most of changes in the assessed values.

Moreover, XGBoost model is better as the difference between its training dataset and test dataset are smaller than the linear regression model and the random forest model.

## Tuning Performance

Line chart

Description automatically generated with medium confidenceChart, line chart

Description automatically generated

Table

Description automatically generated

Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters. In random forest model, it could help to determine the ideal number of trees and minimum number for a node to split (minimum number of data points in a node that are required for the node to be split further). Graphs above show that when min\_n = 5, trees = 175, the model has lowest average RMSE and highest average RSQ.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

The Bayesian model-based optimization is to build a probability model of the objective function and use it to select the most promising hyperparameters to evaluate in the true objective function. In the XGBoost model, it could help to optimize the learning rate, tree depth and number of trees. As shown above, the model performs the best at a learning rate of 0.006, tree depth of 5, and 2000 of trees.

The rule of learning rate: if the learn rate is near 0, learns slower and may underfit; if the learn rate is near 1, learns faster but may overfit. Therefore, the XGBoost model might be underfitted as its learning rate is near 0.

## Recommendations

* The city of Boston should not use year remodeled and owner-occupied as estimators in determining assesses value for houses. Instead, they should use living area square footage, median income of the residence of that zip code, legal area, overall condition, interior condition, home age, number of fireplaces, exterior condition, building style and population as effective estimators as they are the top 10 important variables in the model.
* The city of Boston is recommended to use the XGBoost model to estimate assessed value for regular and average houses built after 1920 in Hyde Park and Cambridge.
* The city of Boston should not use the XGBoost model to estimate assessed value for larger and superior houses that were built before 1915 in Jamaica Plain and Dorchester.
* The distribution of assessed value is right skewed and remodeled year has many missing values, in order to further optimize our model, the city of Boston should update and collect more information on expensive and superior houses in Jamaica Plain and Dorchester.

## Top 10 & Bottom 10 Predictions

Best Estimates

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Based on our best estimate cases, the model is good at predicting assessed value for houses where were built after 1920, year remodeled = 0, living area is relatively normal (1107 - 2274), number of floors between 1.5 to 2.0, colonial building style, hip or gable roof type, wood shake or vinyl structure exterior finish, 6 – 8 total rooms, 3 – 4 bedrooms, semi-modern or modern bath style and kitchen style, hot water heat type, 1 fireplace, average overall and interior condition, zip code of 02136 and 02132, small population, average population density and median income, and cities of Hyde Park or Cambridge.

Worst Estimates

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

According to the worest estimate cases, the model is bad at estimating assessed value for houses where were built before 1915, and has been remodeled, living area is relatively large (1592 - 4062), number of floors at least 2.0, colonial or conventional building style, mansard roof type, frame/clapboard or vinyl structure exterior finish, more total rooms (8 – 14), more bedrooms (3 - 9), no remodeling bath style and kitchen style, hot water or forced air heat type, 0 to 1 fireplace, good overall and interior condition, zip code of 02130 and 02124, larger population, larger population density and median income, and cities of Jamaica Plain or Dorchester.

In comparison, the model performs better in estimating assessed value for newer, regular and average houses in Hyde Park or Cambridge, while it fails to accurately estimate assessed value for older, larger and fancier houses in Jamaica Plain and Dorchester.

However, parcel’s land area in square feet, structure class, number of kitchens, air conditioning, residential interior finish, and residential view do not necessarily influence the estimate of assessed value as in either case (best or worst), they are pretty much the same.

# Kaggle Submission

Kaggle Name: Y1 Ding

Kaggle reported score: 53480.14203

Kaggle reported position at time of submission: #52

<https://www.kaggle.com/competitions/challenge-3-regression-updated>