BOSTON-HOUSE PRICE PREDICTION PROJECT

# INTRODUCTION

This project aims to improve the performance of an already existing model to predict the assessed value of properties for taxing purposes in Boston. The current model has a root mean square error of $ 57,854 which is an estimate of the standard deviation of the prediction errors (deviation of the predicted house price from the actual house price).

To achieve this objective, we will build three models: linear regression, random forest, and extreme gradient-boosted trees. A linear regression model uses a straight line to estimate the relationship between the independent variable (in our case different attributes of properties) and the dependent variable (in our case the assessed value of the property). Random forest on the other hand, as the name suggests is a prediction model made up of a bunch of decision trees. A decision tree is a graphical representation of all the possible solutions (a bunch of business rules) to a decision (in our case which patterns influence house price) based on certain conditions. The underlying idea is that building a collection of decision trees is expected to be superior to just a single decision tree. Contrary to the random forest model, which combines decision trees in parallel, gradient-boosted trees combine decision trees in series. Thus, successive decision trees learn from the mistakes of the previous decision trees to reduce prediction errors. The underlying idea here is that combining many sequentially connected decision trees may result in a stronger decision tree.

In the following sections, we outline the problem statement and the strategies for meeting the objective set here.

# Problem Statement

The current model for predicting the assessed value of properties has a high root mean squared error. In other words, the current model has high variability of prediction errors which means that there is high uncertainty associated with the predicted house prices. This problem has hindered the ability of the City of Boston to forecast revenue from property taxes precisely within a specified margin of error for effective budgeting and planning. Also, the assessed value of some properties is underestimated such that the City of Boston is losing tax revenue. The situation is critical and warrants immediate action.

# Metrics

To evaluate model performance, we will examine root mean squared error (RMSE), mean absolute error (MAE), and r-square.

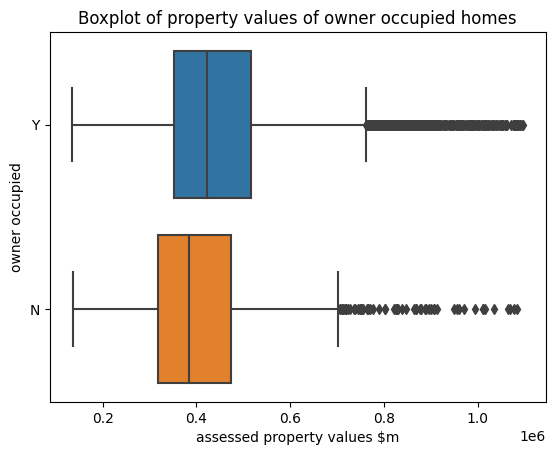
As already mentioned, root mean square error measures the standard division of the prediction errors (thus the extent to which the predicted house prices deviate from the actual house prices). Mean absolute error on the other hand measures the average magnitude of absolute prediction errors (the absolute difference between the predicted house price and the actual house price). Lastly r-square indicates the percentage of the variance in house prices that the independent variables explain collectively. It measures the strength of the relationship between the model and the dependent variable on a convenient 0 – 100% scale.

In business speak, mean absolute error is more interpretable out of these three metrics. For instance, if our model has a mean absolute error of $10,000 it means that predicted house prices deviate from the actual house price by plus or minus $10,000.

# Exploratory Analysis

1. **Do owner-occupied homes have higher assessed values?**

From the chart below, owner-occupied homes tend to have higher assessed values



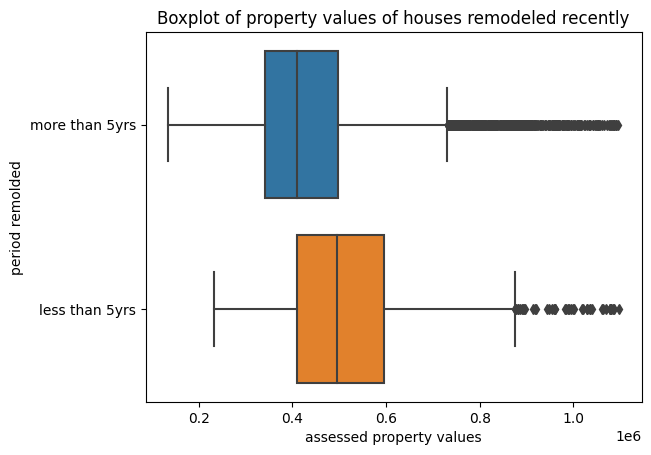
1. **Do houses built in the 1990s have higher assessed values?**

From the boxplot below, there seems to be no significant difference between the assessed value of houses built in the 1990s and other years.



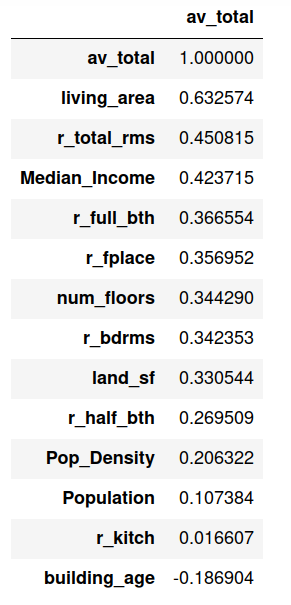
1. **Do houses remolded recently (6 years from 2016 ) have higher assessed values?**

From the chart below, houses remolded recently tend to have higher assessed values.



1. **What other variables may influence the assessed value of properties in boston?**

From the correlation matric plot below, numeric variables like living area, total rooms and median income tend to have somewhat positive correlation with the assessed value of the property.



**Data Dictionary Of Predictors Used in the model**

| **Variable** | **Description** |
| --- | --- |
| r\_overall\_cnd | Residential overall condition: |
| city\_state | City Name and State |
| median\_income | Median Income of the residence of that zip code |
| r\_int\_cnd | Residential interior condition: |
| pop\_density | People per square mile |
| r\_bldg\_style | Residential building style |
| r\_fplace | Total number of fireplaces in the structure |
| r\_total\_rms | Total number of rooms in the structure |
| own\_occ | One-character code indicating if owner receives residential exemption as an owner-occupied property |
| r\_full\_bth | Total number of full baths in the structure |
| num\_floors | # of levels in the structure located on the parcel |
| r\_ext\_cond | Residential exterior condition: |
| land\_sf | Parcel’s land area in square feet (legal area) |
| living\_area | Living area square footage of the property |
| r\_bdrms | Total number of bedrooms in the structure |

**All other dimensions of the dataset below were not used are either likely to be customer identifiers or turn to overfit.**

| **Variable** | **Description** |
| --- | --- |
| pid | Unique 10-digit parcel number |
| zipcode | Zip code of parcel |
| STRUCTURE\_CLASS | Structural classification of commercial building: |
| yr\_built | Year property was built |
| yr\_remold | Year property was last remodeled |
| R\_ROOF\_TYP | Structure roof type: |
| R\_EXT\_FIN | Structure exterior finish: |
| R\_HALF\_BTH | Total number of half baths in the structure |
| R\_KITCH | Total number of kitchens in the structure |
| R\_KITCH\_STYLE | Residential kitchen style: |
| R\_HEAT\_TYP | Structure heat type: |
| R\_AC | Indicates if the structure has air conditioning (A/C): |
| R\_INT\_FIN | Residential interior finish: |
| R\_VIEW | Residential view: |

# Methodology

1. **Data partitioning**
   1. The data was split into 70% for training and 30% for testing
2. **Data preprocessing**
   1. Formula

av\_total ~ r\_overall\_cnd + city\_state +median\_income + r\_int\_cnd +

pop\_density + r\_bldg\_style + r\_fplace + r\_total\_rms + own\_occ + r\_full\_bth +

num\_floors + r\_ext\_cond + land\_sf + living\_area + r\_bdrms

* 1. Numerical predictor preprocessing
     1. replaced missing numerical variables using median
  2. Categorical predictors preprocessing
     1. Replaced missing values with “unknown”
     2. Dummy encoded categorical variables to 1s and 0s using one-hot encoding

1. **Model specification**
   1. Linear regression
      1. use L1 normalization to use only significant predictors
   2. Random Forest - tuned
      1. Use cross-validation to select the best values hyperparameters
      2. The model parameters are as follows:

{ 'max\_depth': 15,

'min\_samples\_leaf': 2,

'min\_samples\_split': 5,

'n\_estimators': 1000}

* 1. XGBoost - tune
     1. Use cross-validation to select the best values hyperparameters
     2. The model parameters are as follows:

{ 'learning\_rate': 0.01,

'max\_depth': 5,

'n\_estimators': 1800,

'random\_state': 20,}}

# Model Metrics & Evaluation

## **Model Summary**

**Linear regression**

| **Dataset** | **Root mean squared error** | **mean absolute error** | **R-Square** |
| --- | --- | --- | --- |
| training | 60,170.10 | 42,585.68 | 0.8329 |
| testing | 61,468.74 | 43,316.26 | 0.8299 |

**Random Forest**

| **Dataset** | **Root mean squared error** | **mean absolute error** | **R-Square** |
| --- | --- | --- | --- |
| training | 29,551.77 | 21,227.79 | 0.9597 |
| testing | 55,750.42 | 39451.0 | 0.8601 |

**XGBoost**

| **Dataset** | **Root mean squared error** | **mean absolute error** | **R-Square** |
| --- | --- | --- | --- |
| training | 39,491.87 | 29,306.08 | 0.928 |
| testing | 50,583.58 | 36,062.24 | 0.8848 |

**Model evaluation - statistical metrics**

We have trained three models using the dataset. From the results reported above, the best model on all metrics: RMSE, MAE, and r-square is the XGBoost model. This model has RMSE of $50,583.83, MAE of $36,062.24, and an r-square of 88%. The RMSE of $50,583.83 is the standard deviation of the prediction errors, MAE of $36,062.24 means that on average the predicted house price is plus or minus $36,062.24, and the r-square of 88% means that all the independent variables used in the model account for 88% of the variance in housed prices.

**Model evaluation - business perspective**

Compared to the existing model, the XGBoost model has greater predictive power because it has an RMSE of $50,583 which is lower than the RMSE of $ 57,854 for the existing model. Effectively, the current model has reduced the variance of the prediction error by about 13%.

Thus the current model is expected to reduce the uncertainty around forecasting assessed property values by about 13%

**Description of Top 10 best predictions**

| **Predictor** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| land\_sf | 5826.4 | 1560.63 | 4000 | 4768.25 | 5557.5 | 6287.25 | 8775 |
| living\_area | 1508.9 | 253.76 | 1152 | 1306 | 1501 | 1714.5 | 1919 |
| num\_floors | 1.55 | 0.44 | 1 | 1.125 | 1.5 | 2 | 2 |
| r\_total\_rms | 6.3 | 0.82 | 5 | 6 | 6 | 6.75 | 8 |
| r\_bdrms | 3.1 | 0.57 | 2 | 3 | 3 | 3 | 4 |
| r\_full\_bth | 1.2 | 0.42 | 1 | 1 | 1 | 1 | 2 |
| r\_half\_bth | 0.4 | 0.52 | 0 | 0 | 0 | 1 | 1 |
| r\_kitch | 1 | 0.00 | 1 | 1 | 1 | 1 | 1 |
| r\_fplace | 0.5 | 0.71 | 0 | 0 | 0 | 1 | 2 |
| Population | 35440.1 | 2459.51 | 28488 | 36314 | 36314 | 36314 | 36314 |
| Pop\_Density | 12283.3 | 2289.74 | 6207 | 13251 | 13251 | 13251 | 13251 |
| Median\_Income | 73818.8 | 5246.20 | 58890 | 75446 | 75446 | 75446 | 75730 |
| av\_total | 433260 | 112640.55 | 354600 | 375050 | 395200 | 423900 | 735400 |
| error | 97.725 | 68.04 | 16.28125 | 34.4765625 | 98.890625 | 146.9453125 | 202.25 |

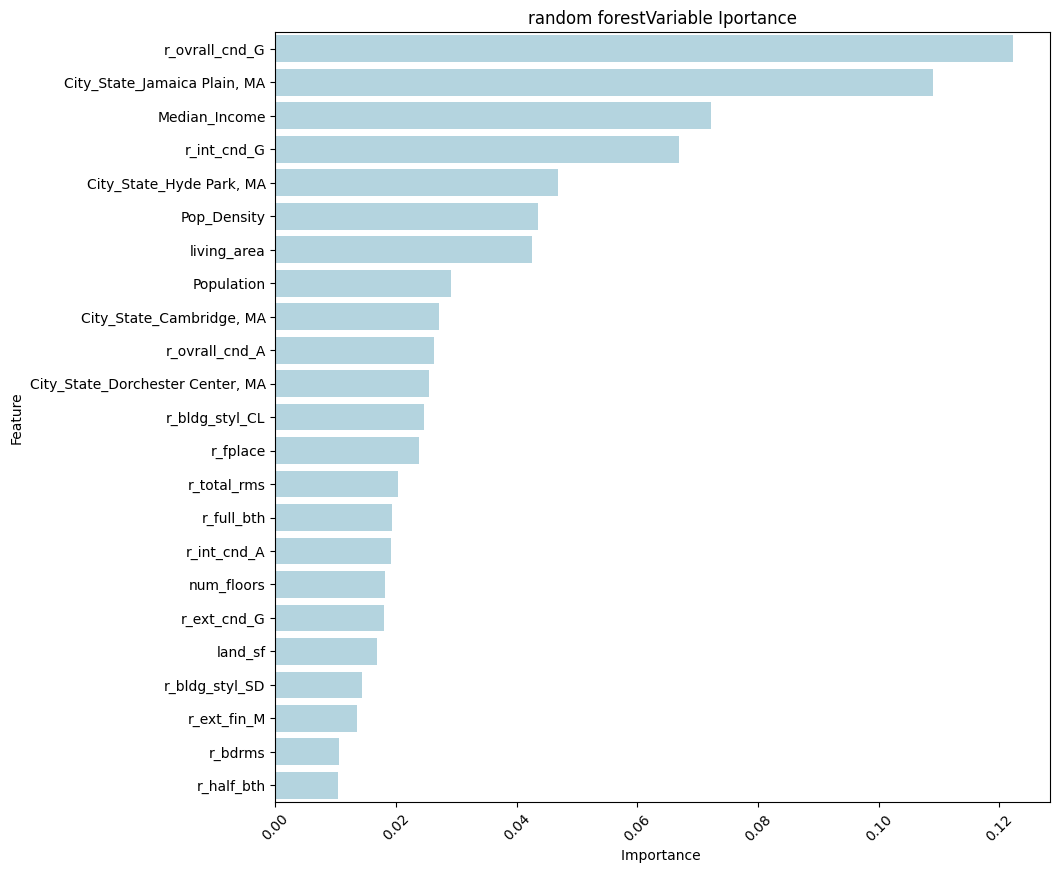
**Description of Bottom 10 worst predictions**

| **Predictor** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| land\_sf | 9742.7 | 12838.88 | 2074 | 4096.25 | 5277 | 9000 | 45522 |
| living\_area | 3457.1 | 2040.62 | 2124 | 2331.25 | 2580 | 3415 | 8623 |
| num\_floors | 2.2 | 0.35 | 2 | 2 | 2 | 2.375 | 3 |
| r\_total\_rms | 9.9 | 3.78 | 6 | 7.25 | 8.5 | 10.75 | 17 |
| r\_bdrms | 4.1 | 1.60 | 2 | 3 | 3.5 | 5 | 7 |
| r\_full\_bth | 2 | 0.82 | 1 | 2 | 2 | 2 | 4 |
| r\_half\_bth | 0.9 | 0.57 | 0 | 1 | 1 | 1 | 2 |
| r\_kitch | 1.1 | 0.32 | 1 | 1 | 1 | 1 | 2 |
| r\_fplace | 0.9 | 1.10 | 0 | 0 | 0.5 | 1.75 | 3 |
| Population | 37186.1 | 5986.57 | 28488 | 35401 | 35401 | 35401 | 47783 |
| Pop\_Density | 11235.9 | 2822.78 | 6207 | 10618 | 10618 | 10618 | 15913 |
| Median\_Income | 68668.2 | 11694.88 | 48841 | 63100 | 75730 | 75730 | 75730 |
| av\_total | 860890 | 237878.23 | 337000 | 720675 | 970650 | 1016100 | 1060100 |
| error | 263827.2625 | 39346.27 | 231036.8125 | 239953.1563 | 255551.375 | 266736.1094 | 366694.375 |

Observation on comparing top 10 predictions to bottom 10 predictions

Variability of the predictors for the top 10 predictions tend to be lower compared to the variability of predictors for the bottom 10 predictions. In addition, houses in the bottom ten predictions has extreme values for living area per square footage, total number of rooms, and total number of bedrooms. The effect of this observation is certain homes warrant differentiated property taxing.

# RECOMMENDATION



From our analysis, the overall condition of the house, the location of the house, population density, living area, and the residential interior of the house has high predictive power in estimating the assessed house value. Houses with good overall condition turn out to have higher assessed values. Likewise, houses with good overall condition and large living area per square footage tend to have higher assessed values. In neighborhoods where the population density is high, home prices tend to be lower.

**Based on our understanding of the results, we recommend the following;**

First, we recommend that the City of Boston adopts the XGBoost model because it reduces the variance of prediction errors of the current model by 13%. In effect, the XGBoost model will improve certainty around property tax revenue forecast to a margin of error of around plus or minus $36,000.

Secondly, we recommend that the City of Boston consider differentiated property tax rates for certain properties with extreme features. For instance, if the blanket property tax rate is 10% of the assessed value, properties with rooms in excess of 10 rooms or living area of more than 2000 area square footage may attract an additional 2% property tax rate. Because these properties have extreme features, their value tends to be underestimated causing the City of Boston to lose tax revenue. By using differential property tax rates, the city may mitigate against such a risk

Lastly, if the City of Boston wants to increase its tax revenues, it may adopt social policies to reduce population density in certain neighborhoods and increase the overall median income of citizens. The city may also give tax incentives to encourage homeowners to own properties with larger living area per square footage.

# Kaggle Submission

Kaggle Name: Elvis Agbenyega

Kaggle reported score: $51,788

Kaggle reported position at the time of submission: #6

(Note: this will change as others post)

<https://www.kaggle.com/t/4b2fe393564a4fd59ce9db748c8f2f3c>