# Executive Summary

*Author: Emma Wang*

## Business Background

Utilizing the feature data we have on the houses, we built machine learning models to predict house values, which will help boost Boston's tax management skills.

## Evaluation Metrics

**RMSE (Root Mean Square Error):** The root of the average squared error between predicted values and observed values. The smaller it is, the more accurate our prediction is, because our predictions will be closer to the true value. If we can find a model with a relatively small RMSE, that means we can assess taxes better, which will improve the tax management ability of the City of Boston. A previous consultant left a model with an RMSE of $57854, and our goal is to beat this record.

**R-square:** We will also compare R-square when we choose the best model. R-square is a number between 0 and 1, which represents the variability of response (tax assessment in this case) that can be explained by our model. Therefore, an R-square closer to 1 represents a relatively better result, if we only consider this aspect.

By choosing those two metrics, we can find our best model which can explain the most variability in house values with the smallest prediction error. Therefore, those two metrics can help us with predicting house prices and further improve the tax management ability of Boston.

## Data dictionary

| Name | Description | Role |
|---|---|---|
| PID | Unique 10-digit parcel number | Unique ID, Rejected |
| ZIP CODE | Zip code of the parcel | |
| OWN_OCC | One-character code indicating if the owner receives residential exemption as an owner-occupied property | |
| AV_TOTAL | The assessed value for the property | Response, target |
| LAND_SF | Parcel's land area in square feet (legal area) | |

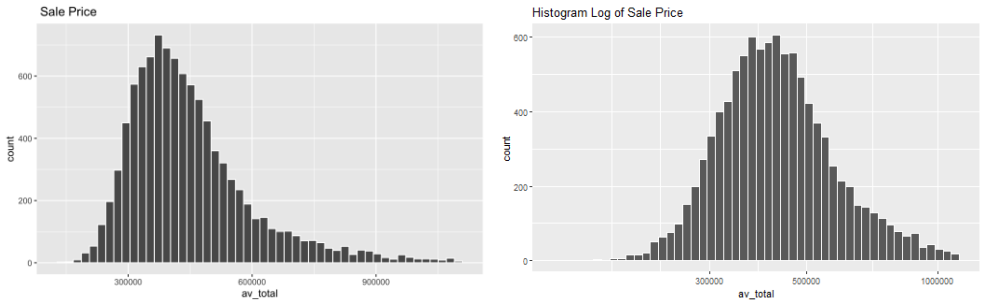
| | | |
|---|---|---|
| YR_BUILT | Year property was built | Used to get house age |
| YR_REMOD | Year property was last remodeled | Used to get house age |
| LIVING_AREA | Living area square footage of the property | |
| NUM_FLOORS | # of levels in the structure located on the parcel | |
| STRUCTURE_CLASS | Structural classification of a commercial building: | |
| R_BLDG_STYL | Residential building style: | |
| R_ROOF_TYP | Structure roof type: | |
| R_EXT_FIN | Structure exterior finish: | |
| R_TOTAL_RMS | Total number of rooms in the structure | |
| R_BDRMS | Total number of bedrooms in the structure | |
| R_FULL_BTH | Total number of full baths in the structure | |
| R_HALF_BTH | Total number of half baths in the structure | |
| R_BTH_STYLE | Residential bath style | |
| 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 : | |
| R_FPLACE | Total number of fireplaces in the structure | |
| R_EXT_CND | Residential exterior condition: | |
| R_OVRALL_CND | Residential overall condition: | |
| R_INT_CND | Residential interior condition: | |
| R_INT_FIN | Residential interior finish: | |
| R_VIEW | Residential view: | |
| POPULATION | Population of people in the ZIP code | |
| POP_DENSITY | People per square mile | |
| MEDIAN_INCOME | Median Income of the residence of that zip code | |
| City_State | City Name and State | |
| Home_age | Calculated using the year of built and remodeled | |

Above showing all the data we have on hand, and we excluded pid, which is the unique identifier of all data points. Other than that, we used the year of building and year of remodeling to calculate a new variable called "home_age", which represents the number of years after building or the last remodeling. Av_total is our target variable, i.e., which we are trying to

predict, and it represents assessed home value.

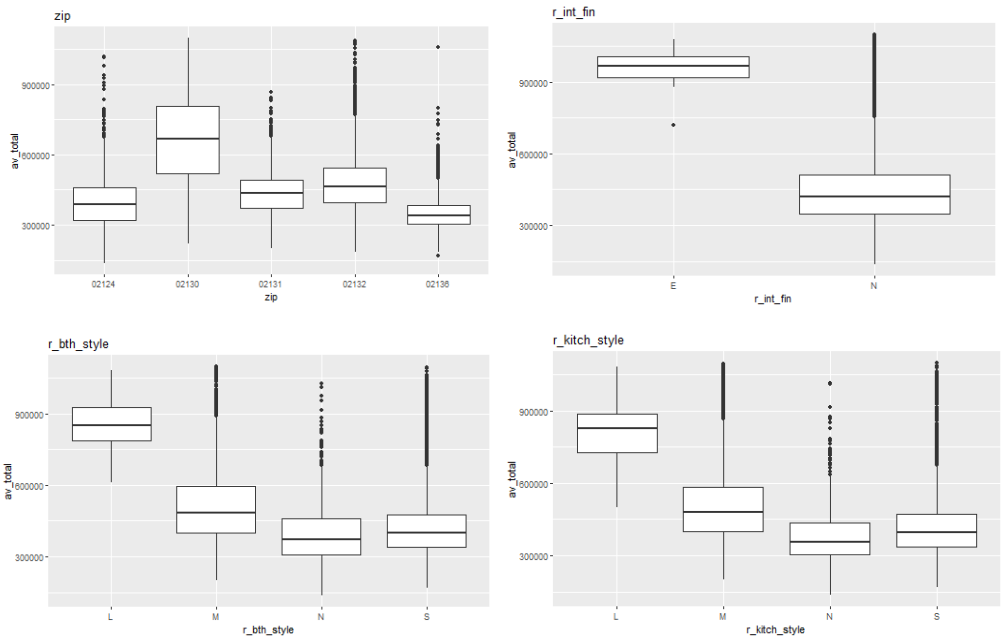
## Supporting Exploratory Analysis

- Target Exploration



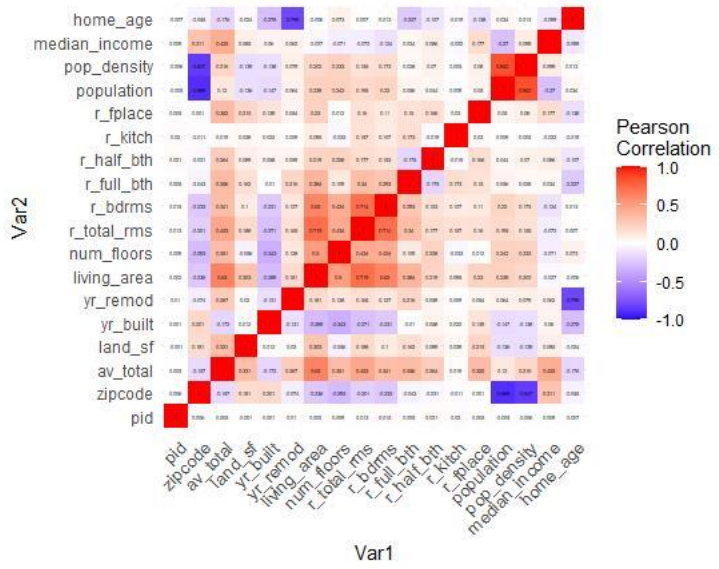

From above, we can see that the house value is right-skewed, more data is concentrated on the left side of the chart.

- Categorical Predictors Exploration




We selected four categorical variables shown above. We can see that the distribution of assessed value is different among different categories in each graph, meaning all of those could be powerful predictors of house prices.

- Numeric Variables Exploration




We created a correlation matrix as shown above, from which we can see that the living area, the number of total rooms, and median income have a stronger relationship with our target variable, which indicates they might be useful predictors.

## Methodology

- Data partitioning
  i. Split the data into 70/30 train/test split using random sampling
  ii. Split the training data into 5 groups
- Data preprocessing
  i. Formula
    - We incorporated all variables into models except pid, which is the unique identifier for houses, and zipcode, which is the same as zip.
  ii. Numeric Predictor Pre-Processing
    - Replaced missing numeric variables with the median
    - Centered and scaled numeric predictors to have a mean of 0 and standard deviation of 1
  iii. Categorical Predictor Pre-Processing
    - Replaced missing categorical variables with unknown
    - Assigned a previously unseen factor level to a new value

- Dummy encoded categories with 1s and 0s

- Model Specification
  i. Train a Linear Regression Model
  ii. Train a Random Forest Model combined with GRID search
    - First we use GRID search to find a relatively good combination of parameters of random forest within a given range.
    - Random Forest is a **bagging method** used to improve the model's performance.
    - Bagging algorithm fits multiple models on different sub-datasets of a training set, at the same time by **performing parallel work**, and it will combine the predictions from all models to generate its final prediction.
  iii. Train an XGBoost Model combined with Bayes tuning.
    - First we use Bayes tuning to find the parameter combination of XGBoost model.
    - XGBoost is a **boosting method** that also can be used to improve the model's performance.
    - Boosting algorithm is an ensemble technique in which the predictors **are trained sequentially**. Each model of the subset will learn from the "mistake" of the previous predictors.

## Model Metrics and Evaluation

| | R-square | RMSE | MAE |
|---|---|---|---|
| Linear Regression | 0.8271 | 61546.24 | 44062.79 |
| Random Forest | 0.8530 | 58352.98 | 42061.29 |
| XGBoost | 0.8706 | 53276.95 | 37811.19 |

Judging by the metrics above, we choose the XGBoost model to be our final one, since it has the highest R-square 0.8706, (which means 87.06% of the variability in house value can be explained by this model), the lowest RMSE, (which means the root of the average squared error is the lowest among all models), and the lowest MAE, (which means the absolute error between the generated value by this model and the true value is also the smallest). In a word, through all these three metrics, we can see that this model interprets and predict house value better and more accurately, with the smallest error.

## Top & Bottom 10 Predictions

- Top 10 best predictions

| yr_built | num_floors | city_state | r_full_bth | r_total_rms | abs_error |
|---|---|---|---|---|---|
| 1945 | 1.5 | Hyde Park, MA | 2 | 6 | 25.8125 |
| 1924 | 2 | Cambridge, MA | 2 | 6 | 32.84375 |
| 1940 | 1.5 | Cambridge, MA | 1 | 6 | 57.28125 |
| 1935 | 2 | Dorchester Center, MA | 1 | 7 | 76.8125 |
| 1945 | 1.5 | Hyde Park, MA | 1 | 7 | 80.84375 |
| 1968 | 2 | Hyde Park, MA | 1 | 7 | 86.9375 |
| 1987 | 1 | Hyde Park, MA | 2 | 6 | 99.4375 |
| 1915 | 2 | Dorchester Center, MA | 1 | 10 | 124.78125 |
| 1950 | 1 | Hyde Park, MA | 1 | 6 | 156.1875 |
| 1920 | 2 | Cambridge, MA | 1 | 8 | 226.03125 |

The above shows the best 10 predictions we have. We can see that most of them are built between 1920-1950, with less than or equal to 2 floors, 1 full bathroom, and the number of total rooms is below 7. Other than that, 40% of them are located in Hyde Park.

- The most overestimated predictions

| yr_built | num_floors | r_bldg_styl | r_total_rms | r_full_bth | city_state | error |
|---|---|---|---|---|---|---|
| 1900 | 2 | CL | 7 | 3 | Jamaica Plain, MA | 374096.625 |
| 1890 | 2.5 | CL | 10 | 1 | Jamaica Plain, MA | 349724.9375 |
| 1910 | 2.5 | CL | 8 | 2 | Jamaica Plain, MA | 312543.375 |
| 1890 | 2.5 | CL | 8 | 2 | Jamaica Plain, MA | 281452 |
| 1909 | 2.5 | VT | 11 | 2 | Dorchester Center, MA | 276433.4375 |
| 1915 | 2 | CL | 9 | 2 | Jamaica Plain, MA | 252864.6875 |
| 1900 | 2 | CL | 11 | 2 | Jamaica Plain, MA | 247581.25 |
| 1910 | 2 | CL | 8 | 2 | Jamaica Plain, MA | 237350.4375 |
| 1900 | 2 | CL | 10 | 2 | Dorchester Center, MA | 221511.8125 |
| 1899 | 2 | VT | 8 | 2 | Dorchester Center, MA | 218561.625 |

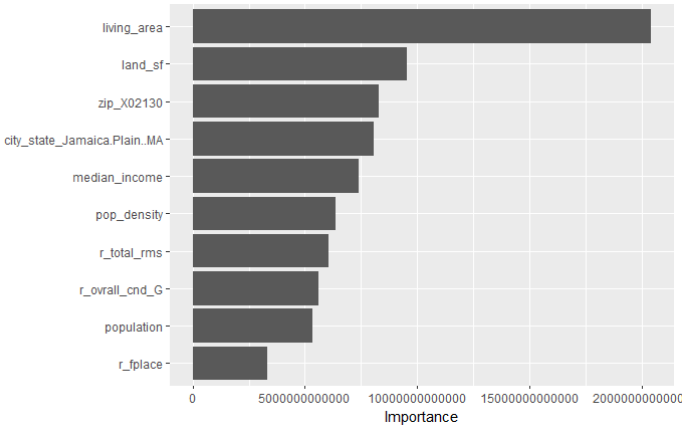
The above shows the most overestimated 10 predictions we have. We can see that most of them are built before 1910, with at least 2 floors, at least 2 bathrooms, and the number of total rooms is above 8. Other than that, all of them are located in Jamaica Plain or Dorchester Center, especially Jamaica Plain, which accounts for 70%. Besides, 80% of them have a building style of Colonial, out of the 16 unique styles we have on record.

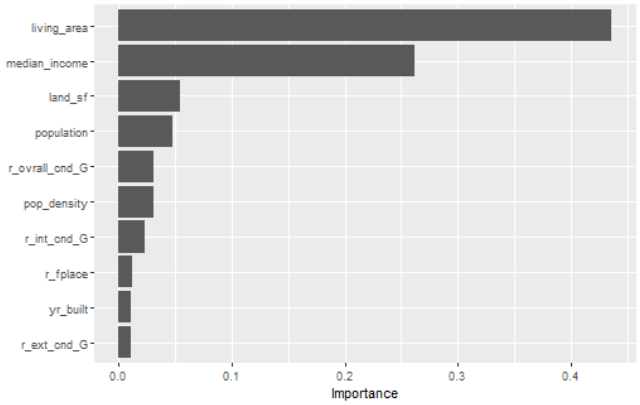
- The most underestimated predictions

| r_bldg_styl | city_state | r_ext_fin | error |
|---|---|---|---|
| CL | Jamaica Plain, MA | W | 284168.8 |
| CL | Jamaica Plain, MA | F | 282901.3 |
| RM | Jamaica Plain, MA | B | 263174.3 |
| SD | Jamaica Plain, MA | F | 236156.2 |
| CL | Jamaica Plain, MA | F | 235090.4 |
| CL | Dorchester Center, MA | M | 205314 |
| CL | Dorchester Center, MA | F | 201513.6 |
| CL | Jamaica Plain, MA | B | 197493.6 |
| CP | Jamaica Plain, MA | W | 193912.1 |
| RN | Cambridge, MA | W | 190262.8 |

The above shows the most underestimated 10 predictions we have. We can see that most of them are located in Jamaica Plain or Dorchester Center, especially Jamaica Plain, which accounts for 70%. Besides, 60% of them have a building style of Colonial, out of the 16 unique styles we have on record. And the exterior finish is more varies compared to our best results and most overestimated results which are mostly vinyl.

## Key Insights



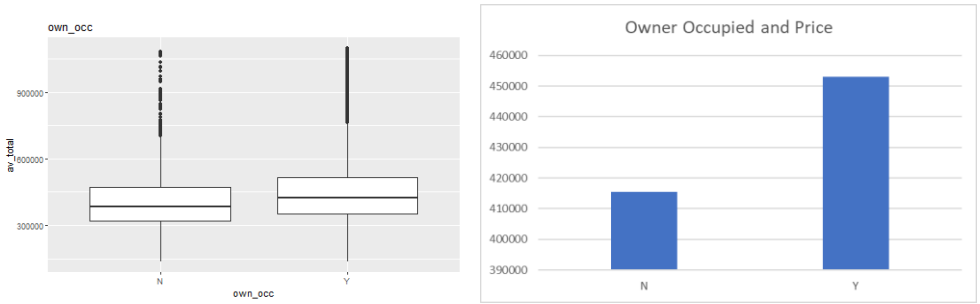

The above shows the top 10 important predictors from our random forest model and XGBoost model. Combine with all the analysis we conducted above, we have the major findings as follows:

- City-related features (population, median income, etc.) have a significant influence on house prices. In both charts, city-related features account for about 50%.
- Except for city-related features, the variables related to house price most closely are living area and land area.
- Whether the house is owner occupied, the year of building and remodeling all have a significant influence on our houses' prices. (Further analysis is provided down below)

- Zip code 002130 seems to have the highest house price, which is associated with Jamaica Plain.
- Jamaica Plain possesses the most overestimated predictions and most underestimated predictions at the same time.
- Houses in Hyde Park seems to fit our model the best.

## Key Business Questions

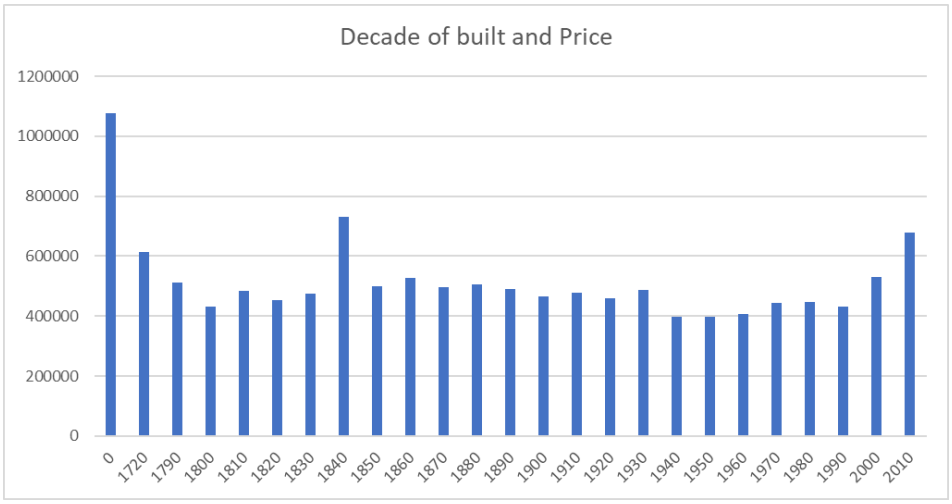
- **Owner-occupied houses have higher assessed value.**




Based on the chart above, we can see that the median value, first and third quantile value of owner-occupied houses is higher than those not occupied. Based on the calculation, the average value of houses is higher for those occupied by owners ($452990 compared to $415437).

Other than that, we also performed additional analysis by putting the owner-occupied variable into the regression model and the p-value of own_occ_N is less than 0.001 with an estimate of -35125.26. Therefore, if we only consider this variable and all other conditions are the same, the houses with owner-occupied will be $35000 more expensive than those without.

- **Houses built in the 1990s do not seem to have a higher value.**

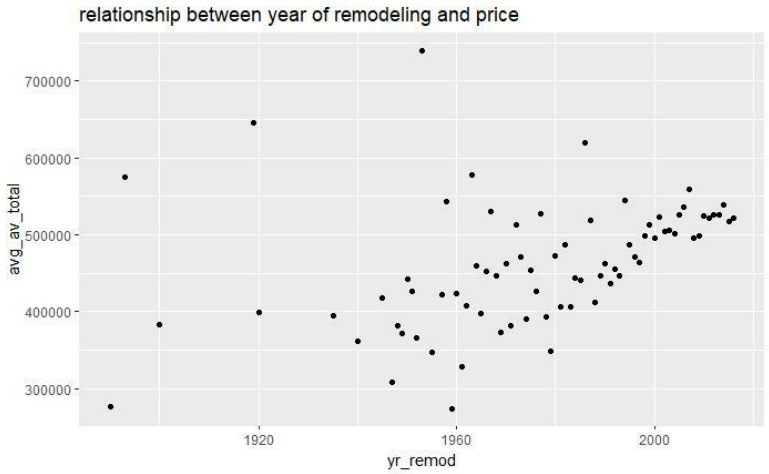



Based on the chart above, we can see that the houses built in the 1990s do not

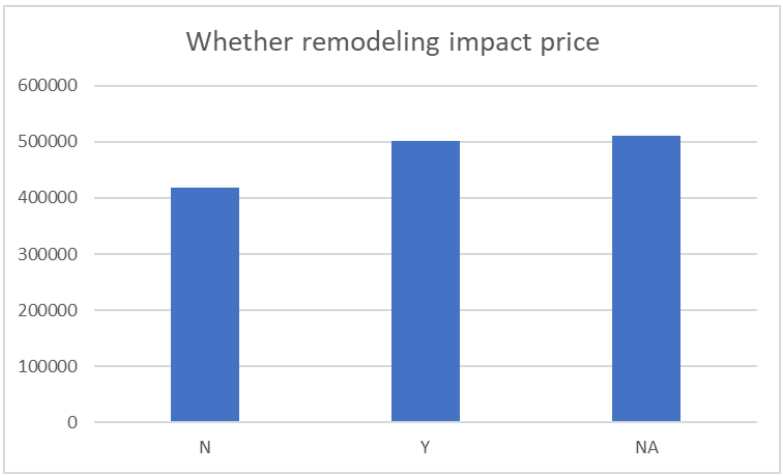
have a higher value than many other decades.

Moreover, in our linear regression model, the yr_built has a negative estimate of -7235.44 and the p-value is smaller than 0.001. That analysis proves that, in general, the house price and year of building are negatively related.

- **Homes that have been recently remodeled tend to have higher home values.**




The above shows the relationship between the year of modeling and the home price for all the houses which have done modeling. We can see the trend that the closer the houses are modeled, the higher the home price.




Whether modeling in general has an impact on price? Yes, based on the chart above, the remodeled houses have a higher average total value.

## Actionable Recommendations

- There's obviously something unique about Jamaica Plain, we should investigate, analyze, and collect appropriate data about it. After that, we can separate it from other cities and build a different model to predict the value better.
- From a practical view, encourage the citizens to remodel the houses so they will be worth

more money, which will increase the City of Boston's tax revenue. Especially houses built before 1910, which tend to be overestimated, we should conduct research on those houses' features and encourage citizens owning those houses to remodel them.

- Besides remodeling, encouraging citizens to decorate their houses with luxury styles can also increase the value of the house which will ultimately increase the government's tax revenue.
- By utilizing other analytical skills such as clustering, we might be able to define "luxury" houses, i.e., those with luxury decorations and larger living areas. By doing so, we can perform different predictive analyses on "luxury houses" and "normal houses", which will further improve our predictability, since the houses with a higher number of floors, higher number of total rooms, houses with luxury kitchens or bathrooms styles seem to fit our model worse.

*Kaggle Submission:*
*Kaggle Name: Emma Wang (Jiawen)*
*Kaggle reported a score: of 52129.47*
*Kaggle reported the position at the time of submission: #4*