Challenge 3 Report

1. Introduction

1.1. Problem Statement

Boston property tax has established a model with an RMSE of 57854 to predict assessed values of property in Boston. The primary purpose of this project is to try to establish better models (with lower RMSE) by means of linear regression, random forest, and XGBoosting, because the government is very concerned about taxing house prices. In addition, to find better hyperparameters for these models, Bayes tuning is employed.

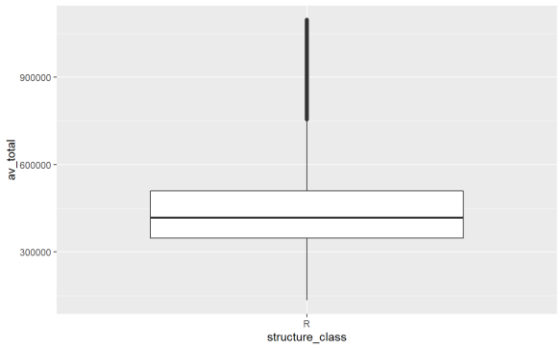
At the same time, three key variables, 'If owner receives residential exemption as an owner-occupied property', 'Year property was built', 'Year property was last remodeled', are required to be paid more attention by Boston, because they asserted that these variables are strong related to the assessed value of houses.

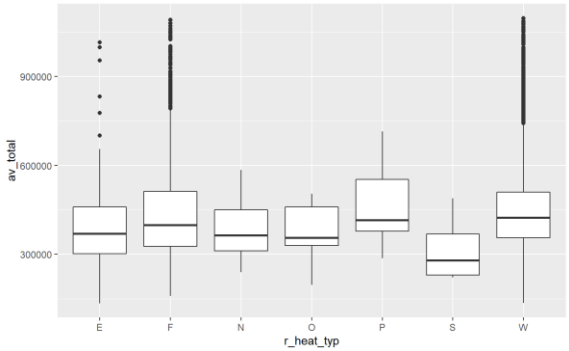
1.2. Data dictionary

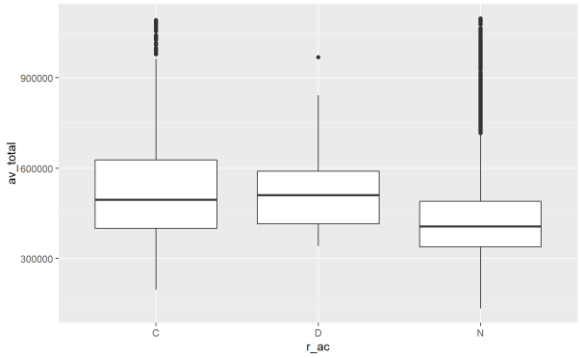
Name	Description	Length
PID	Unique 10-digit parcel number	10
ZIPCODE	Zip code of parcel	5
OWN_OCC	One-character code indicating if owner receives residential exemption as an owner-occupied property	1
AV_TOTAL	**Assessed value for property i.e. what you are predicting**	**13**
LAND_SF	Parcel's land area in square feet (legal area)	6
YR_BUILT	Year property was built	4
YR_REMOD	Year property was last remodeled	4
LIVING_AREA	Living area square footage of the property	8
NUM_FLOORS	# of levels in the structure located on the parcel	10
STRUCTURE_CLASS	Structural classification of commercial building:	1
	A Struct Steel / **C** Brick/Concrete / **E** Metal	
	B Reinforced Concrete / **D** Wood/Frame / **R** Residential	
R_BLDG_STYL	Residential building style:	10
	BL Bi-Level / **DX** Duplex / **SL** Split Level	
	BW Bungalow / **L** Tri-Level / **TF** Two-Family Stack	

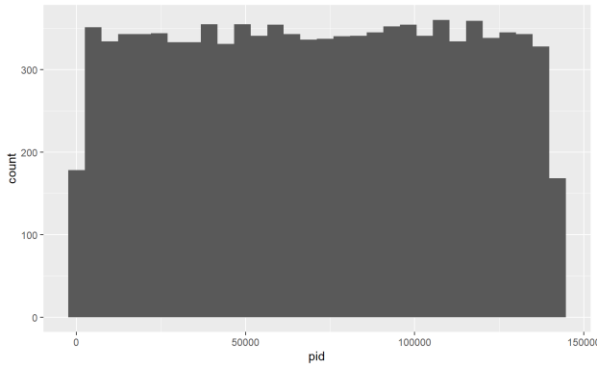
	CL	Colonial	**OT**	Other	**TD**	Tudor	
	CN	Contemporary	**RE**	Row End	**SD**	Semi-Detached	
	CP	Cape	**RM**	Row Middle	**VT**	Victorian	
	CV	Conventional	**RN**	Ranch			
	DK	Decker	**RR**	Raised Ranch			
R_ROOF_TYP	Structure roof type:						10
	F	Flat	**L**	Gambrel	**S**	Shed	
	G	Gable	**M**	Mansard			
	H	Hip	**O**	Other			
R_EXT_FIN	Structure exterior finish:						10
	A	Asbestos	**K**	Concrete	**U**	Aluminum	
	B	Brick/Stone	**M**	Vinyl	**V**	Brick/Stone Veneer	
	C	Cement Board	**O**	Other	**W**	Wood Shake	
	F	Frame/Clapboard	**P**	Asphalt			
	G	Glass	**S**	Stucco			
R_TOTAL_RMS	Total number of rooms in the structure						10
R_BDRMS	Total number of bedrooms in the structure						10
R_FULL_BTH	Total number of full baths in the structure						10
R_HALF_BTH	Total number of half baths in the structure						10
R_BTH_STYLE	Residential bath style						1
	L	Luxury	**M**	Modern			
	N	No Remodeling	**S**	Semi-Modern			
R_KITCH	Total number of kitchens in the structure						10
R_KITCH_STYLE	Residential kitchen style:						1
	L	Luxury	**M**	Modern			
	N	No Remodeling	**S**	Semi-Modern			
R_HEAT_TYP	Structure heat type:						10
	E	Electric	**O**	Other	**W**	Hot Water	
	F	Forced Air	**P**	Heat Pump			
	N	None	**S**	Space Heater			
R_AC	Indicates if the structure has air conditioning (A/C):						1
	C	Central A/C	**D**	Ductless A/C	**N**	None	
R_FPLACE	Total number of fireplaces in the structure						10

R_EXT_CND	Residential exterior condition:						1
	A	Average	**E**	Excellent	**F**	Fair	
	G	Good	**P**	Poor			
R_OVRALL_CND	Residential overall condition:						1
	A	Average	**E**	Excellent	**F**	Fair	
	G	Good	**P**	Poor			
R_INT_CND	Residential interior condition:						1
	A	Average	**E**	Excellent	**F**	Fair	
	G	Good	**P**	Poor			
R_INT_FIN	Residential interior finish:						1
	E	Elaborate	**N**	Normal	**S**	Substandard	
R_VIEW	Residential view:						1
	A	Average	**E**	Excellent	**F**	Fair	
	G	Good	**P**	Poor	**S**	Special	









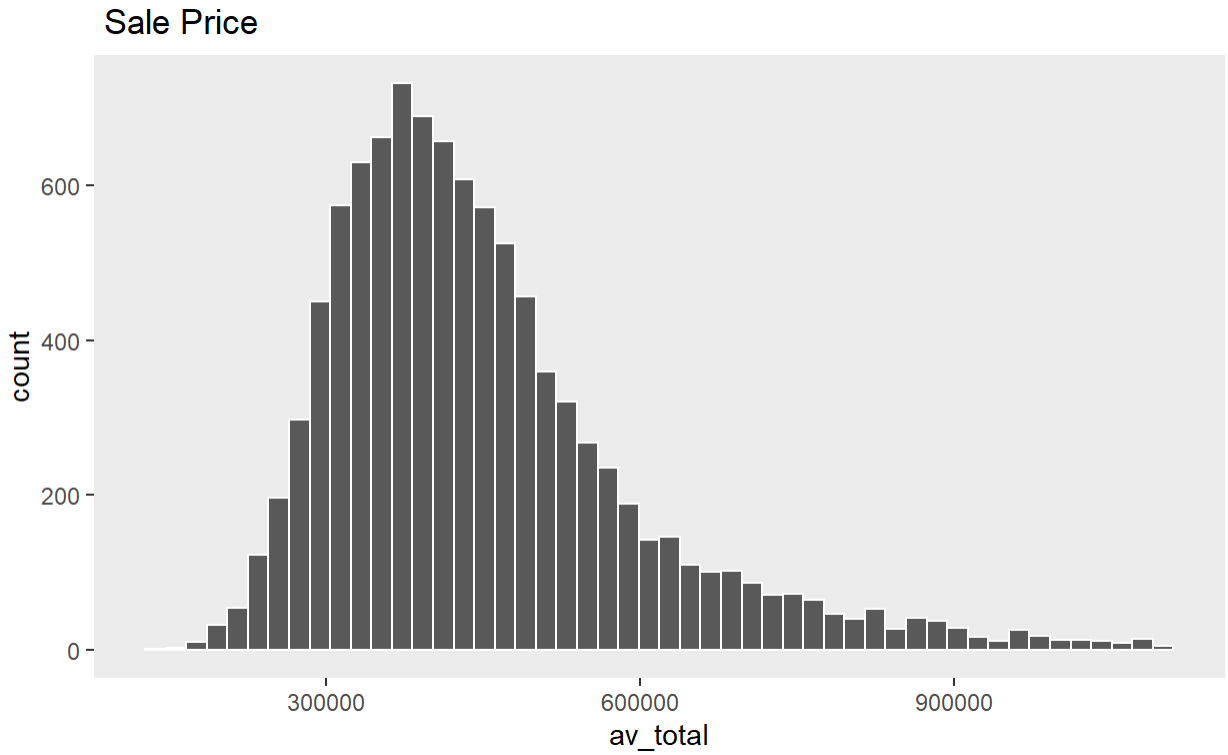
By plotting variables, we found that STRUCTURE_CLASS only has one type and different levels of R_HEAT_TYPE and R_AC have no obvious difference. Therefore, they are excluded, and PID as an identifier was removed as well.

In addition, HOME_AGE was created by subtracting the year the house was built or the year the house was renovated from 2022. This data transformation can be regarded as a new variable to make predictions.

```
boston_transform = boston %>%
  select(!zip )  %>%
  mutate(home_age = if_else(yr_remod > yr_built,2022 - yr_remod,2022 -
yr_built))
```

2. Exploratory analysis

2.1. Exploring target variable



The graph shows us that the data is right-skewed; There are more cheap houses than expensive houses. When modeling this type of result, a strong argument can be made that the price should be logarithmically transformed.

3. Methodology

3.1. Preparing data

3.1.1. Partitioning data into 70/30

3.1.2. Define the recipe for the following models

3.2. Establishing models

3.2.1. Establishing linear regression model and evaluating it

3.2.2. Establishing random forest model and evaluating it

3.2.3. Establishing the XGBoosting model

3.2.3.1. K-fold

```
# Kfold cross validation
kfold_splits <- vfold_cv(train, v=5)
```

K-Folds cross-validator provides train/test indices to split data into train/test sets. Split the dataset into 5 consecutive folds (without shuffling by default).
Each fold is then used once as validation while the 4 remaining folds form the training set.

3.2.3.2. Bayes tuning is used to perform hyperparameter tuning

```
xgb_search_res <- xgb_wflow %>%
  tune_bayes(
    resamples = kfold_splits,
    # Generate five at semi-random to start
    initial = 5,
    iter = 50,
    # How to measure performance?
    metrics = metric_set(rmse, rsq),
    control = control_bayes(no_improve = 5, verbose = TRUE)
  )
```

The tuning algorithm optimizes its parameter selection in each round based on the results of the previous round, which is the primary distinction between Bayesian search and the other approaches. As a result, the algorithm optimizes the choice rather than picking the following set of parameters at random, and it probably gets to the optimum parameter set faster than the previous two techniques. This strategy eliminates the ranges that are most likely not going to produce the best solution and only selects the relevant search space. As a result, it can be useful when you have a lot of data, slow learning, and want to shorten tuning time. 'Initial' is the number of initial results greater than the number of parameters being optimized, which is set as 5. And the max iteration number is 50.

4. Evaluation Model and interpretation

4.1. Metrics

The core of regression model evaluation is to use the difference between the predicted value of the model and the true value. At the same time, because regression models need to estimate continuous variables rather than categorical variables, we use the following metrics to evaluate regression models.

- **MAE:** The MAE stands for the mean of the absolute differences between the dataset's actual and anticipated values. It calculates the dataset's residuals' average.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

- **RMSE:** Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- **R-squared:** The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

4.2. Model performance and explanation

Model	Partition	RMSE	RSQ	MAE
Model 1	train	63385.38	0.8156457	45185.31
Model 2	train	50685.73	0.8935507	37926.59
Model 3	train	49592.50	0.8872969	36508.86

Model	Partition	RMSE	RSQ	MAE
Model 1	test	62621.40	0.8212370	44733.87
Model 2	test	60026.17	0.8466678	43028.21
Model 3	test	52859.15	0.8629171	38935.94

Less RMSE and MAE indicate better models, and higher RSQ means better models. However, generally speaking, each metric has its own weakness, so all three metrics should be considered simultaneously. Model 3 has both the lowest RMSE and MAE, despite not having the highest RSQ. Overall, model 3 is regarded as the best prediction model, which is generated from XGBoosting model.

The RMSE (the most important indicator) is obtained after rooting the MSE, and the root operation makes the error value consistent with the unit of the target variable. For the target variable, house price, the unit of RMSE is house price, maintaining the consistency of dimensions. The RMSE of the testing dataset of model 3 is about **53,000**, which means that the average error between predicted values and true values is about 53000. At the same time, the average house price is **448,564** and the standard deviation is **147,761.** Comparatively, this RMSE is acceptable and less than the benchmark, **57,854.**

4.3. Discussion about the difference between Random Forest & XGBoosting

In machine learning, a random forest is a classifier that contains multiple decision trees (the core is bagging technology), and its output categories are determined by the mode of the categories output by individual trees. The subsets of samples each form a learner, and they are uncorrelated during training.

XGBoost is one of the boosting algorithms. The idea of Boosting algorithm is to integrate many weak classifiers together to form a strong classifier. Unlike bagging, each training of this algorithm is based on the previous results.

Each has its own advantages and disadvantages because of the different ways of processing samples and learners.

The goal of Random Forest is to reduce variance, so it has strong stability and is simpler to operate than boosting algorithms. But when dealing with regression models, it is prone to overfitting.

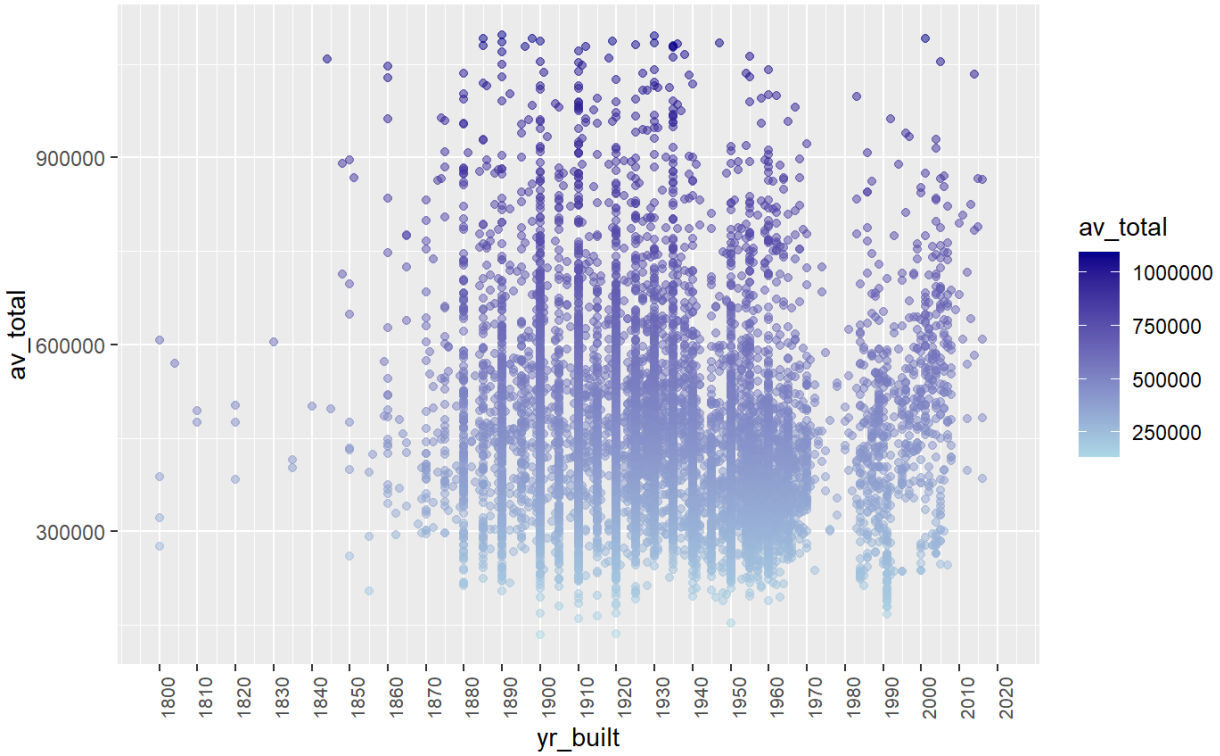
The XGBoosting algorithm can guarantee higher accuracy, but one of its problems is a large amount of computation and is time-consuming. At the same time, the boosting algorithm is also more likely to cause overfitting problems. Overfitting means that a model can get a better fit on the training data than other hypotheses but does not fit the data well on the data set outside the training data (because some features are overly believed to be useful for prediction.).

It can also be seen from our evaluation metrics that the RMSE of XGBoosting is lower. So, if a high-accuracy model is required, XGBoosting is considered to be a better model compared to random forest.

5. Discussion

5.1. Discussion about 3 key variables

```
Coefficients: (4 not defined because of singularities)
                              Estimate    Std. Error t value             Pr(>|t|)
(Intercept)               62952370.7957  1293494.3932  48.668 < 0.0000000000000002 ***
zipcode                     -29459.8916      607.2808 -48.511 < 0.0000000000000002 ***
own_occY                     12038.1618     2220.3203   5.422   0.0000000609298904 ***
land_sf                          7.3893        0.2654  27.845 < 0.0000000000000002 ***
yr_built                      -107.7209       23.2283  -4.637   0.0000035896175267 ***
yr_remod                         2.1443        0.8649   2.479             0.013193 *
living_area                     97.8722        2.3604  41.464 < 0.0000000000000002 ***
num_floors                 -18562.0252     4517.6777  -4.109   0.0000402252079379 ***
```



According to the City of Boston, properties that are owner-occupied were built in the 1990s, or have recently undergone remodeling typically have higher assessed values.

Based on the analysis of linear regression, the variable OWN_OCC is strongly associated with assessed house price (p-value is closer to 0) and YR_REMOD is comparatively related to the target variable (p-value is smaller than 0.05). In addition, the coefficients are both positive, which means that the houses that are owner-occupied and were recently remodeled have higher assessed prices.

According to the scatterplot, the prices of houses that were built between 1990 and 2000 are not higher than in other decades obviously. Therefore, the property built in the 1990s has no higher assessed values.

5.2. Discussion about the top 10 and bottom 10 predictions

The best predictions are selected by identifying 10 rows with the least absolute difference between predictors and real values.

A tibble: 10 × 32

population <dbl>	pop_density <dbl>	median_income <dbl>	city_state <fctr>	home_age <dbl>	error <dbl>	abs_error <dbl>
29826	11505	66735	Roslindale, MA	103	-16.62500	16.62500
29826	11505	66735	Roslindale, MA	20	-23.40625	23.40625
36314	13251	75446	Cambridge, MA	95	-26.53125	26.53125
28488	6207	58890	Hyde Park, MA	87	-88.43750	88.43750
29826	11505	66735	Roslindale, MA	122	108.40625	108.40625
36314	13251	75446	Cambridge, MA	72	-135.28125	135.28125
29826	11505	66735	Roslindale, MA	98	-172.09375	172.09375
28488	6207	58890	Hyde Park, MA	102	198.34375	198.34375
29826	11505	66735	Roslindale, MA	21	-218.56250	218.56250
29826	11505	66735	Roslindale, MA	14	-241.78125	241.78125

The worst predictions are selected by identifying 10 rows with the largest absolute difference between predictors and real values.

A tibble: 10 × 31

r_view <fctr>	population <dbl>	pop_density <dbl>	median_income <dbl>	city_state <fctr>	error <dbl>	abs_error <dbl>
A	47783	15913	48841	Dorchester Center, MA	403408.1	403408.1
A	29826	11505	66735	Roslindale, MA	310786.2	310786.2
A	35401	10618	75730	Jamaica Plain, MA	284320.3	284320.3
A	47783	15913	48841	Dorchester Center, MA	-278189.9	278189.9
A	35401	10618	75730	Jamaica Plain, MA	272230.9	272230.9
A	28488	6207	58890	Hyde Park, MA	271527.1	271527.1
A	35401	10618	75730	Jamaica Plain, MA	-268029.2	268029.2
A	36314	13251	75446	Cambridge, MA	257228.4	257228.4
A	35401	10618	75730	Jamaica Plain, MA	256675.2	256675.2
G	35401	10618	75730	Jamaica Plain, MA	256446.1	256446.1

5.3. Discussion about the top predictors

Variable <chr>	Importance <dbl>
living_area	0.38192596143
median_income	0.33553870613
land_sf	0.07126653591
r_ovrall_cnd_G	0.05819177894
r_int_cnd_G	0.04124157242
r_fplace	0.02349533803
yr_built	0.01140438763
r_kitch_style_M	0.00847436049
r_bldg_styl_CL	0.00831113716
r_ext_cnd_G	0.00803311656

The above predictors are regarded as the most important variables to predict assessed prices. Among them, living area, the median income of a place, Total number of fireplaces in the structure, GOOD condition of property, MODERN style of kitchen have positive relationship with prices of property. And year built has a negative relationship with target variable. Therefore, the property that has larger areas, are built in cities with a higher median income of citizens, have more fireplaces, have good condition, has modern kitchens, and is built recently can have higher assessed prices.

6. Recommendations

In order to predict Boston housing prices more accurately, the following suggestions are provided:

- The boosting algorithm should be used, especially the XGBoosting used in this project, compared to the bagging algorithm, because this type of algorithm guarantees more accuracy.
- From the aspect of the condition of the house, the living area, overall condition, interior decoration conditions, the number of fireplaces, the age of the house, and the kitchen style of the house are all important predictable house price indicators.
- From the geographical location of the house, the income of the population in the city is an important predictor.

7. Kaggle Submission

Kaggle Name: Charlene Wei

Kaggle reported score: 52823.43387

Kaggle reported position at time of submission: #52

(Note: this will change as others post)

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