Boston Housing one more time.

# Regression

Yes, this is like our final in 6005 but with some added twists! You recently joined Boston Consulting Group, and you have been hired to clean up a mess left by a previous consultant for the city of Boston. They built a model with an RMSE of $ 57854 the City of Boston is considering awarding your firm a large consulting contract if you can beat a RMSE of predictions on property tax assessments of $ 57854. To-do this you’ll need to analyze and build some models to assess and predict the av\_total (assessed value) of properties in the greater Boston area.

Your task is to train and compare a Linear Regression, Random Forest and XGBoost model. You must use either GRID search or RANDOM search PLUS k-fold cross validation to tune model hyperparameters. Your report should compare their performance to find the most effective model at predicting AV\_TOTAL. The City of Boston believes that owner-occupied homes have a higher assessed value, homes built in the 1990s, and homes that have been recently remodeled tend to have higher home values.

You have been provided three datasets

* Boston Zips
* Boston Training
* Boston Holdout

In addition to your model report (executive summary) the city would like to understand the differences between Random Forest and XGboost. They don’t quite understand machine learning so your explanation of your model methodology and steps that you take are important for them to understand. You need to explain why you chose a specific metric to evaluate your models with and what it means to the city of Boston.

## Deliverables for Grading

For this project, you will complete and submit the following.

1. **Executive Summary** – like all our previous projects you need an executive summary, key findings, assessment of assertions, methodology, evaluation and most importantly recommendations.
2. **Notebook** – I’ll want your notebook used to produce results in the report. Your code should be appropriately commented so I can tell what is going on – ideally, I should be able to repeat your analysis.
3. **Predictions file** – We want you to apply your model to the boston\_holdout.csv. You will submit this file to Kaggle to see if you beat the benchmark of $57854.

## Required Tasks

You will need to write a Model Report(executive summary) which will contain two major sections an Executive Summary and Detailed Analysis. And you will need to submit your predictions to Kaggle

## Executive Summary (20 points)

A good executive summary will leave the reader with **actionable** takeaways that they can remember and regurgitate at the next meeting. Your executive summary should be just that a ***summary***. What problem are you challenged with? What were 3 or 4 key findings (things you found interesting that influenced the model). What was result of your model, and 2-3 recommendations that you’d make.

* State the problem
* 3 or more Key findings
* Methodology Model tuning analysis
* Model Performance & Interpretation of it.
* 3 or more actionable Recommendations
* Table of top 10 predictions and bottom 10 predictions
  + What can you identify in common with each other?

Helpful hint: do not attempt to draft an executive summary until after you’ve done the analysis.

**Grading:**

This project is worth a total of 25 points

* Executive Summary is worth 20 points
* Kaggle up to 5 points, RMSE above $57854 is worth 3
* Plus, we want your code

Your grade will be based on the following:

* Rigorous application and documentation of the machine learning process.
* Clear, concise, and accurate discussion of the results.
* Clear explanation of how the results can be used to address the business problem at hand
* Beating the benchmark Random Forest model RMSE $57854.

### Kaggle Instructions

Use your WFU email, register on Kaggle, the URL below is a link to the competition.

<https://www.kaggle.com/t/0a304425a50744e9a0a23fe6c3b6e0a2>

**Housing 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 |  |

PLUS

A file(zip.csv) of zip code demographics that may or may-not be predictive of Assess Value.

|  |  |
| --- | --- |
| Name | Description |
|
| ZIP | ZIP CODE – should join to ZIPCODE |
| 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** |