Capstone 3 Project

Michael Ruston

Intro

My Capstone 3 project used real estate listing data from kaggle.com.

Two goals for this project:

-Predict property price using regression modeling, and

-Create a dashboard on Tableau Public

EDA and Data Cleaning

Exploratory Data Analysis

Read in the .csv file.

There are 12 columns and over 500,000 rows.

The ‘status’ columns is over 99% ‘for_sale’ and a small portion are ‘ready_to_build’.

The ‘ready_to_build’ rows are dropped and then this column is no longer needed.

The ‘state’ column is examined and only the top 8 states are retained.

The ‘state’ column now contains just:

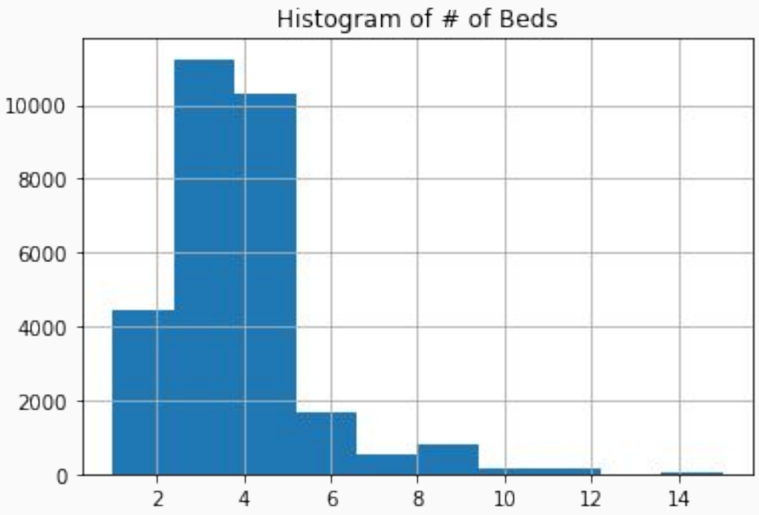
'Massachusetts’, 'Connecticut', 'New Hampshire', 'Vermont', 'Maine', 'New York', 'Rhode Island', 'New Jersey'

Exploratory Data Analysis

Minor cleaning is done on the ‘city’ column.

The ‘bed’ column is examined. Only rows with bed < 16 are retained in order to remove apartment buildings from the analysis.

This resulted in the following distribution for the ‘bed’ column:



Exploratory Data Analysis

The ‘house_size’ column is explored.

Some rows with erroneous values of house_size are removed from the data set.

Duplicate rows are removed

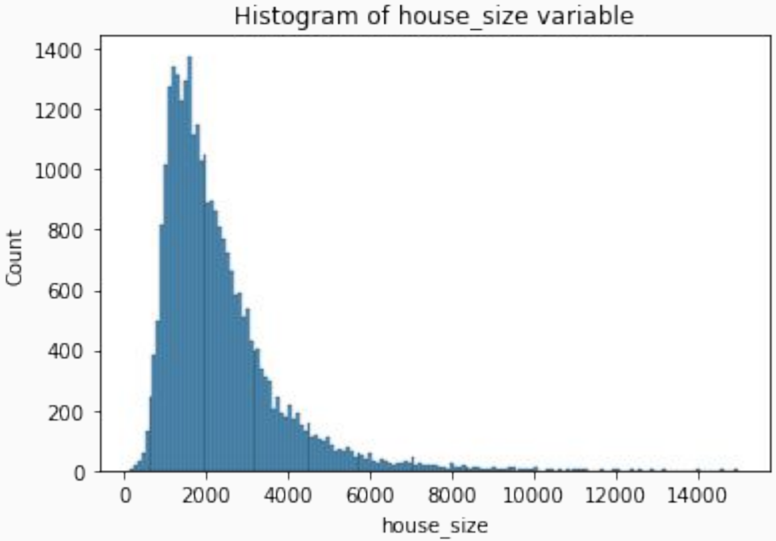
One row with a missing zip code is removed

Exploratory Data Analysis

A ‘prop_type’ column is created with default value of ‘house’ and a value of ‘apt’ if the full_address contains ‘unit’ or ‘apt’.

Next, data is restricted to rows with house_size < 15000 to remove the outlier-sized houses

Most houses are under 6000 (sq ft.).



Exploratory Data Analysis

Some rows have a 0 for the ‘acre_lot’ variable. This is replaced with the next lowest value in the data: 0.01.

Some 123 rows having a missing value for ‘bath’. These rows are dropped.

A ‘sold_flag’ column is created based on whether or not the ‘sold_date’ column is missing.

Exploratory Data Analysis

Some rows have duplicate values for ‘full_address’.

Looks like there could be some errors in the data here, so the rows with duplicate ‘full_address’ values are removed.

After EDA and data cleaning, there are some 29,075 rows of data.

Modeling

Two Models

Two models were created:

-Linear Regression

-Random Forest Regression

Linear Regression RMSE on Training Data: $1,363,000

Random Forest RMSE on Training Data: $405,500

Linear Regression RMSE on Test Data: $1,030,000

Random Forest RMSE on Test Data: $955,900

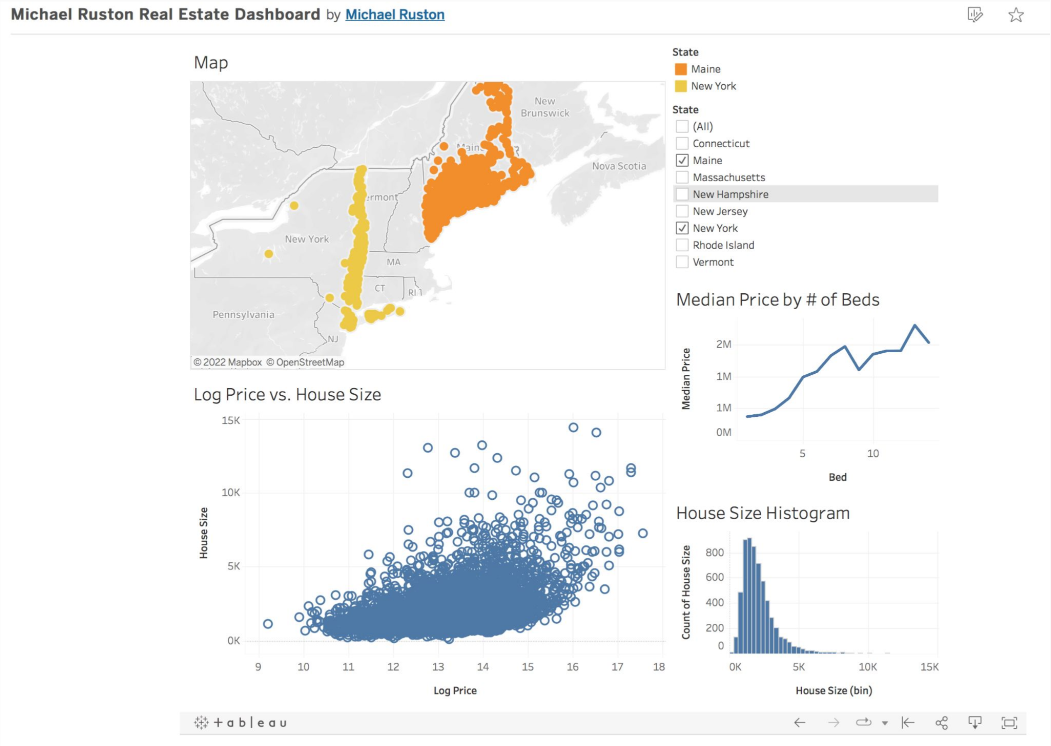
The Random Forest model has lower RMSE on the Test Data, but results indicate overfitting of the Random Forest model.

Dashboard

Dashboard

A dashboard was created using this data on Tableau Public.

Dashboard Link:
https://public.tableau.com/app/profile/michael.ruston/viz/MichaelRustonRealEstateDashboard/Dashboard1



Michael Ruston Real Estate Dashboard by Michael Ruston

State
🟧 Maine
🟨 New York

State
□ (All)
□ Connecticut
☑ Maine
□ Massachusetts
□ New Hampshire
□ New Jersey
☑ New York
□ Rhode Island
□ Vermont

Thanks!

Contact us:

Michael Ruston

michaelruston@gmail.com