Predictive model : Residence sale price in District of Columbia region.

Author : Pavan Kumar Kulkarni

Date : September 23, 2018

1. Background and problem statement.

Importance of housing market - According to NAHB housing market contributes 15-18% of GDP. For most individuals and families their house is the biggest asset, source of wealth and savings. A healthy housing market if of paramount importance in overall health of economy. Many economist attribute housing bubble as the catalyst for 2008-10 great recession.

Is it possible to build machine learning model which can predict the residential sale price using the attributes made available to public by DC government via these datasets?

Computer Assisted Mass Appraisal (CAMA) database. The dataset contains attribution on housing characteristics for residential properties, and was created as part of the DC Geographic Information System (DC GIS) for the DC Office of the Chief Technology Officer (OCTO) and participating D.C. government agencies.

2. Potential clients.

Anyone who is interested in residential property sale can use the model built. This includes

- **Residential agents**. - To estimate the sale price to get maximum value for their clients.
- **Homeowners** - Who are looking to sell residential properties. Getting know estimated sale price will help get maximum price and timely sale.
- **Home buyers** - Who are in market looking for residences. Estimated sale price based house attributes will help in zeroing on reasonable budget and expected residence attributes.
- **Public offices/Government agencies** -
  who need residential properties price for planning and making public policies.

3. Datasets.

The Dataset is made available to public by DC government.
http://opendata.dc.gov/datasets/computer-assisted-mass-appraisal-residential.

The dataset has around 107K records with 39 attributes.

4. Approach:

Language chosen to complete this project was R as it has many libraries and support for machine learning and statistics. R Markdown file format is used as it supports dynamic documents with R. These are fully reproducible.R Markdown document is written in markdown (an easy-to-write plain text format) and contains chunks of embedded R code.

It helps both creator and user of the application to concentrate on one document for both code and comments/explanations.

In this project also the associated RMD file and output in HTML file format has detailed commentary for each module of the code, graphs and results output along with actual R code.

In this section, a summary of project and reasoning for some of the decisions made in the project will be listed.

The project has below sections.

- 0.1 Introduction. - Background of the project.
- 0.2 Set up. -
    - All the libraries used in the project.
    - Reusable functions - Mainly for graphs/plots.
    - Read data.
- 0.3 Exploratory Data Analysis.
    - Each attribute/group of attributes are analysed via visualization to understand the distributions and outliers.
    - Tread missing values.
    - Grouping the values and coding.
- 0.4 Feature Engineering.
    - Create derived attributes such as age, remodelled, total bathrooms etc.
    - Near zero variance column treatment.
    - Converting to factors.

    - Dummy variable creations and correlations.
- 0.5 Prepare the dataset.
    - Outliers, scale the attributes.
    - Split dataset to train and test.
- 0.6 Run models.
    - Run various regression models such as linear regression, random forest, regularised regression and gradient boosting.
    - Perform cross validation to get best hyper parameters.
    - Compare the performance of each of the models on RMSE and R^2
- 0.7 Final Model.

    Based on runtime, RMSE, R^2 Lasso is the best model. However the residual vs fitted plot has complimentary profiles compared to random forest. Lasso has overestimate at lower price range and underestimate at higher price range. Random Forest has opposite i.e underestimate at lower price and over estimate at higher price range. Hence it is decided to create an ensemble model where 60% weight is given Lasso and 40% to random forest model.

    The final model has 84% R^2 on test data.

5. Next steps :

Possibly the extensions to the project can be

1. Create API interface using Plumber package where user can input data and get estimates.
2. User can get explanations as to 'why' this estimate using Lime package.
3. Include other regions in USA and housing market is very region specific also.