DeepRE.ml DATATHON

Challenge Description

October 9, 2018

1 Overview

The city of Boston conducts an annual tax assessment of all the properties in the Boston jurisdiction. As the city has to send officers to almost 200,000 properties every year, they are wondering whether automation can help them assess properties cheaper and faster, or even predict growth in certain areas or buildings, so they don't have to send an officer there.

Thus, they are asking for your Data Science help. The assignment is provide suggestions and methodologies to help drive better business decisions using data analysis.

2 Technical Details

2.1 Datasets

You are given 11-years-worth of tax assessments of Boston properties. Each year, roughly 160,000-170,000 properties are assessed for tax purposes in Boston, adding up to $\sim$ 1.8M entries in total provided. The datasets are accessible upon agreeing to the official terms and conditions of the competition via this link. (*Please note that we kindly ask each team member, should teams be formed, to review and agree with the terms of the competition.*)

We have provided 2 *versions of the data for you*:

1. Pre-cleaned datasets ('train_clean', 'test_clean') for each year with unified column names, data types, imputed missing values across all 11 years. Missing values were imputed using attribute-specific and common-sense techniques, such as k-means-based imputation, mean and most-frequent (for categorical types) imputation, and placeholder '0' indicating the given attribute is not applicable for the type of property at hand.

2. Basic datasets ('train_raw', 'test_raw'), where only unification of column names and data types were performed. This version contains missing values, should you be inclined to impute them yourself or if you are not satisfied with the way they were imputed for you in the preprocessing steps.

You will be by no means penalized for the choice of the dataset. The purpose of this challenge is not data cleaning, and this is why the pre-processing step was done for you.

You also have a **data key** for all 54 attributes (plus 'LatLong' standing for Latitude/-Longitude of the property) at your disposal as well as **property occupancy codes**.

Additionally, there are many additional APIs and open data sources you could leverage, including other real estate APIs or demographics and economic indicators (such as Census, FRED, etc). There is a big caveat to this discussed in subsection 2.2.

Each property has a unique identifier Parcel ID ("PID"). About 40,000 unique properties were selected at random from 2018 (as well as the previous years) and held out as a test set. This challenge is about predicting value, thus column **'AV_TOTAL' representing the total assessed tax value (dependent variable to be predicted)** as well as significant confounding variables ('GROSS_TAX', 'AV_BLDG', 'AV_LAND') were removed from the test set and will be used for model performance evaluation on DeepRe.ml side (*with focus on the most recent tax year - 2018*).

2.2 Important rules to note

We allow groups of 3 at maximum to collaborate on the assignment. You can look for your team in the DeepRe.ml Slack channel.

This dataset can be found in pieces on various open-source government websites. You are welcome to explore various outside data sources, however, ***you must not*** download and submit the true tax assessed value for PIDs that could in theory to be found online - this would be considered cheating. We require you to submit your code with the report, so we would find this out anyway. Just a side note: in real life, most likely you won't be so lucky to be able to overfit to the test set.

2.3 Deliverables

There will be two final deliverables for this datathon.

1. Short PDF report - MAX 4 pages of text (excluding space taken up by visualizations/tables) explaining your findings, methodology and recommendations for Boston city officials how to improve their tax valuation using data science.

2. Code base, properly commented, showing your work.*

*Note: Without the code base, we won't be able to consider you for the prize money award.

2.4 Deadline

Submissions need to be submitted via email to deepredatathon@gmail.com with a copy to mrh932@g.harvard.edu by **October 17, 2018, 11:59PM**. No submissions after this deadline will be considered.

2.5 Evaluation

Prediction of the tax-assessed value 'AV_TOTAL' for 2018 *for the held-out test set* is the main benchmark for the competition, for which the test data can be found in 'test_clean/test_clean_2018.csv' (or in 'test_raw/test_raw_2018.csv' should you choose to perform your own data imputation). We encourage you to hold out a validation set out of your training set for accuracy testing purposes.

The metric used for evaluation was chosen to be root mean square error (RMSE) defined below:

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

where y_i represents true 'AV_TOTAL' assessed tax value, $\hat{y}_i$ the assessed tax value based on your estimator for property i and N is the number of test observations.

In general, however, **more emphasis** is given on finding **creative ways how to identify growth in value** in either sub-areas or properties themselves. Thus, creative ways to identify valuation growth potential going above and beyond will receive extra attention.

After the submission deadline, the submitted solutions will be reviewed and winner (or the winning team) announced within 5 business days. The prize will be then paid out to the winner and the winning solution will be described to all of the competition participants to maximize the learning experience. Finally, DeepRE.ml will reach out to authors of select additional promising solutions regarding future work possibilities.

2.6 Helpful tips:

There are many ways you can approach this problem and be successful.

Below are some (non-exhaustive) suggestions:

- Time series analysis of similar properties and valuation trends
- Geospatial analysis in the year 2018 without the temporal component
  - Area (e.g., census tract) valuation growth prospects
  - Property level valuation growth prospects

Any general/clarifying questions can be answered in the general section of the DeepRE.ml Slack channel, which is accessible here.

Good luck and have fun!