Zillow Data Analysis

Martin Skarzynski

October 14, 2017

1 Introduction

The Zillow Prize is a Kaggle competition that aims to inspire data scientists around the world to improve the accuracy of the Zillow "Zestimate" statistical and machine learning models.

My goal is to compete for the Zillow prize and write up my results.

2 Methods

2.1 Data

The data were obtained from Kaggle website and consist of the following files:

- properties_2016.csv.zip
- properties_2017.csv.zip
- sample_submission.csv
- train_2016_v2.csv.zip
- train_2017.csv.zip
- zillow_data_dictionary.xlsx

The `zillow_data_dictionary.xlsx` is a code book that explains the data. This data will be made available on figshare to provide an additional source if the Kaggle site data become unavailable.

2.2 Analysis

Data analysis was done in Jupyter Notebook (formerly known as IPython Notebook) [1] Integrated Development Environment using the Python language [2] and a number of software packages:

- NumPy [3]
- Pandas [4]
- Scikit-learn (Pedregosa et al. 2011)

2.3 Visualization

The following packages were used to visualize the data:

- Matplotlib [5]

- Seaborn
- r-ggplot2
- r-cowplot

The use of `R` code and packages in a `Python` environment is possible through the use of the `Rpy2` package.

2.4 Prediction

Machine learning prediction was done using the following packages:

- scikit-learn (Pedregosa et al. 2011)
- xgboost
- r-caret

2.4.1 Reproducibility

Reproducibility is extremely important in scientific research yet many examples of problematic studies exist in the literature (Couzin-Frankel 2010)[6].

The names and versions of each package used herein are listed in the accompanying `env.yml` file in the `config` folder. The computational environment used to analyze the data can be recreated using this `env.yml` file and the `conda` package and environment manager available as part of the Anaconda distribution of Python.

Additionally, details on how to setup a Docker image capable of running the analysis is included in the `README.md` file in the `config` folder.

The code in the form of a jupyter notebook (`01_zillow_MWS.ipynb`) or Python script (`01_zillow_MWS.py`), can also be run on the Kaggle website (this requires logging in with a username and password).

More information on the details of how this project was created and the computational environment was configured can be found in the accompanying `README.md` file.

This Python 3 environment comes with many helpful analytics libraries installed It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python (a modified version of this docker image will be made available as part of my project to ensure reproducibility). For example, here's several helpful packages to load in

3 Results

There are several columns which have a very high proportion of missing values. I will remove features that have more than 80% missing values.

3.1 Feature Importance by Random Forest and Xgboost

4 Conclusions

In Progress

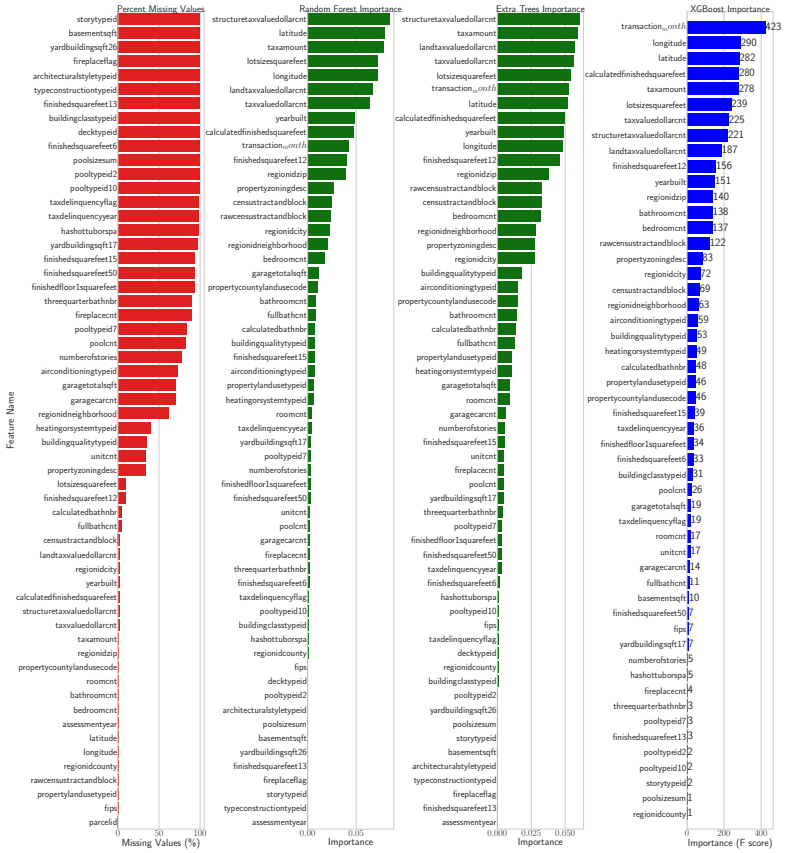

Figure 1. Missing Values and Importance: Criteria for removing Features

References

[1] F. Pérez and B. E. Granger. Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9:2129, 2007.

[2] F. Pérez, B. E. Granger, and J. D. Hunter. Python: an ecosystem for scientific computing. *Computing in Science & Engineering*, 13:1321, 2011.

[3] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13:2230, 2011.

[4] W. McKinney. Data structures for statistical computing in python, 2010.

[5] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9:9095, 2007.

[6] J. Couzin-Frankel. Cancer research. as questions grow, duke halts trials, launches investigation. *Science*, 329:6145, 2010.