Zillow Data Analysis

Martin Skarzynski

October 25, 2017

1 Introduction

1.1 Zillow Prize

The Zillow Prize is a data science and machine learning competition organized by the Zillow real estate company. The competition is host by Kaggle, a company that provides datasets and computational kernels for data science challenges. Kaggle was purchased by Google earlier this year. The Zillow Prize competition has recently declared to be the second largest data science challenge on Kaggle. The purpose of the Zillow prize is to inspire data scientists around the world to work on improving the accuracy of the Zillow "Zestimate" [**?**] home price estimate algorithm.

1.2 Zestimate algorithm

Zillow's Zestimate home valuation was first released in 2006 [**?**] and has since had a major impact on the United States real estate industry. The Zestimate algorithm created a new standard of providing free, publicly available housing data and home price estimates. The Zestimate algorithm relies on "7.5 million statistical and machine learning models" that have been refined over the years to have a margin of error on only 5%.

The real estate industry is a major contributor to the U.S economy and was worth roughly 30 trillion dollars in 2016. Homeownership equity is a major form of wealth that Americans hold, while mortgages are a major type of American private debt. It is extremely important for U.S. homeowners and for the U.S. economy that home prices are estimated correctly. Without proper estimates, lender and borrowers cannot confidently monitor their assets and liabilities. Overevaluation of real estate assets and the potential returns on real estate debt for lenders has been linked to the great recession. [**?**]

1.3 My Goal

My goal in working with the Zillow Prize data was to answer the question of which features in the dataset are the most important determinants of the major evaluation metric for the Zillow Prize. This metric is called "logerror" in the data and is defined as the difference between the log of Zestimate price and the log of the actual sales price. To achieve this goal, I cleaned the provided housing valuation data and employed two different machine learning methods, random forest and xgboost, to calculate importance scores for each of the features in the cleaned dataset.

2 Methods

2.1 Data

The dataset provided for the Zillow challenge consisted of 58 features. I first determined the percent representation of missing for each of these features (Figure 1). I then used median imputation to fill in the missing values for the remaining features. Next, I used two prediction methods: Random Forest and XGBoost to obtain importance scores for each of the remaining features (Figure 2). I also compared the Random Forest importance scores to the importance scores obtained using the Extra Trees Regression method (Supplemental Figure 1). I made a final list of features by removing any features that had an importance of less than 0.001 from Random Forest model or an XGBoost F score importance of less than 10. Of the initial 58 features, now 39 features remained in final list.

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) [**?**] Integrated Development Environment using the Python language [**?**] and a number of software packages:

- NumPy [**?**]
- Pandas [**?**]
- Scikit-learn [**?**]

2.3 Visualization

The following packages were used to visualize the data:

- Matplotlib [**?**]
- Seaborn [**?**]

2.4 Prediction

Machine learning prediction was done using the following packages:

- scikit-learn (Pedregosa et al. 2011)
- xgboost [**?**]

2.5 Reproducibility

Reproducibility is extremely important in scientific research yet many examples of problematic studies exist in the literature [**?**].

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

3.1 Missing values

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

3.2 Feature Importance by Random Forest and Xgboost

```
/Users/marskar/anaconda3/lib/python3.6/site-packages/sklearn/cross_validation.py:41: Deprecati
  "This module will be removed in 0.20.", DeprecationWarning)
```

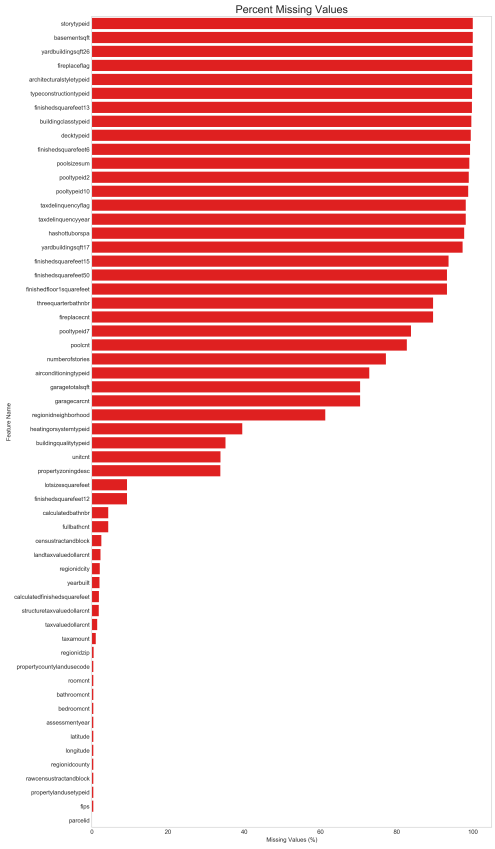
```
Out[20]: (2985217, 58)
```

```
Out[21]: (90275, 3)
```

```
Out[28]: 39
```

4 Conclusions

I used the


Percent Missing Values
storytypeid
basementsqft
yardbuildingsqft26
fireplaceflag
architecturalstyletypeid
typeconstructiontypeid
finishedsquarefeet13
buildingclasstypeid
decktypeid
finishedsquarefeet6
poolsizesum
pooltypeid2
pooltypeid10
taxdelinquencyflag
taxdelinquencyyear
hashottuborspa
yardbuildingsqft17
finishedsquarefeet15
finishedsquarefeet50
finishedfloor1squarefeet
threequarterbathnbr
fireplacecnt
pooltypeid7
poolcnt
numberofstories
airconditioningtypeid
garagetotalsqft
garagecarcnt
regionidneighborhood
heatingorsystemtypeid
buildingqualitytypeid
unitcnt
propertyzoningdesc
lotsizesquarefeet
finishedsquarefeet12
calculatedbathnbr
fullbathcnt
censustractandblock
landtaxvaluedollarcnt
regionidcity
yearbuilt
calculatedfinishedsquarefeet
structuretaxvaluedollarcnt
taxvaluedollarcnt
taxamount
regionidzip
propertycountylandusecode
roomcnt
bathroomcnt
bedroomcnt
assessmentyear
latitude
longitude
regionidcounty
rawcensustractandblock
propertylandusetypeid
fips
parcelid
Feature Name
0
20
40
60
80
100
Missing Values (%)

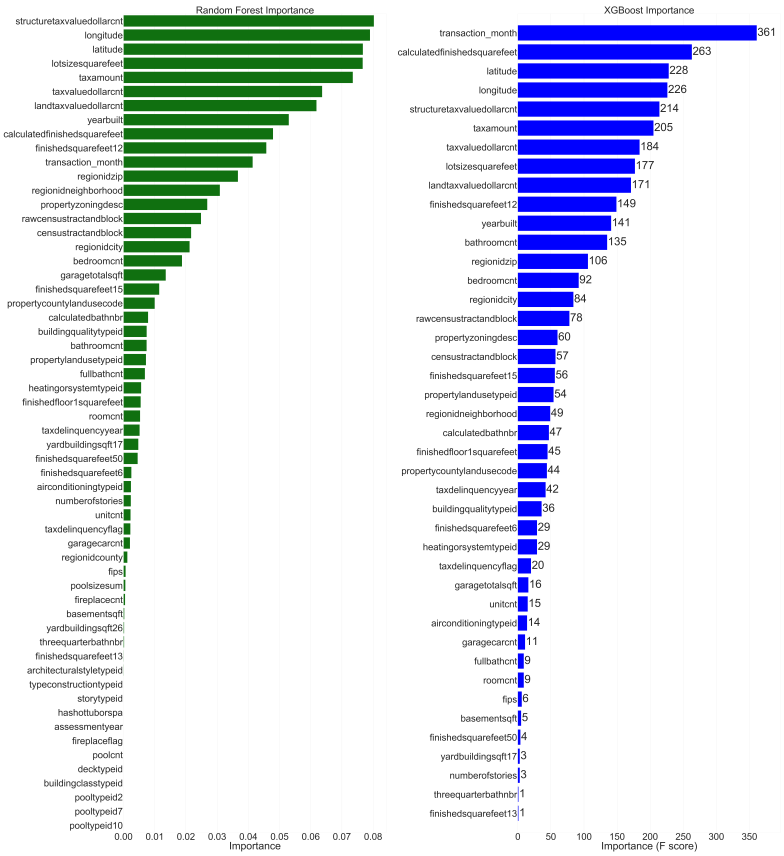


Figure 1. Feature Importance. Criteria for removing Features