zillow_writeup

September 21, 2017

1 Zillow Challenge

1.1 Primary objective

This is an ipnyb file that describes my efforts to put together a decent entry in the Zillow challenge. This is my **primary objective**.

1.2 Secondary objectives

After signing up for the challenge, I decided to try things out in **BOTH** R and Python. This lead me to outline several secondary objectives:

1. Compare Machine Learning in Python and R
2. Compare JupyterLab and RStudio as IDEs for
3. Compare Pandas and dplyr packages
4. Compare Scikit-Learn and Caret packages

The Secondary Objectives are outlined in Table 1 below.

1.3 Secondary objectives

After signing up for the challenge, I decided to try things out in **BOTH** R and Python. This lead me to outline several secondary objectives:

1. Compare Machine Learning in Python and R
2. Compare JupyterLab and RStudio as IDEs for
3. Compare Pandas and dplyr packages
4. Compare Scikit-Learn and Caret packages

The Secondary Objectives are outlined in Table 1 below.

1.3.1 Table 1.

|Subject
Comparison Area

Language	Python	R
IDE	JupyterLab	RStudio

File type	ipynb	R Markdown
Wrangling	Pandas	dplyr
Machine Learning	Scikit-Learn	Caret

1.4 Zillow Prize Final Project

This document describes my plan for the Advanced Data Science I (140.711.01) final project.

For my project, I will compete for the Zillow prize and write up my results.

The data provided for the challenge are described at Zillow prize site.

The challenge entails training a machine learning algorithm that can beat Zillow's proprietary Zestimate at predicting home values.

1.5 First steps

1. Create Kaggle and GitHub accounts
2. Create a GitHub repo for the Advanced Data Science I (140.711.01) final project.
3. Download the data files and put in the repo
4. Add the data files to .gitignore except for zillow_data_dictionary.xlsx
5. Split the training data into training and test sets.
6. Try different algorithms using the caret package and measure performance
7. Select the top algorithm(s)
8. Assess opportunities to improve performance of the top algorithm(s)

1.6 Last week

Last week I made a Kaggle account and downloaded the data files. I added the data files to .gitignore except for zillow_data_dictionary.xlsx, which is a useful code book that explains the data.

1.7 Next steps

My next task is to explore the data, figure out what to do about missing data, and split the data into training and test sets.

Specifically, I plan to split the data into two groups randomly, where 2/3 of the data will be used for training and 1/3 will be used for testing. I am not sure if I want to set up a cross-validation experiment. Perhaps a 10-fold or 5 * 2 cross-validation.