Zillow®

1301 Second Avenue
Floor 31
Seattle, WA 98101
Phone (206) 470-7000
Fax (206) 470-7001

Exercise for Data Scientist

ZILLOW CONFIDENTIAL
Please delete this document and all associated files that may accompany it once you have completed the exercises. Thank you.

Exercise 1: Modeling Home Values

Objective

Construct a valuation model to predict a home's current market value and score a validation data set with this model. You should return to Zillow an R or Python script file that upon execution will, without error, read the training data, train relevant model(s), read the validation data, score the validation with the model(s), and output relevant diagnostics. You should also provide an explanation for your choice of model.

Details

1. The code should be written in R or Python.
2. The code should read the training data from the `training_ZILLOW_CONFIDENTIAL.csv` file that is provided and build a predictive model to estimate the value of the field "transvalue."
3. The code should apply the model derived from the training data to the validation data contained in the `validation_ZILLOW_CONFIDENTIAL.csv` that is provided. All records in the validation data should be scored by the model regardless of the presence of missing values on any of the variables in the model.
4. The code should produce relevant diagnostics about the performance of the model on the validation data (i.e., a comparison of "transvalue" in the validation set to the predicted value of this field).
5. You may use any R/Python contributed package. Your code should load any necessary packages/modules.
6. Both training and validation data are transactions from King County, WA.
7. You may use other data you can find from the public domain to improve the accuracy of your model.
8. You should report your results in terms of median absolute error and percent of estimates within 5%, 10% and 20% of the sales price.

Description of fields contained in the training and validation data sets

- `propertyid`: Unique ID for home
- `transdate`: Date of current sale
- `transvalue`: Price of current sale
- `transdate_previous`: Date of previous sale of home (if any)

- `transvalue_previous`: Price of previous sale of home (if any)
- `bathroomcnt`: Number of bathrooms in home
- `bedroomcnt`: Number of bedrooms in home
- `builtyear`: Year home was constructed
- `finishedsquarefeet`: Finished square footage of the home
- `lotsizesquarefeet`: Lot size of property in square feet
- `storycnt`: Number of stories for the home
- `latitude`: Latitude of the home * 1,000,000
- `longitude`: Longitude of the home * 1,000,000
- `usecode`: Type of home (all homes in both training and validation are single-family homes)
- `censustract`: Census tract in which home is located
- `viewtypeid`: Nominal variable indicating the type of view from the home (blank or NULL value indicates no view)

Exercise 2: Next Steps

Objective

How would you extend and improve your solution to modeling home values in exercise 1? Your discussion should include mention of data processing, alternative model/algorithm choices, feature creation and use of additional data. Take into account the cost/benefit tradeoff for your choices, including the ability to scale to the entire country.