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

Objective

Construct a model to predict a home's current market value and score a test data set with this model. We have not provided the actual transaction values in the test data set but will use them to calculate the accuracy of the model you build. The accuracy of this model is an important factor we take into account when making hiring decisions. However, just as important is the quality of the write-up that accompanies your predictions as we are looking to understand how and why you arrived at the solution you submit. Include as much detail as you think is needed to help us understand what information you evaluated during your modeling process and how you arrived at the final model that generated predictions. You should return two things to Zillow: a CSV file of predicted transaction values and an explanation for how you arrived at the final model you chose as a separate PDF or Markdown file.

Details

- You should set-aside about 5-10 hours to complete this exercise but you don't need to complete it all in one go.
- You highly recommend you do modeling in **R or Python** as that is what we use at Zillow
- You may use any library/package available to the public. Please document which ones you used them in your write-up.
- Both training and test data are home sale transactions from 2015 in King County, WA
- You may use other data from the public domain if you feel it might improve your predictions – we have already provided some data from the 2013 American Community Survey. **WARNING**: merging in external data can be very time consuming!
- The code should read the training data from the `training_ZILLOW_CONFIDENTIAL.csv` file that is provided to build a predictive model to estimate the value of the field "SaleDollarCnt."
- The code should apply the model developed via the training data to the test data contained in the `test_ZILLOW_CONFIDENTIAL.csv` that is provide. All records in the test data should be scored by the model regardless of the presence of missing values on any of the variables in the model.
- Please provide a CSV file of ZPIDs and home price predictions in a column named "SaleDollarCnt" for all homes in the test set – we will use these predictions to evaluate the accuracy of the model you built. The file should have the following structure:

```
PropertyID,  SaleDollarCnt
  48649606,          88729
  48649606,         572270
  48649674,         230063
```

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

- We will evaluate the accuracy of your model via the Average Absolute Percent Error (AAPE). This is our preferred internal metric for comparing model performance. We report our Median Absolute Percent Error (MAPE) on Zillow.com nationally and for King County should you want compare accuracies. They can be calculated as:
  - AAPE: $mean\left(\frac{abs(predicted-actual)}{actual}\right)$
  - MAPE: $median\left(\frac{abs(predicted-actual)}{actual}\right)$

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

Description of fields contained in the training and test data sets

- `PropertyID`: Unique ID for home
- `TransDate`: Date of current sale
- `SaleDollarCnt`: Price of current sale
- `BathroomCnt`: Number of bathrooms in home
- `BedroomCnt`: Number of bedrooms in home
- `BuiltYear`: Year home was constructed
- `FinishedSquarefeet`: Finished square footage of the home
- `GarageSquareFeet`: Size of protected garage space if any
- `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 test are single-family homes)
- `ZoneCodeCounty`: The intensity of use or density the lot is legally allowed to be built-up to
- `viewtypeid`: Nominal variable indicating the type of view from the home (blank or NULL value indicates no view)
- `censusblockgroup`: The FIPS code for the census block group this property is located in. You can derive the census tract FIPS by truncating the rightmost digit.
- `BGMedHomeValue`: The median home value in the block group
- `BGMedRent`: The median rent value in the block group
- `BGMedYearBuilt`: The median year structures in the block group were built
- `BGPctOwn`: Percentage of homes that are owner-occupied in the block group
- `BGPctVacant`: Percentage of housing that is vacant in the block group
- `BGMedIncome`: Median income of households residing in the block group
- `BGPctKids`: Percentage of households with children under 18 years present at home
- `BGMedAge`: Median age of residents of the block group