CSE 519 - Data Science Fundamentals
Homework 2: Exploratory Data Analysis in IPython

Charuta Pethe
111424850

1 Baseline Model

I implemented the baseline model as a simple linear regression on the following variables: 'bathroomcnt', 'bedroomcnt', 'calculatedfinishedsquarefeet', 'fips', 'latitude', 'longitude', ‘rawcensustractandblock', 'regionidcounty', 'roomcnt', 'yearbuilt', 'taxvaluedollarcnt', 'logerror'

It gave the following result for a train-test split of 95-5%:

```
Mean Squared Error =  0.0232334388791
          R2 score =  0.00385372119725
```

2 Improvement Efforts

I tried various improvements to the baseline model:

1. I added and removed different variables from the list.
2. I changed the handling of missing values in the training as well as test data, replacing the empty fields with mean or mode as appropriate.
3. I tried k Nearest Neighbors, decision tree, gradient boosting regressor and SGD regressor. However, none of these models gave a better performance than linear regression. The mean squared error was significantly high, and the R2 score was significantly less or even negative.
4. I included the month in transaction date as an attribute in the training set, and used the month in the column header in the test set. However, this gave a bad score on Kaggle.
5. I added columns with squares of some variables, which improved the model performance on a train-test split, but gave a bad score on Kaggle.

3 Improved Model

3.1 Attributes selected:

The following columns are taken as attributes in the training dataset:

'bedroomcnt', 'calculatedfinishedsquarefeet', 'decktypeid', 'airconditioningtypeid', 'roomcnt', 'yearbuilt', 'taxamount', 'calculatedbathnbr', 'fireplacecnt', 'poolcnt', 'rawcensustractandblock', 'unitcnt', 'buildingqualitytypeid', 'fullbathcnt', 'latitude', 'longitude', 'regionidcounty', 'logerror'

3.2 Handling of missing values:

Attributes having missing values in both training and test set:

Attribute	Handling
Unit count	replaced with mode
Building quality type ID	replaced with mean
Deck type ID	replaced with zero, since only one other value exists
Calculated bath number	replaced with mean
Full bath count	replaced with mean
Fireplace count	replaced with zero, since only one other value exists
AC type ID	replaced with mode
Pool count	replaced with zero (assuming that empty field means no pool)

Attributes having missing values in the test set:

Bedroom count	replaced with mode
Total area	replaced with mean
Room count	replaced with mode
Year built	replaced with mode
Tax amount	replaced with mean
Raw census tract and block	replaced with mean
Latitude	replaced with mean
Longitude	replaced with mean
Region ID county	replaced with mode

3.3 Handling of categorical variables

The three variables 'decktypeid', 'airconditioningtypeid' and 'regionidcounty' are categorical. Dummy variables are generated for each of these, and the first one is dropped for better model performance.

3.4 Prediction

The model performs linear regression to predict test values. Regression coefficients of dummy columns of categorical variables are the highest.

3.5 Performance

This model is an improvement to the previous model in the following ways:

1. Variable selection has been done. Only those variables which reduce the mean squared error have been included as attributes in this model.
2. Missing values have been handled. Empty fields have been replaced with the mean or mode of the column as appropriate.
3. While encoding categorical variables, the first dummy variable has been dropped to improve the model accuracy.

This model gave the following result for a train-test split of 95-5%:

```
Mean Squared Error =  0.0187423989857
          R2 score =  0.00593852948568
```

4 Experience

Over the course of exploring the dataset and trying out various prediction models, I had the following observations:

1. Simpler is better. Of all the complex models I implemented, linear regression gave the best score.
2. Data preprocessing is important, as it drastically improves the model's performance.
3. I had thought that including transaction month as an attribute would improve the score on Kaggle, as the submission has separate columns for different months. However, this did not happen, which came as a surprise to me.
4. In some cases, models which gave a better score on a train-test split gave a worse score on Kaggle.

5 Possible Improvements

The model can possibly be improved by:

1. Using non-empty columns to predict empty fields in other columns (e.g. using latitude and longitude to predict region ID county)
2. Using evolution of prices as a time series to predict monthwise log error
3. Using a better selection of variables
4. Trying different combinations of handling missing values for each variable differently.