Goal:

Given the historical housing prices dataset, our attempt was to predict the future housing prices trend for the two urban regions – Seattle and DMV. In order to accomplish this, we applied Machine Learning techniques on our dataset.

Dataset:

Data has median estimated home value per square feet across two urban regions – Seattle and DMV. The data was granular at the county level. The original dataset we acquired from Zillow has data since 1992. To ensure the highest quality and cleanliness, we chose to utilize the data points since 2000.

To summarize, the data set included categorical features like “State” and “County”; with “Housing Price Per SqFt” and “Year” being the only numerical elements.

Approach:

Model:

As a first step in designing a predictive model we chose the Simple Linear regression model to find if there is a linear relationship between the Housing Prices and Year over Year trend of the geographical boundaries.

Scikit-Learn Python packages are used to create, fit and evaluate the model. Pandas and Matplotlib libraries are used to load, process and visualize the datasets.

Why Linear Regression?

We have a dataset which contains the continuous variables that a Linear Regression would use to find a statistical relationship.

Alternative models:

Other models such as logistic regression, decision tree etc. were considered. However, due to the limited set of features available in this dataset we could not use an alternative model.

Methodology:

Data Pre-processing:

* 1. Data Split for Train and Test
  2. Removal of outliers and non-numeric values
     1. Outliers plot
     2. Quantiles to get the lower and upper bounds
  3. Handling missing data, required backfilling for two years

The data was missing for two years, it was backfilled linearly

* + 1. Missing data plot
    2. After backfill plot
  1. How (dis)similar are my train and test data?

Reshaping Train and Test

Model Training:

Fit and Evaluate

Model Testing:

Residual Analysis Chart

Model Tuning:

Outcome:

Historical Data Trend Chart1

Model Prediction Trend Chart2

Trend shows a Linear

Very few values are less than actual.

Difference is consistent across counties.

Conclusion:

Met the goal?

Given the scope of the dataset, although a 100% accuracy is not achieved, the model was able to give a reasonable prediction that shows a linear trend.

What could have helped for this model?

Adding more variables such as technology company growth, employment rate, occupational inflow to urban cities etc.

Training with more extensive historical data with more number of years.

Including external factors such as Dollar value inflation, Overall GDP rate to the model.

What’s next?