STAT 5606 **Dhruvkumar Patel**

Project: House Price Prediction using Machine Learning Algorithm

Data:

I decided to scrape data for this project instead of finding on Kaggle.com. The data contains Beds, Bath, Sqft, Address, Price as important features. The data is scrapped from Zillow.com of Philadelphia area. The biggest challenge was collecting data from Zillow as it only allowed to scrape first 9-10 house data per page and it was only 20 pages. So I have collected data for around 15-20 days. After scrapping data was cleaned in Excel where it was more easy to find duplicates values and ordering of rows together. There are 1072 rows with 5 important features.

Problem:

House Price Prediction using Machine Learning Algorithm.

Approach:

Firstly, most important thing is cleaning data where few rows were dropped which was scrapped and not will be used. Then I did data cleaning on columns by finding unique values and assigning some values to null value.

I did Exploratory Data Analysis where I found that most houses price are between 100,000 and 400,000 with average of 3 Bedroom and 3 Bathroom with area between 1000 and 1800 sqft. The area with Zipcode 19103, 19106, 19119,19123, 19130, 19146, 19147.

I did Feature Engineering by normalization of features and factorize. The data has 1015 rows and 5 features. The data was split into 75% train and 25% test data and applied various Machine Learning models like Linear Regression, Ridge Regression, Support Vector Regression, Random Forest Regression, Decision Tree and Gradient Boosting.

Results:

The best model is Gradient Boosting with 86% accuracy and 0.62 R squared value and 0.008 MSE. There were other model with better accuracy rate but it didn’t have good R squared values. Decision Tree had accuracy of 99% but it has only 0.10 R squared value. Random Forest had accuracy of 92% but it has 0.55 R squared value. All other has accuracy less than 60% and even worst R squared values.