Names: Manuel Duran, Adonis Shareef, Tushar Muley

Assignment: Milestone 5 -Final Project Paper

Date: August 14, 2021

Milestone 5 - Final Paper

Executive Summary

The Austin,Texas Metro area has seen an increase in population of 34% compared to the 2010 Census confirmed by the most recent Census data released in May of 2021. These new settlers are looking for affordable housing as they move into the newest tech hub outside of the traditional hubs like Silicon Valley and Seattle. Our team was tasked with building a machine learning model that will take our housing data and train against that data to make predictions on housing prices in the Austin area.

We were given a data set with over 15,000 rows all containing various features and information that Zillow contains for analysis. Our goal was to accurately predict the prices of homes in the Austin market for families and individuals coming to work in the rapidly growing area.

In the beginning we had to analyze the data and begin our prepping for the model we built. We updated features from human language to model ready input. With further analysis we then began to split and train the data trying to add or delete irrelevant features from the training data. To ensure we were getting the most accurate predictions of the home prices. We ended up trying three different models starting at 36% accuracy and ending with our most accurate model at 58% which was very close to our initial target of 60%.

From our analysis, we believe we have created a stepping stone to provide value to our customers when searching for a home in the Austin Metro area. This model could be used in other markets to provide value,insight and drive more views to Zillow.com. New users could look for a new home as well as analyze markets that are rapidly increasing in price.

Abstract

Recent arrivals of families and technology related individuals have driven up the prices of homes in the Austin,Texas area in a brief amount of time. Our team is looking into existing home sales and trying to predict how quickly home prices will change. Anyone wanting to make a decision to move to the Austin area would want to consider living expenses and where home prices are headed.

The team chose to analyze home prices from Zillow.com which provided 47 different features to predict the direction of home prices. We reviewed histograms and a coefficient matrix along with reviewing average home sizes and prices from the various sources like Zillow, RedFin and Realtor.com.

We did find a strong correlation with features of homes like number of bedrooms, bathrooms and living space. We did not realize that other features like school ratings and distance to school or the number of schools would also play a role in prices of homes. We believe we are missing data concerning new home build prices that might influence existing home prices. Our data also had pure land sales that needed to be reviewed for accuracy.

We used Multivariable Linear Regression (MLR) to predict home prices based on 44 different features to predict the latest home prices. We expected the model to be able to predict home prices with an accuracy between 80 percent and 60 percent. We have to take other variables into consideration; however, we believe Austin market home prices will keep going up.

Introduction

The team is proposing to build a predictive model that can assist new and future settlers to the Austin, Texas area who are planning to purchase a home and work in the new technology hub outside of the traditional Silicon Valley. The focus of our predictive models is on existing homes that will appeal to families or individuals that would be employed in the Austin market. We have further narrowed our scope to traditional family homes. We made the decision to focus on homes that families would move into because many of the technology companies that are sprouting up in this new hub tend to lure families. Features that might not have been as important before, like the number of bedrooms or bathrooms are trending to be more important to our analysis. We found that a lot of the data being actual properties it was easy to do minor Google searches to see the properties. This being Texas we found plenty of empty lots and open land.

This affects our analysis quite a bit because we now must look at the data a whole lot differently, not only as empty lots, but also the smallest things like air conditioning not being on a property. Due to the hot Summer weather driving the price down dramatically even though the square footage is greater or equal to that of a home with air-conditioning. We collaborated on these new decisions and methods after first

working independently and developing our models. We then began to diagnose our findings and review each other's work and found the best model and the methods we chose to develop.

Methods

Our method of choice for solving our price prediction issue was using the Decision Tree Regressor model. Before beginning the process of building the model we reviewed and adjusted a few characteristics within our dataset. For data preparation purposes our team adjusted all booleans features within our dataset to integers to assist in making our model creation easier. Secondly, we adjusted our dataframe to exclude any houses priced above one million dollars and less than 4,000 square feet within our dataframe. The team decided the selection demographic is now more family based, and we wanted to be sure that the home prices we were predicting for fell into a range that an average technology employee could afford. Also, the homes a million dollars and above were so few that they caused a lot of skewing of our pricing.

Figure 1: Scatterplot of Living area and Price of home

The above scatterplot provides the full scope of our data. The largest amount of our data is below with the 4,000 square foot living area and pricing under one million dollars.

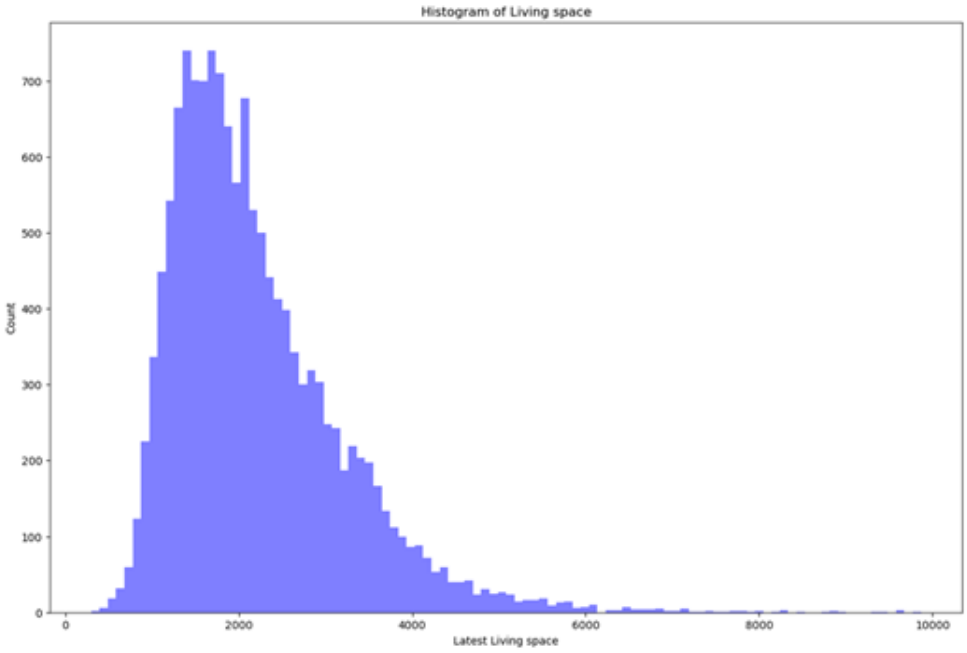

Figure 2: Histogram of Living Area

Figure 2 shows a histogram showing how the majority of the homes are under 4,000 square feet.

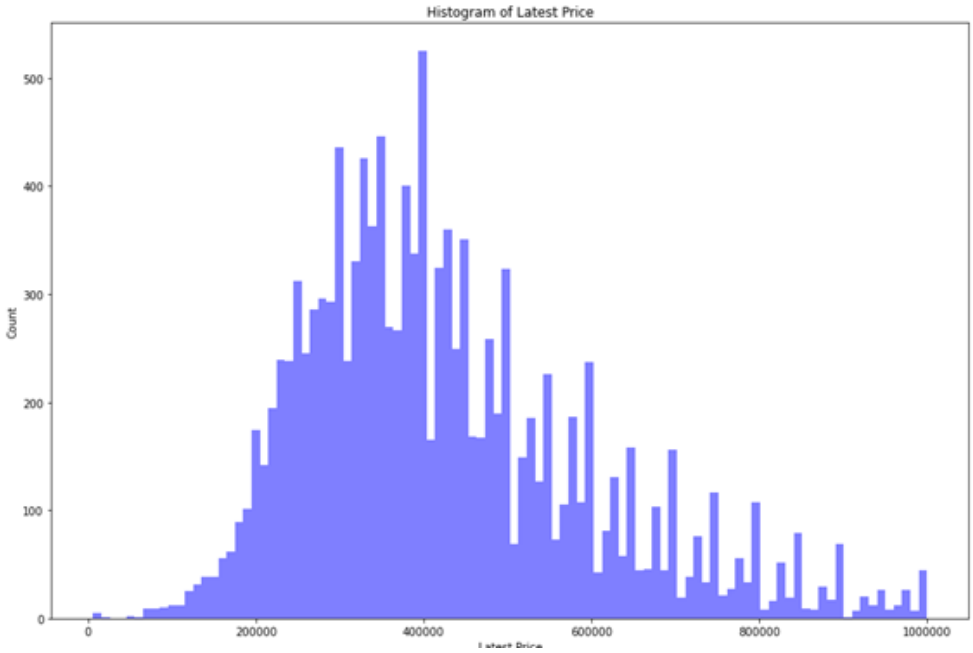

Figure 3: Histogram of Home Prices

Figure 3 shows a histogram of the home prices and you can see the majority are between 200,000 and 800,000 with peeks in certain price areas.

After our adjustments our team began the model creation process. We began by selecting ‘latestprice’ as our independent variable. As for the dependent variables, we selected many features that we believe would affect the pricing in the Austin area.

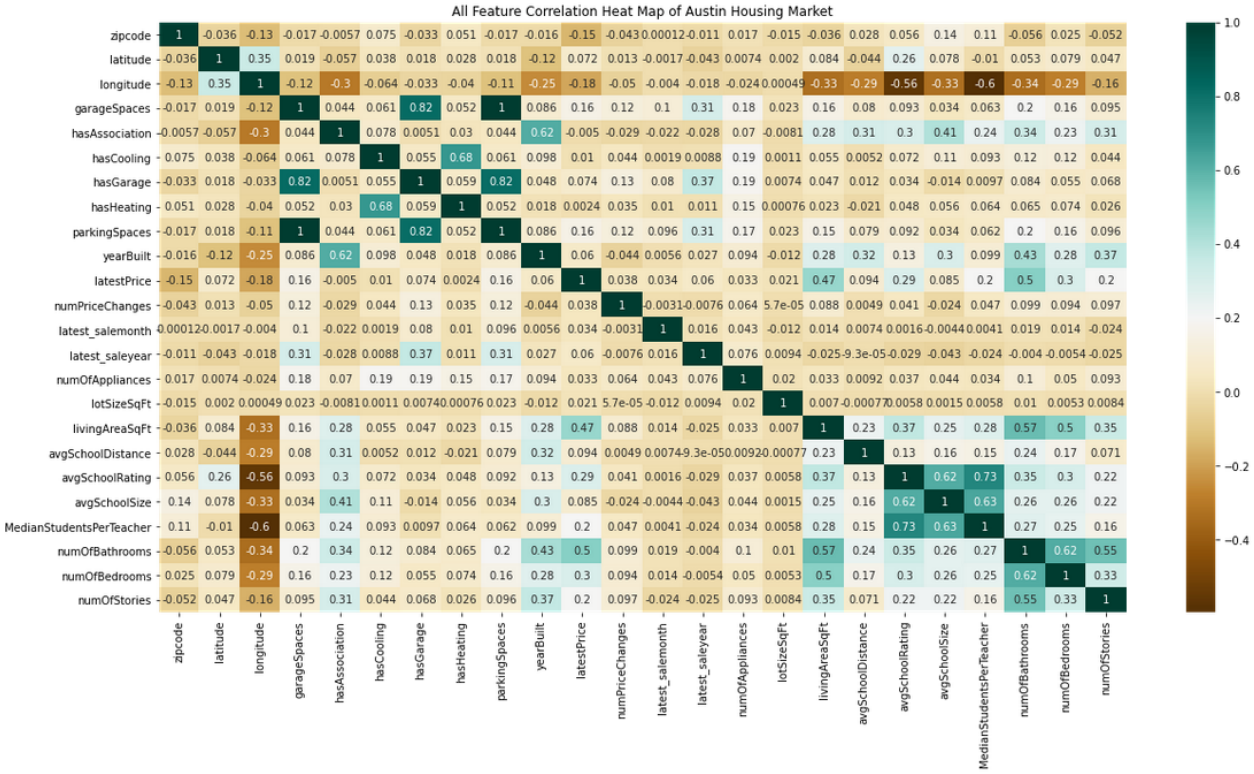

Figure 4: Original Correlation Matrix to select Dependent Variables

The above figure shows the correlation matrix used to select the dependent variables.

After the variable selection we were then able to split the data into training and testing sets and insert those sets into a model fit to perform our linear regression. For our next step we created a coefficient dataframe to show which dependent variables have a positive or negative correlation between the independent variables. An example can be viewed in Figure 5 below of one of the many combinations of the model that were run. Lastly, we created a dataframe in order to present our predicted housing prices from our model along with a variance column that displays the difference between the actual housing price.

	Coefficient
garageSpaces	32864.661887
hasAssociation	-161118.390524
hasCooling	-136560.262041
hasGarage	-43803.057673
hasHeating	177430.174709
lotSizeSqFt	0.000452
numOfBathrooms	239032.267320
numOfBedrooms	5851.132821
numOfStories	-59464.494782

Figure 5: Example of the many Coefficient dataframes reviewed

Results

Our team did run some different Multivariable Linear Regression models using different combinations of dependent variables. Our team also ran the Ordinary Least Squares and Decision Tree Regressor models. Each member contributed their analysis to improve the overall single model. We settled on the combination which returned the best accuracy. Looking at Figure 6 the R-Squared value was lower than we expected.

Model Type	R2 Value
Multivariable Linear Regression	0.5291
Ordinary Least Squares	0.5130
Decision Tree Regressor	0.5843

Figure 6: R squared of all Models run

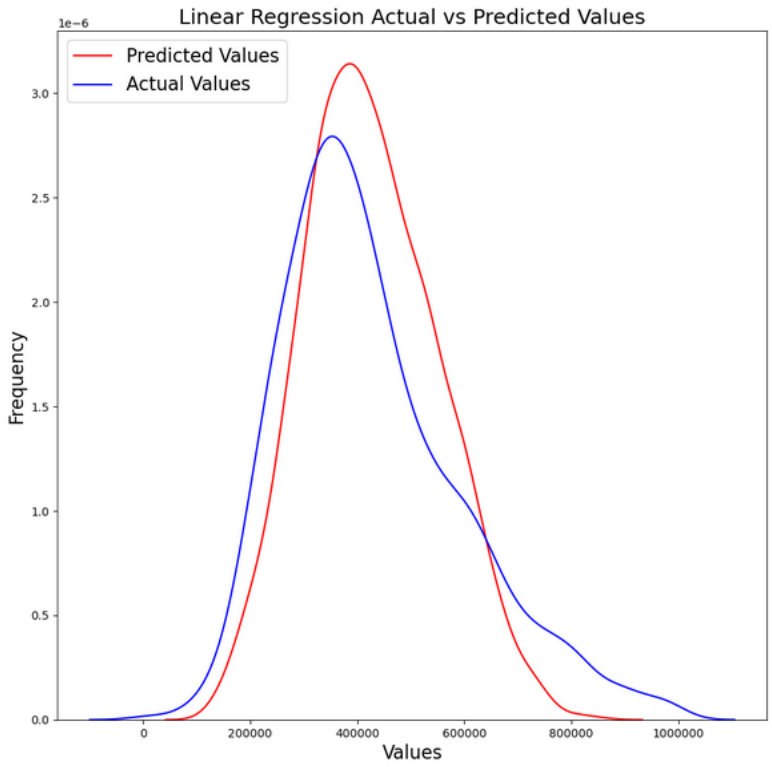

Figure 7: Actual compared to Predicted values for Multivariable Linear Regression Model

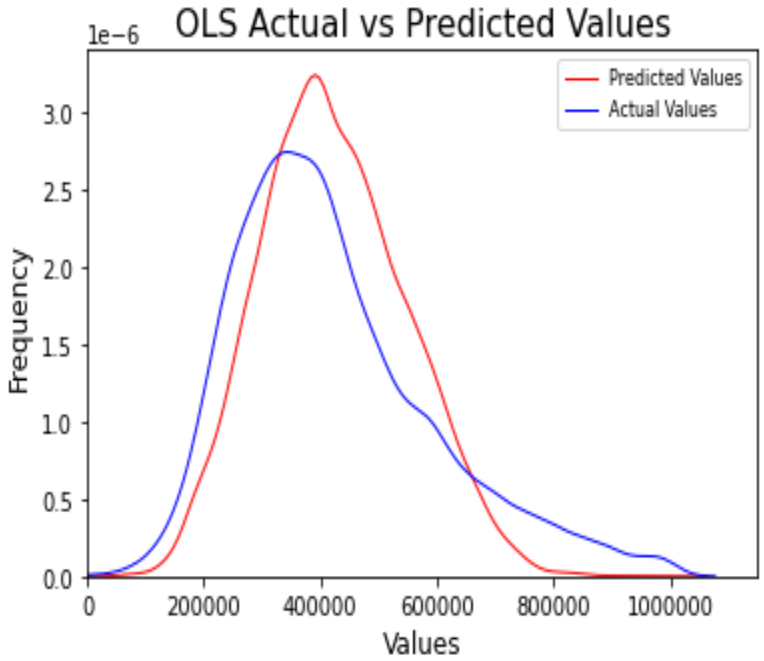

Figure 8: Actual compared to Predicted values for Ordinary Least Squares Model

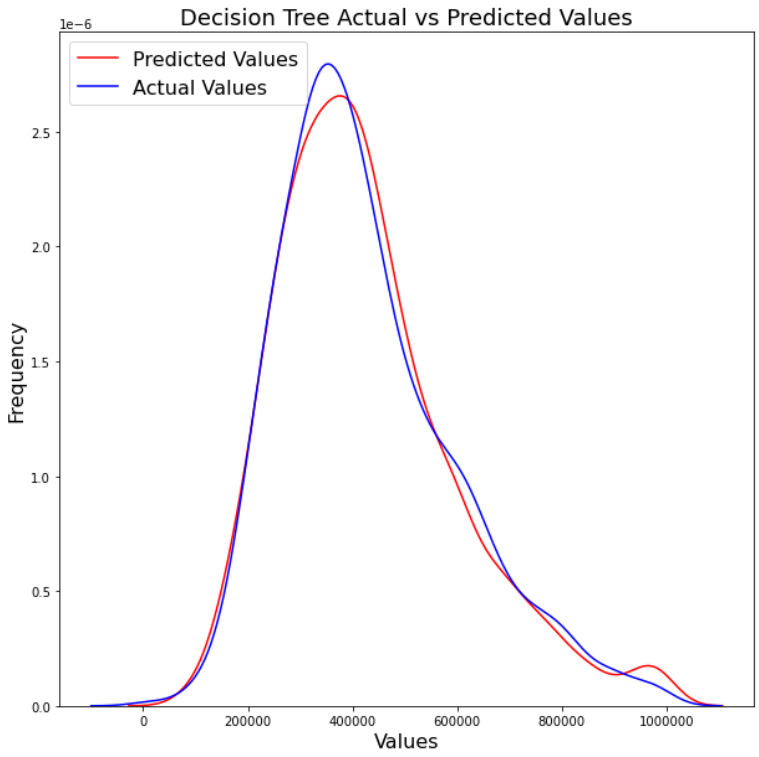

Figure 9: Actual compared to Predicted values for Decision Tree Regressor Model

The predicted values are close in some sense but as the values go closer to the one million dollars mark the variance increases for Multivariable Linear Regress, OLS model. The Decision Tree Regressor model predicted values proved to be closer to the actual home sale prices.

ID	Real Values	Predicted Values	$ Variance	% Variance
14489	649,900	999,000	(349,100)	-53.72%
2779	635,000	515,000	120,000	18.90%
2079	205,000	221,900	(16,900)	-8.24%
10943	349,900	348,000	1,900	0.54%
10993	260,000	240,000	20,000	7.69%
4096	260,000	295,000	(35,000)	-13.46%
12412	329,900	335,000	(5,100)	-1.55%
4433	179,500	320,000	(140,500)	-78.27%
12427	375,000	374,900	100	0.03%
6175	975,000	699,000	276,000	28.31%
12565	335,000	330,000	5,000	1.49%
6780	275,000	419,550	(144,550)	-52.56%
3128	350,000	330,000	20,000	5.71%
8903	313,000	250,000	63,000	20.13%
9849	140,000	265,000	(125,000)	-89.29%
4908	745,000	685,000	60,000	8.05%
6613	399,900	275,000	124,900	31.23%
10795	360,000	359,900	100	0.03%
10771	725,000	619,900	105,100	14.50%
602	230,000	225,000	5,000	2.17%
12625	449,900	595,000	(145,100)	-32.25%
13212	689,000	620,000	69,000	10.01%
14255	749,000	699,000	50,000	6.68%
508	539,000	545,000	(6,000)	-1.11%
8826	689,900	875,000	(185,100)	-26.83%

Figure 8: Sample of Actual compared to Predicted values with Variance for Dollars and Percentage

Figure 8 shows a sample of 25 rows of data comparing actuals to predicted values. For some homes the predictions are very close. For example, home IDs 12427 and 10795 the dollar variance is hundred dollars. Home site 9849 has a variance of negative 89.29% which is large.

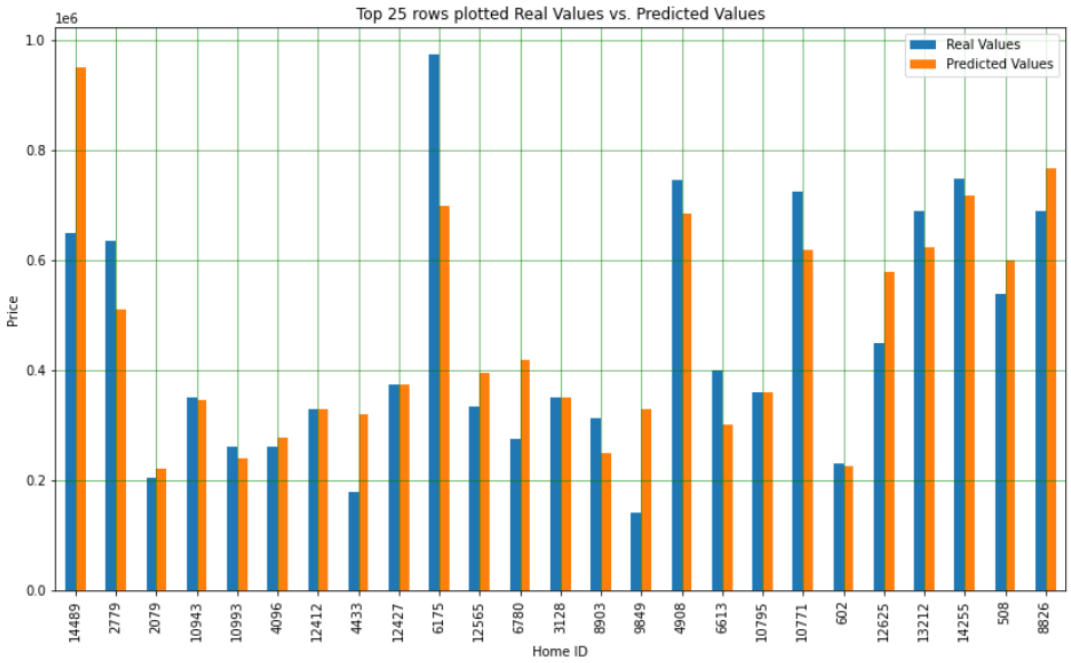

Figure 9: Top 25 Home ID plotted to show the difference between predicted and actual sales prices

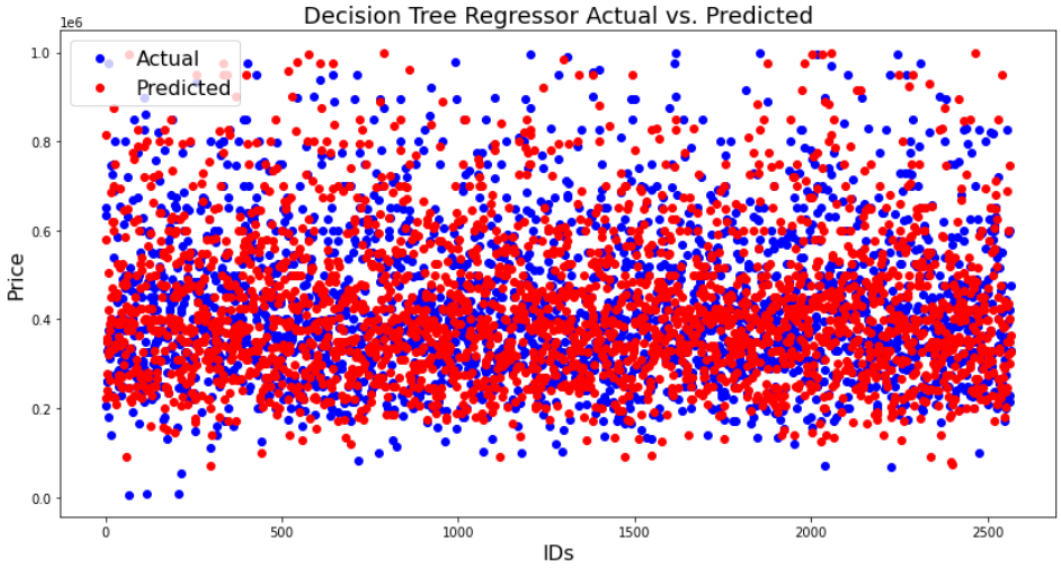

Figure 10: Scatter plot showing Actuals and Predicted values for the full data set

Figure 9 shows a bar plot of the first 25 home IDs with predicted and actual sale price values. Figure 10 is a Scatter plot of all predicted and actual home prices. In this plot as you can see about 50% of the predicted prices are close to actual prices.

Discussion/Conclusion

Having run a few different models and changing the dependent variables to get a mix of variables that would give us increasing accuracy. We settled on 44 features compared to the 47 we originally started with in the Zillow dataset. We ran various versions of Multivariable Linear Regression, Ordinary Least Squares and Decision Tree Regressor models. The best of those outcomes are displayed in this paper. The accuracy of our model when we first ran the model to get a baseline was a R-square value of 0.36. From there we reviewed the dependent variables and made changes to increase the R-square value to 0.46. We then iterated over feature pruning to get out a more accurate model, the Decision Tree Regressor with an R-squared value of 0.58, slightly missing our target of 60 percent accuracy on home prices.

Our goals for the future will be to see if we can take our Decision Tree Regressor model further, if not within a reasonable time. We think if we had a little more time we would have split the category of homes by a different type. Maybe deriving a type that could classify homes based on price, neighborhood, living space, lot size and top five features. We can break the homes into four or five classes. Allow the Decision Tree Regressor to provide more accuracy prices. That would take more than the 10 weeks we have for this class.

Acknowledgments

We want to thank our Professor Alsaleem for providing us guidance through the project. We also want to thank Eric Pierce for gathering and uploading Zillow Austin housing market data into Kaggle.com.

Reference

1. *Austin, TX House Listings*. (2021, April 12). Kaggle.

   https://www.kaggle.com/ericpierce/austinhousingprices?search=home+prices+&fileType=csv

2. Gates, B (2021, April 15). *Austin's median home price jumps nearly 29%, highest increase in nation.* From kxan.com

   https://www.kxan.com/news/local/austin/austins-median-home-prices-jump-nearly-29-highest-increase-in-nation/

3. *Startup Software Engineer Salary in Austin, Texas*. From Salary.com

   https://www.salary.com/research/salary/posting/startup-software-engineer-salary/austin-tx

4. *Austin, Texas Average Home Prices*. From Realtor.com

   https://www.realtor.com/realestateandhomes-search/Austin_TX/overview

5. Chukwu, D. (May 2021). *Austin named fastest growing major metro in the US.* From KVUE.com

   https://www.kvue.com/article/money/economy/boomtown-2040/austin-population-growth-census-data/269-c1e8725e-3489-4445-9bb5-fc340887cc43