Economic Indicator

Prediction of price for future houses

Domen Demser
Computer Science
Florida State University
Tallahassee Florida USA
dd19o@fsu.edu

Ibrahim Samad Afridi
Computer Science
Florida State University
Tallahassee Florida USA
isa21a@fsu.edu

ABSTRACT

The price of housing is a major concern for many people, both those who are looking to buy a home and those who already own one. The factors that affect housing prices are complex and varied, including economic conditions, interest rates, and supply and demand. In this paper, we will use data science to predict the price of future houses and see if the time is the only factor that affect the prices of houses.

In this paper we will talk about the importance of Data Science process as well as the steps we took to successfully predict the economic indicator for future US housing prices. First, we follow the data science process of asking an interesting question, where we propose two questions: With the passage of time the value of houses increase but is it time or other factors which affect the value of houses? Can we predict the prices of houses for the future?

Second, we dive into ways of preprocessing our data so that we can gather useful information, using correlation coefficient with heatmaps and other scatter plot graphs on our dataset that will later be used in development of machine learning models. We will also show how linear regression, decision trees, and random forest regression have different prediction R2 scores which determine the accuracy of each model.

Finally, we review the scores of each model, compare and discuss each model accordingly so that we can finally use the model that is the best-fit for our project. We will also discuss on how we could improve our models and any future work that could be done in order to get better predictions and accuracy.

CCS CONCEPTS

• Regression Models → Logistic regression, Decision Tree, Random Forest Regression

KEYWORDS

Data Science, machine learning, data preprocessing, algorithms, logistic regression, decision tree, random forest regression.

ACM Reference format:

Domen Demser, Ibrahim Afridi. 2023. Economic Indicator: Prediction of price for future houses. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 8 pages.* https://doi.org/10.1145/1234567890

1. Motivation

The motivation for this project is to create an economic indicator for future housing prices for individuals and for businesses to make a well informed decision before buying the house. The current recession has change the market value of the houses and causes instability in the market for buyers. This project will help byers to a make a clear decision.

We believe that data science can be a powerful tool for making sense of complex data and providing insights that can help people make better decisions. We hope that this project will help people understand the housing market better and make more informed decision about their homes during this difficult time.

Overall this project seeks to leverage the power of data science to provide a more valuable insights and help people make better decisions in the face of uncertainty.

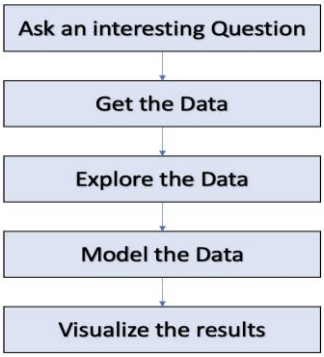
2. Data Science Process

Data science process starts by first asking an interesting question after the idea comes to the mind then comes the second step which is getting the data, after having the idea and getting the idea comes exploring the data, and modeling the data and lastly visualize the results and to see if our question makes sense or not if does not make any sense then we can have to change our first step which is asking an interesting question and then again following up the steps.

[*]Article Title Footnote needs to be captured as Title Note
[†]Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

The question that came to our mind was can we predict the housing prices, then we start getting the data using the google dataset search and after getting the data we implemented different techniques and models so that we can visualize our results.



2.1 Asking an Interesting Question

The process of every data science project starts with asking an interesting question. We always need to have a good reason to use the data we will be using, and it is important to keep in mind that it will be used to help solve every day's problem. In this project we are dealing with economic indicator where we want to achieve and see if we could figure out the factors that affect the price of real estate. Therefore, we propose two questions that are important. Since it might be self-explanatory that with the passage of time the value of houses and other real estate increases, it is worth asking ourselves if it is the time or there might be other factors that affect the value of houses. Another question that came up before gathering the data was if we can predict the prices of houses for the future based on the data. After these two questions we had been proposed, we knew exactly what to look for.

2.2 Getting the Data

Now that we know what to look for we started to look for suitable datasets that would help us answer our question. First, we head over to Google Dataset Search engine, using the keywords such as, US housing dataset. We wanted to keep the datasets limited to the one country due to the scope of our project. Finally, we find a dataset that is suitable for our project! Great, we download it from Kaggle's website but the job is not done just yet. The dataset that we found had 22 columns with both categorical and numerical values, which opposed another question of how are we going to handle these values. We have also noticed that there is a big potential in information we can gather from this dataset since it contains the main factors a renter would consider when deciding whether the price is a fair price for what you get.

From the picture below we can see some of the factors that will later be used when making the regression models but more about that later. The main factors we wanted to focus on are if the houses come with parking, the amount of bedrooms and bathroom they come with, what region are they in, the type of the listing (i.e. apartment, condo, house, other), how big is it (in sqfeet), and what is the price of the housing.

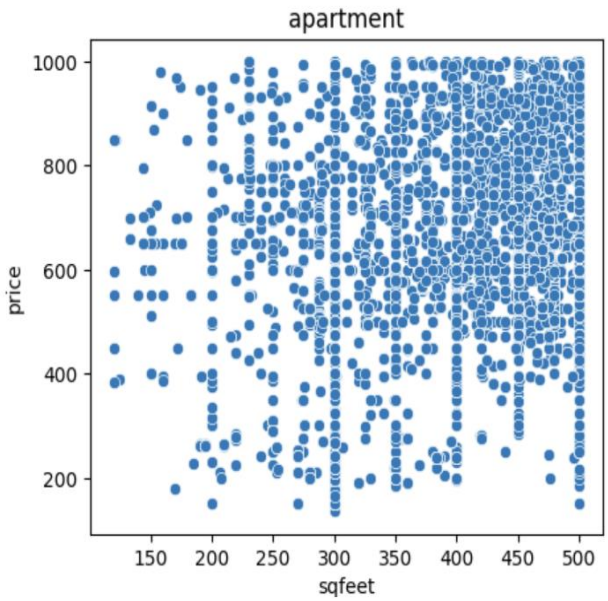
```
print(file.head())
```

```
           id                                                url  \
0  7049044568  https://reno.craigslist.org/apa/d/reno-beautif...
1  7049047186  https://reno.craigslist.org/apa/d/reno-reduced...
2  7043634882  https://reno.craigslist.org/apa/d/sparks-state...
3  7049045324  https://reno.craigslist.org/apa/d/reno-1x1-fir...
4  7049043759  https://reno.craigslist.org/apa/d/reno-no-long...

        region                  region_url  price       type  sqfeet  beds  \
0  reno / tahoe  https://reno.craigslist.org   1148  apartment    1078     3
1  reno / tahoe  https://reno.craigslist.org   1200      condo    1001     2
2  reno / tahoe  https://reno.craigslist.org   1813  apartment    1683     2
3  reno / tahoe  https://reno.craigslist.org   1095  apartment     708     1
4  reno / tahoe  https://reno.craigslist.org    289  apartment     250     0

   baths  cats_allowed  ...  wheelchair_access  electric_vehicle_charge  \
0    2.0             1  ...                  0                        0
1    2.0             0  ...                  0                        0
2    2.0             1  ...                  0                        0
3    1.0             1  ...                  0                        0
4    1.0             1  ...                  1                        0

   comes_furnished  laundry_options  parking_options  \
0                0      w/d in unit          carport
1                0      w/d hookups          carport
2                0      w/d in unit  attached garage
3                0      w/d in unit          carport
4                1  laundry on site              NaN

                                           image_url  \
0  https://images.craigslist.org/01616_daghmBUvTC...
1  https://images.craigslist.org/00V0V_5va0MkgO9q...
2  https://images.craigslist.org/00t0t_erYqC6LgB8...
3  https://images.craigslist.org/00303_3HSJz75zlI...
4  https://images.craigslist.org/01616_fAL4WFV8z0...
```

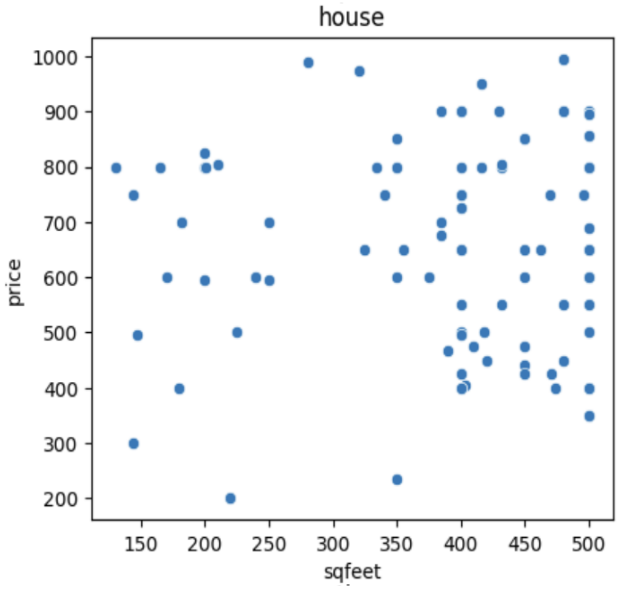
2.3 Exploring the Data

Now that we have the dataset it is important to see and look what the data is trying to tell us. Exploring Data Analysis or EDA, encompasses the "explore data" part of the data science process. EDA is crucial for every data science project but unfortunately is often overlooked. Therefore, if our data is bad, our results will be bad and conversely, understanding your data well can help us create smart and appropriate models.

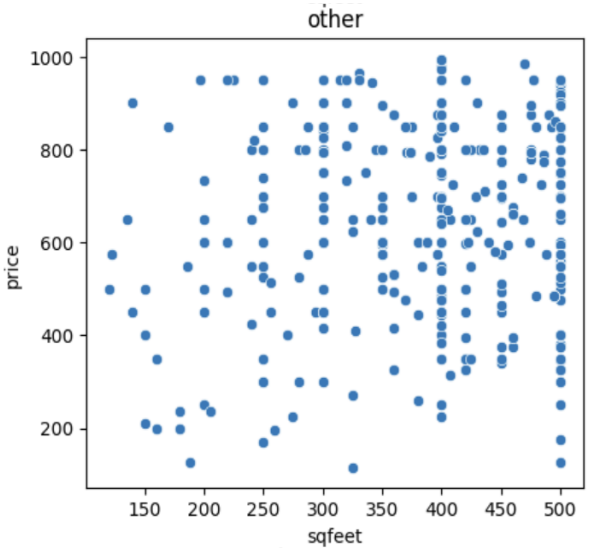
Since we knew what EDA is from the lectures we learned in glass, we knew we will have to explore global properties with using scatter plots and aggregation functions to help us summarize the data. Therefore, the first part of exploring the data was to make few graphs and observe them.



First, we plot the graph that shows the relationship between the price and sqfeet for the apartment. Here, we were able to see that most of the apartments range between 400 and 1000 (per month) with having a sqfeet size between 350 and 500. Then we plot another graph for houses.

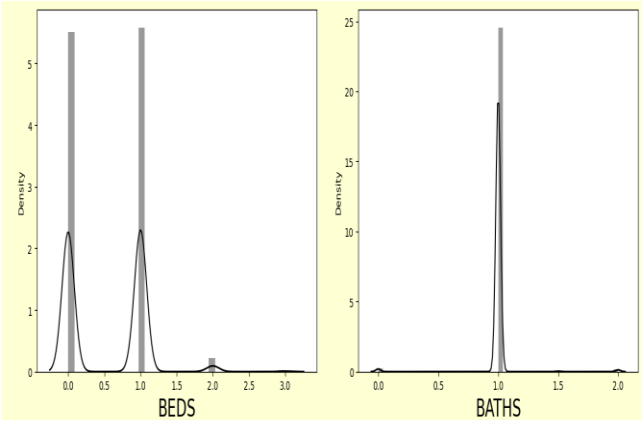


From this graph we are able to see that there might not be as many houses as there are apartments, as well as, the plots on the graph are widely spread out so we moved to make another and last graph.



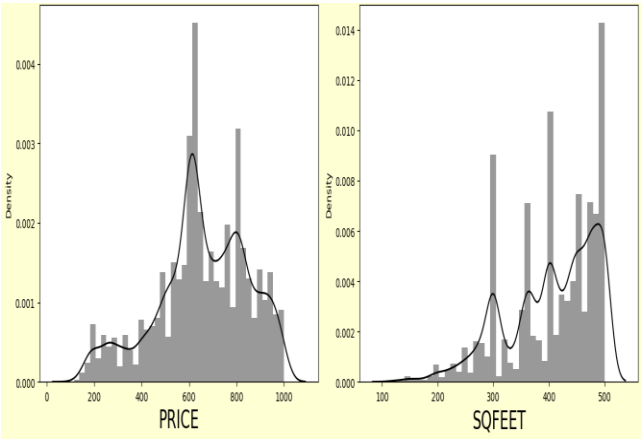
This graph shows "other" so you might question what this could be. This graph represents any kind of other housing listing, for example condos. Now that we have all three graphs at hand, we were able to make a conclusion that the most common type of housing in US are apartments, houses, and other. With the range between 120 to 500 square feet of area, the most popular are apartments which are followed by the "other" category with having the houses as a least popular area. This information was crucial to the development of this project and will help us especially when creating the models later.

After gathering this data, we also wanted to see what is the most common number of bedrooms and bathrooms that renters are interested in. We show this in the graph below.



From our dataset, we can see that the most popular are apartments with either 0 or 2 bedrooms which means that they are either condos or 1 bedroom apartments that come

with one bathroom only. Next, we show the price range and square feet area for the housing.



From our dataset, we can see that there is a peak range of price at $600 per month which means that most of the housings are close to 600$ per month as well as having the most of the square feet area laying close to 500. Now that we know what our data is trying to tell us and with all the information we have gained above, we are ready to move to the next step which is data preprocessing.

2.4 Data Preprocessing

Data preprocessing is another crucial step in the data science process. It helps us "clean" the data to prevent any misleading results and inaccurate results of the machine learning models. In our project, the first step of data preprocessing was to determine null values. Since there was not a lot of missing data in our data set, we decided to drop the rows that contain the missing values since we did not want to replace them with the mean values of that specific row. The reason we decide to that is because we wanted the dataset to be as original as it can possibly be so that we can gather the most accurate results from this dataset.

The next step of the data preprocessing was to perform the dimensionality reduction of the entire dataset. With that being said, we drop any columns we don't need, i.e. region URL, id and many more, to speed up the computational power of the entire program and to get more accurate predictions.

	id	url	\
4	7049043759	https://reno.craigslist.org/apa/d/reno-no-long...	
43	7048997207	https://reno.craigslist.org/apa/d/reno-new-pri...	
53	7049016213	https://reno.craigslist.org/apa/d/carson-city-...	
72	7047408531	https://reno.craigslist.org/apa/d/reno-remodel...	
73	7037948367	https://reno.craigslist.org/apa/d/reno-give-us...	

	region_url	price	sqfeet	beds	baths	cats_allowed	\
4	https://reno.craigslist.org	289	250	0	1.0	1	
43	https://reno.craigslist.org	800	400	1	1.0	1	
53	https://reno.craigslist.org	825	483	2	1.0	0	
72	https://reno.craigslist.org	289	250	0	1.0	1	
73	https://reno.craigslist.org	800	400	1	1.0	1	

	dogs_allowed	smoking_allowed	...	sd	tn	tx	ut	va	vt	wa	wi	wv	wy
4	1	1	...	0	0	0	0	0	0	0	0	0	0
43	1	1	...	0	0	0	0	0	0	0	0	0	0
53	0	0	...	0	0	0	0	0	0	0	0	0	0
72	1	1	...	0	0	0	0	0	0	0	0	0	0
73	1	1	...	0	0	0	0	0	0	0	0	0	0

After drop the unnecessary columns, we had to convert the categorical values to numerical values, and we do that by using One-Hot Encoding techniques. This was a perfect technique to use for our dataset since we had columns such as region. The region and many other columns contained rows of data, such as "reno / tahoe" for region. This would be really hard without One-Hot Encoding because we needed to split the rows into two different columns using the encoding mentioned above. This produced two additional columns in the data set, reno and tahoe, marking the rows with either 0 or 1 accordingly for each housing listing. We perform the same steps on every column in such format.

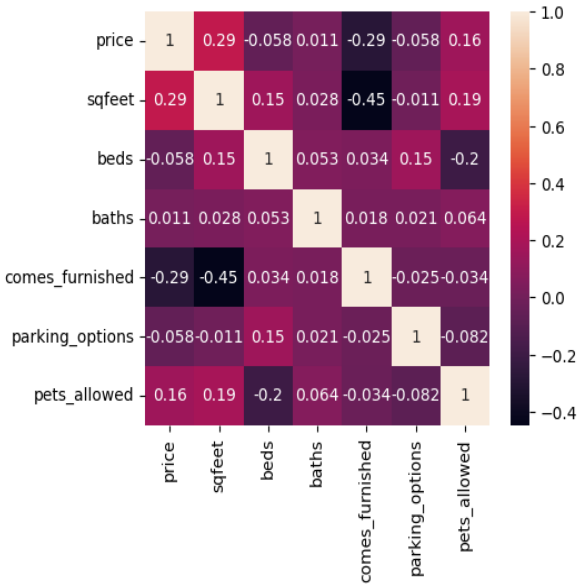
Another step we took towards the data preprocessing was to use clustering on the column with the name parking. This column had different kinds of parking options, such as, valet parking, no parking, garage parking, etc. Because this column was a crucial factor for a renter when deciding on a housing listing we had to use cluster to make additional columns for every single option and mark them with either 0 or 1. After the use of the methods like One-Hot Encoding and Clustering, our dataset comes to the total number of 64 columns.

Lastly, we change the continuous values to discrete values and remove the outliers from the data. Removing the outliers was also an important step so we can easily process our data because the dataset came with over 1 million rows of data and with ending up with 64 columns, our computer were just not powerful enough to process this large amount of data.

Finally, after the data preprocessing was wrapped up, we were able to move to the next step which was determining the correlation coefficient for our data.

2.5 Correlation Coefficient

Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship.



In the picture above we can see the relationship between the numerical variables. For this heatmap, the numerical variables used were: price, sqfeet, beds, baths, comes_furnished, parking_options, and pets_allowed. From the correlation heatmap that our data produced using the numerical variables listed above, we can notice that the highest coefficient with the price is square feet and right after is the pets_allowed. We can also see the lowest coefficient having a relationship between the price and comes_furnished.

It is important to know that correlation between two variables does not imply a casual relationship and that it can also be determined using a scatter plot between these two variables. Therefore we used heatmaps to make the visualization more readable.

Now that we have all the information and have our data ready for further steps in this project, it's time to get to the fun part where we create certain machine learning models that will make the predictions.

2.6 Modeling the Data

Before we made the models, we had to ask ourselves which models would fit our dataset the best so that it can produce the most accurate results. Therefore, we decided to create three different models and see which one will be the best at predicting the future prices for the US housing. We make three different models: logistic regression, a Decision Tree model, and Random Forest Regression.

2.6.1 Logistic Regression

For the first model we choose to select Logistic Regression as a first model for our data. You might ask what this model is so let me first tell you a little about it. Logistic regression is a classification algorithm It is used to predict a binary outcome based on a set of independent variables. There are three main types of logistic regression: binary, multinomial and ordinal. In our project we use the logistic regression of a binary type since it is the most suitable one to the problem we are trying to solve.

But why do we choose logistic regression as being one of our models? That is because it is used when our dependent variables are binary which just means that a variable has only two outputs. In our case, we wanted to showcase if the price of the housing in the future will increase or decrease. Therefore, the goal of binary logistic regression is to train a classifier that can make a binary decision about the class of a new input observation.

In our dataset, out of all the variables, price is the dependent variable, and we take square feet, beds, baths, comes_furnished, and pets_allowed as our independent variables. After we determine the desired variables we split the data into two different sets: testing and training. We set the training set to be 70% and testing set to be 30%. After applying logistic regression to the testing set, we got the following results:

```
Logistic Regressor
Mean Absolute Error 75.72808251289264
Mean Squared Error 18888.783872480075
Root Mean Squared Error 137.43647213341907
R2 score 0.4942348400465103
```

From the above picture, we can see that the accuracy of the model is about 50%, looking at the R2 score. We thought this was a moderate result but since we didn't want to stop just here and be satisfied with this result, we wanted to see how our data will perform on other machine learning models. That's why we went ahead and created another model, called Decision Tree.

2.6.2 Decision Tree

A decision tree model is another regression model in which the final outcome of the model is based on a series of comparisons of the values of predictors against threshold values. Decision tree build regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes, where leaf nodes show the final predictions based on the decision (parent) nodes.

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). We use standard deviation to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogenous its standard deviation is zero. Additionaly, we perform standard deviation reduction which is based on the decrease in standard deviation after a dataset is split on an attribute since constructing a decision tree is all about finding attribute that returns the highest standard deviation reduction.

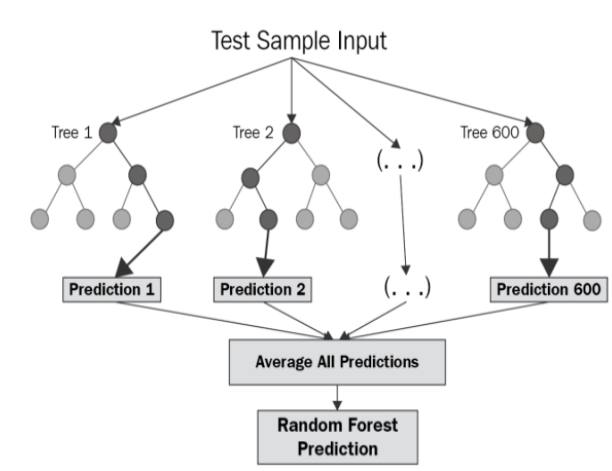
In our dataset, out of all the variables, price is the dependent variable, and we take square feet, beds, baths, comes_furnished, and pets_allowed as our independent variables. After we determine the desired variables we split the data into two different sets: testing and training. We set the training set to be 70% and testing set to be 30%. After applying Decision Tree to the testing set, we got the following results:

```
DecisionTree
Mean Absolute Error 54.53973322035225
Mean Squared Error 13431.812432532613
Root Mean Squared Error 115.89569635035036
R2 score 0.640350442396524
```

From the above picture, we can see that the results are much better when using a decision tree as our model for the dataset. Decision tree scored much better accuracy of around 64% which confirmed our speculations on building a better model. In fact, we tried to display the decision tree as a graph of decision nodes and leaf nodes but unfortunately, we did not have any luck creating one of those within the code itself. Nevertheless, using this model was definitely a step in the right direction but we didn't just stop here. Our curiosity drove us further into creating another model, called Random Forest Regression model.

2.6.3 Random Forest Regression

Random Forest regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It operates by constructing a multitude of decision trees at training time. To help you understand what we mean by all these words, we have provided the this picture below:



In our dataset, out of all the variables, price is the dependent variable, and we take square feet, beds, baths, comes_furnished, and pets_allowed as our independent variables. After we determine the desired variables we split the data into two different sets: testing and training. We set the training set to be 70% and testing set to be 30%. After applying random forest regression to the testing set, we got the following results:

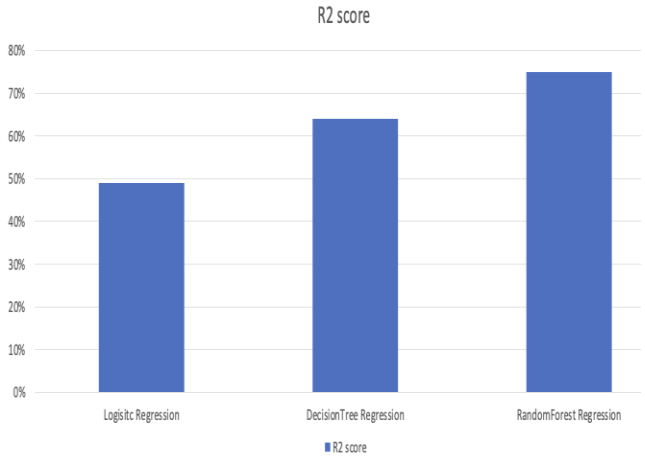
```
RandomForest Regressor
Mean Absolute Error 49.16114177879336
Mean Squared Error 9356.295226816757
Root Mean Squared Error 96.72794439466166
R2 score 0.7494762932378383
```

From the picture above, we can see that using Random Forest regression was the model we were looking for. The accuracy of 75% is by far the best one and this immediately makes it a top model to be used in our project. Again, we wanted to showcase the procedural tree but unfortunately we had no luck of creating the graph for it. Since we found our model, it was time to review the final results and draw a conclusion to the models itself.

2.7 Visualization

Visualizing the results of a data science project is important because it can help us make the results more understandable and accessible to a wider audience, in our case, to you readers. It is used to communicate complex information in a clear an concise way, and it can help to identify trends and patterns that would be difficult to see from just looking at the raw data.

Up to this point, we have seen many visualizations when we were reviewing the data. After creating our models, we wanted to show the complete conclusion of the models and how they compare based on the R2 score or accuracy for each model.



As we mentioned before, Random Forest regression was our top pick. Based on the numbers and percentages it is sometimes hard to imagine how much of a difference between the models there actually is. Here, we show that the difference between the random forest regression and logistic regression model is more than ¼ of a difference. Even selecting decision tree model would not be as bad if we would not know how to make random forest regression since it performed much better than the logistic regression model. Since we must pick a model that would be the best-fit for our project, the final decision once again lands on picking random forest regression model.

3. Conclusion

3.1. Summary

We first ask an interesting question which was the predication of the housing prices and how it changes. For that we gather our data using different online sources like google dataset search engine and then after that we get the data from Kaggle. After getting the data we clean the data using different techniques like removing the columns and then removing the outliers of the dataset. Also, we use different techniques like one hot encoding technique to change the column to different column so that I can have important columns. I also use the clustering technique to change the whole column's different values into one cluster. Then I implement different models and looking at the results of different models, we cannot predict which model is better to implement or using one model of data science can help us better predict the results. By implementing each model we come to know which model is a better fit for our dataset. We first used logistic regression thinking it will help us to predict the better accuracy, but its accuracy was only 50 percent. Then we come up with another model which is decision tree which gives us far better result than the logistic regression when we implement it and also it was computationally easy for us to run giving us the accuracy of 64% which was better model for our dataset then the logistic regression. Then finally we implement Random Forest regression model which was not computationally easy as compare to the decision tree but it gives the best of all the other 2 models. The accuracy of this model was almost 75% which is about 25% higher than the logistic regression.

3.2. Limitation and Future Work

The unforeseen issues we faced during our project was that we did not expect that it will be computationally hard for us to run the large dataset. Even after cleaning the dataset and removing the outliers it was still hard for us to compute the big data. The reason for this much of big data was using the one Hot Encoding technique which transforms the whole column into different columns. The reason for using this technique was that we did not want to waste the important column of the dataset which was type string so we could

not use the logistic regression and other techniques so for that reason we use one Hot Encoding.

For future work, we will try to use strong computational machines which can implement such models on big data. Also, we would need to implement higher models which can find the better accuracy of our dataset.

ACKNOWLEDGMENTS

We would like to acknowledge and give warmest thanks to my supervisor Dr. Guang Wang who made this work possible. His guidance carried me through all the stages of the project. We would like to thank our group partners Domen Demsar and Ibrahim Afridi, for their valuable contribution to this project. Ibrahim Afridi who was responsible for cleaning the data and then implementing the different models. Domen Demsar who was responsible for collecting the data and then cleaning the data.

REFERENCES

[1] https://www.kaggle.com/code/flavioakplaka/rental-property-price

[2] https://datasetsearch.research.google.com/search?src=0&query=usa%20housing%20dataset&docid=L2cvMTFqOWI5c2JubQ%3D%3D

[3] https://www.kaggle.com/code/madhantm/house-price-using-linearregression-decisiontree

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[5] https://scikit-learn.org/stable/modules/tree.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting.