Project Report

**Used Cars Price Prediction**

Heli Patel *- ID 015267122*

*Harshala Doiphode - ID 015222480*

Devangi Doshi *- ID 014520584*

**SECTION 1. Introduction**

**Motivation**

Vehicle’s potential resale value is an important factor while buying a vehicle. It is important to make an informed decision while purchasing a vehicle. There are various factors like standard colors, mileage which affects the resale value of the vehicle.

**According to CNN reports, the demand for used cars has been soaring mainly, because of the shortage of computer chips and other components. Also, partly because of the pandemic that forced people to travel in their safety bubble rather than using public transport. Hence, to better understand the importance of used cars and factors driving the value of a car, our purpose is to analyze the quality of used cars and predict its resale value.**

**Objective**

Craigslist is the world's largest collection of used vehicles for sale, yet it's very difficult to collect all of them in the same place.

**The main objective of our project is to analyze the attributes that are important while buying a car, cars that are more in demand and also a perfect price for a particular car.**

**Literature Review**

According to CBS news, there has been a 10% increase in purchase of used cars alone, in April. This has happened due to the shift in consumer behavior. As many were leery of using public transport and found it imperative to possess a private vehicle, this led to increase in car demand and also in price. On top of that, due to the pandemic, car production had slowed down due to the microchip shortage. An overall trend of increased demand for used cars resulted in shortage of supply, causing inflation. Hence, our motive is to observe this shift in consumer behavior.

**SECTION 2. System Design & Implementation Details**

**Algorithms Selected**

***Linear Regression***

Linear Regression, a Scikit-Learn Library was used because it fits a Linear model with coefficients to minimize the residual sum of squares between the observed targets in the datasets, and the targets predicted by the linear approximation.

***Ridge Regression***

Ridge Regression from the Scikit-Learn Library was taken into consideration as this model solves a regression model where the loss function is the linear least squares function and regularization is given by the I2-norm. This estimator has built-in support for multivariate regression (i.e., when y is a 2d-array of shape (n\_samples, n\_targets)).

***XGBRegressor***

XGBRegressor from the Scikit-Learn Library was used because it is used to produce a predictive model from an ensemble of weak models.

***Lasso Regression***

Lasso Regression from the Scikit-Learn Library was used because the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients(also called the l1-norm).

***Random Forest Regressor***

Random forest regressor from the Scikit-Learn library was used because it is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

***KNeighborsRegressor***

KNeighborsRegressor from the Scikit-Learn library was used because the target is predicted by local interpolation of the targets associated with nearest neighbors in the training set.

**Technologies & Tools**

***Python***

Python 3 was the programming language chosen for this project because it is a popular language to use for data mining. It supports multiple libraries like SkLearn, Pandas, etc.

***Jupyter Notebook***

Jupyter Notebook was used to write the code for the project because it is an interactive web tool that allows us to write code and get visualizations step by step.

***Google Colaboratory***

Google Colaboratory was used to write code for the project. It provides pre-installed data mining libraries and also provides free GPUs and TPUs from google servers which accelerates working and training models.

**Component Details**

***Step 1. Data Visualization***

In this step, graphs were created to visualize and better understand the data.

***Step 2. Data Preprocessing***

In this step, the data got cleaned and the features for best predictions got selected. From this, a preprocessed dataset was created.

***Step 3. Final Model Evaluation and Prediction***

In this step, the models got fitted to get final predictions.

**SECTION 3. Experiments / Proof of Concept Evaluation**

**Dataset and Preprocessing**

The dataset used for this project is scrapped every few months. It contains most all relevant information that craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude and 18 other categories. This dataset can be found at https://www.kaggle.com/austinreese/craigslist-carstrucks-data. The dataset has a size of 1.35 GB and is composed of 24 columns and 426880 rows. Most of these values are categorical.

For predicting the susceptIble price of used cars. We have done some data preprocessing on data as a part of preprocessing we checked the unwanted columns, null and duplicate values in the rows. We also drop a few unwanted columns which include: Id - This column has the entry id of the vehicle. VIN (Vehicle Identification Number) - This column serves as a car fingerprint that no two vehicles in operation have the same VIN. As per our understanding this is insignificant from the analysis perceptive. Image-url - This column has the url of images of reselling cars. Description - This column has a description about the condition of used cars. Size - This column gives the information about the size of the car that is medium, large, small. Paint\_color - This column gives the information about the color of the car. Model - This column gives the information about the manufacturer of the car. Further as a part of preprocessing we have drop row level duplicates values in data and we also drop columns having null values more than 60%.

For the column having null values less than 60%. We have imputated those features based on their types. To fill the missing values of categorical data we have use mode and for numerical data we have use mean. We also did Categorical encoding using LabelEncoder.

Before going further for analysis. We are going to check whether the data had outliers or not. An outlier is an observation that lies an abnormal distance from other values in a random sample from the population. The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the end of the distribution. We saw that there were no outliers in the data.

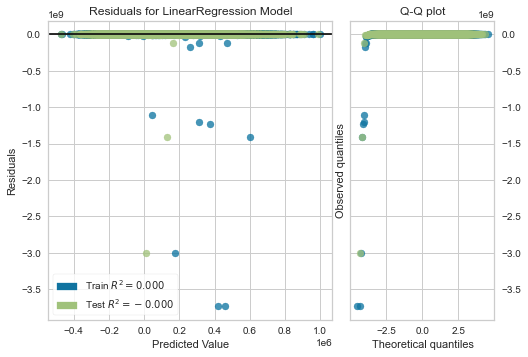
**Methodology**

To predict the price of the used car, the preprocessed dataset was used into training and testing datasets. The size of the training dataset was 0.7 while the size of the test dataset was 0.3. After that the models got fitted to get final predictions.

**Comparison of Algorithms**

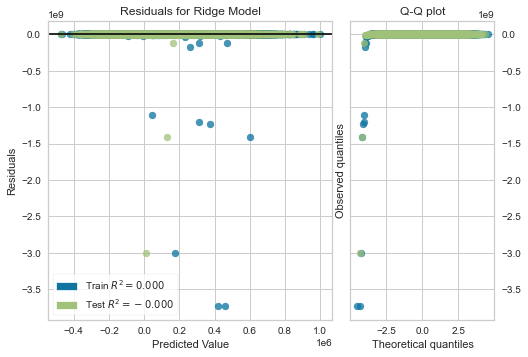
The algorithms were evaluated based on their R2 score and plotted using residual plot for better visualization.

Linear Regression:



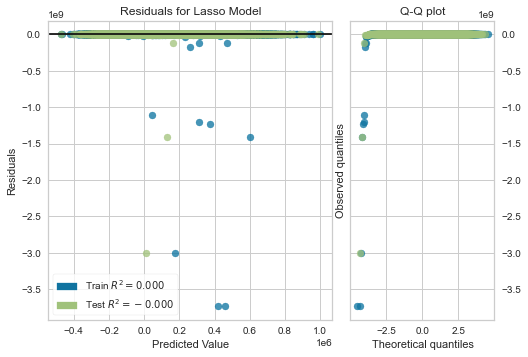
As we see from above figure, linear regression performs better on the training data than on the testing data. The r2 score for testing data being negative and zero depicts that the data is non-linear and does not follow any particular pattern/ trend.

Ridge Regression:



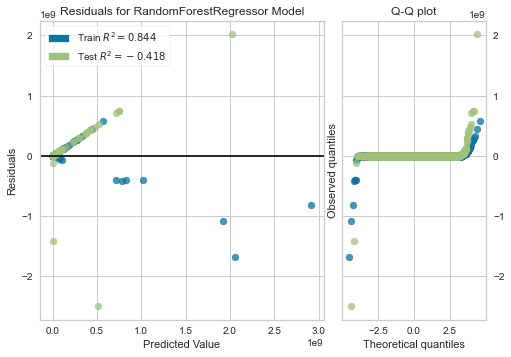
Just like linear regression, ridge regression also depicts no correlation between the attributes. This is possible since the data is not normally distributed across the horizontal black line.

Lasso Regression:



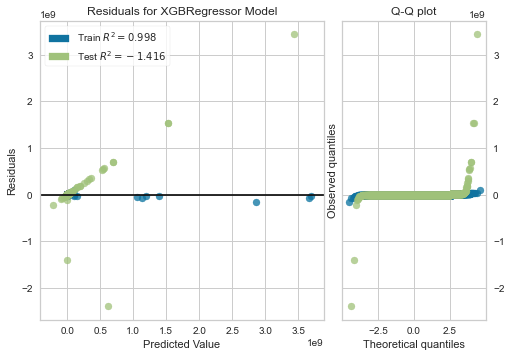
Lasso Regression has also not shown any promising results for the used cars dataset.

Random Forest Regression:



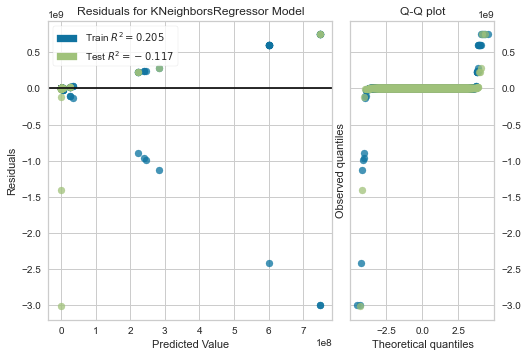
Here, random forest regression has performed well on the training data with an r2\_score of 0.844, whereas performance has worsened on the test set. Also, from the Q-Q plot, it is seen that at the ends of the horizontal line the data points curve off, reflecting the extreme values in the datasets.

XGBRegressor:



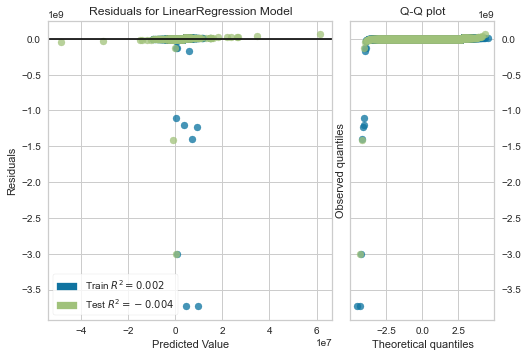
XGBRegressor has performed extremely well with an r2\_score= 0.998 on the training data, however, it has not been able to perform on the new data points coming from the testing data. Also, it can be observed from the Q-Q plot that the curve at the ends of the horizontal axis is not sharp compared to the Random Forest Regressor.

KNeighborsRegressor:



From the above Q-Q plot of KNeighborsRegressor, it can be seen that the data points are further distant on the extreme ends, as compared to what was observed in Random Forest Regressor and XGBRegressor.

Polynomial Features of degree 3 with Linear Regression



From the above figure, we see that the training data points are further down the negative axis. It means that the predicted values are much more than the actual values, hence the data point moved in the negative direction.

**Analysis of Result:**

|  |  |
| --- | --- |
| **Model** | **R2\_score** |
| Linear Regression | -0.00020313676915373335 |
| Ridge Regression | -0.00020313514345393635 |
| Lasso Regression | -0.00020313514345393635 |
| Random Forest Regressor | -0.4176867532673787 |
| XGBRegressor | -1.4157070624737873 |
| KNeighborsRegressor | -0.11731544080405865 |
| Polynomial Features deg=3 with Linear Regression | -229.9049989079235 |

As we see from the above table, Polynomial Feature transformation of degree=3 with Linear Regression has performed the worst in this dataset. Also, it is important to note that all the r2 scores have been negative, clearly indicating that the model does not perform well on the testing data.

**SECTION 4: Discussion & Conclusion**

**Decisions Made**

**Difficulties Faced**

* The dataset was huge: it had many rows and columns. Columns had to be selected by importance in the prediction.
* Rows had too many null and duplicate values. They were unknowingly affecting the overall accuracy of the models. Also Identifying what columns to remove, considering the amount of columns was difficult.
* Identifying areas to visualize was a tedious task since it was necessary to provide meaningful insights. Identifying such areas needed a deeper study into the dataset.

**Things that Worked**

* Managed to give a few meaningful insights from the data by providing good visualizations.
* We also Managed to preprocess the data properly, so that the model could be trained properly. It improved the overall prediction of trained models in most algorithms.

**Future Work**

* The dataset is limited to US cars, a world wide dataset would benefit the public around the world. Also, an inclusion of electric cars and scooters would be an add-on.
* Apart from the attributes in the dataset, attributes like safety, comfort, infotainment, seating capacity, trunk capacity, ground clearance can also be considered for used car price prediction.

**Conclusion**

Various types of machine learning algorithms were implemented for prediction. We used various techniques to ensure our data is clean. We made decisions about dropping a column by analysing values in those columns. Focused more on providing various charts and figures for better understanding. Implemented LinearRegression, Ridge Regression, XGB Regression, Lasso Regression, Random Forest Regression, KNeighbor Regression.

**SECTION 5. Project Plan / Task Distribution**

|  |  |  |
| --- | --- | --- |
| **Component** | **Person assigned** | **Person who did the task** |
| Data Visualizations | Harshala Doiphode | Harshala Doiphode |
| Data Preprocessing | Heli Patel | Heli Patel |
| Model Training and Predication | Devangi Doshi | Devangi Doshi |
| Project Proposal | Everyone | Everyone |
| Project Presentation | Everyone | Everyone |
| **Project Report - Section 1** |  | |
| Motivation and Objective | Heli Patel | Heli Patel |
| Literature Review | Harshala Doiphode | Harshala Doiphode |
| **Project Report - Section 2** |  | |
| Algorithms Selected | Heli Patel | Heli Patel |
| Technologies & Tools | Heli Patel | Heli Patel |
| **Project Report - Section 3** |  | |
| Visualizations | Harshala Doiphode | Harshala Doiphode |
| Dataset and Preprocessing, Methodologies | Heli Patel | Heli Patel |
| Comparison of Algorithms, Analysis of Results | Devangi Doshi | Devangi Doshi |
| **Project Report - Section 4** |  | |
| Decisions Made | Devangi Doshi  Harshala Doiphode | Devangi Doshi  Harshala Doiphode |
| Difficulties Faced and Things that Worked | Heli Patel | Heli Patel |
| Conclusion | Harshala Dophode | Harshala Dophode |