Used Cars Database: Price Prediction

SPRING BOARD DSC CAPSTONE PROJECT 2 BY SNEHA RANI

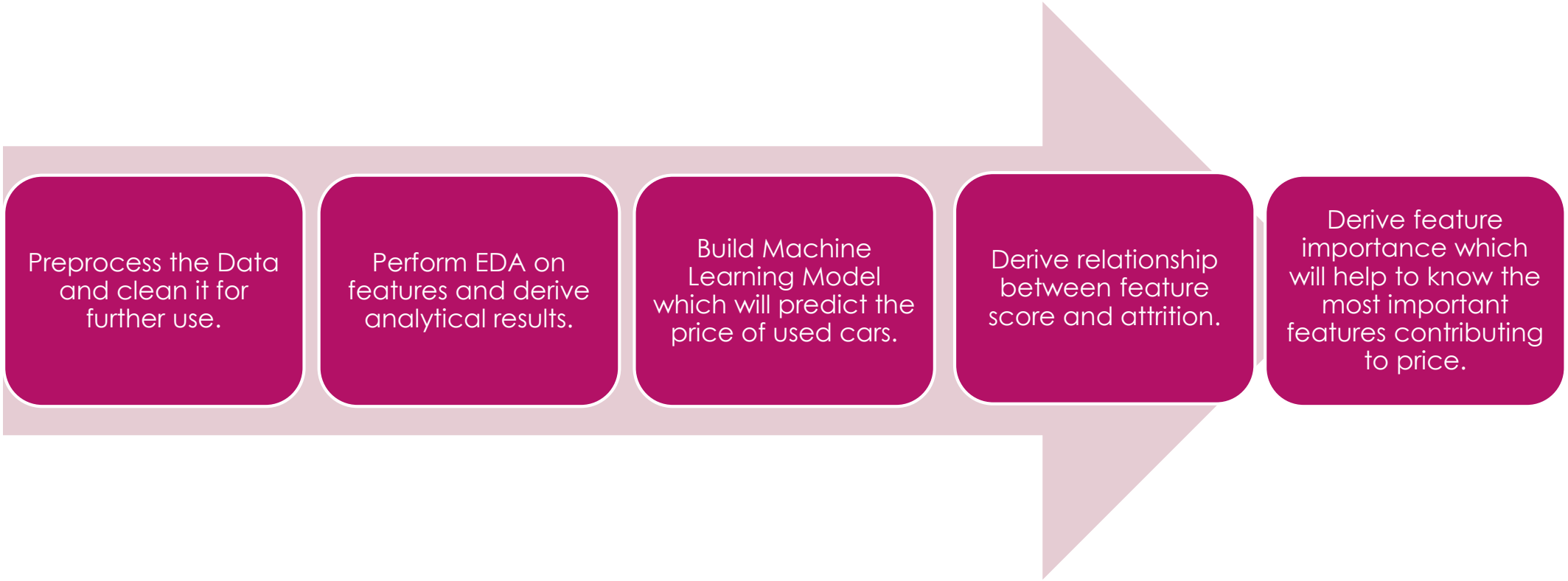
Introduction

- Used cars are of great value as they have higher life and last longer with good maintenance. Unfortunately, purchasing a used vehicle can also be just as complicated as figuring out which new car would suit one’s needs.
- The value of used cars depends on a number of factors. For example, the most important ones are usually the age of the car, its mileage, its horsepower etc.
- Machine learning techniques can work through those factors and will make it easy for customers to purchase pre-owned cars with the best budget.

Problem and Client

- Client : Ebay Classifieds
- Client would like to do better valuation of used car sale price derived from different features.
- With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. And providing a better pricing model for used car will help to win customers and enhance sale on used car.

Approach


Preprocess the Data and clean it for further use.
Perform EDA on features and derive analytical results.
Build Machine Learning Model which will predict the price of used cars.
Derive relationship between feature score and attrition.
Derive feature importance which will help to know the most important features contributing to price.

- **Pandas:** Loading the data, data wrangling, and manipulation
- **Scikit-learn:** Libraries for Regressors, Model evaluation, Feature Importance
- **Data Visualization:** MatplotLib, and Seaborn

Data Source

- Data was taken from the source : https://www.kaggle.com/orgesleka/used-cars-database
- Dataset has over 370,000 used cars scraped with Scrapy from Ebay-Kleinanzeigen. The content of data is in German-English so it does not follow American English words.
- Dataset consists of 20 features and consist of (Int and Object) data type

Data Cleaning/Wrangling

Exploratory Data Analysis

What is the average price of the car based on vehicleType?

vehicleType	price
andere	720695.185737
bus	10452.253687
cabrio	15292.173537
coupe	26703.163520
kleinwagen	5826.302574
kombi	7912.791616
limousine	11359.258957
not-available	22345.811762
suv	13430.022687

The table showed the average price of each vehicle type, among which 'andere' vehicle type having (720695.185737 Euro) was a bit costlier than others (for instance, bus, cabrio, coupe, kleinwagen, kombi, limousine, suv).

Exploratory Data Analysis (Continued…)

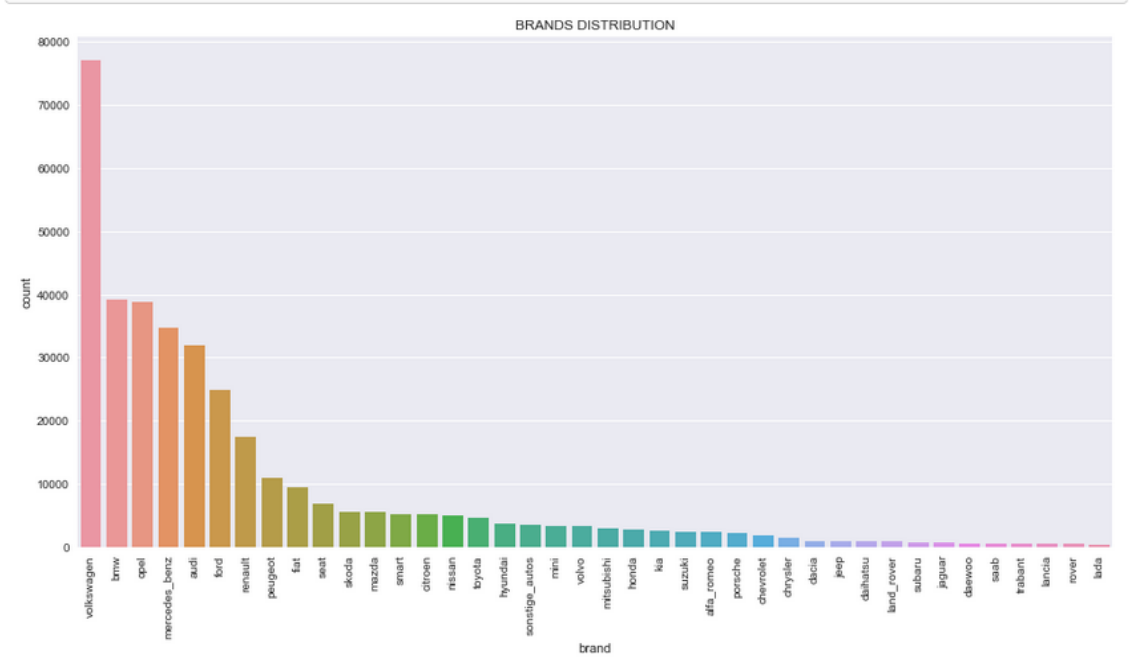
Can we find out the min, max and average price of the car of different brands?

	price_min	price_max	price_mean
brand			
alfa_romeo	1	74185296	36999.409713
audi	1	99999999	16306.013191
bmw	1	99999999	15263.303998
chevrolet	1	999999	7655.222841
chrysler	1	37500	4117.356264
citroen	1	27322222	9089.353743
dacia	1	19990	5905.268539
daewoo	1	4200	1034.998124
daihatsu	1	12850	1761.402581
fiat	1	12345678	5503.193059
ford	1	99999999	8708.254527
honda	1	48500	3946.573153
hyundai	1	35999	5496.463808

As shown in the table, car prices varied from 1 to 99999999 and had varied average price. Prices as low as 1 and as high as 999999999 did not define well for the model. So, it was further cleaned(discussed later) and filtered for better price range data.

Exploratory Data Analysis (Continued…)

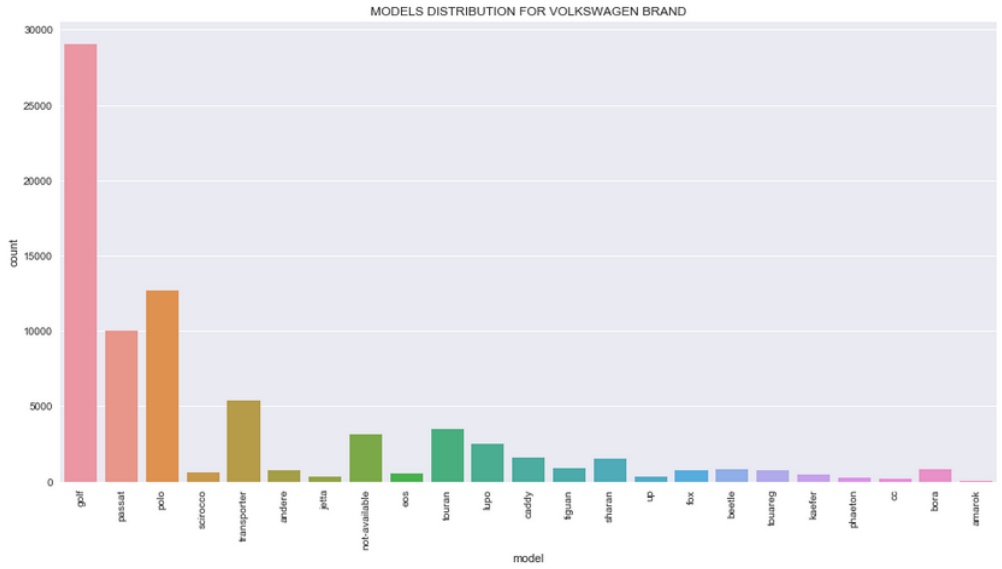
Which is the most popular brand among used car in the market?



Based on the bar graph, this column had 40 different brands and Volkswagen was listed as the most popular brand. BMW was the next competitor.

Exploratory Data Analysis (Continued…)

Which is the most running model in Volkswagen?



From the Graph shown, ‘Golf’ was found to be the most popular and running model in Volkswagen. Polo seems to be the second most popular model.

Exploratory Data Analysis (Continued…)

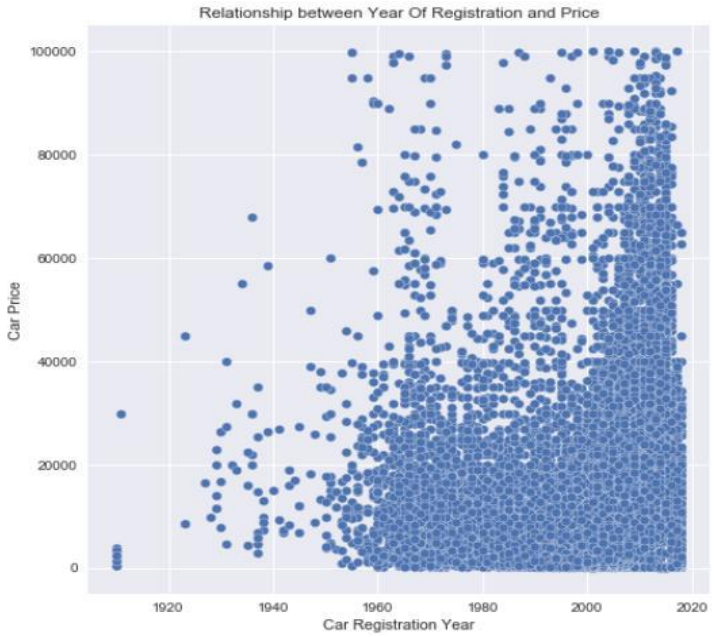
What is the min, max and average horse power of different vehicle type?

	powerPS_min	powerPS_max	powerPS_mean
vehicleType			
andere	0	12684	102.846910
bus	0	12512	114.009271
cabrio	0	16312	145.921410
coupe	0	20000	174.278791
kleinwagen	0	15020	68.992328
kombi	0	19312	136.838470
limousine	0	19211	132.636122
not-available	0	16011	73.757670
suv	0	17322	166.731899

As shown in Table, Minimum powerPS of a car is **0** and maximum is **20000** powerPS which is not a valid value. So, I have put a valid range for powerPS values to have better data understanding in order to avoid inconsistent data.

Exploratory Data Analysis (Continued…)

Is there any variance in the price of the car based on year of registration? Does vehicle of various type impact differently on car price and year of registrations?



As shown in the scatter plot, based on year vs. price, I saw that newly registered cars seemed to be more expensive than the older ones. There were also few old cars which were expensive too.

- Scatter Plot shows that there exists a positive relationship between year of registration and price.
- On further grouping the data points by Vehicle Type they still showed a positive correlation.
- Also Pearson coefficient between Year of Registration and Price was 0.35 which shows a weak relationship, but for specific Vehicle Type ‘Kombi’ it had a strong coefficient of 0.70.

Modeling the Data

Is My Data ready to apply Machine Learning Algorithms?

Steps that I followed to get my Data Ready:

- Pre-cleaned the data for missing value in some columns by introducing a new Category 'Not Available'.
- Certain columns were deemed insignificant and so were not used in the project.
- Some of data range was not logically possible and so data was filtered based on possible range values only.
- Created Dummy Variables for categorical variables for easy interpretation.
- After all these, final dataframe was created and it had 67 columns.

Modeling the Data

Training and Test Datasets

When Fitting Models, I wanted to ensure two things:

- ✓ Finding the best Model(in terms of Model Parameters)
- ✓ Model is highly likely to generalize i.e. perform well on unseen data.

Approach:

- I will try first by building linear regression base line model using 'L1' and 'L2' regularization by :
- Splitting the data into a training and test (hold-out) set
- Train on the training set, and test for accuracy on the testing set

Modeling the Data with Linear Regression

I first built Linear Regression Base model and calculated the variance score and Mean Squared error(as shown below):

Linear Regression Train and Test Scores

Item	Train_Score_Values	Test_Score_Values
Mean absolute error	2860.6473	2867.7233
Mean squared error	24678214.3804	24670192.4396
Root Mean squared error	4967.7172	4966.9097
Variance Score	0.6259	0.6206

Variance score of 0.6259 and low MSE (24670192.4396) quantifies the quality of the model trained on the training set.

Modeling with Ridge

- I started off with Ridge Regression technique for computing R-squared scores over a certain range of alphas as shown in (Table on Left).
- Keeping alpha value = 0.05, I then computed Ridge Regression train and test scores based on Mean absolute error, Mean squared error, Root Mean squared error, Variance Score (Table on Right).

```
Alpha: 0.05      0.620337638841
Alpha: 0.1       0.61850754865
Alpha: 0.5       0.583651670405
Alpha: 5.0       0.296858182842
Alpha: 10.0      0.190293973132
```

Ridge Regression Train and Test Scores

Item	Train_Score_Values	Test_Score_Values
Mean absolute error	2829.0751	2833.9009
Mean squared error	24740711.8781	24689268.0231
Root Mean squared error	4974.0036	4968.8296
Variance Score	0.6250	0.6203

Modeling with Lasso

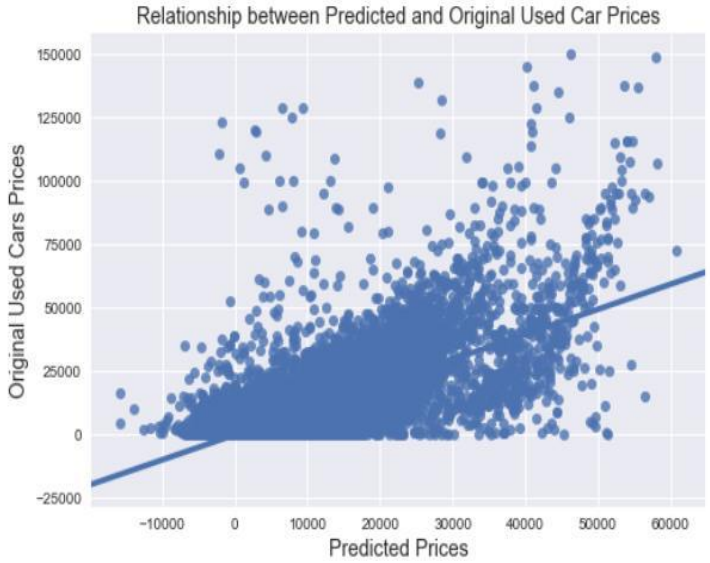
- I applied the same with Lasso Regression techniques and calculated the test scores.

```
Alpha score: 0.05        0.62018171113
Alpha score: 0.1         0.618932600607
Alpha score: 0.5         0.600822665333
Alpha score: 5.0         0.34978309345
Alpha score: 10.0        -1.0002767401e-06
```

Lasso Regression Train and Test Scores

Item	Train_Score_Values	Test_Score_Values
Mean absolute error	2852.8866	2860.0955
Mean squared error	24709049.7275	24699407.9301
Root Mean squared error	4970.8198	4969.8499
Variance Score	0.6255	0.6202

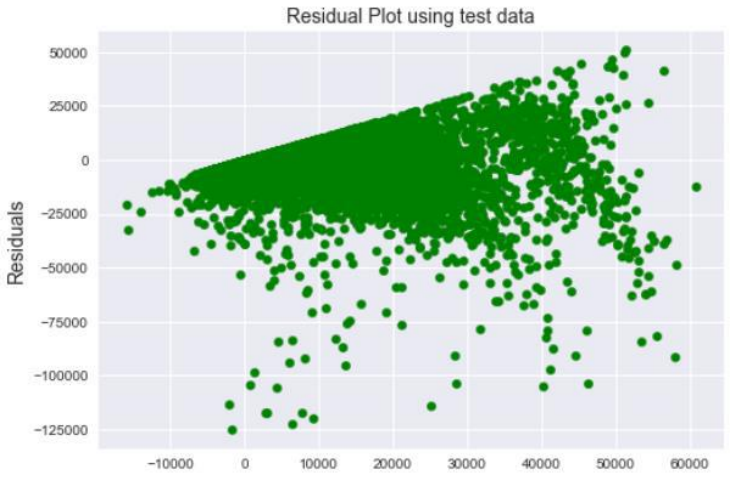
Conclusion with Base, Ridge and Lasso Regressions

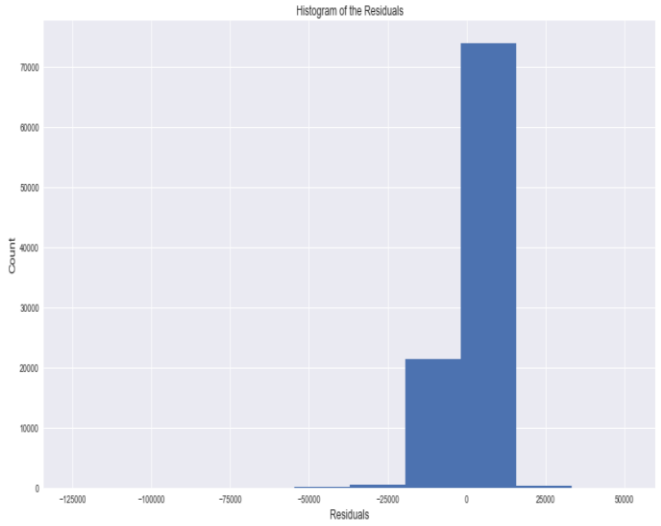


- Variance score (0.6203) of the Lasso model is almost same as compared to the plain Linear and Ridge models
- From all the analysis done till now, Linear Regression model is predicting better as it has a good R-squared score with minimum MSE (24670192.4396) compared to both Lasso and Ridge regression.
- Considering Linear Regression as a better model, I plotted scatterplot between the predicted prices and the original prices to evaluate how well regression model predicts the price.

Continued…

I further plotted the residual plot and histogram graph on test data.





It can be seen that the variance of error terms (residuals) was not constant. It was spreading out as the predicted value increased thereby unevenly scaling down the model's performance. The plot shows the non-linearity in the data which was not captured by the model.

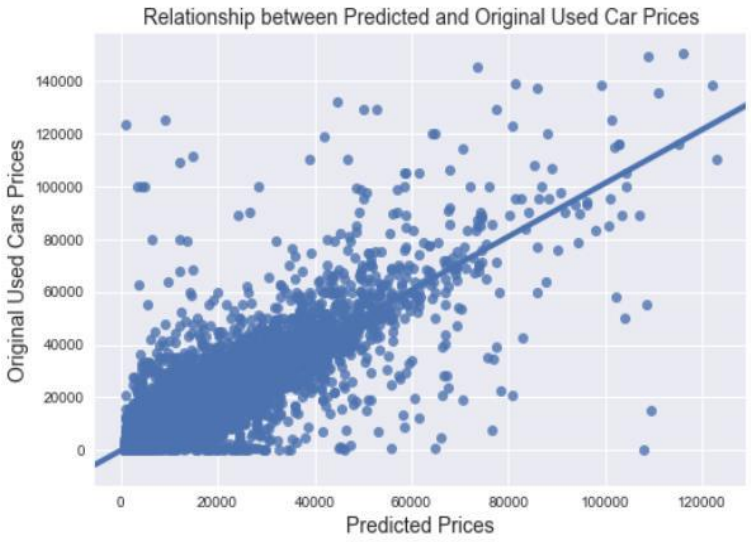
Modeling with Random Forest

- In order to get better model, I applied hyper-parameter tuning with Random Forest algorithm in order to make sure that our model is general and it works beyond our data sets.
- From the table, we can see that we have increased the R-squared value 0.62 to 0.83 with minimum MSE (11202998.2023) score by removing some of the outliers initially and tuning the model by selecting the best set of hyper-parameters which clearly shows the best fit from all the above models.

Random Forest Test Scores

Item	Test_Score_Values
Mean absolute error	1646.6879
Mean squared error	11202998.2023
Root Mean squared error	3347.0880
Variance Score	0.8277

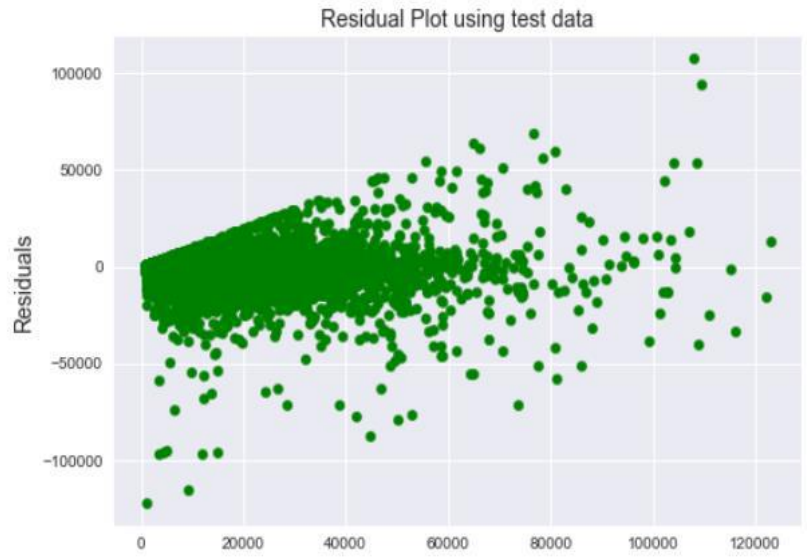
Continued....

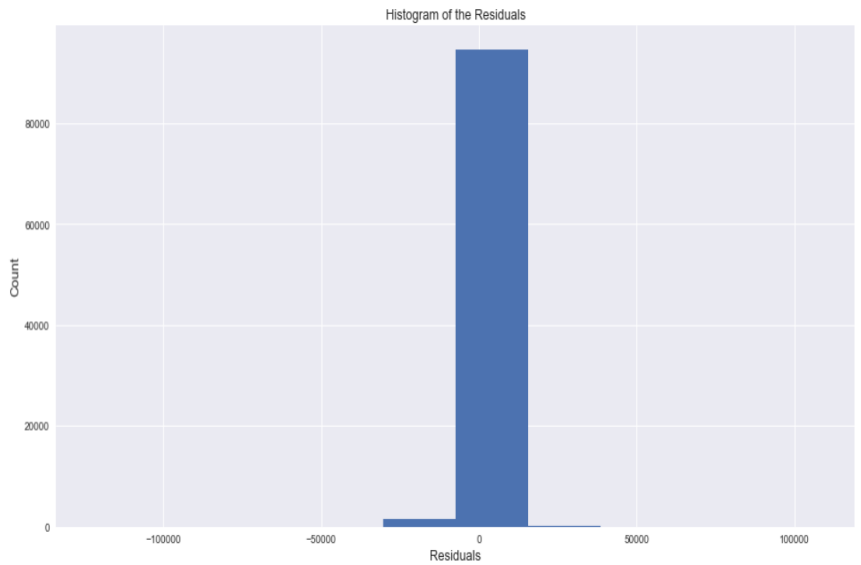


As shown in graph, we see strong correlation showing the upward trend. Also, the data points were not too dispersed and lied on the 45-degree line which tends to be a perfect model.

Continued…

I further plotted the residual plot to assess observed error is consistent with random error. As seen, variance of error terms(residual) is moderately constant, its not spreading out.





Feature Importance

```
features
powerPS                0.402250
yearOfRegistration     0.359745
kilometer              0.118886
brand_porsche          0.042547
Name: positive, dtype: float64
```

- I computed the relative importance of each attribute by finding the highest coefficient value.
- As seen in the table, features as powerPS, yearOfRegistration, kilometer, brand_porsche with the highest score which makes sense and are somewhat linked to the features one might think influence customers to buy a used car with better information.

- I was able to develop a Random Forest Regressor Model with a good R-squared value of 0.83 and minimum MSE (11202998.2023) which will help the client to use machine learning model to predict the price of used cars.
- With exploratory data analysis I am able to suggest different useful findings on different features which will help to understand used car prices in a better way.

Recommendations for the Client

- ✓ ***Client should implement this model to predict better pricing of used cars which will be of great value to the customer.***
- ✓ ***Also, Feature Importance and Exploratory Data Analysis will help client to know important features.***
- ✓ ***This model will be really beneficial to both Buyer and Seller.***

Further Research

- I tried choosing a specific range of values for tuning the model using Random Regressor which is not a good approach to start with. I would rather try to tune multiple parameters using Grid Search CV for my future work.
- I will try to use advanced techniques as **Gradient Boosting Regression** for my car dataset which seems to be a high-performing tool to enable better predictions.

THANK YOU


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data