CAR PRICE ANALYSIS AND PREDICTION

ISM 6136 Data Mining

Dr. Kiran Garimella

University of South Florida

Group Project on Full Data Mining Project Analysis

By

Aishwarya Kulkarni

Keerthana Yelchuri

Kshitija Nandkishor Dound

Saroni Sinha

Background

Cars are life and blood of the commutation. Making it affordable is the biggest challenge encountered by the stakeholders. There are several factors that affect price of the car. Each of the factors plays unique role. These attributes spread across fuel type, car body, engine size, horsepower, no. of cylinders, etc. Prediction of the price based on these factors can be a game changer. Stakeholders can draw useful conclusions from such predictions. For example, manufacturers can use such predictions to reduce the total cost based on certain manipulations in these factors and maximize the total profit. Another good example can be, vendors using such predictions to capture the market and boost the sales. Moreover, such predictions can be a valuable aid to the customers while ranking their preferences and corresponding them with the monetary value they are willing to spend. Such insights can contribute towards globalization as well. This can be untangled to a scenario where companies enter the markets of foreign countries by increasing their delivered quality at an affordable value.

Motivation

The prices of new cars within the business are fastened by the manufacturer with some extra prices incurred by the government in the form of taxes. So, customers shopping for a brand new automotive can be assured of the cash they invest to be worthy. In auto sales, several factors contribute towards the total cost of the cars. For example, fuel type, car's body, engine size, horsepower, no. of cylinders, etc. With the sales of the cars increasing and changing rapidly worldwide, what affects the units of new car sales has become a topic of mass interest. Based on the existing data, we aim to use the data mining algorithms to predict the price of the cars based on the factors that affect the price.

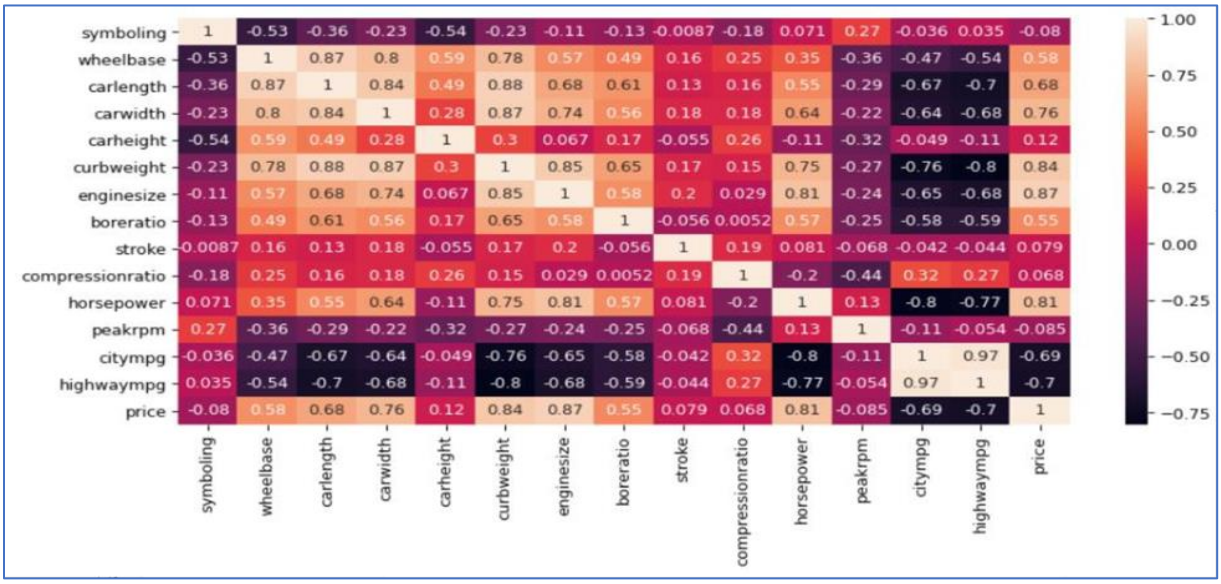
Solution Methodology and Metrics

In an effort to forecast the prices of car and evaluate the metrics that make better predictions we have used automobile data set from UCI Machine Learning Repository, which consists of car prices based on its technical features. Using the algorithms mentioned below we will predict the *price* variable. We preferred cleaning the dataset first. After extracting the cleaned data, we have divided and trained the data in 80% training set - 20% testing set mechanism.

We are considering to train the data using following algorithms: -

- Linear Regression (Ordinary Least Squares) having seed value as 6136.
- Linear Regression (Online Gradient Descent) having seed value as 6136.
- Boosted decision tree regression using maximum number of leaves 20/10, minimum number of samples from per leaf node from 10/ 5, learning rate from 0.2/ 0.1 and total number of trees constructed from 100/ 50.

Evaluation Metrics



Considering above metrics Price is highly and positively correlated with wheelbase, carlength, carwidth, curbweight, enginesize and horsepower. Price is negatively correlated to citympg and highwaympg (-0.70 approximately). This suggest that cars having high mileage may fall in the 'economy' cars category, and are priced lower. These cars are designed to be affordable by the budget buyers who value more fuel efficiency mileage over powerful engine.

Description of Dataset

Dataset consists of prices of cars based on its technical specifications such as car manufacturer, its engine capacity, fuel efficiency, body-type, etc. The dataset contains 205 rows and 26 columns.

This data set consists of three types of entities:

- The specification of an auto in terms of various characteristics,
- Its assigned insurance risk rating.
- Its normalized losses in use as compared to other cars.

The second rating corresponds to the degree to which the auto is riskier than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is riskier (or less), this symbol is adjusted by moving it up (or down) the scale. A value of +3 indicates that the auto is risky, −3 that it is probably pretty safe.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size and represents the average loss per car per year.

Dataset source: https://archive.ics.uci.edu/ml/datasets/Automobile

The above dataset consists of data taken from 1985 Ward's Automotive Yearbook. Here's the list of original sources of the data:

1. 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
2. Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038.
3. Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037.

Sr No.	Attribute	Attribute Information
1	Car_ID	Unique id of each car (Interger)
2	Symboling	Assigned insurance risk rating; a value of +3 indicates that the car is risky; -3 suggests that it is probably a safe car (Categorical)
3	carCompany	Name of car company (Categorical)
4	fueltype	fuel-type i.e. petrol or diesel (Categorical)
5	aspiration	Aspiration used in a car (Categorical)
6	doornumber	Number of doors in a car (Categorical)
7	carbody	Body-type of a car (Categorical)
8	drivewheel	Type of drive wheel (Categorical)
9	enginelocation	Location of car engine (Categorical)
10	wheelbase	Weelbase of car (Numeric)
11	carlength	Length of car (Numeric)
12	carwidth	Width of car (Numeric)
13	carheight	Height of car (Numeric)
14	curbweight	The weight of a car without occupants or baggage (Numeric)
15	enginetype	Type of engine (Categorical)
16	cylindernumber	Number of cylinders placed in the car engine (Categorical)
18	fuelsystem	Fuel system of a car (Categorical)
19	boreratio	Bore ratio of car (Numeric)
20	stroke	Stroke or volume inside the engine (Numeric)
21	compressionratio	Compression ratio of an engine (Numeric)
22	horsepower	Power output of an engine (Numeric)
23	peakrpm	Peak revolutions per minute (Numeric)
24	citympg	Mileage in city (Numeric)
25	highwaympg	Mileage on highway (Numeric)
26	price(Dependent variable)	Price of a car (Numeric)

Data Cleaning

```
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64
 14  enginetype        205 non-null    object
 15  cylindernumber    205 non-null    object
 16  enginesize        205 non-null    int64
 17  fuelsystem        205 non-null    object
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64
 22  peakrpm           205 non-null    int64
 23  citympg           205 non-null    int64
 24  highwaympg        205 non-null    int64
 25  price             205 non-null    float64
```

Our Car dataset has a total of 26 columns as we can see in the above screenshot. To clean our dataset, we used python. Here is the Link to our colab document that shows the data cleaning codes.

Below are the reasons and methods used to clean our dataset before we can deploy it on the Azure Studio.

4. The first step is to remove the car_ID column from the dataset. Whenever we deploy our dataset on azure, we always try to remove the index columns and include only the columns that have our main information.
5. When we see the column “CarName”, it has few errors. Some of the cars have names as “vw” and “vokswagen”. Instead it should be “Volkswagen”. Similarly, “porcshce” should be “Porsche”, “toyouta” should be “toyota”, “Nissan” should be “nissan” and “maxda” should be “mazda”
6. We can also see that s lot of columns in the dataset are of type Object. We need to convert them into int datatype so that we can use them in our analysis. So, the columns like 'doornumber', 'cylindernumber’ which have the values in words need to be converted into numbers. (refer screenshot below).

	doornumber	cylindernumber
0	two	four
1	two	four
2	two	six
3	four	four
4	four	five

→

cylindernumber	doornumber
4	2
4	2
6	2
4	4
5	4

Similarly, columns like carbody have 5 different categories. We converted each category into different columns.

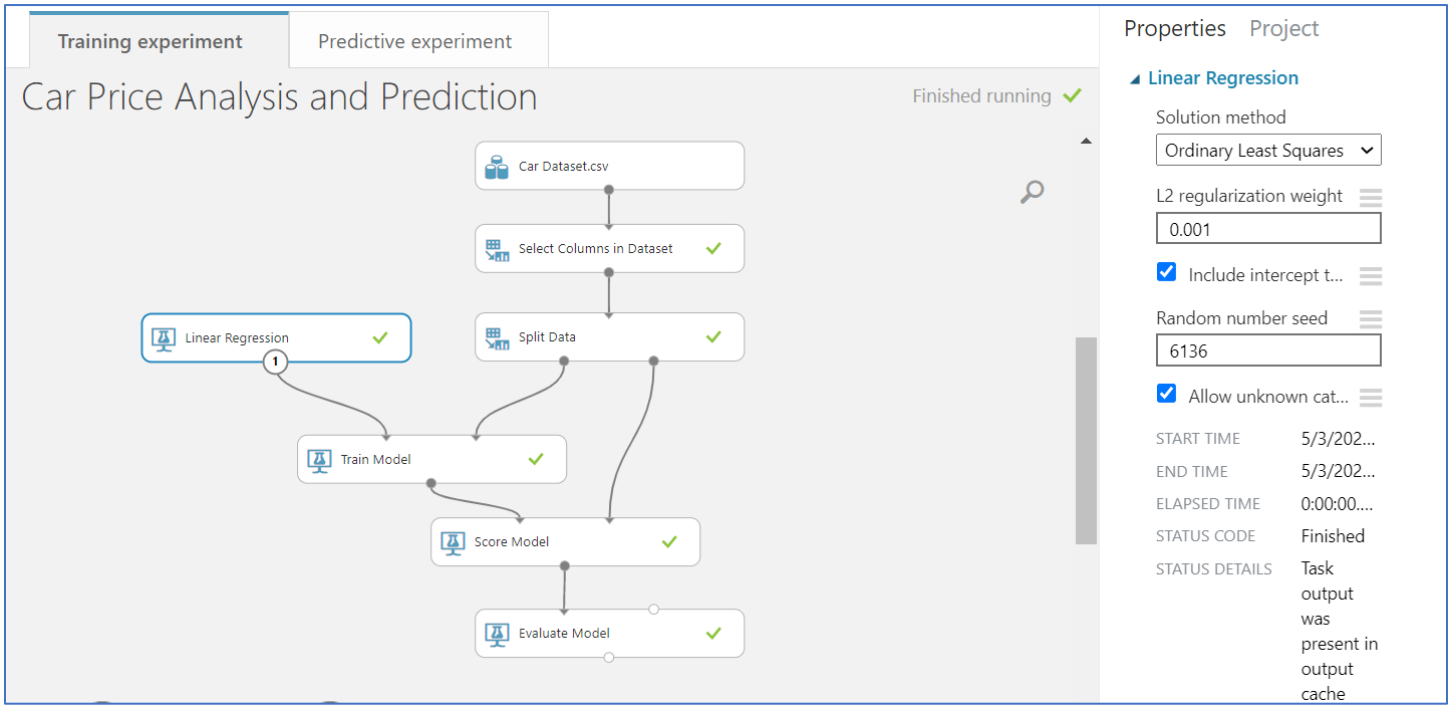
Predictive model in Azure ML : Comparative analysis of algorithms

In order to predict the price of car, we have built two predictive models using:

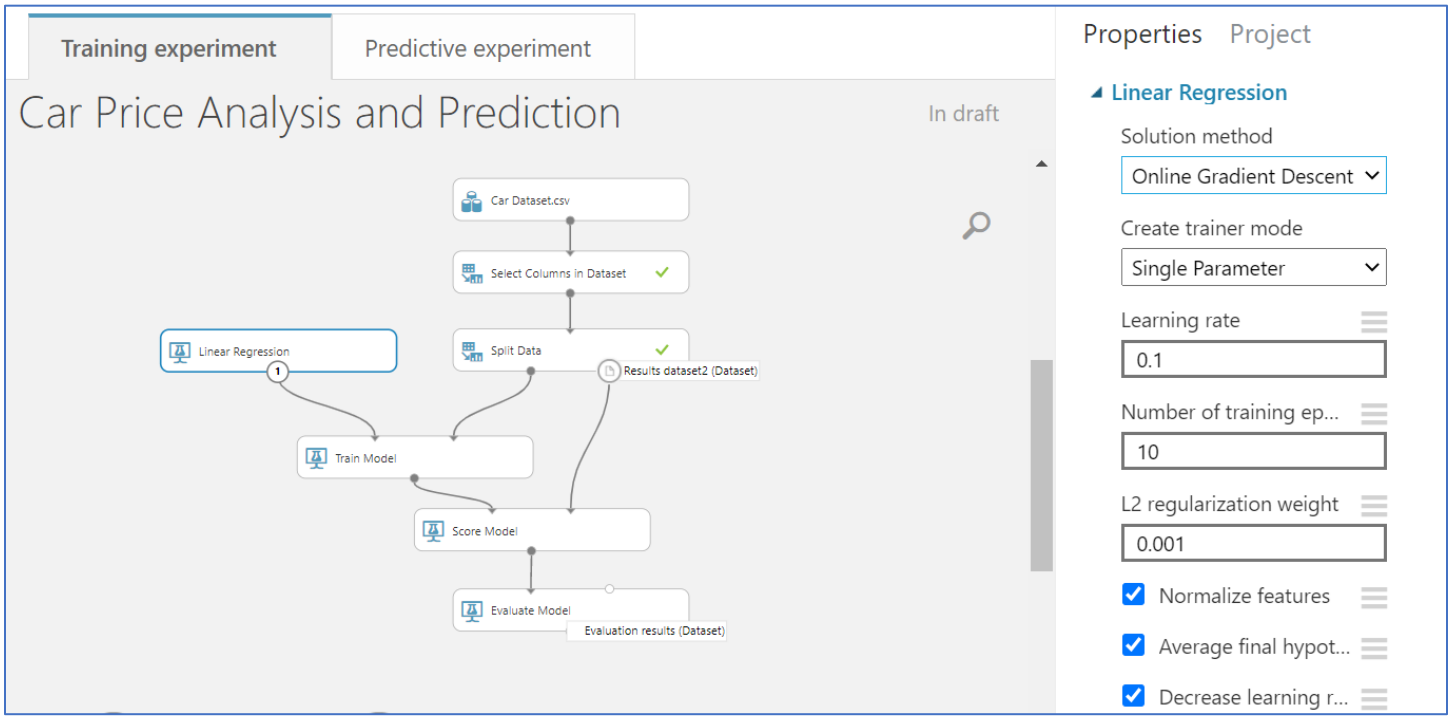
- Linear Regression
- Boosted Decision Tree Regression

Linear Regression:

Model 1:



Model 2:



In our Model 1, we are using the Ordinary Least Squares as the Linear Regression solution method, whereas, in Model 2, we are using Online Gradient Descent.

Explanation of Tasks:

Column Selection: We have selected all the columns from the dataset to predict the dependent variable Price.

Split Data: Splitting the data into training data and testing data (Train=80%,Test=20%). We have used random seed 6136.

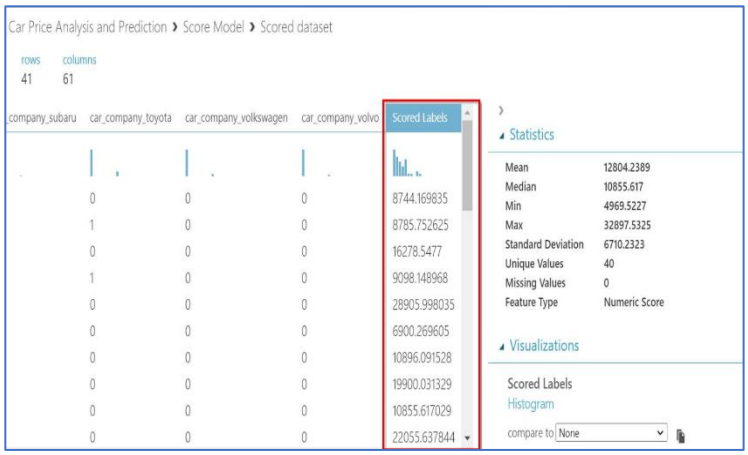
Train Model: We are training the model on the Price variable.

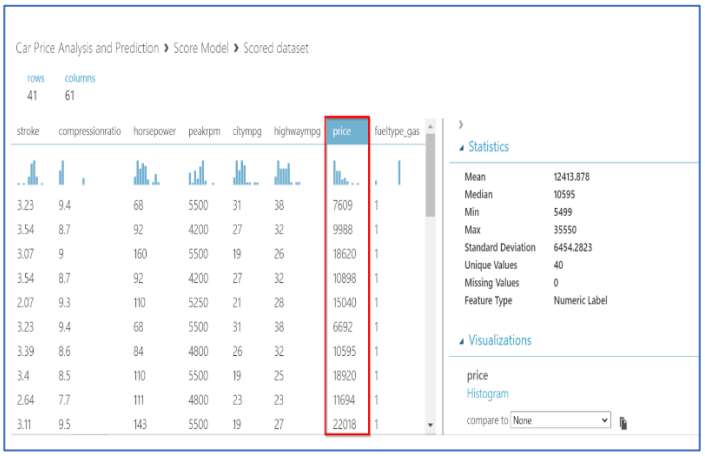
Score Model: We are testing the model and predicting the car price variable.

Result Interpretation:

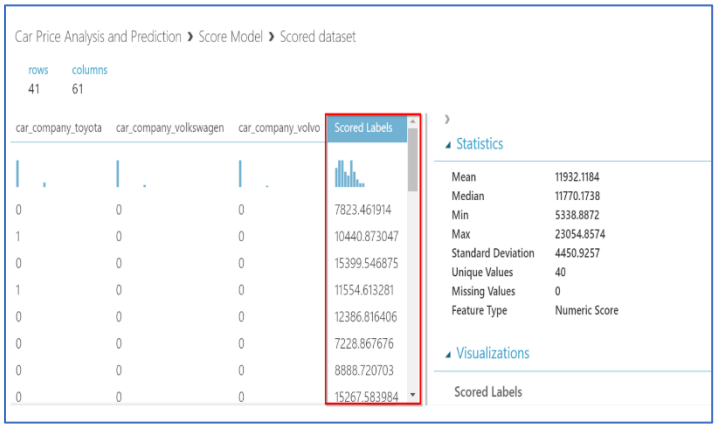
The scored labels column is the result column in this model. The numbers are the model-predicted price of the car, which can be compared with the actual price.

Model 1 (Ordinary Least Squares):



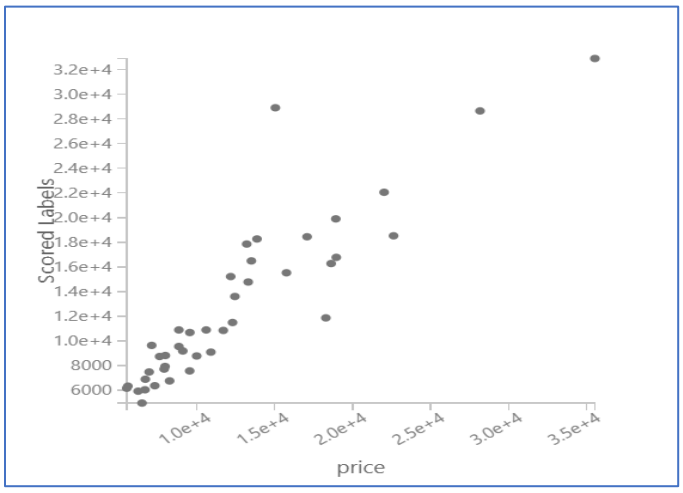


Model 2 (Online Gradient Descent):

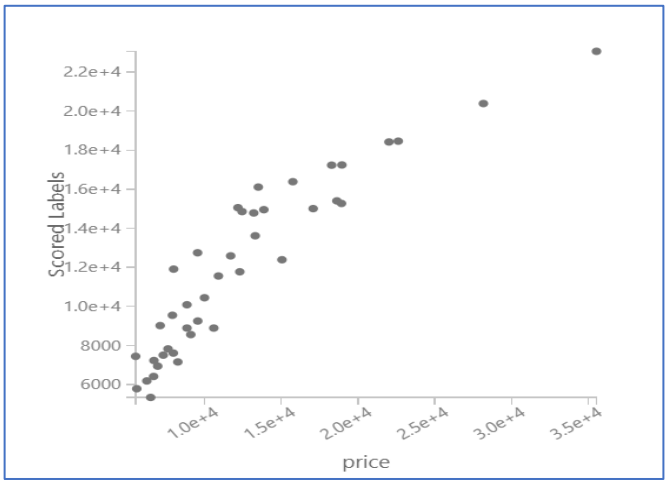


We can visualize how good our model fit is by the comparing the linear regression model which plots the actual vs the predicted values.

Model 1:

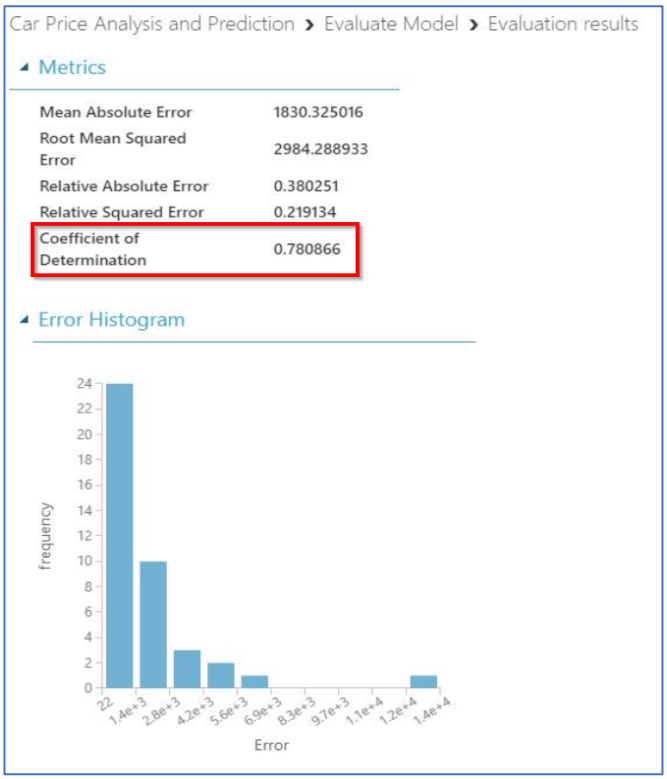


Model 2:

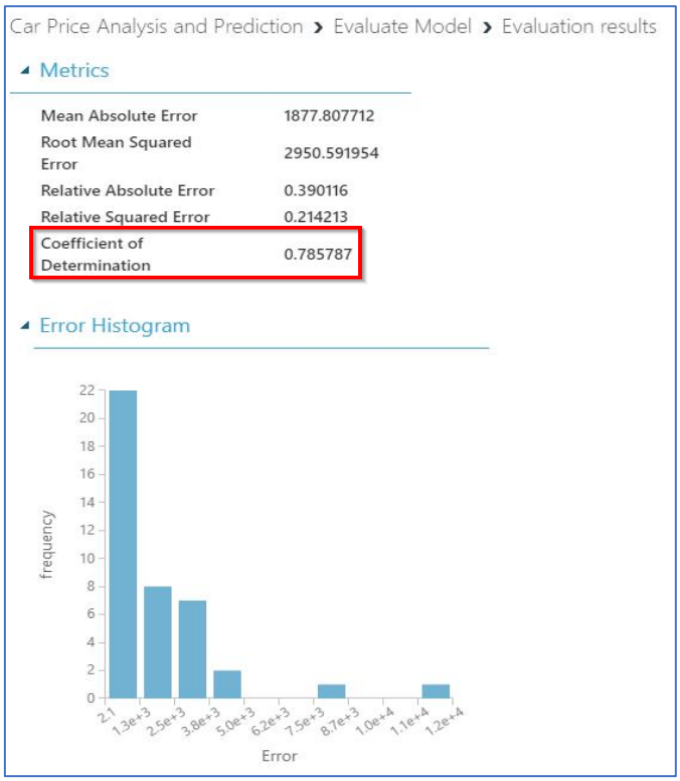


Evaluate Model:

Model 1:



Model 2:

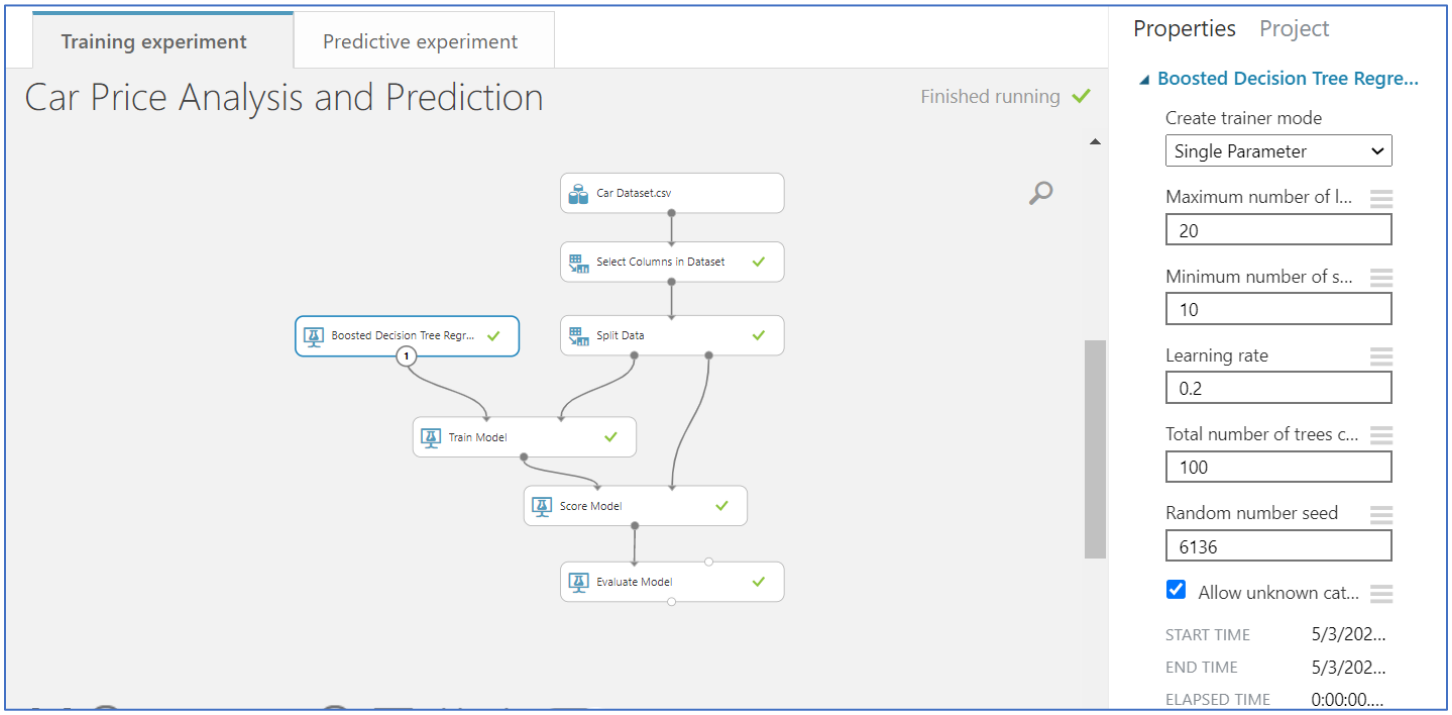


Both the linear regression models seem to be a good fit based on the metrics that we see.

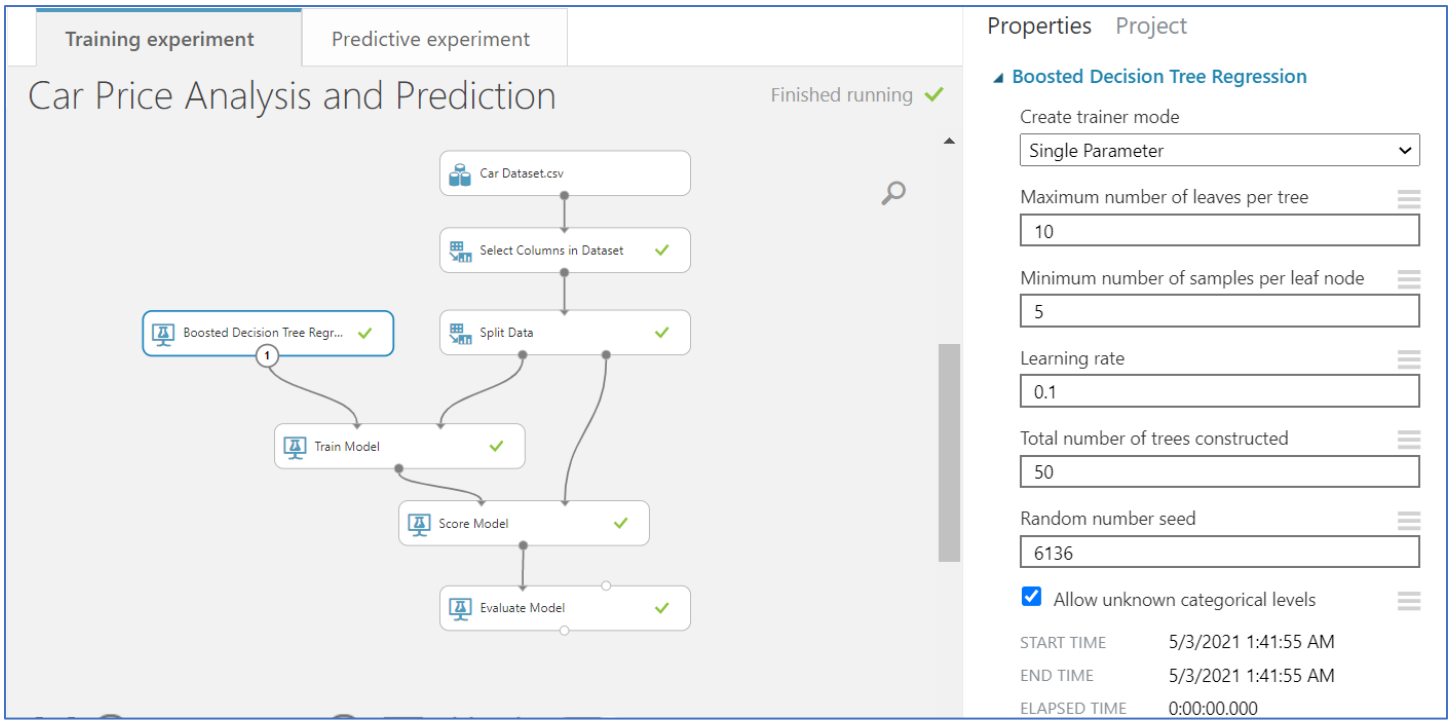
They have high coefficient of determination which specifies the measure of how much the variability of price variable can be explained by the independent variables. Thus, it signifies goodness of fit.

Boosted Decision Tree Regression:

Model 3:



Model 4:



In our Model 3, we are having the default values for a boosted decision tree regression model. Whereas, in Model 4, we change the maximum number of leaves from 20 to 10, minimum number of samples from per leaf node from 10 to 5, learning rate from 0.2 to 0.1 and total number of trees constructed from 100 to 50.

Explanation of Tasks:

Column Selection: We have selected all the columns from the dataset to predict the dependent variable Price.

Split Data: Splitting the data into training data and testing data (Train=80%,Test=20%). We have used random seed 6136.

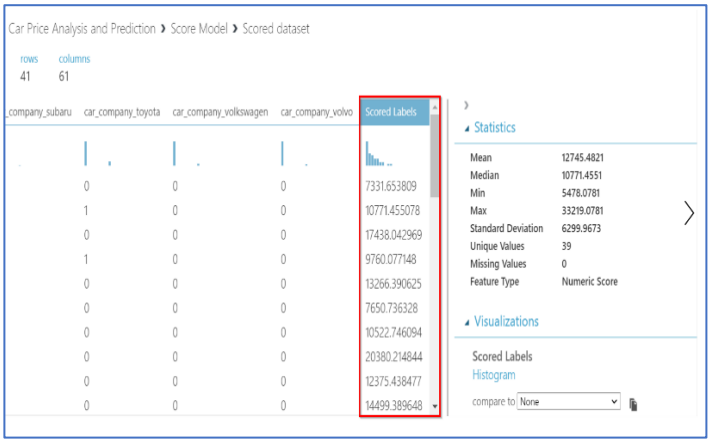
Train Model: We are training the model on the Price variable.

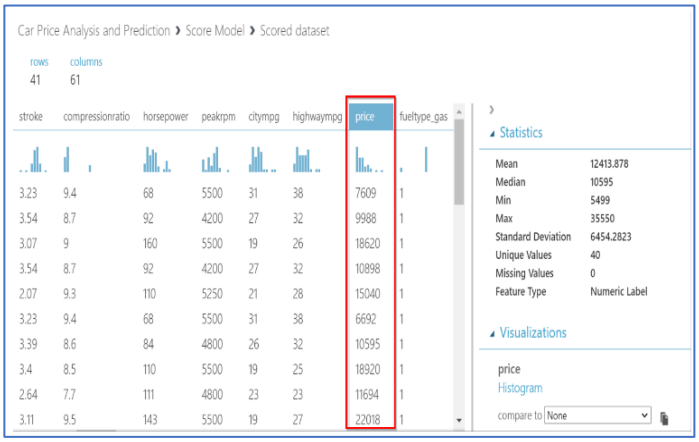
Score Model: We are testing the model and predicting the car price variable.

Result Interpretation:

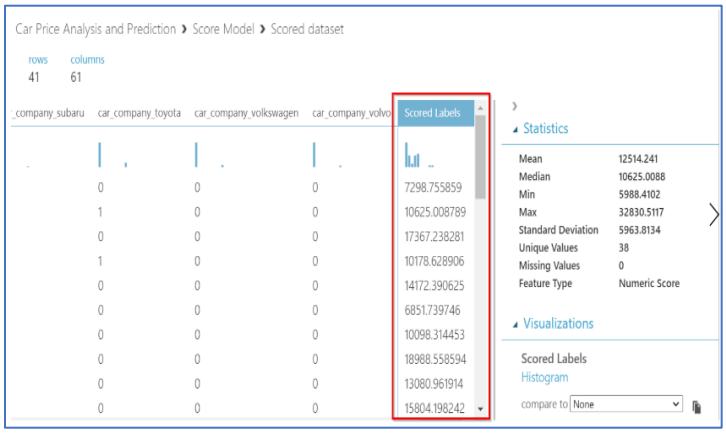
The scored labels column is the result column in this model. The numbers are the model-predicted price of the car, which can be compared with the actual price.

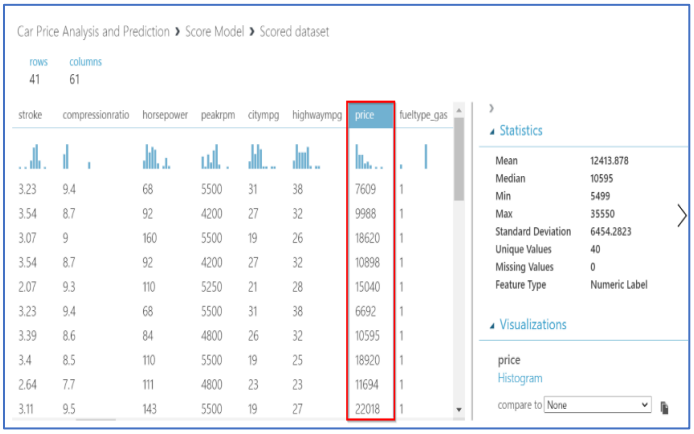
Model 3:





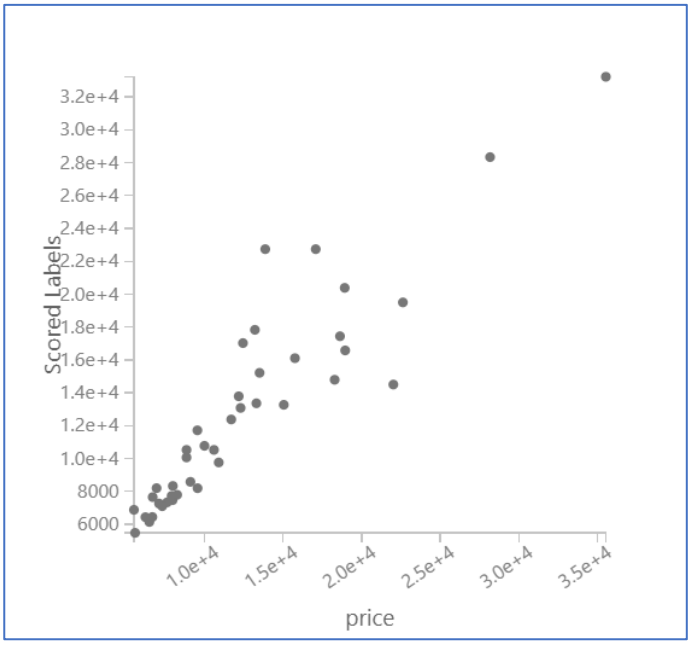
Model 4 :



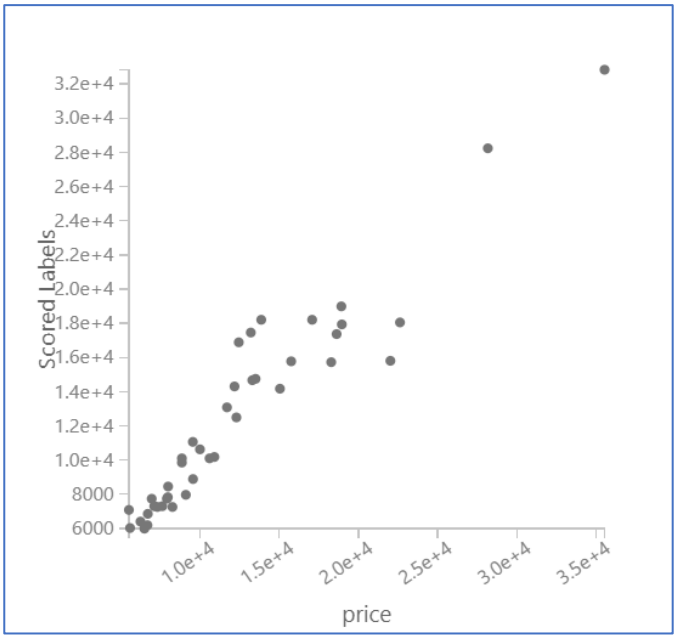


We can visualize how good our model fit is by the comparing the linear regression model which plots the actual vs the predicted values.

Model 3:



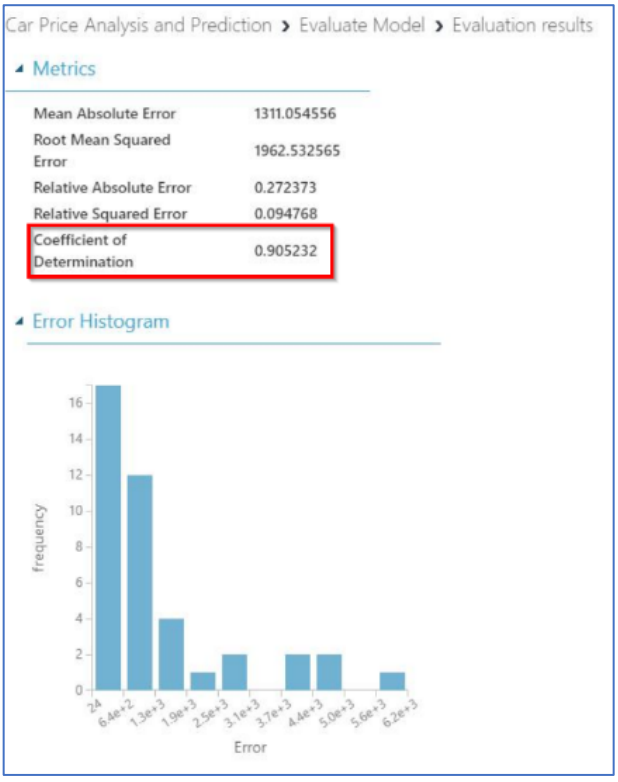
Model 4:



Evaluate Model:

Model 3:

Model 4:



This Model 3 is a good fit, as we can see from the coefficient of determination and root mean square error metrics. The high coefficient of determination proves the model's goodness of fit in predicting price of car.

In comparison to Model 3, Model 4 is a comparatively stronger fit, since the coefficient of determination is greater as well as the root mean square error is less.

Summary Sheet

Model	Algorithm	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Model 1	Linear Regression (Ordinary Least Squares)	1830.325016	2984.288933	0.380251	0.219134	0.780866
Model 2	Linear Regression (Online Gradient Descent)	1877.807712	2950.591954	0.390116	0.214213	0.785787
Model 3	Boosted Decision Tree Regression	1655.813822	2588.197839	0.343997	0.164825	0.835175
Model 4	Boosted Decision Tree Regression	1311.054556	1962.532565	0.272373	0.094768	0.905232

Recommendations for company

Whenever the company wants to predict the price of any car, the web service deployed on azure based on our data mining model will help in predicting the price based on the input parameters as you see in the below image

This will help the management to understand how exactly the price vary with the independent variables. Thus, they can accordingly manipulate the design of the cars and the business strategy to meet certain price levels so as to maximize the profit.

Input to the web service:

Enter comma-separated values below:

symboling,doornumber,wheelbase,carlength,carwidth,carheight,curbweight,cylindernumber,enginesize,boreratio,stroke,compressionratio,horsepower,peakrpm,citympg,highwaympg,price,fueltype_gas,aspiration_turbo,carbody_hardtop,carbody_hatchback,carbody_sedan,carbody_wagon,drivewheel_fwd,drivewheel_rwd,enginelocation_rear,enginetype_dohcv,enginetype_l,enginetype_ohc,enginetype_ohcf,enginetype_ohcv,enginetype_rotor,fuelsystem_2bbl,fuelsystem_4bbl,fuelsystem_idi,fuelsystem_mfi,fuelsystem_mpfi,fuelsystem_spdi,fuelsystem_spfi,car_company_audi,car_company_bmw,car_company_buick,car_company_chevrolet,car_company_dodge,car_company_honda,car_company_isuzu,car_company_jaguar,car_company_mazda,car_company_mercury,car_company_mitsubishi,car_company_nissan,car_company_peugeot,car_company_plymouth,car_company_porsche,car_company_renault,car_company_saab,car_company_subaru,car_company_toyota,car_company_volkswagen,car_company_volvo
3,2,88.6,168.8,64.1,48.8,2548,4,130,3.47,2.68,9,111,5000,21,27,13495,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

Predicted price of the car as output (using our best fit model- boosted decision tree):

Conclusion

Prediction of car prices episodes a pertinent problem for the stakeholders in the car market. This project demonstrates a relevant solution to such problems. The project further shows the ways in which rigorous data modelling contributes towards ultimate delivery of reliable predictions with a decent accuracy. The experiment further shows 78% of the variance in the price (dependent variable) is explained collectively by the independent variables using linear regression; and boosted decision tree minimizes the overall prediction error by giving the prediction accuracy of approximately 90%. On comparing both the algorithms, we can observe that the boosted decision tree algorithm is more effective in forecasting the price of the cars.

Both Intra and Inter country perks of such predictions make them useful to both global and local companies. For instance, it can be helpful for a multiplayer like Toyota. Toyota has a strong customer base in several developed countries. However, it fails to do the same in several other developing countries like India. Thus, it can use such predictions and develop a strong base in these countries.