Verzeo Machine Learning Batch April-May ML041B11

Major Project Report

Predicting car prices using given data-sets

Contributers

Anoushka Sinha: Data cleansing and exploratory data analysis.

Srashti Gangil: Data cleansing and exploratory data analysis.

Mohit Kamal: Data cleansing and exploratory data analysis.

Mahaan Mitra: Model Implementation.

Sarthak Srivastava: Model Implementation.

Vikash Tiwari: Model Implementation.

Kalpit Gupta: Model Implementation.

Nishant Athawale: Project Report.

Problem Statement

Predict the costs of used cars given the data collected from various sources and distributed across various locations in India.

List of features and their description

* **Name:**The brand and model of the car.
* **Location:** The location in which the car is being sold or is available for purchase.
* **Year:**The year or edition of the model.
* **Kilometres Driven:**The total kilometres driven in the car by the previous owner(s) in KM.
* **Fuel Type:**The type of fuel used by the car.
* **Transmission:**The type of transmission used by the car.
* **Owner Type:** Whether the ownership is First-hand, Second hand or other.
* **Mileage:**The standard mileage offered by the car company in kmpl or km/kg
* **Engine:**The displacement volume of the engine in cc.
* **Power:**The maximum power of the engine in bhp.
* **Seats:**The number of seats in the car.
* **New Price:** The price of a new car of the same model.
* **Price:**The price of the used car in INR Lakhs.

Python Modules Used

* Pandas
* Numpy
* Matplotlib
* Seaborn
* Scikit
* Turicreate

Approach

Following is the approach for the project

1. Cleaning and organizing data.
2. Exploratory Data Analysis
3. Developing a Model

Note: Refer Model 1 for the following.

Cleaning and organizing data.

Training and Testing data was provided

Following steps were carried out on both testing and training data.

The data provided was unorganised, unstructured and not processed and thus unsuitable for use in developing the model. Following steps were taken :

* The data was analysed to get an idea of what features were and what were their types.
* Data types of features were analysed.
* Number of null values were calculated for each feature.

The above steps gave us an idea about the structure of data and what modifications were to made, in order to make it useful for the prediction model.

It was observed that many features having numeric values had object data type. So in order to make those numeric values suitable for exploratory data analysis, the values have to be converted to float type(to handle the NaN values). Also some of the data was missing, hence had NaN values too.

Following steps were taken:

* For features mileage, power and engine capacity all the character parts were truncated and numeric parts were converted to float.
* NaN values were replaced as per following convention
* For features with continuous values, NaN values were replaced with mean value of that feature, with the same car company.
* For features with discrete values, NaN values were replaced with mode value of that feature, with the same car company.

Statistical data for the datasets were obtained by using “.describe()” function. From it was realised that there were some cars having some feature values = 0. For any car such case is not possible. Thus were replace the zero value by following the convention mentioned above.

Handling Outliers

Exploratory Data Analysis

Following steps were carried out on training data

A correlational matrix was plotted in form of a heatmap to analyse the relation between the different features.[Refer Cell No: 77]

A negative correlation implies as the value of one feature increases, that of other features decreases. A positive correlation implies that as value of one feature increases, that of other also increases.

Following were the significant observations:

* Kilometres Driven on the car negatively correlated with the year car was purchased in.
* Power developed by the car engine strongly positively correlated with its engine capacity.
* Mileage negatively correlated with Engine capacity and power developed by the car.

Further analysis was made using different plots and charts.

**Brands vs Price**

Two plots were plotted :-

1. Price vs Brand(Bar plot) [Refer Cell No: 78]
2. Count vs Brand (Count plot)[Refer Cell No: 79]

From these plots two observations were made:

1. Bentley and Lamborghini were the most expensive cars in the dataset.
2. Cars with highest prices are least in number.

**Locations vs Price**

A swarm plot was plotted in order to understand variations of price with location[Refer Cell No: 81]

It was observed that there weren’t any price deviations of vehicles with respect to their location.

**Year vs Price**

Three plots were plotted :-

1. Price vs Year(Scatter plot) [Refer Cell No: 83]
2. Year vs Price (Line plot)[Refer Cell No: 82]
3. Count vs Year (Count plot)[Refer Cell No: 84]

It was observed that newer cars had higher price.

**Kilometres Driven vs Price**

A scatter plot was plotted in order to understand variations of price with kilometres driven.[Refer Cell No: 85]

It was observed that that cars which have been driven less are sold for a higher price.

**Fuel Type vs Price**

Two plots were plotted :-

1. Count vs Fuel Type(Count plot) [Refer Cell No: 86]
2. Price vs Fuel Type (Box plot)[Refer Cell No: 87]

Following Observations were made :

1. Diesel and petrol cars constitute the majority of dataset.
2. Diesel and Petrol have a broad price range. But Electric cars have the highest mean price, followed by Diesel and Petrol

**Transmission vs Price**

Two plots were plotted :-

1. Count vs Transmission(Count plot) [Refer Cell No: 88]
2. Price vs Transmission (Bar plot)[Refer Cell No: 89]

It was observed that manual transmission cars were more popular that automatic, which is explained by the observation that Manual transmission cars significantly cost less.

**Owner type vs Price**

Two plots were plotted :-

1. Count vs Transmission(Cat plot) [Refer Cell No: 90]
2. Price vs Transmission (Violin plot)[Refer Cell No: 91]

It was observed price depreciates with change in ownership of cars. First hand cars are higher priced than 2nd hand and above.

**Dependant Features vs Price**

A pair plot was plotted with target(Price) vs Features(Kilometres Driven, Engine Capacity, Power, Mileage).[Refer Cell No: 92]

As observed in correlational matrix, price decreases with increase in Kilometres driven and mileage, while it increases with increase in Engine Capacity and Power.

**Seats vs Price**

Two plots were plotted :-

1. Count vs Transmission(Count plot) [Refer Cell No: 94]
2. Price vs Transmission (Cat plot)[Refer Cell No: 95]

It was observed that 5 seater cars were the most popular cars. It is explained by the observation that 5 seaters cars are pretty cheap.

Implementing the Model

Model 1

In machine learning, it is easier to develop any model if its features have an numerical value. In our data, we have many categorical values for features of brand, locations, fuel type, transmission and owner type.

So for each feature, each category is converted into column with a value 0/1(No/Yes).

In end each of those features are replaced by the category columns, according to their index.

After, one-hot encoding, Model is trained using Linear regression algorithm. This model achieved 77% accuracy.

Now to achieve even better accuracy, the model is fitted within a pipeline in which it is scaled in order to help the algorithm train them model efficiently.

After training our pipelined model, this new model achieves 94.56% accuracy

Now we predict the car prices of the testing set using this model.

Note: Refer Model 2 for the following.

Model 2

One hot encoding is done for categorical features. The categories in each feature are replaced by a numeric value. This one hot encoding is done for both testing and training set.

A correlational matrix was plotted to analyse the multicollinearity between different features and price of the car. Features having near zero correlation are excluded from the dataset as they wont affect the price of car as such and would help us save time in training the dataset.

Since turicreate is being used, dataframe is converted into SFrame(its equivalent in turicreate). Also a train and validation split is made to measure the performance of the model.

The model is then implemented using Lasso (or L-1) regularisation which is an intrinsic feature selection task and penalises unimportant feature weights to reduce to zero depending on the L-1 penalty value chosen. We use a range of L-1 penalty values over which we loop to see the RMSE in each case and also see the number of non-zero coefficients. Finally we choose that value of the L-1 penalty which had produced the minimum RMSE.

Through the process mentioned above, 88.59 was obtained as the best penalty value according to the RMSE score. So the final model is trained using this value and car prices are predicted.

Accuracy of model is obtained via R2 score.

For the validation set, R2 score is 84.53.

For entire training set, R2 score is 93.72.