Prediction of used car values

Muthukumar Kadhirvel

2022-08-12

**Introduction**

**Problem Statement**

Companies always have a tough time pricing their products so that they stay competitive and get more customers. This job is even harder for companies that sell used products like cars/laptops/mobile phones etc. The reason is that for new products, the companies would have a complete control of the various components that go into building the product and so we can add the requisite profit margin and price the same. But with used products there are a lot of unknowns and hence more difficult to price them competitively and gain an advantage over the competition.

**Justify why it is important/useful to solve this problem**

Companies can make use of their historical data and price them but there are multiple variables to choose from and it is a highly time-consuming process. We also need to keep updating the historical data year by year and price the products that come to the companies accordingly. Here is where Data Mining would help these companies a lot as once the model is trained and evaluated, we can use that to price the incoming products a lot easier and without wasting a lot of time.

**How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?**

As we have seen before there are a lot of positives that we can gain when using data mining to predict and price incoming products. These companies can use their employees more effectively as if we do not use data mining then they must be involved in manually pricing each product that would take a lot of valuable time away from them.

**Explain where you obtained your data**

In this project we will work with a dataset that has prices of used cars from India.

https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv.

Using the properties of various used cars like Location/Year/Driven Kilometers/Fuel Type/Transmission Type/Owner Type/Mileage/Engine power/Seats/Price etc. we can come with a model that can predict the price of any car that can come into the system. This would be useful for any business that deals with used cars like a car sales company/website that lists the used cars from various companies for sale.

**Milestone Summary**

**EDA**

Used car companies always need to stay ahead of the competition and one way of doing that is to price their inventory competitively and without wasting lot of time. This is where EDA - Exploratory Data Analysis comes in as the first step as we will clean and transform the data to use in building a suitable model.

In our dataset we have a lot of categorical variables and that needs to be converted to useful variables like numeric or dummy as we cannot use categorical variables in building a model.

The original dataset has 6019 rows and fourteen columns, and we create the below visualizations to make sense of the original data.

Histogram of Year against Count Histogram of Seats against Count

Chart, histogram

Description automatically generatedA picture containing chart

Description automatically generated

Bar chart of Transmission against Count Bar chart of Owner Type against Count

A picture containing logo

Description automatically generatedChart

Description automatically generated

We can see that most of the inventory is that of from the recent 10 years and five seaters.

We can see that most of the inventory is that of Manual transmission and First owner cars.

**Data preparation**

We will first drop the less useful features in building the model.

Name – Cannot be grouped and unique for most of the inventory

Location – Can be grouped but have a lot of values for it and hence would not be usable

New Price – Not all the rows have values and hence can be dropped

Then we transform the numerical variables into having proper float values

Mileage – It has values in both km/kg and kmpl and we will convert them to a unique metric of kmpl but converting 1 km/kg to 1.4 kmpl

Engine – It has values that end with CC that is removed to store them in float format

Power - It has values that end with bhp that is removed to store them in float format

Then we take care of the null values by replacing missing values with the median for the below columns.

Mileage/Engine/Power/Seats

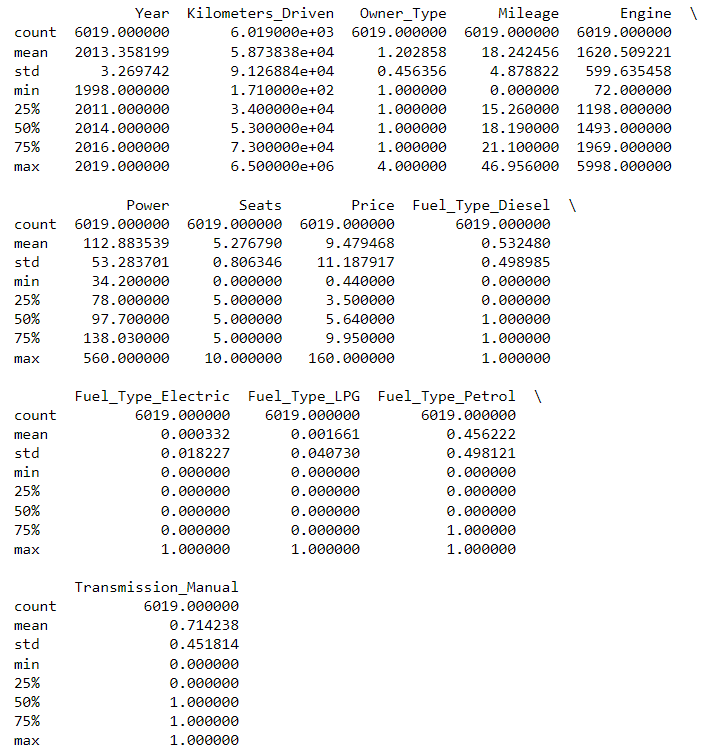
Then we create dummy variables for the below categorical variables

Fuel Type – Has values of Diesel/Petrol/CNG/LPG/Electric

Transmission – Has values of Manual and Automatic

Owner Type – Has values of First/Second/Third/Fourth & Above that are converted to numerical values of 1/2/3/4

As we can see below all the variables are now converted into useful numeric variables and we can go ahead with building and training various models to use that would do the prediction with a greater accuracy.



**Model building and evaluation**

As the data is now prepared, we can split the data into training dataset and test dataset in the ratio of 80% to 20% where the target is the Price column, and the features series has the rest of the columns. As ours is a regression problem we will use R2 score as the evaluation metric and not accuracy which is used for classification problems.

First, we fit and transform using Linear Regression and we get a R2 score of 68% for Training data and 74% for Test data. This is an incredibly small number and if we use Linear Regression as our model we will end up with a poor model.

Then we fit and transform a PCA on training data with retaining 90% variance and Linear Regression and we get a R2 score of 65% for Training data and 69% for Test data. As we can see the R2 score decreased because PCA reaps benefit only when we have a greater number of features but in our example, we just had 12 features that got reduced to 8 features.

Then we fit and transform using Random Forest Regression and we get a R2 score of 98% on Training data and 90% on Test data. As we can see the R2 score increased because of the possibility of nonlinear relationship in the data.

Therefore, we can see that Random Forest Regression model works the best for our project and we can continue using that model for future predictions.

**Conclusion**

**What does the analysis/model building tell you?**

The final model we chose was Random Forest Regression model and the higher R2 score than Linear Regression clearly indicates that we have nonlinear relationship in our data.

**Is this model ready to be deployed?**

Since we have a 98% and 90% R2 score on training data and test data respectively, our model is ready to be deployed.

**What are your recommendations?**

Once we have a substantial amount of new data into the system, we can include that and retrain our model and make it a recurring mechanism to keep the model in line with the current data.

**What are some of the potential challenges or additional opportunities that still need to be explored?**

Though the higher R2 score indicates that we have a more stable model there are a couple of issues with this model.

When the data becomes larger, Random Forest Regression model becomes more time consuming and difficult to train the model

Random Forest Regression model is overly sensitive to changes so if any parameter, noise, environmental change occurs then the model losses its effectiveness.