FLIP ROBO

Car Price Prediction

Submitted by:

NIRAV MEHTA

ACKNOWLEDGMENT

The following research papers assisted me in developing my model and making predictions by helping me comprehend the issue of car prices, the numerous aspects influencing the pricing of a certain car, volatility in car prices, and their causes:

Over the previous ten years, the number of cars produced has gradually increased; in 2016, more than 70 million passenger cars were created. The used automobile market, which has grown into a thriving business on its own, has resulted from this. The recent emergence of internet marketplaces has enabled the requirement for both the buyer and the seller to be better informed on the trends and patterns that define a used car's market value. We will attempt to create a statistical model that can forecast the price of a used automobile based on historical consumer data and a predetermined set of variables using machine learning algorithms like Lasso Regression, Multiple Regression, and Regression Trees.

This study attempts to develop a model to forecast the reasonable costs of used automobiles based on a number of variables, such as engine size, fuel type, gearbox, road tax, and vehicle miles. In the used car market, this model can be advantageous to vendors, purchasers, and automobile manufacturers. Based on the data that users submit, it can eventually output a price prediction that is reasonably accurate. Machine learning and data science are utilised during the model development process. The dataset was obtained via scraping used car listings.

To attain the highest accuracy, a variety of regression techniques, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression, were used in the study.

Given the variety of elements that influence a used car's market pricing, determining if the quoted price is accurate is a difficult undertaking. The goal of this research is to create machine learning models that can precisely forecast a used car's price based on its attributes so that buyers can make educated decisions. On a dataset made up of the sale prices of various makes and models across American cities, we put several learning

techniques into use and evaluated their effectiveness. Our findings demonstrate that, although computationally intensive, the Random Forest model and K-Means clustering with linear regression produce the best outcomes. Traditional linear regression also produced good results, with the benefit of requiring substantially less training time than the previous method.

INTRODUCTION

- ## Business Problem Framing

The impact of COVID 19 on the market has caused significant changes in the auto industry. Now, certain cars are expensive because they are in high demand, while others are less expensive because they are not. An important and intriguing challenge is estimating the cost of used vehicles. It is difficult to estimate a car's resale value. It goes without saying that a variety of factors affect how much second-hand cars are worth. The car's age, make, model, place of production, mileage (the distance it has travelled), and horsepower are typically considered to be the most significant factors (amount of power that the engine produces). One of our clients deals with small merchants who market second-hand automobiles. Due to the influence of COVID 19 on the market, our customer is having issues with their old machine learning models for valuing car prices. They are therefore searching for fresh machine learning models using fresh data. A model for valuing automotive prices must be created.

- ## Conceptual Background of the Domain Problem

With certain additional costs imposed by the Government in the form of taxes, the manufacturer sets the prices of new cars in the market. Customers who purchase a new car can be sure that their investment will be worthwhile. Used car sales are rising globally, though, as a result of rising new car prices and consumers' inability to purchase new cars due to a lack of cash.

Predicting car prices using machine learning is a particularly fascinating subject because there are numerous elements that affect an automobile's second-hand market pricing. A used car price prediction system is required to accurately assess the value of the vehicle utilising a range techniques features.

Therefore, it is in their best interests as business owners if they can accurately forecast the salvage value (or residual value) of automobiles. If the seller first underestimates the residual value, the clients' payments will be larger and they'll almost surely choose a different seller. The clients' monthly payments will be cheaper if the residual value is exaggerated, but the vendor may struggle to sell these expensive used cars at the overestimated residual value. As a result, it is clear that determining the used car market value is also of paramount commercial significance.

There are websites that provide this service, but they might not always use the most accurate prediction algorithm. Knowing their true market value is crucial when purchasing and selling.

By implementing machine learning models, we are attempting to assist the client in working with tiny traders who sell used automobiles to comprehend the pricing of the used autos. These models would aid buyers and sellers in their understanding of the used automobile industry, which would then enable them to successfully sell used cars.

- ## Review of Literature

Used automobile prices have been predicted using a variety of studies and related works utilising various methodology and approaches, with findings ranging in accuracy from 50% to 90%.

Galarraga et al. (2014) employed the European labelling system, which categorises cars based on their relative fuel consumption levels, as a novel alternative indication for energy efficiency for light cars. To estimate the pricing functions for automobiles and determine the marginal price of highly ranked cars in terms of energy efficiency, they used the hedonic price approach.

Dastan (2016) sought to identify the variables influencing used automobile prices. In order to achieve this, horizontal cross-sectional data was collected from online ads for used cars. Many features, including the front view camera, brand, model, age, traction, mileage, gear, fuel type, torque, width, fuel tank volume, ABS, panoramic glass roof, rear window defroster, power steering, start/stop, sunroof, and cooled torpedo, have been proven to have an impact on the price of a car.

• Motivation for the Problem Undertaken

In this project, I have to create a model that uses readily available independent variables to estimate the cost of second-hand autos. The management will then utilise this model to determine exactly how the prices fluctuate depending on the variables.

This is the second assignment I've had to create a machine learning model for during my internship. Additionally, we had to use a variety of online scraping methods to collect the data for the model.

By working on this project, I gained experience with site scraping, data analysis, and model construction. Using ML models, I was able to estimate the best used car prices. The model will also help management better grasp the pricing trends in the "used automobile" industry.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

I started by scraping the necessary information from the website cardekho.com. I downloaded data for several sites and saved it in csv and excel formats.

My target column in this particular instance is "Price," and it was a continuous column. Since it is obviously a regression problem, I must design the model using all regression procedures.

The dataset had no null values.

We performed feature engineering to extract the necessary feature format because the original data was out of format because it was scraped from the cardekho.com website.

I have employed plotting techniques like distribution plots, bar plots, regular plots, pie plots, scatter plots, and count plots to gain a better understanding of the features. I was better able to comprehend the relationship between the features thanks to these charts.

Additionally, I discovered outliers and skewness in the dataset, so I eliminated them using the z-score approach for outliers and the yeo-johnson method for skewness.

I built models using all available regression algorithms, then modified the best model and saved it. Finally, using the saved model, I was able to anticipate the car price.

- Data Sources and their formats

The information was gathered in csv and excel formats from the cardekho.com website. Selenium was utilised to scrape the data. The dataset is saved as an excel file and a csv file once the necessary features have been removed.

In addition, my dataset comprised 6 columns, including target, and 480 rows. The following features information is provided.

Brand: The car's brand name;
Fuel: The kind of fuel the engine uses;
KMS driven: The distance travelled;
Whether it's automatic or manual Model: The make/model of the car;
Manufacture Year: The year that the car was made; Cost: The cost of the car

- Data Preprocessing Done

I started by using Selenium to scrape the necessary information from the cardekho.com website.

After importing the dataset that was saved in csv format and the necessary libraries, I performed all the statistical analysis, including checking the dataset's shape, columns, data kinds, numeric values, value counts, and other factors.

All of the columns were of the object data type when I checked the data types of the columns, therefore I adjusted the data types of the columns to the appropriate data types.

Using feature extraction methods, I was also able to extract the year the cars were manufactured from the brand column.

There were no empty values in the dataset, and when I checked for null values, I didn't find any in the dataset.

A visualisation of the data using univariate, bivariate, and multivariate analysis. Plotted a number of categorical and numerical plots, including pie plots, count plots, bar plots, distribution plots, box plots, and pair plots, using the seaborne and matplotlib libraries to visualise each feature.

The object data type columns were then all encoded using Label Encoder to enable model creation and viewing easier.

Box plots were used to identify outliers, and the Z Score Method was used to remove them from columns. The data frame was then saved as "df usedcars" once the outliers had been eliminated.

Used the power transformation approach to detect and reduce skewness in numerical columns (yeo-johnson).

To prevent any form of data biasness, features and target variables were separated, and feature scaling was carried out using the Standard Scaler approach.

If present, checked Variance Inflation Factor (VIF) eliminated high multicollinearity problems.

- ## Data Inputs- Logic- Output Relationships

A target and other features make up the dataset. The target variable changes as the values of our independent variables change since the features are independent and the target is dependent.

I used EDA to examine the relationship between the features and the target, and to do so, I used a variety of plots, including bar plots, reg plots, scatter plots, pie plots, count plots, pair plots, and others.

I used a heat map and a bar plot to assess the association between the target and the features and found both a positive and a negative correlation between the label and the features.

The following significant factors both positively and negatively affect Price:

Characteristics with a strong positive link to the target: Manuf Year.

Features that negatively correlate strongly with the target: Energy & Variant

Hardware and Software Requirements and Tools Used

The following hardware, software, and tool requirements are necessary to construct machine learning projects.

Hardware necessary:
Core i5 or higher processor, 8 GB or more RAM, and 250 GB or more ROM/SSD are required.

Software needed: Anaconda, using Python 3 as the primary language.

Libraries Used:

- import numpy as np
- import pandas as pd
- import seaborn as sns
- import matplotlib.pyplot as plt
- from sklearn.preprocessing import LabelEncoder
- from sklearn.preprocessing import StandardScaler

- from statsmodels.stats.outliers_influence import variance_inflation_factor
- from sklearn.linear_model.LinearRegression
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.neighbors.KNeighborsRegressor
- from sklearn.svm.SVR
- from sklearn.ensemble import RandomForestRegressor
- from xgboost import XGBRegressor
- from sklearn.ensemble import GradientBoostingRegressor
- from sklearn.ensemble import ExtraTreesRegressor
- from sklearn.model_selection import cross_val_score
- from sklearn.model_selection import GridSearchCV

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

To change the data types and extract values from columns in the dataset, I used feature extraction & conversion procedures. I used Label Encoding to encrypt the categorical data.

I have applied the Z Score approach to eliminate outliers. Additionally, I used the to remove skewness technique of Yeo-Johnson.

To determine the relationship between dependent and independent features, use the Pearson's correlation coefficient. Additionally, I've used standardisation. Model construction using all regression algorithms is the next step.

- Testing of Identified Approaches (Algorithms)

This particular challenge was a regression problem because Price was my objective and a continuous column. Additionally, I built my model using all regression procedures.

I discovered ExtraTreesRegressor to be the best model with high scores by examining the r2 score and cross validation score. Additionally, in order to find the optimum model, we must test a number of them. Cross validation is also necessary in order to avoid overfitting's confusion.

- Run and Evaluate selected models

The Linear Regression model gave us an R2 Score of 28.03 %.

The Decision Tree Regressor Model gave us a R2 Score of 100 %.

The KNearest Neighbors Regression Model gave us a R2 Score of 93.68 %.

The Lasso Regressor Model gave us a R2 Score of 30.08 %.

The Ridge Regressor Model gave us a R2 Score of 28.09 %.

The Decision Tree Regressor Model gave us a R2 Score of 100 %.

The Gredient Boosting Regressor Model gave us a R2 Score of 99.87 %.

After Hyper Parameter Tuning, we have got a better R2 score of 97.72 %.

The graphic displays the linear relationship between the projected and actual used-car prices. Actual values are provided by the best-fitting line, while anticipated values are shown by green dots.

- Key Metrics for success in solving problem under consideration

I used an r2 score, which indicates the accuracy of our model.

I've utilised mean absolute error, which provides the extent of the discrepancy between an observation's predicted value and its actual value.

One of the most used metrics for assessing the accuracy of forecasts is root mean square deviation, which I have utilised in this analysis.

- Visualizations

I performed a univariate analysis of the attributes using count plots and pie charts. For the univariate analysis of the target variable, I utilised a distribution plot.

Maruti makes the majority of used automobiles, with Hyundai, Honda, Renault, Mahindra, Tata, and Toyota following. There aren't many used luxury brand automobiles.

Petrol is the most common fuel type, followed by Diesel, and then CNG, which is utilised in the fewest used cars.

The majority of used vehicles are the manual variety (80.4 percent). The remainder is automatic (19.6 percent)

- Interpretation of the Results

I have used distribution plots and count plots in univariate analysis to show the counts in numerical variables and pie plots to show the counts in categorical variables.

To examine the relationship between the target and the other features in a bivariate study, I used bar graphs. I then used pair plots to examine the pairwise relationship between the characteristics and scatter plots for multivariate analysis.

Skewness and outliers were detected using box plots and distribution plots. Respectively. And I noticed that some of the features were skewed both to the right and to the left. I was able to comprehend the relationship between the two using the heat map and bar plot. Independent and dependent characteristics. Heat maps also assisted in identifying the multicollinearity problem and feature importance.

Scaling both the train and test datasets has a positive effect in that it prevents the model from failing.

To acquire the best model out of a train dataset, we must apply numerous regression techniques when creating different models. And in order to choose the optimal model, we must consider a number of indicators, including mae, mse, rmse, and R2 Score.

ExtraTreesRegressor, which had an R2 Score of 99.97%, was the model I found to be most accurate. Then, I ran hyper parameter tuning to increase the best model's accuracy, which caused the R2 Score to change significantly to 99.97 percent.

CONCLUSION

- Key Findings and Conclusions of the Study

In this project report, we demonstrate how to anticipate used automobile prices using machine learning techniques. After this project was finished, we gained insight into the data collection, pre-processing, analysis, cleaning, and model-building processes.

In this work, we employed many machine learning models to forecast the used automobile sales price. By doing feature engineering on the data, we have gone through the analysis process, discovering the relationship between the features and the target through visualisations. and gathered the crucial information. We then built machine learning (ML) models using these features to forecast the pricing of cars.

To avoid overfitting, we verified the CV score after training the model.

Performed hyper parameter optimization on the top model, which resulted in the top model's R2 score increasing and reaching 99.97%. We also have successful pricing predictions for automobiles.

- Learning Outcomes of the Study in respect of Data Science

I gained a lot of knowledge about automotive features and other car selling platforms while working on this project. Building machine learning models has made it possible to estimate the price of used cars, which has improved understanding of the various benefits and drawbacks of selling used vehicles.

Due to the variety of data it contains and the necessity of utilising Selenium to scrape the data from the cardekho.com website, I thought this problem to be rather intriguing to manage.

We were able to better grasp the data by using a graphical representation thanks to data visualisation. It has helped me comprehend what the data is attempting to convey.

This study is an exploratory attempt to estimate house costs using 8 machine learning algorithms, then compare the findings.

Finally, by projecting the sale price of used vehicles and developing a model for car price forecasting that might assist clients in understanding the future pricing of used automobiles, our goal was accomplished. Research on used automobile prices can make advantage of new machine learning analytical approaches.

• Limitations of this work and Scope for Future Work

LIMITATIONS:

The first issue is that scraping the data is a dynamic process, and the second was that the data wasn't the right data type.

More skewness and outliers will also lower the accuracy of our model, which will come next. It appears to be fairly good that we have obtained an accuracy of 99.97 percent even after dealing with all these flaws because we have done our best to deal with outliers and skewness.

FUTURE WORK:

In terms of future work, we want to anticipate automobile prices using more data and more sophisticated methods like artificial neural networks and genetic algorithms.

Future versions of this machine learning model might integrate with different websites that can supply real-time data for price prediction.