|  |
| --- |
| CAR PRICE PREDICTION ANALYSIS |

|  |
| --- |
| DATA SCIENCE (DSCI) |

|  |  |
| --- | --- |
| SPRING 23 |  |



Contents

￼

￼

[Descriptive Statistic of Dataset 3](file:///C:\\Users\\ardia\\Desktop\\TEACHING\\SPRING%2022\\DSCI6002%20Data%20Science\\Projects\\Technical%20report%20Template.docx" \l "_Toc96341553)

[Exploratory Data Analysis 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

Data visualization………………………………….

Data Processing…………………………………….

Feature Importance……………………………..

Model Evaluation with Flask API……….

￼

Conclusion and business recommendations

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| Car Price Prediction |

|  |  |  |
| --- | --- | --- |
| Executive Summary "We have obtained a dataset from the car search website CarDekho.com, which includes information on various vehicles such as their selling prices and current prices. This data can help us in making informed decisions while purchasing a new car, like determining the appropriate price for a particular type of car. After conducting exploratory data analysis, we have developed a model using the Random Forest method, which is a supervised machine learning model, to identify the factors that impact a car's price. The model has been deployed into the Flask terminal using a PICKLE file, enabling users to enter car parameters and determine its price." | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Sai Karthik Navuluru**  **Sowmya Reddy Baddam**  **Thriyogya Kokirala**  **Naga amrutha SaranyaNareddy** | **Questions?**  Contact : snavu3@unh.newhaven.edu |  |

|  |
| --- |
|  |

|  |  |
| --- | --- |
| **Descriptive Static Dataset**   * A descriptive static dataset is a collection of data that provides a snapshot of a specific population or phenomenon at a particular point in time. * This type of dataset includes data that has already been collected and is typically used for descriptive analysis, which involves summarizing and presenting data using statistical measures. * They can be used to describe the characteristics of a population, such as age, gender, income, education level, or geographic location. They can also be used to describe the distribution of a particular variable, such as the number of customers who purchase a certain product or the frequency of a particular behavior. * Examples of descriptive static datasets might include census data, which provides a snapshot of the population of a particular country or region at a particular point in time, or financial data, which summarizes the financial performance of a company or industry over a particular period. * Descriptive static datasets are commonly used in statistical analysis and data science to provide a baseline understanding of a particular phenomenon or to support hypothesis testing and modeling. They can also be used as inputs to machine learning algorithms, which can learn patterns and relationships in the data and make predictions based on that learning. |  |
| Table  Description automatically generated |  |
| Table  Description automatically generated |

Data Encoding

* Data encoding refers to the process of converting data from one form to another for the purpose of transmission or storage. This is necessary because different systems and devices may use different methods to represent and store data.
* The most common forms of data encoding include:

1. **ASCII Encoding:** This involves encoding text characters using a 7-bit binary code. It is widely used for representing text in computer systems.
2. **Unicode Encoding:** This is a standard for representing characters from different languages and scripts using a single encoding system. It supports more than 100,000 characters from over 100 scripts, including Latin, Greek, Cyrillic, Chinese, Japanese, and Korean.
3. **Binary Encoding**: This involves representing data using a sequence of 0's and 1's. It is used extensively in computer systems and is the basis for all digital communication.
4. **Base64 Encoding:** This is a method of encoding binary data using only printable ASCII characters. It is often used to encode data for transmission over email or other text-based communication channels.
5. **Huffman Encoding:** This is a form of variable-length data compression that assigns shorter codes to more frequently occurring data values. It is used to compress data for efficient storage and transmission.

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, table

Description automatically generated

Table

Description automatically generated

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of analyzing and understanding data to uncover patterns, relationships, and anomalies. It involves using various statistical and visualization techniques to gain insights into the data and to develop an understanding of its characteristics.

The main goals of EDA are to:

1. Summarize the primary features of the data
2. Understand the relationships between different variables in the data
3. Identify any outliers or anomalies in the data
4. Determine any patterns or trends in the data
5. Formulate and test hypotheses about the data
6. Prepare the data for further analysis.

Some common techniques used in EDA include:

1. **Data visualization:** This involves creating graphs, charts, and other visual representations of the data to help understand its characteristics.
2. **Correlation analysis:** This involves analyzing the relationship between different variables in the data using correlation coefficients.
3. **Hypothesis testing:** This involves formulating and testing hypotheses about the data to determine whether certain relationships or patterns exist.
4. **Data cleaning:** This involves identifying and correcting any errors or inconsistencies in the data to ensure that it is accurate and reliable.

EDA is an important first step in any data analysis project, as it helps to identify any issues or challenges that may need to be addressed before further analysis can be conducted. By gaining a deep understanding of the data through EDA, analysts can develop more informed hypotheses and make more accurate predictions.

Chart

Description automatically generated with low confidence

Chart, histogram

Description automatically generated

Methodology:

The goal of this project is to create a good regression model that can accurately predict car prices. For this project execution some past car data is used for which car pricing and other basic attributes are used. The cost of the car is the dependent variable, while the other characteristics are the independent variables.

Random Forest is a regression model that is based on ensemble learning. It employs a decision tree model to build the ensemble model, which, as the name implies, combines numerous decision trees to produce a forecast. The benefit of this model is that the trees are created in parallel and are highly uncorrelated, resulting in good outcomes because each tree is not susceptible to individual tree mistakes. The use of Bootstrap Aggregation or bagging, which provides the randomness required to create robust and uncorrelated trees, helps to assure this uncorrelated behavior. As a result, this model was chosen to accommodate for the enormous number of features in the dataset and to compare a bagging strategy to the gradient boosting.

Chart, radar chart

Description automatically generated

Data Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Correlations of each feature in dataset:

Chart, treemap chart

Description automatically generated

Here we have 4 categorical data features,

Transmission- [Manual, Automatic]

Seller type- [Dealer, Individual]

Fuel type- [Petrol, Diesel, CNG]

Owner- [0, 1, 3]

To calculate the vehicle age, we are adding a new column and difference of current year and vehicle’s year will be the age of vehicle.

Graphical user interface

Description automatically generated with medium confidence

We can see here that these 4 features (Transmission, Seller type, Fuel type, and Owner) are categorical data and impacting the selling price of the car.

Chart, scatter chart

Description automatically generated

The present price of a car directly influences Selling Price prediction. Both are highly correlated and here directly proportional to each other. Car age is affecting negatively as the Selling Price decreases for an older car.

****Data Preprocessing**:**

Data Preprocessing is the process of transforming or encoding data so that it may be easily parsed by the machine. In other words, the algorithm can now easily interpret the data's features.

To produce an actual dummy encoding from a Data Frame, we need to pass drop first=True.

**Feature Importance:**

Here, we are using ExtraTreeRegressor to get the notable features in the dataset.

Chart, bar chart

Description automatically generated

Here we can see Present Price is more important than other features.

Now split the data into training and test for building a model.

Word

Description automatically generated

**Modeling:**

Here I used Random Forest Regressor algorithm for predicting the car price. It is based on Decision trees.

Introductory Section

Graphical user interface, text, application

Description automatically generated

Hyperparameter optimization or tuning is the challenge of selecting a collection of appropriate hyperparameters. A hyperparameter is a value for a parameter used to influence the learning process. Other factors, such as node weights, are, on the other hand, learned.

Text

Description automatically generated

Now fit the data in the model and this will take time to train the model.

Scatter chart

Description automatically generated with medium confidence

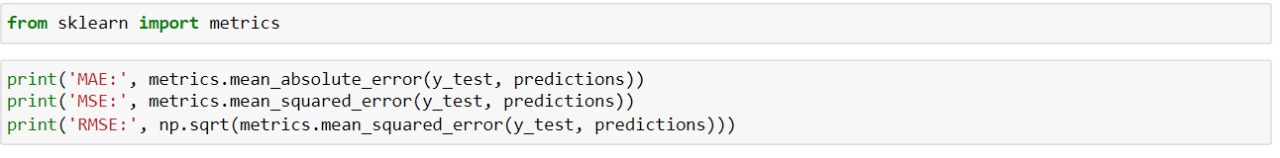
****Model Evaluation**:**

This is the best way to select the model by seeing its performance. And it is an important and integral part of model development.

Chart, scatter chart

Description automatically generated

The three common evaluation metrics for regression problems



****Model Building with Flask API**:**

For building the model first pickle the model

Text

Description automatically generated

Create a new app.py file. Now, import every important module and library to deploy the model. Also load the model in the app.py file.For the web application you need to create a html file for the structure of website, and you can add CSS and JS for styling and other things.

Graphical user interface, text

Description automatically generated

## Result Section:

## **Deployment on Flask**

## A screenshot of a computer screen Description automatically generated with low confidence

A picture containing vehicle, wheel, car, land vehicle

Description automatically generated

## Conclusion and business recommendations:

## It can be tough to determine whether a used car is worth the advertised price while looking at adverts. Mileage, make, model, year, and other characteristics can all have an impact on a car's true value. It might be difficult to price a secondhand car correctly from the standpoint of a seller. Thus, the use of machine learning algorithms will help in developing models for predicting used car prices based on given features of car.

## Contributions/References:

<https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/>

<https://towardsdatascience.com/predicting-car-price-using-machine-learning-8d2df3898f16>

<https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses>

<https://medium.com/geekculture/used-carprice-prediction-complete-machine-learning-project-d25559cf2d2a>