Project Report – Used Car Price Prediction

Email: Dillon.l.constantine@aib.ie

# GitHub URL

# Abstract

This project involved the implementation of various data pre-processing and modelling techniques to analyse a dataset of used car prices. The dataset was initially cleaned by addressing null values and converting data types where necessary. Exploratory data analysis was then performed, the data was split into test and training sets and three modelling techniques were applied.

# Introduction

As a person with a fascination with cars, this project on analysing the used car prices has allowed me to combine my interests with data analysis and modelling techniques which has been a rewarding experience.

# Dataset

The dataset is a collection of used cars for sale in Toronto, Canada in 2023. Web Scraping was used by the dataset author to scrape the data from Autotrader.ca using Selenium Script via Chrome Driver. This dataset is readily available on Kaggle.com. The dataset contains a total of 24,198 entries and 16 columns.

Below is a description to outline each column contained in the Used Car Dataset.

1. Year: The year in which the car was produced.
2. Make: The brand or manufacturer of the car.
3. Model: The specific model of the car.
4. Kilometres: The distance travelled by the car in kilometres over its lifetime.
5. Body Type: The body style or configuration of the car (e.g. sedan, SUV, hatchback).
6. Engine: The engine specifications of the car.
7. Transmission: The type of transmission system in the car (e.g. automatic, manual).
8. Drivetrain: The drivetrain configuration of the car (e.g., front-wheel drive, all-wheel drive).
9. Exterior Colour: The colour of the car's exterior.
10. Interior Colour: The colour of the car's interior.
11. Passengers: The maximum number of passengers the car can accommodate.
12. Doors: The number of doors the car has.
13. Fuel Type: The type of fuel used by the car (e.g., gasoline, diesel).
14. City: The estimated city fuel efficiency of the car per 100km.
15. Highway: The estimated highway fuel efficiency of the car per 100km.
16. Price: The price of the car.

The dataset contained a combination of numeric and categorical columns. Some of the columns have missing values which need to be addressed when cleaning the data for modelling purposed.

# Implementation Process

The following steps were used in preparing the data, exploring relationships, and implementing different modelling techniques to predict the price of used cars.

1. Package Import: Imported the necessary packages and libraries required for data manipulation, analysis, and modelling.
2. Data Import: Imported the Used car dataset and performed an initial exploration to get a sense of the data structure and content.
3. Null Value Check: Checked for any null values in the dataset to identify any missing data points.
4. Data Manipulation:
   1. Model Column Conversion: Converted the values in the "Model" column to lowercase for consistency and ease of data manipulation.
   2. Null Value Handling: Addressed null values in each column using various techniques:
      * Kilometre Column: Converted the "Kilometre" column to a numeric format and dropped remaining null values.
      * Body Type Column: Filled null values based on similar car models and dropped remaining null values.
      * Drivetrain Column: Categorized the "Drivetrain" column into four types and populated null values based on similar car models.
      * Exterior Colour Column: Populated null values based on similar car models.
      * Interior Colour Column: Assigned six core colours and populated null values based on similar models.
      * Passengers Column: Removed this column as it had a high percentage of null values, Over 50% of the data was missing.
      * Doors Column: Extracted the number of doors from the column and populated null values based on similar models and body types.
      * Fuel Type Column: Categorised fuel types into four types and populated null values based on car models.
      * City & Highway Columns: Extracted the numeric values from the strings and converted them to appropriate data types.
      * Transmission Column: Created a new column for "Transmission Type" and populated the null values based on similar car models.
      * Engine Column: Created a new "Cylinder" field and extracted the number of cylinders for each car.
5. Column Removal: Dropped an unnecessary columns from the data frame ("Engine" and "Transmission").
6. Outlier Removal: Identified fields that correlated with car prices, such as "Kilometre", "City" and "Highway". Removed the outliers based on predefined quartiles to ensure more accurate modelling.
7. Exploratory Data Analysis: Used the Seaborn library to generate various graphs and visualisations to gain insights into the data.
8. Data Split: Split the dataset into training and testing data, with the testing data comprising 30% of the total dataset.
9. Modelling Techniques: Applied three different modelling techniques to predict used car prices.
   * + Decision Tree Modelling: Built a decision tree regression model to predict car prices.
     + Random Forest Regression: Utilised random forest regression to predict car prices.
     + Linear Regression: Applied linear regression to model the relationship between car prices and other variables.

# Results

The Results section of this report presents the outcomes and findings obtained from performing Exploratory Data analysis and implementing various modelling techniques on the dataset of used car prices.

## Exploratory Data Analysis (EDA)

The below chart shows the top 15 used car makes for sale in 2023. The bar chart provides a visual representation of the number of vehicles for sale for the top 15 car makes within the data set. From the below chart it can be seen the most common used car for sale in Canada is Ford with approx. 2100 cars. Honda, Jeep, Hyundai, and BMW making up the top 5, all have a similar number of cars available for sale in 2023 approx. 1700.

A picture containing text, screenshot, plot, line

Description automatically generated

The below graph depicts the used car prices based on varies body types. From the graph of the most common body types, the coupe is the most expensive averaging around 80,000 Canadian dollars. Whereas hatchback cars are the lowest value of approx. 30,000 Canadian dollars.

A picture containing screenshot, diagram, rectangle, plot

Description automatically generated

The below Line graph shows the price distribution of used cars listed for sale in 2023 that were manufactured in the last 5 years, 2018 – 2023. From the graph, the cars manufactured in the 2022 are being listed for sale for the most followed closely by 2023.

A graph showing the price of used cars

Description automatically generated with low confidence

## Modelling

### Linear Regression

The linear regression model performed poorly on the dataset of used car prices. This is evident from the evaluation metrics obtained from the model below.

Linear Regression Results:

Mean Squared Error (MSE) : 5400006350374.541

Mean Absolute Error (MAE): 151901.29861979125

R2 Score: -9942206061419.477

The MSE indicates the models’ predictions differed significantly from the actual prices resulting in large errors on average. Similarly, the MAE suggests that there are massive differences between he predicted price and the true price. Also, the R2 indicates that the linear regression model failed to variation in used car prices accurately.

The results above and the graph below show that the linear regression model was not suitable in the prediction of used car prices dataset.

A picture containing text, screenshot, line, plot

Description automatically generated

### Decision Tree

The Decision tree performed well on the used car prices dataset as the evidence shows below.

Decision Tree Results:

Mean Squared Error (MSE): 0.058466905046927316

Mean Absolute Error (MAE): 0.12886585969852277

R2 Score: 0.8923538270080482

The MSE indicates that, on average, the squared difference between the predicted prices and the actual prices was relatively small. Similarly, the MAE suggests an average difference between the predicted prices and the true prices. The R2 score shows that the decision tree model was able to explain approx. 89.2% of the variance in the used car prices. This tells us that it’s a relatively good fit of the model to the data, suggesting that the decision tree was able to account for large portion of the dataset’s patterns and relationships.

Overall, based on the results and the graph below it suggests that the decision tree model was effective in predicting the used car prices.

A picture containing line, plot, diagram, screenshot

Description automatically generated

### Random Forest

The random forest model produced the excellent results below on the used car prices dataset.

Random Forest Regression Results:

Mean Squared Error: 0.03322253525206475

Mean Absolute Error: 0.1042450485047047

R2 Score: 0.9388324253848472

The MSE and MAE indicate relatively low errors in predicting the prices. Additionally, the R2 score demonstrates that the random forest model explains approximately 93.9% of the variance in the used car prices.

A picture containing line, plot, screenshot, diagram

Description automatically generated

In summary, the random forest model performed the better of all three models on the used car price dataset. The random forest model had the lowest MSE and MAE values, indicating smaller errors in price prediction. Also, the R2 score of the random forest model was the highest, indicating a better fit to the data and a higher percentage of explained variance meaning a more accurate prediction of used car prices.

# Insights

* Ford is the most common make of car up for sale in Canada in 2023. The second and third are Honda and Jeep.
* The most expensive body type in the top 6 most common body type is the Coupe, where the average price is greater than 80,000 Canadian dollars.
* The most expensive cars to buy are from 2022, where the average price of used cars is just under 70,000 Canadian dollars.
* A Linear Regression model is not a suitable choice for predicting used car prices in the dataset. The model is not able to adequately capture the underlying patterns and relationships in the data, resulting in large errors and low explanatory power.
* The Decision Tree model performed reasonable well returning an R2 Score of 0.89235 suggesting the model explains around 89.2% of the variances in target variables but the Random Forest model had an improved accuracy in predicting the price of cars. Random Forest had an R2 Score of 0.9388 explaining approximately 93.9% of variances in car prices. Therefore Random forest shows lower errors and a higher R2 score meaning better predictivity performance.