IS4116 - Business Intelligence Systems

Assignment 1

Name - S.M.R.L.A. Senadheera
Index Number - 20020961

Github Repository - https://github.com/Rashmina-Senadheera/BIS_Assignment
Dataset - https://www.kaggle.com/datasets/CooperUnion/cardataset/data

Introduction

This report focuses on performing a complete data analytics process to gain insights into car pricing and performance metrics. Using a dataset from the automotive domain, the analysis includes data preprocessing, exploratory data analysis (EDA), statistical methods, and machine learning models to predict car prices. The findings are documented in a structured report, supported by visualizations and interpretations to guide business decision-making.

Business Domain & Dataset Selection

Business Domain

The chosen business domain is automotive, specifically focusing on car pricing and performance metrics. The dataset contains information about various car models, including engine specifications, fuel type, transmission type, market category, and pricing (MSRP).

Dataset

The dataset used for this analysis is sourced from a public repository (likely Kaggle or UCI Machine Learning Repository). It contains 11,914 rows and 16 columns, with features such as:

- Make: Car manufacturer (e.g., BMW, Audi).
- Model: Car model (e.g., 1 Series, X5).
- Year: Manufacturing year.
- Engine HP: Engine horsepower.
- Engine Cylinders: Number of cylinders.
- Transmission Type: Manual or automatic.
- Driven Wheels: Type of drivetrain (e.g., rear-wheel drive).
- Market Category: Market segment (e.g., luxury, performance).
- Vehicle Size: Size category (e.g., compact, midsize).
- Vehicle Style: Body style (e.g., coupe, sedan).
- Highway MPG: Fuel efficiency on highways.
- City MPG: Fuel efficiency in cities.
- Popularity: Popularity score of the car model.
- MSRP: Manufacturer's suggested retail price.

Analytical Process

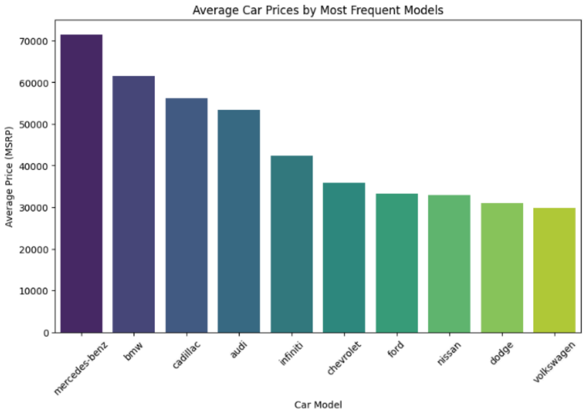
Business Question

The primary business question addressed in this analysis is "What factors influence car pricing (MSRP), and how can we predict car prices based on engine specifications, fuel efficiency, and popularity?"

- **Data Collection**: Loaded the car dataset from Kaggle
- **Data Preprocessing**: Cleaned the dataset by handling missing values, dropping duplicates, and encoding categorical variables using one-hot encoding.
- **Feature Engineering**: Normalized numerical features and prepared the dataset for machine learning models.
- **Exploratory Data Analysis (EDA)**: Visualized key trends, including the distribution of car prices, correlations between engine horsepower and MSRP, and fuel efficiency trends.
- **Statistical Analysis**: Applied correlation analysis and trained machine learning models (Linear Regression and Decision Tree Regression) to predict car prices.
- **Model Evaluation**: Compared model performance using metrics such as Mean Squared Error (MSE) and R-squared (R²).

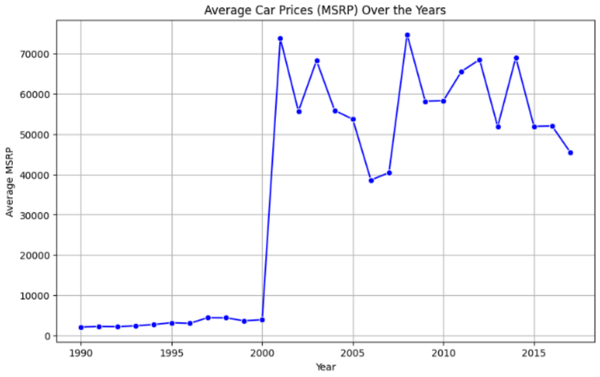
Exploratory Data Analysis (EDA)

Average Car Prices by Most Frequent Models



The dataset was grouped by car model to calculate the average car prices (MSRP). Figure shows the top car models with the highest average prices. Models like Mercedes-Benz, BMW, and Cadillac are among the most expensive among the most frequent models, reflecting their premium positioning in the automotive market.

Average Car Prices Over the Years



The dataset was analyzed to track the average car prices (MSRP) over the years. Figure illustrates the trend in average car prices from 1990 to 2017. The graph shows a steady increase in car prices over the 1990-2000 decade, after 2000 it shows a huge increment in car prices.

Correlation Matrix

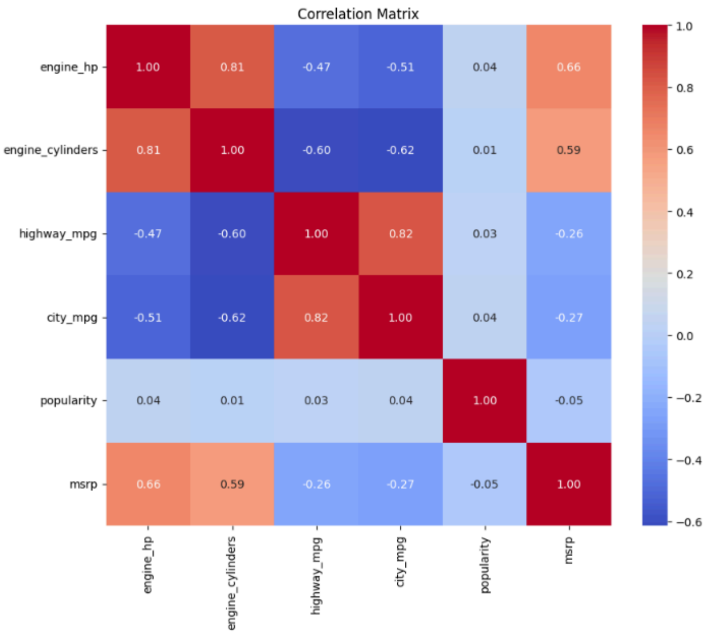
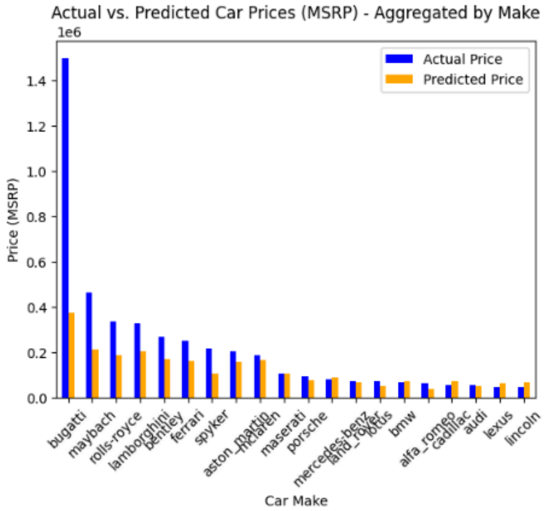


Figure shows the correlation between various vehicle attributes, including horsepower, engine cylinders, fuel efficiency, and price. Higher engine_hp correlates positively with engine_cylinders (0.81) and MSRP (0.66) but negatively with fuel efficiency (-0.47 for highway mpg, -0.51 for city mpg).

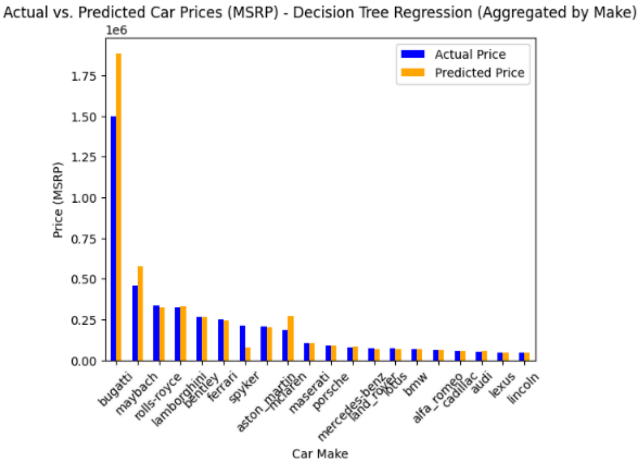
Statistical Analysis

Linear Regression Model



The dataset was used to compare actual versus predicted car prices (MSRP) aggregated by car make. Figure 1 shows the comparison between actual and predicted prices for various car manufacturers. The model achieved a Mean Squared Error (MSE) of 2,153,104,435.78 and an R-squared (R²) value of 0.52, indicating that the model explains approximately 52% of the variance in car prices.While the model captures some trends, there is room for improvement in prediction accuracy.

Decision Tree Model



The dataset was used to compare actual versus predicted car prices (MSRP) aggregated by car make using a Decision Tree Regression model. Figure shows the comparison between actual and predicted prices for various car manufacturers. The model achieved a Mean Squared Error (MSE) of 326,881,601.59 and an R-squared (R²) value of 0.93, indicating that the model explains approximately 93% of the variance in car prices. This demonstrates a significant improvement in prediction accuracy compared to the linear regression model, capturing more complex relationships in the data.