Analysis of Models for Predicting the price of Used Cars in the Second-hand Market

Introduction

The 'Market for Lemons' theory is a crucial concept that explores the complexities of asymmetric information in markets, particularly highlighting adverse selection. In markets where sellers possess more information than buyers, such as the second-hand car market, sellers may take advantage of this information gap by predominantly offering 'lemons'—inferior quality goods—to unsuspecting buyers, thus skewing market dynamics in favor of sellers.

Therefore, it is essential for buyers to be able to estimate the prices of vehicles in these second-hand markets. However, this process is complex and influenced by various factors. Attributes such as brand, model, age, and other relevant features significantly impact the value of a used car.

To address this, we employ machine learning techniques to implement three regression models: Linear Regression, K-Nearest Neighbors Regression, and Decision Tree Regression. We then evaluate the performance of these models and assess their accuracy in estimating the MSRP (Manufacturer's Suggested Retail Price) of used cars.

The findings of this project could have significant implications for various stakeholders in the automotive industry, including car dealers, buyers, and sellers. A reliable model for estimating used car prices could help streamline the buying and selling process, leading to more transparent and efficient transactions in the second-hand market.

Dataset

The dataset utilized for this project is sourced from Kaggle (https://www.kaggle.com/datasets/CooperUnion/cardataset). This dataset provides comprehensive automotive information and consists of 11,914 data points. Each data point includes 16 distinct features, encompassing details such as make (brand), model, year, engine fuel type, engine horsepower, engine cylinders, transmission type, driven wheels, number of doors, market category, vehicle size, vehicle style, highway mpg (miles per gallon), city mpg (miles per gallon), popularity, and MSRP.

Our analysis of the dataset began with the 'year' feature. To gain a visual understanding, we plotted a bar graph to illustrate the distribution of cars across manufacturing years. Figure 1 demonstrates that the majority of cars in the market were manufactured between 2015 and 2017, which aligns with expectations given that the dataset was last updated in 2017. Additionally, we explored the relationship between manufacturing year and MSRP. Figure 2 reveals an interesting trend, with a noticeable increase in the average price of cars after the year 2000. Subsequently, prices stabilized post-2000.

Figure 1

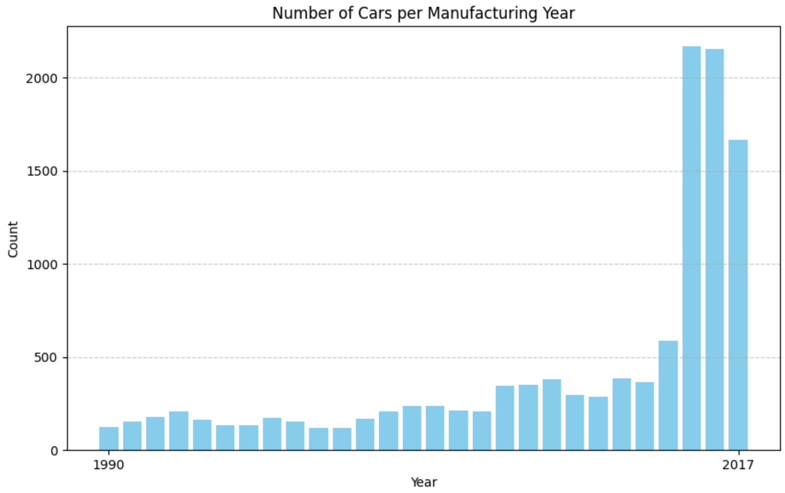
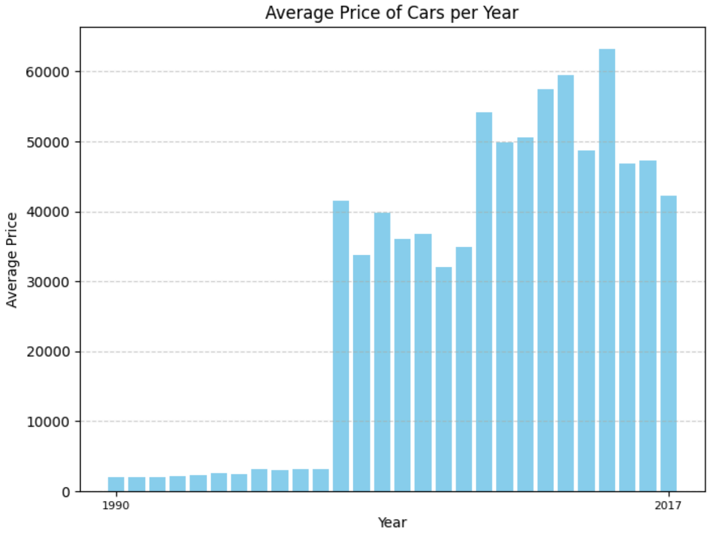


Figure 2



Extending our analysis to the 'Make' feature, which denotes the brand of each car, we plotted similar graphs. Figure 3 depicts the distribution of cars across different brands, while Figure 4 illustrates the average MSRP for each brand in the dataset. Notably, we observed an outlier in the dataset, 'Bugatti,' a luxury Italian car brand, which had an MSRP exceeding $750k with only 3 data points.

Figure 3

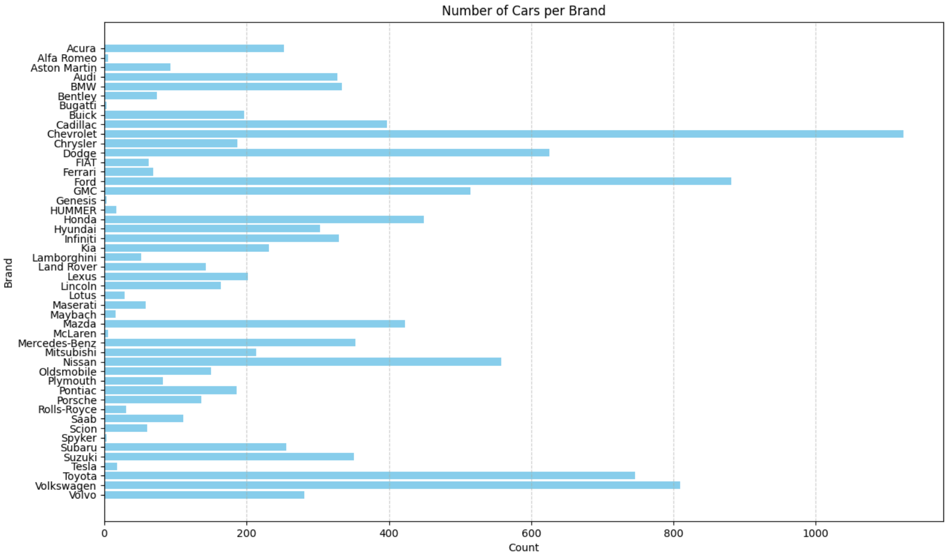
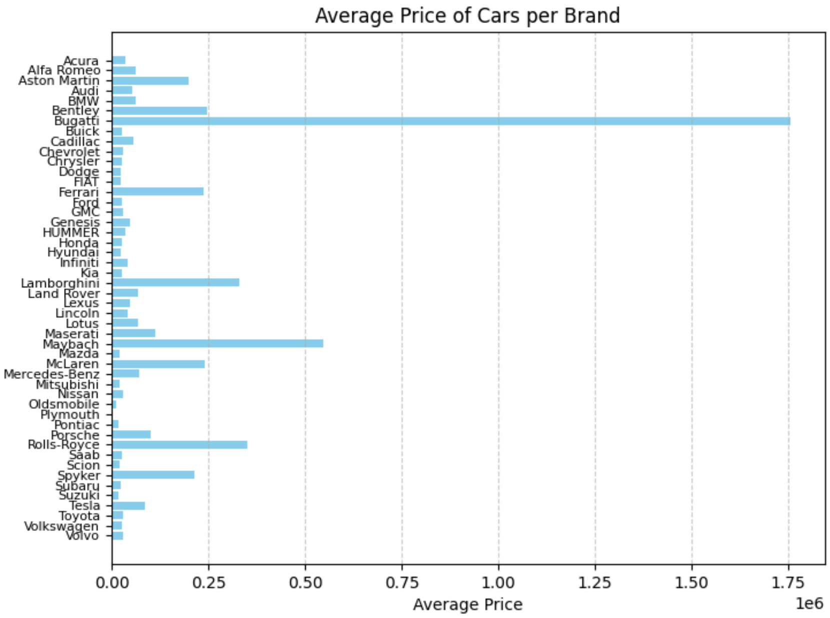


Figure 4



Preprocessing

To prepare the dataset for analysis, several preprocessing steps were implemented. Firstly, the 'Bugatti' cars were identified as outliers due to their exceptionally high MSRP values and were subsequently removed from the dataset.

Next, the 'year' feature was transformed into the 'age' of the car by calculating the difference between the manufacturing year and the year of the MSRP (2017), providing a more relevant and informative feature for analysis.

For qualitative features such as 'Make', 'Engine Fuel Type', 'Transmission Type', and 'Driven Wheels', min-max scaling was applied to normalize the values between 0 and 1, ensuring that each feature contributes equally to the analysis.

Additionally, one-hot encoding was utilized for quantitative features including 'Age', 'Engine HP', 'Engine Cylinders', 'highway MPG', and 'city MPG'. This technique converts categorical variables into a binary matrix, enabling the machine learning model to interpret these features correctly.

It is important to note that the 'Model' feature of the car was not included in the analysis since each brand has very distinct models, which would create too many features from the one-hot encoding. Instead, features such as 'Engine HP' and 'Engine Cylinders' were used to consider the performance of the car, which indirectly captures the feature ‘Model’.

Furthermore, the 'Popularity' column was dropped from the dataset as it represents the popularity of the brand, which is subjective and may be correlated with the 'Make' feature, potentially introducing bias into the analysis.

Models & Training

We selected 3 regression models from the class CSCI1420 for this task: linear regression, K-nearest neighbors regression, and Decision Tree Regression. Each model was chosen for its specific characteristics and suitability for the task at hand.

Linear Regression: Linear regression was included to establish a foundational understanding of the dataset and provide a baseline for comparison with more complex models. It assumes a linear relationship between the input features and the target variable, making it a suitable starting point for regression tasks.

K-Nearest Neighbors (KNN) Regression: KNN regression was chosen for its ability to capture local patterns and nonlinear relationships in the data. This model predicts the target variable by averaging the values of its k nearest neighbors, making it suitable for datasets with complex structures.

Decision Tree Regression: Decision tree regression was included for its ability to handle both numerical and categorical data, as well as its capacity to capture complex interactions among features. Decision trees partition the data into subsets based on the input features, allowing them to model intricate relationships.

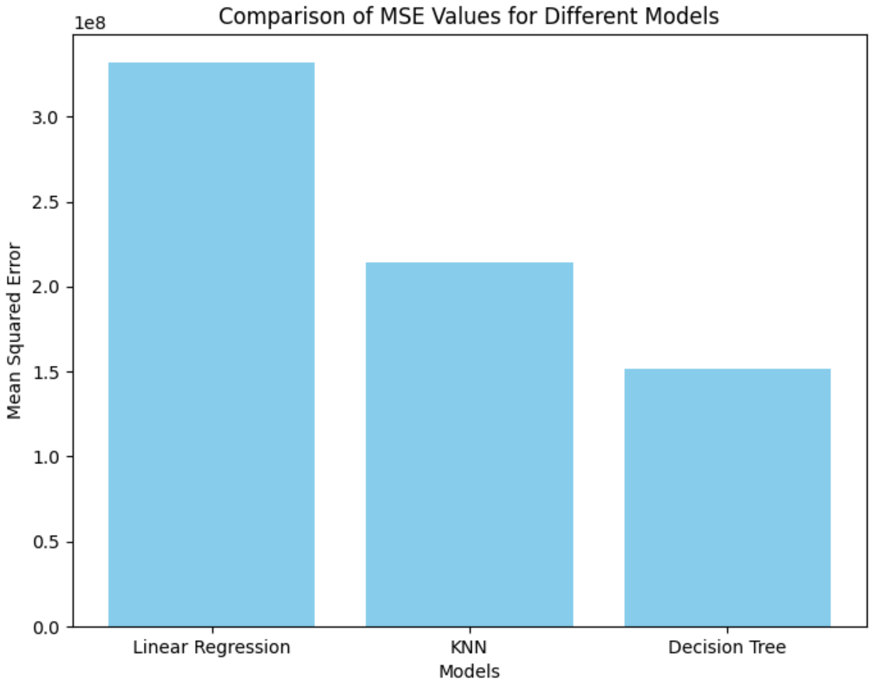
To train these models, the dataset was split into training and testing sets using an 80-20 split. Each model was then trained on the training set and evaluated on the testing set using mean squared error (MSE) as the evaluation metric.

Results

After training and evaluating the three regression models on the dataset for predicting the MSRP of used cars in the second-hand market, Figure 5 displayed the MSE values of each model, which are:

Linear Regression	K-Nearest Neighbors	Decision Tree Regression
331995999.9286485	214637214.95353758	151999263.60338786

Figure 5



Additionally, the actual MSRP and predicted MSRP were plotted for each model. Figures 6, 7, and 8 show the actual MSRP from the training dataset versus the predicted MSRP from the linear regression model, KNN model, and decision tree model, respectively.

To further analyze the accuracy of the predictions, we calculated the slope of the regression lines for each model. Ideally, a perfectly accurate prediction would result in a slope of 1. The linear regression model had a slope of 0.9, indicating slight underestimation, while both the KNN and Decision Tree Regressor models had slopes of 1.0.

Interestingly, the data points in the Decision Tree graph are more clustered along the regression line, indicating a tighter fit to the predicted values, whereas the data points in the KNN model exhibit more dispersion.

It's noteworthy that both the KNN and Decision Tree models demonstrate high accuracy for predicting MSRP values below $10,000. However, their accuracy diminishes significantly for MSRP values above $30,000.

Figure 6

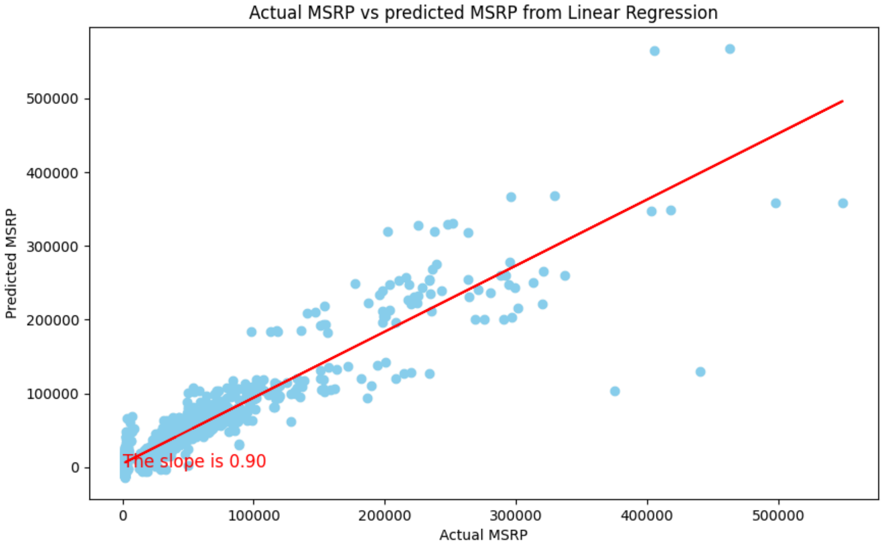


Figure 7

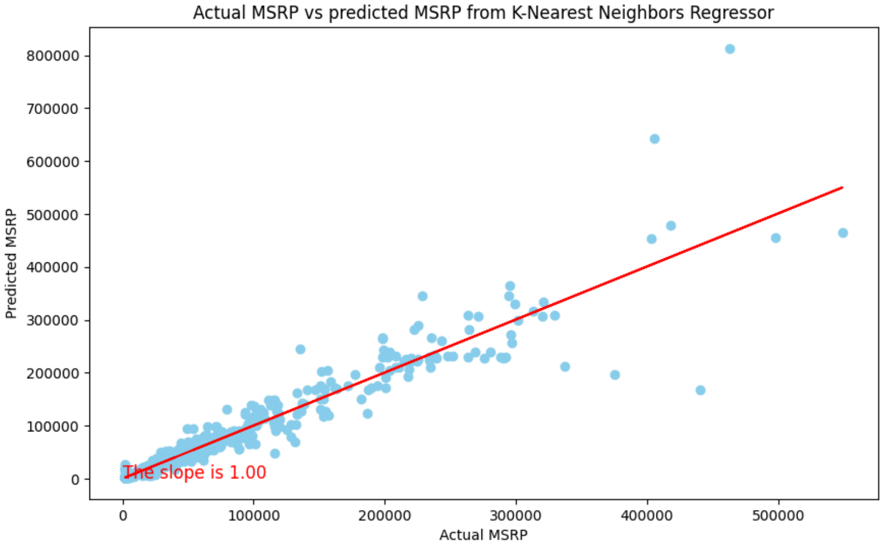
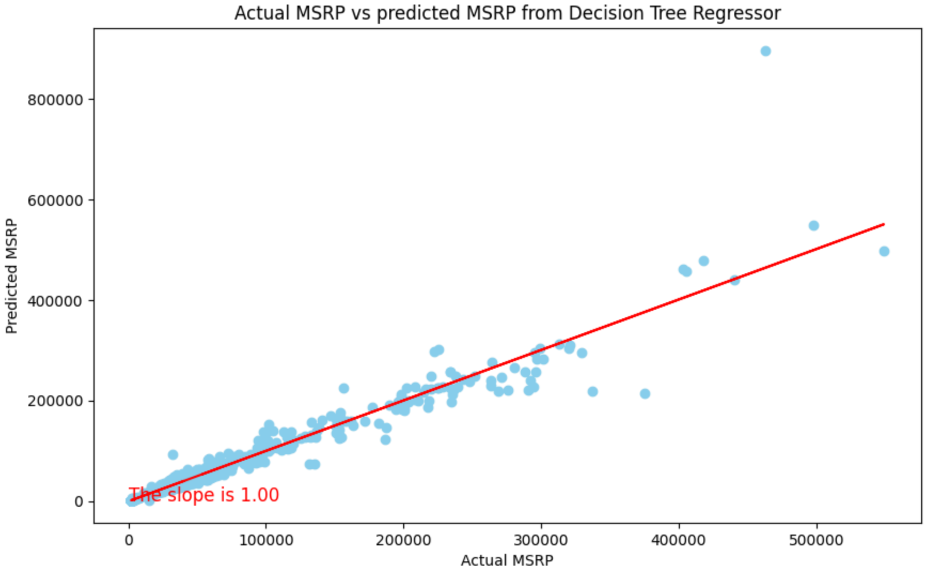


Figure 8



Conclusion

In conclusion, our analysis indicates that the Decision Tree Regressor outperforms the Linear Regression and K-Nearest Neighbors models in predicting the MSRP of used cars in the second-hand market. The Decision Tree Regressor model exhibited the lowest Mean Squared Error and demonstrated a higher level of accuracy in predicting prices. The K-Nearest Neighbors model also performed well but was slightly less accurate than the Decision Tree Regressor.

However, it is important to note some limitations of our study. The dataset used for analysis could be improved by including additional features related to the car's history, such as mileage and accident history, which were lacking in the current dataset. These features could potentially improve the accuracy of the models by providing more comprehensive information about each car's condition.

Despite these limitations, the dataset provided valuable insights into each car model's attributes, such as engine type and fuel efficiency (mpg), which were instrumental in our analysis. Future studies could benefit from incorporating additional features to further enhance the predictive accuracy of models for estimating the prices of used cars in the second-hand market.