Analysis of Second-Hand Car Sales Data: Improved Predictive Models and Clustering Algorithms

Kingsley Ebube Onoh

December 18, 2023

Abstract

This paper presents a comprehensive analysis of second-hand car sales data using supervised and unsupervised learning techniques. The dataset includes crucial features such as manufacturer, model, engine size, fuel type, year of manufacture, mileage, and price. The analysis aims to identify key factors impacting car prices and develop accurate prediction models. Additionally, clustering algorithms are utilized to segment the data based on inherent patterns.

1 Regression Models for Car Price Prediction

1.1 Single Numerical Input Features

The study compares three linear and polynomial regression models based on engine size, year of manufacture, and mileage. The "year of manufacture" emerges as the most impactful predictor with the lowest mean squared error (MSE) of $105,993,894.20$. Polynomial regression, particularly with a degree of 2, outperforms linear models, achieving MSE values of $105,993,894.20$, $132,678,999.95$, and $162,468,566.87$.

1.2 Multiple Numerical Input Features

Combining all three numerical features into a single regression model reduces the MSE to $89,158,615.76$, signifying the benefit of incorporating multiple input variables.

1.3 Regression Model Using Numerical and Categorical Variables

A Random Forest Regressor incorporating both categorical and numerical variables demonstrates significant improvement in predictive accuracy, achieving an MSE of $400,036.65$. This surpasses models relying solely on numerical features.

1.4 Artificial Neural Network (ANN) Model for Car Price Prediction

An engineered ANN model utilizing Keras with a sequential architecture exhibits commendable performance. Featuring an input layer (128 neurons, ReLU activation), a hidden layer (64 neurons, ReLU activation), and an output layer (single neuron for regression), the model achieves an MSE of $27,128,108.00$.

1.5 Model Selection and Performance Analysis

Comparison across linear regression, polynomial regression, Random Forest, and ANN models using the MSE metric reveals the supremacy of the Random Forest model. It boasts the lowest MSE of $27,128,108.00$, indicating its superior predictive power.

2 Clustering Algorithm Comparison

2.1 K-means Clustering

Implementing k-means clustering with 'Price' and 'Engine size' as features identifies two optimal clusters with an inertia of $254,403,525,780.34$ and a silhouette score of 0.5680.

2.2 DBSCAN Clustering

Employing DBSCAN clustering on the dataset using 'Engine size' and 'Year of manufacture' as features yields optimal parameters and a remarkable silhouette score of 0.9998.

2.3 Comparison and Evaluation Metrics

Comparison of silhouette scores (0.5680 for k-means and 0.9998 for DBSCAN) forms the basis for evaluating clustering performance. The DBSCAN algorithm outperforms k-means, showcasing a higher silhouette score and forming clusters with exceptional coherence.

3 Conclusion

The Random Forest model emerges as the most effective predictor for second-hand car prices, achieving the lowest MSE. For clustering endeavors, the DBSCAN algorithm surpasses k-means, evidenced by its higher silhouette score and the formation of well-defined clusters. These findings guide the selection of optimal models and parameters for enhancing the accuracy of second-hand car price predictions and gaining insights into the data structure.

4 Future Work

Further investigations could explore the inclusion of additional features, such as car condition, maintenance records, and accident history. Additionally, investigating more complex machine learning models and ensemble methods could potentially yield further improvements in predictive accuracy.

5 Limitations

The analysis is limited by the scope and size of the dataset. Larger datasets and more diverse models could provide more generalizable results.