PORTFOLIO OF WORK

**Analysis Of Second Hand Car Sales Data**

**with Supervised and Unsupervised Learning Models**

horizontal line

# Introduction

Predicting the price of a car is a challenging task due to the variety of factors influencing it, such as the car’s year of manufacture, mileage, engine, size and more. Different machine learning techniques can be employed to model these relationships, ranging from linear and non-linear regression models to more sophisticated algorithms like Artificial Neural Networks (ANNs) and clustering techniques. This report aims to analyze the effectiveness of different models and techniques for predicting car prices using a dataset with various numerical and categorical features. The study will compare the performance of models using single and multiple features, as well as evaluate the impact categorical variables. Additionally, clustering algorithms will be explored to identify potential groupings within the data.

1. **Single Numerical Feature Regression Models**

The initial step in this analysis involves evaluating regression models that use a single numerical feature to predict car prices. Three features were considered: year of manufacture, mileage and engine size. For each feature, both linear and polynomial regression models were applied, and their performance was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Adjusted R-squared (R²).

**1.1 Results**

| **Features** | **Model** | **MSE** | **MAE** | **Adjusted R²** |
| --- | --- | --- | --- | --- |
| Year of Manufacture | Linear | 130,915,500 | 7,016.34 | 0.5163 |
|  | Polynomial | 104,171,300 | 5,397.39 | 0.6151 |
| Mileage | Linear | 161,242,800 | 8,002.01 | 0.4043 |
|  | Polynomial | 129,144,000 | 6,449.06 | 0.5228 |
| Engine | Linear | 228,402,700 | 10,780.46 | 0.1562 |
|  | Polynomial | 228,209,700 | 10,771.20 | 0.1567 |

**1.2 Discussion**

The results indicate that polynomial regression models outperform linear models across all three features. The year of manufacture emerged as the best predictor, showing the highest adjusted R² (0.6151) when using a polynomial model. This suggests that the relationship between the year of manufacture and car price is better captured by a non-linear model, highlighting the importance of accounting for non-linear relationships in predictive modeling.

### 2. Multiple Numerical Features Regression Models

Building on the analysis of single-feature models, multiple numerical features were combined to assess whether this improves the predictive accuracy of the models. The Gradient Boosting Regressor (GBR) was used as the primary model for this analysis due to its ability to handle complex interactions between features.

**2.1 Results**

| **Model** | **MSE** | **MAE** | **R-squared** | **Adjusted R²** |
| --- | --- | --- | --- | --- |
| Gradient Boosting Regressor | 15,568,152.84 | 2,079.19 | 0.9426 | 0.9426 |
| AdaBoost  Regressor | 40,222,395.56 | 4,910.48 | 0.8518 | 0.8517 |
| LinearSVR | 344,644,997.71 | 10,415.50 | -0.2701 | -0.2704 |
| Random Forest Regressor | 24,747,441.56 | 2,940.07 | 0.9088 | 0.9088 |

#### **2.2 Discussion**

The results show that the Gradient Boosting Regressor (GBR) outperforms other models in terms of both MSE and R-squared metrics, indicating its strong predictive capabilities when multiple numerical features are considered. The Random Forest Regressor also performs well, but the LinearSVR model underperforms, demonstrating a negative R-squared value, which indicates that the model fails to fit the data better than a simple mean-based model.

**3. Incorporating Categorical Variables**

In addition to numerical features, categorical variables such as the car's make, model, and fuel type are likely to influence its price. A Random Forest Regressor model was trained using both numerical and categorical features to assess the impact of including all relevant variables.

**3.1 Results**

| **Model** | **MSE** | **R-squared** |
| --- | --- | --- |
| Random Forest Regressor | 403,673.49 | 0.9985 |

##### **3.2 Discussion**

The Random Forest Regressor model, which incorporated both numerical and categorical features, showed a dramatic improvement in performance compared to models using only numerical features. The MSE decreased from 19,442,810 to 403,673.49, and the R-squared value increased to 0.9985. These results highlight the importance of including categorical variables in the model, as they provide additional information that significantly enhances predictive accuracy.

#### **4. Artificial Neural Network (ANN) Model**

Given the success of regression models, an Artificial Neural Network (ANN) model was developed to predict car prices using all available information from the dataset. The ANN model's architecture consisted of an input layer with 64 neurons, three hidden layers with 64, 32, and 16 neurons, respectively, and an output layer with one neuron.

##### **4.1 Model Architecture**

* Input Layer: 64 neurons (ReLU activation)
* Hidden Layers: 64, 32, 16 neurons (ReLU activation)
* Output Layer: 1 neuron (no activation, regression problem)

##### 

##### 

##### **4.2 Results**

| **Model** | **MSE** | **MAE** | **R-squared** | **Adjusted R-squared** |
| --- | --- | --- | --- | --- |
| Artificial Neural Network | 42,715.74 | 130.79 | 0.9998 | 0.9998 |

##### 

##### 

##### 

##### **4.3 Discussion**

The ANN model significantly outperformed all other models in this study, achieving the lowest MSE and MAE and near-perfect R-squared and Adjusted R-squared values. The architecture of the ANN allowed it to capture complex non-linear relationships between the features and the target variable, resulting in highly accurate predictions. Hyperparameter tuning, including adjustments to the learning rate, number of hidden layers, and batch size, contributed to the model's superior performance.

#### 

#### **5. Clustering Analysis Using k-Means**

In addition to regression models, k-Means clustering was employed to identify potential clusters within the car sales data. Various combinations of numerical variables were used as input features, and the optimal number of clusters (k) was determined based on evaluation metrics such as the Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI).

##### 

##### 

##### 

##### **5.1 Results**

| **Feature Combination** | **Silhouette Score** | **CHI** | **DBI** |
| --- | --- | --- | --- |
| Year of Manufacture and Price | 0.678 | 102,440.368 | 0.528 |
| Mileage and Year of Manufacture | 0.570 | 138,569.757 | 0.539 |
| Mileage and Price | 0.532 | 117,151.984 | 0.575 |
| Engine Size and Price | 0.678 | 102,440.439 | 0.528 |
| Engine Size, Mileage, Price and Year of Manufacture | 0.677 | 102,442.311 | 0.528 |

##### 

##### **5.2 Discussion**

The k-Means clustering algorithm identified three optimal clusters based on the combinations of engine size, price, and year of manufacture. These clusters were characterized by high Silhouette scores, significant CHI values, and low DBI values, indicating that the clustering results were well-defined and meaningful. From the results, we see that these combinations had the highest performance: Engine size and Price, Year of Manufacture and Price, and Engine size, Price and Year of Manufacture.

#### 

#### **6. Comparing Clustering Algorithms**

To further explore the clustering results, the BIRCH clustering algorithm was applied to the same dataset, and its performance was compared to the k-Means results.

##### 

##### 

##### **6.1 BIRCH Clustering Results**

| **Feature Combination** | **Silhouette Score** | **CHI** | **DBI** |
| --- | --- | --- | --- |
| Engine Size and Price | 0.685 | 85,034.062 | 0.473 |
| Engine Size and Mileage | 0.567 | 137,053.656 | 0.536 |
| Engine Size and Year of Manufacture | 0.561 | 168,550.893 | 0.533 |

##### 

##### **6.2 Discussion**

The BIRCH clustering algorithm produced slightly better results than the k-Means algorithm, particularly for the combination of engine size and price, which achieved a higher Silhouette Score (0.685) and lower DBI (0.473). The CHI value, however, was lower than that of k-Means, suggesting that while BIRCH formed tighter clusters, k-Means may have produced more distinct clusters overall. These results suggest that BIRCH may be more effective for specific variable combinations, but k-Means remains a strong contender for general clustering tasks.

#### **7. Conclusion**

This study explored a range of regression models and clustering algorithms to predict car prices and identify patterns within the data. Polynomial regression models outperformed linear models for single-feature predictions, with the year of manufacture emerging as the best predictor. The inclusion of multiple numerical features significantly improved model accuracy, with the Gradient Boosting Regressor performing best among traditional regression models. Incorporating categorical variables further enhanced predictive accuracy, as demonstrated by the Random Forest Regressor. However, the ANN model surpassed all other models, achieving near-perfect predictive performance. Clustering analysis using k-Means and BIRCH algorithms revealed distinct patterns in the data, with BIRCH demonstrating a slight edge in specific scenarios. Overall, the combination of advanced regression models, ANNs, and clustering algorithms provides a robust framework for predicting car prices and understanding the underlying structure of the data.

### References

1. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. [Link](https://web.stanford.edu/~hastie/ElemStatLearn/)
   * This book provides a comprehensive introduction to various supervised and unsupervised learning models, including regression, decision trees, and clustering techniques used in the analysis.
2. **Murphy, K. P.** (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
   * This reference offers a detailed explanation of machine learning models, including artificial neural networks, and their applications in predictive analytics.
3. **Breiman, L.** (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
   * This paper presents the Random Forest algorithm, explaining its advantages in handling both numerical and categorical features in predictive modeling.
4. **Bishop, C. M.** (2006). *Pattern Recognition and Machine Learning*. Springer.
   * This resource covers a range of machine learning techniques, including the clustering algorithms k-Means and BIRCH, used to identify patterns in the car sales data.
5. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & others**. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830. Link
   * This paper describes the Scikit-learn library, which was utilized for implementing various machine learning models and evaluation metrics in the study.