**Car Sales Dataset**

**Introduction**

**The car sales dataset consists of various attributes of car listings, such as brand, model, year, transmission type, mileage, fuel type, tax, miles per gallon (MPG), engine size, and price. This report provides a complete workflow for preparing the dataset, conducting exploratory data analysis (EDA), and building predictive models using machine learning. The goal is to predict car prices accurately based on various features.**

**1. Data Preprocessing**

**Preprocessing is essential to ensure that the data is clean, consistent, and ready for analysis. This step involves:**

* **Handling missing values.**
* **Correcting data inconsistencies.**
* **Encoding categorical variables.**
* **Scaling numerical features.**
* **Removing duplicates and outliers.**
  1. **Handling Missing Values**

|  |  |
| --- | --- |
| **Column** | **Missing Values** |
| **Brand** | **18** |
| **Mileage** | **6** |
| **Fuel Type** | **17** |
| **MPG** | **12** |
| **Engine Size** | **8** |

**Missing values were handled based on the nature of each feature:**

* **Numeric columns like mileage, MPG, and engine size were filled with their respective means.**
* **Categorical columns like fuel type were filled with the mode (most frequent value).**
* **Specific models had missing brand values that were filled with known brand information based on model names (e.g., models like 'Arteon' and 'Beetle' were assigned to 'VW').**

**1.2 Data Cleaning**

* **Stripped extra spaces from the 'model' column to ensure consistency.**
* **Replaced incorrect or inconsistent brand names (e.g., 'bmw' to 'BMW', 'hyundi' to 'Hyundai').**
* **Dropped unnecessary columns such as 'carID' and rows with remaining missing values.**

**1.3 Removing Duplicates**

**Duplicates can skew the analysis and were identified and removed to ensure the dataset is unique.**

**1.4 Encoding Categorical Variables**

**Categorical variables (Brand, Model, Transmission, and Fuel Type) were encoded using Label Encoding to transform them into numerical representations required for machine learning models.**

**1.5 Feature Scaling**

**Feature scaling was applied to ensure all numerical columns (such as Mileage, Engine Size, etc.) are on a similar scale, improving the performance of the models.**

**Final Preprocessed Data**

**The preprocessed dataset is clean, without any missing values or duplicates, and is ready for analysis and machine learning.**

**2. Exploratory Data Analysis (EDA)**

**2.1 Descriptive Statistics**

* **The average mileage of cars in the dataset is around 36,000 miles.**
* **The average price of a car is around £15,000.**
* **The majority of cars have a fuel type of Petrol, followed by Diesel.**

**2.2 Key Observations**

* **The most expensive car in the dataset costs £139,000, while the cheapest car is priced at £695.**
* **The car with the highest mileage has 220,000 miles, and the car with the highest MPG offers 470 MPG.**
* **The dataset includes cars from various popular brands such as BMW, Audi, Ford, and Mercedes.**

**2.3 Grouped Insights**

* **Average Mileage by Brand: Skoda cars tend to have the highest average mileage, while BMW cars have the lowest.**
* **Total Price by Transmission Type: Cars with automatic transmission tend to have a higher cumulative price than manual cars.**

**2.4 Correlation Analysis**

**A heatmap was generated to show the correlation between numerical variables:**

* **Price has a strong positive correlation with Engine Size and Year, indicating that newer and larger-engine cars are generally priced higher.**
* **Mileage has a negative correlation with Price, which suggests that cars with higher mileage are priced lower.**

**3. Predictive Modeling**

**3.1 Machine Learning Models**

**To predict car prices, various machine learning algorithms were used, and their performances were compared based on Mean Absolute Error (MAE) and R² scores.**

**3.2 Models Built**

1. **Linear Regression: A basic model that assumes a linear relationship between features and target (price).**
2. **Random Forest Regressor: An ensemble model that uses multiple decision trees to make predictions.**
3. **Gradient Boosting Regressor: A boosting algorithm that improves weak learners over iterations.**
4. **XGBoost Regressor: An efficient and scalable version of Gradient Boosting.**

**Feature Selection**

**Feature selection was performed using SelectKBest to retain the most important features for prediction, which reduced the feature space while maintaining model performance.**

**3.3 Performance Evaluation**

|  |  |  |
| --- | --- | --- |
| Model | Mean Absolute Error (MAE) | Score |
| Linear Regression | **2114.60** | **0.88** |
| Random Forest | **1576.25** | **0.92** |
| Gradient Boosting | **1595.70** | **0.91** |
| XGBoost | **1540.83** | **0.93** |

**3.4 Best Model**

* **XGBoost outperformed the other models with the lowest MAE (1540.83) and the highest R² score (0.93), indicating that it provides the most accurate price predictions.**
* **Random Forest came in a close second, and both ensemble methods showed significant improvement over Linear Regression.**
* **Gradient Boosting and XGBoost models captured non-linear relationships and interactions between features effectively.**

**3.5 Model Insights**

* **Cars from premium brands such as BMW, Mercedes, and Audi were predicted to have higher prices.**
* **Higher engine size and newer models resulted in higher price predictions, as expected.**

**3.6 Model Visualizations**

**To visualize model performance:**

* **A scatter plot of Actual vs Predicted Prices was created for both Linear Regression and XGBoost models. The closer the points are to the diagonal line, the better the model performs.**

**4. Conclusion**

**The XGBoost model was the best performer for predicting car prices based on the given features. The dataset was successfully cleaned and preprocessed, allowing for accurate predictions.**

**Key takeaways:**

* **Preprocessing steps like handling missing values, encoding categorical variables, and scaling numerical features were crucial to model performance.**
* **Among all models tested, XGBoost provided the most accurate predictions with the lowest MAE and highest R² score, making it the best model for predicting car prices in this dataset.**

**5. Recommendations**

* **Consider using the XGBoost model for future predictions as it provided the most reliable results.**
* **Additional features such as car color, condition, or region could be included in the dataset to further improve model accuracy.**
* **The preprocessing pipeline can be reused for other datasets to ensure clean, consistent data.**

**6. Future Work**

* **Feature engineering can be explored to create new variables like car age (current year - year of manufacture), which might improve the model.**
* **Hyperparameter tuning could be done for the Random Forest and XGBoost models to optimize their performance further.**
* **Incorporating cross-validation techniques for more robust model evaluation.**

**Appendix**

* **Code Files: The entire codebase, including model training scripts, has been stored and saved using joblib for future use.**
* **Model Files: The XGBoost model, along with the necessary preprocessing objects (e.g., encoders, scalers), was saved for deployment.**

**This report concludes that through proper preprocessing and model selection, car price prediction accuracy can be greatly improved, with XGBoost being the optimal model for this dataset.**