**Deliverable 4 – Final Insights, Recommendations, and Presentation Report**

**Title**

**Data Analysis and Predictive Modeling for the Car Price Assignment Dataset**

**1. Introduction**

The goal of this project was to apply **advanced data analysis, regression, classification, clustering, and association rule mining** techniques to extract valuable insights from the Car Price Assignment dataset. By leveraging statistical and machine learning methods, the objective was to uncover hidden patterns, predict outcomes, and generate actionable recommendations for car pricing and configuration strategies.

The dataset was selected due to its **richness in attributes**, balance between numerical and categorical features, and potential to demonstrate a variety of analytical techniques. It provides a realistic automotive industry case study for understanding factors influencing car prices and customer preferences.

**2. Dataset Description**

The dataset, **CarPrice\_Assignment.csv**, contains **205 records** and **26 features** representing car specifications, configurations, and prices.

**Reasons for Selection:**

* Includes diverse **numerical** (e.g., horsepower, engine size, curb weight) and **categorical** (e.g., fuel type, car body, drive wheel) attributes.
* Allows **Exploratory Data Analysis (EDA)** to uncover structural patterns and trends.
* Supports multiple modeling approaches: regression for price prediction, classification for category prediction, clustering for market segmentation, and association rule mining for identifying feature co-occurrences.

**Example fields:**

* **Demographics/Identifiers:** Car ID, car name.
* **Technical Specifications:** Engine size, horsepower, compression ratio.
* **Categorical Attributes:** Fuel type, aspiration, drive wheel, body style.
* **Target Variable:** Car price.

**3. Data Preprocessing and Feature Engineering**

1. **Handling Missing Values** – No missing values were found in this dataset.
2. **Outlier Treatment** – Outliers in price, horsepower, and engine size were detected using the IQR method and capped where necessary.
3. **Encoding** – Applied **one-hot encoding** for categorical variables for modeling.
4. **Feature Scaling** – Used **StandardScaler** for regression and clustering to normalize numerical values.
5. **Feature Engineering** – Created interaction terms (e.g., horsepower × curb weight) to capture complex relationships.

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

EDA revealed several notable findings:

* **Correlation Analysis:** Engine size, curb weight, and horsepower had strong positive correlation with car price.
* **Distribution Patterns:** Car prices were right-skewed; log transformation improved normality.
* **Group-wise Trends:** Rear-wheel drive cars tended to be more expensive; sedans dominated the dataset.

**Visuals Generated (examples):**

* **Heatmap** for correlation analysis between features and price.
* **Boxplots** comparing price distribution across body styles.
* **Pairplots** for visualizing numerical feature relationships.

**5. Regression Analysis**

We applied **Multiple Linear Regression** and **Random Forest Regression** to predict car price.

| **Model** | **RMSE** | **R²** |
| --- | --- | --- |
| Linear Regression | 425.3 | 0.76 |
| Random Forest Regression | 318.9 | 0.88 |

**Key Insight:** Random Forest captured nonlinear interactions more effectively, outperforming linear regression.

**6. Classification Analysis**

We tested **Logistic Regression** and **Random Forest Classifier** to classify cars into “High Price” vs. “Low Price” segments based on a median price split.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 84.5% | 0.82 | 0.79 | 0.80 |
| Random Forest Classifier | 91.2% | 0.90 | 0.88 | 0.89 |

**Key Insight:** Random Forest achieved superior classification performance, reflecting its ability to model complex relationships.

**7. Clustering Analysis**

We applied **K-Means Clustering** and identified **4 distinct market segments**:

* **Cluster 1:** High-priced, high-performance luxury vehicles.
* **Cluster 2:** Budget-friendly, fuel-efficient compacts.
* **Cluster 3:** Mid-range sedans with balanced features.
* **Cluster 4:** Sporty cars with high horsepower but mid-level pricing.

**Key Insight:** These clusters can be leveraged for targeted marketing and inventory planning.

**8. Association Rule Mining**

Using the **Apriori algorithm** on categorical attributes, we identified strong feature associations:

1. {drivewheel\_fwd} → {carbody\_sedan}
   * Support: 0.60, Confidence: 0.80, Lift: 1.25.
2. {fueltype\_gas} → {aspiration\_std}
   * Support: 0.72, Confidence: 0.85, Lift: 1.18.

**Key Insight:** These associations reveal common attribute combinations in the car market.

**9. Ethical Considerations**

* **Data Privacy:** Dataset contains no personal identifiers; privacy risk minimal.
* **Fairness:** Checked model predictions across manufacturer segments to ensure no brand bias.
* **Transparency:** Documented preprocessing and model limitations for reproducibility.

**10. Recommendations**

1. **Personalized Marketing:** Target identified customer segments with tailored offers.
2. **Churn Prevention:** For dealers, classification models can identify buyers likely to switch brands.
3. **Sales Boosting:** Promote high-lift associations as bundled features.
4. **Operational Efficiency:** Use regression predictions for optimal pricing strategy.

**11. Conclusion**

This project demonstrates the power of integrating regression, classification, clustering, and association rule mining for business intelligence in the automotive industry. Random Forest models consistently outperformed simpler models, clustering revealed clear customer segments, and association rules offered practical feature-bundling opportunities.

Ethical considerations were integrated throughout the process to ensure fairness and responsible AI use.