**Car Price Prediction Project Report**

**1. Introduction**

This project aims to predict car prices based on various features such as engine specifications, car dimensions, fuel type, and other attributes. The objective is to help a Chinese automobile company understand key factors influencing car prices in the US market, enabling them to design and price cars competitively.

**2. Dataset Overview**

Various car models available in the US market:

* **Symboling** (Insurance risk rating)
* **Fuel Type** (Petrol or Diesel)
* **Car Body** (Sedan, Hatchback, SUV, etc.)
* **Engine Specifications** (Cylinders, horsepower, engine size, etc.)
* **Mileage** (City and highway fuel efficiency)
* **Weight & Dimensions** (Curb weight, wheelbase, and length)
* **Price**

**3. Data Preprocessing**

* **Handling Missing Values**: Replaced using median/mode.
* **Categorical Encoding**: Converted categorical variables using Label Encoding.
* **Feature Scaling**: Standardized numerical variables using StandardScaler().
* **Train-Test Split**: 80% training, 20% testing.

**4. Model Implementations**

The following regression models were implemented:

1. **Linear Regression**: Assumes a linear relationship, serves as a baseline.
2. **Decision Tree Regressor**: Captures non-linear relationships but prone to overfitting.
3. **Random Forest Regressor**: Reduces overfitting by averaging multiple decision trees.
4. **Gradient Boosting Regressor**: Sequentially corrects errors for better predictions.
5. **Support Vector Regressor (SVR)**: Uses hyperplanes for prediction, works well with high-dimensional data.

**5. Model Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **R²** |
| Linear Regression | 1.23e+07 | 2087.31 | 0.844 |
| Decision Tree Regressor | 9.53e+06 | 2090.70 | 0.879 |
| Random Forest Regressor | 3.40e+06 | 1307.78 | 0.957 |
| Gradient Boosting Regressor | 5.28e+06 | 1597.28 | 0.933 |
| Support Vector Regressor | 8.69e+07 | 5696.71 | -0.100 |

**6. Feature Importance Analysis**

Using **Random Forest Regressor**, the most influential features were:

* **Engine Size**: Strong correlation with car price.
* **Horsepower**: Higher power leads to higher prices.
* **Curb Weight**: Heavier cars generally cost more.
* **Mileage (MPG)**: Fuel efficiency impacts price.

**7. Hyperparameter Tuning**

Performed GridSearchCV on the best model (Random Forest). Optimized parameters improved R² score significantly.

**8. Conclusion**

* **Best Model**: Random Forest Regressor (Highest R²: 0.957)
* **Worst Model**: Support Vector Regressor (Negative R²)
* **Business Insights**:
  + Engine power and weight are major price determinants.
  + Fuel efficiency plays a role but is secondary.
  + Hyperparameter tuning enhances prediction accuracy.