**Report: Predicting Used Car Sale Prices**

**Introduction**

The rising demand for accurate pricing mechanisms in the used car market has led to the development of machine learning models capable of predicting prices based on car characteristics. This project, part of the BUSA8001 course, involved developing forecasting models to predict used car prices using data cleaning, feature engineering, and advanced regression techniques. It also included a competitive Kaggle component to evaluate performance.

**Data Overview and Preprocessing**

The dataset consists of train.csv and test.csv files, containing features like engine power, car dimensions, and mileage. Preprocessing involved:

* **Cleaning numerical features**: Many features, like back\_legroom, were in a string format combining numbers and text. Using regular expressions, numerical values were extracted and converted to float for consistency.
* **Handling categorical data**: Categorical variables with more than five unique values were grouped into the five most frequent categories plus "other." One-hot encoding was then applied.
* **Missing values**: Numerical columns were imputed with mean values to maintain data integrity, while categorical columns with multiple values were simplified.

**Exploratory Data Analysis (EDA)**

EDA provided critical insights:

1. **Summary Statistics**: Key statistics highlighted outliers in variables like mileage and engine displacement.
2. **Correlation Analysis**: A heatmap of selected features revealed strong correlations between power, horsepower, and price. Features with multicollinearity were flagged for potential exclusion during modeling.
3. **Visual Patterns**: Scatter plots and box plots showed clear trends between features like mileage and price, emphasizing their importance in prediction.

**Feature Engineering**

New features were engineered to enrich the dataset:

1. **Engine Efficiency Score**: Combines fuel tank volume, horsepower, and torque.
2. **Car Age**: Derived from manufacturing year to reflect depreciation.
3. **Legroom Difference**: Highlights interior design aspects.
4. **Latitude-Longitude Product**: Explores location-based pricing patterns.
5. **Fuel Tank-to-Width Ratio**: Evaluates design efficiency.

These features provided additional context for predictive modeling and improved model accuracy.

**Model Development**

Three models were developed and evaluated:

1. **Ridge Regression**:
   * Regularized linear regression to handle multicollinearity.
   * Performance on the training set showed RMSE of ~2000, indicating moderate prediction accuracy.
2. **Random Forest Regressor**:
   * Leveraged ensemble learning with parameters like n\_estimators=700 and max\_depth=60.
   * Achieved significant improvements, with RMSE dropping below 1500.
3. **Gradient Boosting Regressor**:
   * Adjusted hyperparameters, such as learning rate (0.55) and estimators (600), for fine-tuned predictions.
   * Delivered an RMSE of ~1400, outperforming other models.
4. **Ensemble Learning**:
   * Combined Ridge and Gradient Boosting predictions.
   * Further reduced RMSE to ~1350, demonstrating the value of combining model strengths.

**Results and Evaluation**

* **Kaggle Submission**: Predictions from the ensemble model were submitted, achieving a competitive leaderboard position.
* **Metrics**: RMSE was the primary metric, supplemented by feature importance analysis from Random Forest and Gradient Boosting models.
* **Feature Insights**: Variables like car\_age and engine\_efficiency\_score emerged as significant predictors.

**Challenges and Limitations**

1. **Data Quality**: Features with incomplete or ambiguous values required extensive cleaning.
2. **Model Complexity**: Gradient Boosting and Random Forests were computationally expensive, limiting hyperparameter exploration.
3. **Kaggle Constraints**: Submission limits restricted iterative improvements.

**Conclusion and Recommendations**

This project highlighted the potential of machine learning in pricing used cars. Key takeaways include:

1. The importance of robust feature engineering in improving predictive performance.
2. Ensemble learning's ability to integrate complementary models effectively.
3. The need for advanced imputation techniques for better handling of missing data.