**Approach and Model Selection for Boeing Data Science Challenge:**

**Introduction:**

In response to the Data Science Challenge presented by Boeing Global Services, I developed a comprehensive solution to predict vehicle trim and dealer listing prices from a dataset of used car listings. The challenge involved utilizing a training dataset comprising various parameters relevant to car sales, including a unique identifier per listing, twenty-six descriptive features, and two target variables: `Vehicle\_Trim` and `Dealer\_Listing\_Price`.

**Data Preprocessing:**

Understanding the critical role of data quality in predictive modeling, I initiated the process with thorough data exploration and preprocessing. This step included handling missing values, particularly for the target variables, to ensure the integrity of the training dataset. Given the presence of categorical data, I employed one-hot encoding for nominal features like `VehFuel` and `VehPriceLabel` to convert them into a machine-readable format. Furthermore, I applied `LabelEncoder` to ordinal categorical variables, including `Vehicle\_Trim`, enabling their effective use in modeling.

To enrich the dataset, I transformed the `VehFeats` column from string representations to lists, counting the number of features per vehicle as an additional predictive variable. Similarly, I extracted the number of previous users from the vehicle's history, adding another dimension to the dataset. Non-informative columns were identified and removed to focus the model training on relevant features.

**Feature Selection:**

Feature selection was conducted using Recursive Feature Elimination (RFE) with a `RandomForestRegressor` utilized to pinpoint the top 10 features contributing to the prediction of dealer listing price, ensuring the model's focus on the most impactful variables.

**Model Development and Evaluation:**

***Regression for Predicting Dealer Listing Price***

In tackling the prediction of the ‘Dealer\_Listing\_Price’, I expanded the range of regression models evaluated to include not only traditional models like Linear Regression and Random Forest Regression, but also advanced algorithms known for their performance in complex datasets:

* *Linear Regression:* A baseline model for its simplicity and interpretability.
* *Random Forest Regressor:* Leveraged for its robustness against overfitting and ability to handle non-linear data.
* *Decision Tree Regressor:* For its interpretability and ease of use.
* *SVR (Support Vector Regression):* Utilized for its effectiveness in high-dimensional spaces.
* *XGBRegressor:* Chosen for its speed and performance, attributable to its gradient boosting framework.
* *LGBMRegressor:* Selected for its efficiency with large datasets and high performance.
* *CatBoostRegressor:* Known for its handling of categorical variables and speed.

Each model was trained on the training dataset and evaluated using the R-squared metric, allowing for a comparison of their ability to predict the dealer listing price accurately.

***Classification for Predicting Vehicle Trim***

For the classification task of predicting ‘Vehicle\_Trim’, a similar approach was adopted, utilizing a mix of basic and advanced algorithms:

* *Logistic Regression:* A foundational model for binary and multiclass classification problems, extended here with an increased iteration limit for convergence in complex datasets.
* *Random Forest Classifier:* For its ensemble approach that combines multiple decision trees to improve prediction accuracy and control overfitting.
* *Decision Tree Classifier:* For a straightforward classification approach, offering clear interpretability.
* *SVC (Support Vector Classifier):* Applied for its effectiveness in defining decision boundaries in multi-dimensional space.
* *XGBClassifier:* Utilized for its performance and speed, driven by gradient boosting.
* *LGBMClassifier:* Chosen for its scalability and efficiency, particularly in handling large data volumes.
* *CatBoostClassifier:* Renowned for its automatic handling of categorical variables and robustness.

The models were assessed based on accuracy and F1-score, ensuring a comprehensive evaluation of their predictive capabilities for vehicle trim.

**Submission Preparation:**

The ensemble models were then applied to the test dataset, and carefully prepared to mirror the preprocessing and feature selection of the training set. The predictions for `Vehicle\_Trim` were mapped back to their original names using the `LabelEncoder` inverse transformation, ensuring clarity in the submission file.