Design Document: Predicting Car Resale Prices Using Machine Learning

# 1. Problem Statement

An online car dealership wants to predict the resale prices of used cars based on several features such as the car’s make, model, year, mileage, engine size, fuel type, and other relevant attributes. The objective of this project is to build a machine learning model that can accurately estimate the resale price of a car given its features.

# 2. Key Design Decisions

## 2.1 Training/Test Data Split

Decision: Split the dataset into training and testing sets using an 80%-20% split.  
Rationale: An 80-20 split is commonly used to provide enough data for model training while retaining sufficient data for testing. This helps evaluate model performance on unseen data and ensures the model generalizes well.

## 2.2 Acceptable Error Rates

Decision: Acceptable error rates are defined using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Based on initial experiments:  
MAE Target: Less than $2500.  
RMSE Target: Less than $3000.

## 2.3 Machine Learning Algorithms Used

Decision: Multiple regression models were trained and evaluated to select the best performer:  
- Linear Regression  
- Random Forest  
- Gradient Boosting

## 2.4 Outlier Removal

Decision: Outliers were detected and handled through Interquartile Range (IQR) for continuous features such as mileage and engine size. Manual investigation of extreme resale prices was also conducted.

## 2.5 Dimensionality Reduction Decisions

Decision: Principal Component Analysis (PCA) was explored but not implemented, as the dataset's key features contributed to the predictive power.

## 2.6 Feature Engineering

New features were engineered, such as the age of the car (year of sale minus year of manufacturing), and categorical features like FuelType and Transmission were one-hot encoded.

## 2.7 Model Versioning

Version control was implemented to track model iterations using Git and MLflow. This enabled tracking of hyperparameter tuning and model performance improvement.

# 3. Model Evaluation

Random Forest was selected as the final model. Below are the evaluation metrics for the model:  
- Mean Absolute Error (MAE): 11411.812674699999  
- Root Mean Squared Error (RMSE): 13291.006460411945  
- R-squared (R²): -0.04844166010442352

# 4. Visual Enhancements

Several diagrams were included, such as train-test split visualization, outlier detection, and feature importance visualization.

# 5. Conclusion

The Random Forest model performed best, with an MAE of 11411.81 and an R² score of -0.04844. Feature importance analysis highlighted that car age, mileage, engine size, and car condition were the most important factors.