**Project Report**

**Project Title:**

Car Dheko - Used Car Price Prediction

**Domain:**

* Automotive Industry
* Data Science
* Machine Learning

**Objective:**

Develop a machine learning model that accurately predicts the prices of used cars based on a set of features (make, model, year, fuel type, transmission type, etc.). The model will be deployed in a user-friendly Streamlit application.

**Project Scope:**

* Use historical data on used car prices from multiple cities.
* Perform data preprocessing, exploratory data analysis, and feature engineering.
* Build, train, and optimize a machine learning model for price prediction.
* Deploy the model as a Streamlit web application for user interaction.

**Approach:**

**1. Data Processing:**

* **Import and Concatenate Datasets:**
  + Import datasets from different cities in unstructured formats.
  + Standardize them into a structured format.
  + Add a new column ‘City’ to distinguish the city-wise data.
  + Concatenate all datasets into a single dataset.
* **Handling Missing Values:**
  + Removed columns with more than 40% missing values.
  + For numerical columns, used median or mode imputation.
  + For categorical columns, used mode imputation.
* **Removing Redundant Features:**
  + Removed features that conveyed the same information (e.g., columns with similar meanings).
* **Data Type Checking and Conversion:**
  + Converted string data to numerical values where necessary (e.g., removing "kms" from KmsDriven).
  + Ensured that categorical data had consistent labels, corrected spelling errors, and checked for repetitive values.
* **Handling Duplicates:**
  + Removed duplicate rows to ensure clean data.
* **Handling Outliers:**
  + Used the IQR method to detect and remove outliers from continuous variables.

**2. Exploratory Data Analysis (EDA):**

* **Feature Selection:**
  + Identified continuous and categorical columns.
    - Continuous columns: ['km', 'price', 'Mileage', 'Engine', 'Max\_Power', 'Torque', 'Length', 'Width', 'Height', 'WheelBase'].
    - Categorical columns: ['city', 'bt', 'transmission', 'ownerNo', 'oem', 'modelYear', 'InsuranceValidity', 'FuelType', 'RTO', 'Top\_features', 'Seats', 'Color', 'NoofCylinder', 'ValuesperCylinder', 'TurboCharger', 'GearBox', 'SteeringType', 'FrontBrakeType', 'RearBrakeType', 'NoDoorNumbers'].
* **Correlation Analysis:**
  + Performed a correlation check to identify the relationships between features and price.

**3.Model Building:**

* Several machine learning algorithms were applied, but **Random Forest Regressor** yielded the best performance:
  + **R² Score**: High accuracy in predicting prices, showing that the model explained most of the variability in used car prices.
  + **Error Metrics**: Low MAE and MSE scores indicated that the model was making relatively small prediction errors, providing reliable estimates.

**4. Feature Importance & Model Refinement:**

* After training the initial model, I focused on **feature importance** to simplify and optimize the model further:
  + Removing non-contributing features like TurboCharger, Seats, and SteeringType improved performance.
  + This pruning process was crucial in reducing noise, improving interpretability, and enhancing model efficiency.
  + **After feature reduction**, the model retained its high predictive power, confirming that the remaining features were the most influential.

**5. Model Optimization & Hyperparameter Tuning:**

* Using **Grid Search** for hyperparameter tuning, I fine-tuned the Random Forest model to enhance its performance.
  + This step increased the model's accuracy by adjusting parameters like n\_estimators, max\_depth, and min\_samples\_split.

**6. Final Results:**

* The final **R² Score** after optimization was excellent, indicating that the model captured most of the variance in the dataset.
* After model tuning and feature reduction, the Random Forest model performed well, showing strong predictive ability.

**7. Streamlit Application Deployment:**

* Deployed the model using **Streamlit** to allow users to input relevant car features and receive real-time price predictions.
  + The app is responsive, user-friendly, and delivers predictions efficiently, making it a practical tool for estimating used car prices.

**Personal Reflections:**

* Throughout this project, my focus was on maintaining data integrity and simplifying the modeling process. By removing redundant features and using the right imputation methods, I was able to create a lean and effective model.
* The transition from raw data to a streamlit web app involved constant iteration, but the final result was both accurate and easy to deploy.
* **Key Takeaway**: Removing unnecessary features and focusing on key variables played a significant role in improving model performance, showing that simpler models can sometimes outperform more complex ones.