

<b>Project Title</b>	<b>Car Dheko - Used Car Price Prediction</b>
<b>Skills take away From This Project</b>	<ol style="list-style-type: none"><li>1. Data Cleaning and Preprocessing</li><li>2. Exploratory Data Analysis</li><li>3. Machine Learning Model Development</li><li>4. Price Prediction Techniques</li><li>5. Model Evaluation and Optimization</li><li>6. Model Deployment</li><li>7. Streamlit Application Development</li><li>8. Documentation and Reporting</li></ol>
<b>Domain</b>	<b>Automotive Industry , Data Science, Machine Learning</b>

### **Problem Statement:**

#### **Objective:**

Imagine you are working as a data scientist in Car Dheko, your aim is to enhance the customer experience and streamline the pricing process by leveraging machine learning. You need to create an accurate and user-friendly streamlit tool that predicts the prices of used cars based on various features. This tool should be deployed as an interactive web application for both customers and sales representatives to use seamlessly.

#### **Project Scope:**

We have historical data on used car prices from CarDekho, including various features such as make, model, year, fuel type, transmission type, and other relevant attributes from different cities. Your task as a data scientist is to develop a machine learning model that can accurately predict the prices of used cars based on these features. The model should be integrated into a Streamlit-based web application to allow users to input car details and receive an estimated price instantly.

## Approach:

### 1) Data Processing

#### a) Import and concatenate:

- i) Import all city's dataset which is in unstructured format.
- ii) Convert it into a structured format.
- iii) Add a new column named 'City' and assign values for all rows with the name of the respective city.
- iv) Concatenate all datasets and make it as a single dataset.

#### b) Handling Missing Values: Identify and fill or remove missing values in the dataset.

- i) For numerical columns, use techniques like mean, median, or mode imputation.
- ii) For categorical columns, use mode imputation or create a new category for missing values.

#### c) Standardising Data Formats:

- i) Check for all data types and do the necessary steps to keep the data in the correct format.

(1) Eg. If a data point has string formats like 70 kms, then remove the unit 'kms' and change the data type from string to integers.

#### d) Encoding Categorical Variables: Convert categorical features into numerical values using encoding techniques.

- i) Use one-hot encoding for nominal categorical variables.
- ii) Use label encoding or ordinal encoding for ordinal categorical variables.

#### e) Normalizing Numerical Features: Scale numerical features to a standard range, usually between 0 and 1. (For necessary algorithms)

- i) Apply techniques like Min-Max Scaling or Standard Scaling.

#### f) Removing Outliers: Identify and remove or cap outliers in the dataset to avoid skewing the model.

- i) Use IQR (Interquartile Range) method or Z-score analysis.

## 2) Exploratory Data Analysis (EDA)

- a) **Descriptive Statistics:** Calculate summary statistics to understand the distribution of data.
  - i) Mean, median, mode, standard deviation, etc.
- b) **Data Visualization:** Create visualizations to identify patterns and correlations.
  - i) Use scatter plots, histograms, box plots, and correlation heatmaps.
- c) **Feature Selection:** Identify important features that significantly impact the car prices.
  - i) Use techniques like correlation analysis, feature importance from models, and domain knowledge.

## 3) Model Development

- a) **Train-Test Split:** Split the dataset into training and testing sets to evaluate model performance.
  - i) Common split ratios are 70-30 or 80-20.
- b) **Model Selection:** Choose appropriate machine learning algorithms for price prediction.
  - i) Linear Regression, Decision Trees, Random Forests, Gradient Boosting Machines, etc.
- c) **Model Training:** Train the selected models on the training dataset.
  - i) Use cross-validation techniques to ensure robust performance.
- d) **Hyperparameter Tuning:** Optimize model parameters to improve performance.
  - i) Use techniques like Grid Search or Random Search.

## 4) Model Evaluation

- a) **Performance Metrics:** Evaluate model performance using relevant metrics.
  - i) Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared.

- b) **Model Comparison:** Compare different models based on evaluation metrics to select the best performing model.

## 5) Optimization

- a) **Feature Engineering:** Create new features or modify existing ones to improve model performance.
  - i) Use domain knowledge and exploratory data analysis insights.
- b) **Regularization:** Apply regularization techniques to prevent overfitting.
  - i) Lasso (L1) and Ridge (L2) regularization.

## 6) Deployment

- a) **Streamlit Application:** Deploy the final model using Streamlit to create an interactive web application.
  - i) Allow users to input car features and get real-time price predictions.
- b) **User Interface Design:** Ensure the application is user-friendly and intuitive.
  - i) Provide clear instructions and error handling.

## Results:

1. A functional and accurate machine learning model for predicting used car prices.
2. Comprehensive analysis and visualizations of the dataset.
3. Detailed documentation explaining the methodology, models, and results.
4. An interactive Streamlit application for real-time price predictions based on user input.

## Project Evaluation metrics:

### 1. Model Performance:

- 1.1. Mean Absolute Error (MAE)
- 1.2. Mean Squared Error (MSE)
- 1.3. R-squared

### 2. Data Quality:

- 2.1. Completeness and accuracy of the preprocessed data.

### 3. Application Usability:

- 3.1. User satisfaction and feedback on the Streamlit application.

### 4. Documentation:

- 4.1. Clarity and completeness of the project report and code comments.

## Technical Tags:

Data Preprocessing, Machine Learning, Price Prediction, Regression, Python, Pandas, Scikit-Learn, Exploratory Data Analysis (EDA), Streamlit, Model Deployment

## Data Set:

The Dataset contains multiple excel files, each represents its city, columns in each excel gives you an overview of each car, its details, specification and available features.

**Data Collected From:** [CarDekho](#)

**Dataset Link:** [Dataset](#)

**Feature Description Link:** [Features](#)

## Data Set Explanation:

- The dataset includes detailed information about used cars listed on CarDekho for different cities.
- Preprocessing Steps: Convert the unstructured data into structured data. Handling missing values, standardizing data formats, encoding categorical variables, and normalizing numerical features.

### **Project Deliverables:**

1. Source code for data preprocessing and model development.
2. Documentation detailing the methodology, models used, and evaluation results.
3. Visualizations and analysis reports derived from the exploratory data analysis.
4. Final predictive model with a user guide.
5. Deployed Streamlit application for price prediction.
6. Should give justification for the method of approach and model selection.

### **Project Guidelines:**

- Follow consistent coding standards (e.g., PEP 8 for Python).
- Use version control (e.g., Git) for tracking changes and collaboration.
- Write clear and concise documentation and code comments.
- Ensure code modularity and reusability.
- Perform thorough testing and validation of models.
- Ensure the Streamlit application is user-friendly and well-documented.

### **Timeline:**

- 10 days

## References:

Project Orientation Link (Tamil)	<a href="#">Car Dheko - Used Car Price Prediction Orientation Session(TAMIL) - Google Docs</a>
Project Orientation Link (English)	<a href="#">Car Dheko - Used Car Price Prediction Orientation Session(ENGLISH) - Google Docs</a>
GitHub Reference	<a href="#">How to Use GitHub.pptx</a>
Capstone Explanation Guideline	<a href="#">Capstone Explanation Guideline</a>

## PROJECT DOUBT CLARIFICATION SESSION ( PROJECT AND CLASS DOUBTS)

**About Session:** The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

**Note:** Book the slot at least before 12:00 Pm on the same day

**Timing:** Tuesday, Thursday, Saturday (5:00PM to 7:00PM)

**Booking link :** <https://forms.gle/XC553oSbMJ2Gcfug9>

## LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

**About Session:** The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

**Note:** This form will Open on Saturday and Sunday Only on Every Week

**Timing:** Monday-Saturday (11:30PM to 12:30PM)

**Booking link :** <https://forms.gle/1m2Gsro41fLtZurRA>