Project Report – CarDekho – Used Car Price Prediction

Table of Contents

[Project Title 2](#_Toc177677091)

[Skills Take Away from this Project 2](#_Toc177677092)

[Domain 2](#_Toc177677093)

[Problem Statement: 2](#_Toc177677094)

[Project Scope: 2](#_Toc177677095)

[Approach: 2](#_Toc177677096)

[Data Processing 2](#_Toc177677097)

[Import and concatenate: 2](#_Toc177677098)

[Handling Missing Values: 2](#_Toc177677099)

[Standardising Data Formats: 3](#_Toc177677100)

[Removing Outliers: 3](#_Toc177677101)

[Exploratory Data Analysis (EDA) 3](#_Toc177677102)

[Model Development 3](#_Toc177677103)

[Deployment 3](#_Toc177677104)

# Project Title

Car Dheko - Used Car Price Prediction

# Skills Take Away from this Project

* Data Cleaning and Preprocessing
* Exploratory Data Analysis
* Machine Learning Model Development
* Price Prediction Techniques
* Model Evaluation and Optimization
* Model Deployment
* Streamlit Application Development
* Documentation and Reporting

# Domain

Automotive Industry, Data Science, Machine Learning

# Problem Statement:

* Develop a machine learning model which can predict prices of the used cars from CarDheko data set according to the features.
* Deploy the Machine Learning model in the Streamlit application.

# Project Scope:

* Analyse the historical data of the used car prices and predict the car prices with the set of features.
* Cleaning the Dataset from CarDekho
* Build a Machine Learning Algorithm for predicting the used car price.
* Deploy the developed machine learning algorithm using the Streamlit application.

# Approach:

## Data Processing

### Import and concatenate:

* Import all city’s dataset which is in unstructured format.
* Convert it into a structured format.
* Add a new column named ‘City’ and assign values for all rows with the name of the respective city.
* Concatenate all datasets and make it as a single dataset.

### Handling Missing Values:

* Identify and fill or remove missing values in the dataset.
* For numerical columns, use techniques like mean, median, or mode imputation.
* For categorical columns, use mode imputation or create a new category for missing values.

### Standardising Data Formats:

* Check for all data types and do the necessary steps to keep the data in the correct format.
* 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.
* Encoding Categorical Variables: Convert categorical features into numerical values using encoding techniques.
* Use one-hot encoding for nominal categorical variables.
* Use label encoding or ordinal encoding for ordinal categorical variables.
* Normalizing Numerical Features: Scale numerical features to a standard range, usually between 0 and 1.( For necessary algorithms)
* Apply techniques like Min-Max Scaling or Standard Scaling.

### Removing Outliers:

* Identify and remove or cap outliers in the dataset to avoid skewing the model.
* Use IQR (Interquartile Range) method or Z-score analysis.

# Exploratory Data Analysis (EDA)

* The data is then evaluated using the statistical methods and finding the correlation matrix/heatmap to understand all the variable relationship with the price.
* The non-important values are then removed from the data frame
* Data Visualization to understand the distribution of the data and heatmap to understand the relationship of each feature with other features and target.
* Main impacting features will be selected at last.

# Model Development

* Cleansed data is then undergo to train, test, split with the ratio of 80:20
* Then model Linear Regression, XGboost, Random Forest, Decision Trees, etc., will be evaluated with the r2 score.
* Search for the scope of hyper tuning.

# Deployment

* Streamlit application will be developed to receive the input as features from the user.
* The received input is then pushed to pickle package and extract the answer from the streamlit application.