**Introduction:**

The aim of this technical project document is to describe the process of building a car price prediction system using machine learning algorithms. The main goal of this project is to develop a system that can predict the price of a car based on several parameters such as car\_name, model, year of manufacture, seller type, max\_power, mileage, engine size, transmission type, and fuel type. The system will use historical data from various sources to train the machine learning algorithms.

**Scope:**

The car price prediction system will use data from various sources such as online marketplaces, car dealerships, and private sellers. The system will analyse the data and use machine learning algorithms to predict the price of a car based on the given parameters. The system will be able to handle a large amount of data and provide accurate predictions with a high level of confidence.

**Objectives:**

The main objectives of this project are as follows:

1. To build a car price prediction system using machine learning algorithms

2. To collect and pre-process the data from various sources

3. To train the machine learning algorithms using historical data

4. To validate the accuracy of the system using test data

5. To optimize the system for improved performance

**EDA :**

Perform EDA on car data by using different visualization technique:- scatterplot, barplot, subplot, countplot, boxplot, histogram, pair plot, heatmap, regplot, joint plot and kdeplot.

* In EDA we’ve separated the categorical columns and numerical columns for better univariate and bivariate analysis.
* For univariate analysis on categorical columns we’ve visualized the count plots which demonstrates different count for all columns in our dataset:

1. Fuel\_type – diesel>petrol>CNG>LPG
2. Seller\_type – Individual>dealer>Trustmark Dealer
3. Transmission – Manual>Automatic
4. Owner – 1st owner> 2nd owner>3rd owner>4th and above owner

* For univariate analysis on numerical columns we’ve visualized the box plot which is used to show the distribution of numerical data along with the symmetry and skewness of the data. Thus from the box plot we saw, most of the features have outliers which were dealt later in feature engineering.
* Through histogram of different features we saw that mileage is normally distributed, while car age, max power and selling price is right skewed.
* The scatter plot between km\_driven and selling price depicts that vehicle which travelled a longer distance tends to mean the vehicle is older and hence the selling price will be lower.
* Selling Price of cars 2 years old would be high and gradually decreases with car of 17 years old
* Selling Price of cars seems to have higher prices when sold by Dealers when compared to Individuals
* It can be observed that Selling Price would be higher for cars that are Automatic.

* Selling Price of cars with Fuel Type of Diesel is higher than Petrol and CNG
* The heat map depicts that selling\_price and max power are in slight correlation with each other.
* Regplot between selling\_price and max power gives the best fit line between the two.

**Data Collection and Pre-processing:**

* The first step in building a car price prediction system is to collect and pre-process the data. The data will be collected from various sources such as online marketplaces, car dealerships, and private sellers. The data will be in different formats, so it will need to be pre-processed before being used for training the machine learning algorithms.
* In feature engineering we do the following:

1. split the "name" column and takes the "manufacturer" name from it.
2. create the function car age from the buying year to current year.
3. modify mileage to corrected mileage and converts all the km/kg to kmpl, thus bringing all the values in one single unit.
4. convert engine size from object data type to float data type by eliminating cc from the string and converting the numerical values to float data type.
5. convert power values from object data type to float data type by eliminating bhp from the string and converting the numerical values to float data type.
6. replacing all the missing values in the dataframe in mileage, max\_power with mean, engine size with median and seat with mode.
7. And finally perform feature encoding on categorical columns. One hot encoding on fuel type, seller type and manufacture and label encoding on transmission type and owner type.

**The pre-processing steps include:**

**Data cleaning**: Remove any irrelevant or duplicate data from the dataset

**Data integration**: Combine data from different sources to create a comprehensive dataset

**Data** **transformation**: Convert data into a suitable format by using some transformation technique like log transformation, cube root transformation, square root transformation for machine learning algorithms.

**Data normalization**: Scale the data to a common range to improve accuracy of the algorithms

**Data Training and Validation:**

Once the data has been pre-processed, the next step is to train the machine learning algorithms. The system will use different algorithms such as linear regression, decision tree regression, and random forest regression to train the model. The algorithms will be evaluated based on their r2\_score, MSE, and MAE.

The system will also be validated using test data. The test data will be a subset of the original data that was not used for training the machine learning algorithms. The system will be evaluated based on its ability to predict the price of the car accurately. The validation process will be repeated several times to ensure the system's accuracy and reliability.

**Optimization:**

The final step is to optimize the system for improved performance. The optimization process will involve fine-tuning the machine learning algorithms and improving the pre-processing steps. The aim is to improve the accuracy of the system.

**Conclusion:**

The car price prediction system will provide a valuable tool for car buyers and sellers. The system will be able to predict the price of a car accurately based on several parameters such as brand or manufacture, year of manufacture, car\_age, mileage, engine size, transmission type, and fuel type. The system will use machine learning algorithms to analyse historical data and provide accurate predictions with a high level of confidence. The system will be optimized for improved performance and reliability.

**Summary:**

With an objective to achieve the generalised model we’ve divided the project into major 5 files that is – App.py, EDA.py, data\_preprocessing.py, training.py and utils.py.

* Eda.py is the class that has all the exploratory data analysis done on our data frame.
* data\_preprocessing file includes class “data\_cleaning” which includes methods that does data cleaning on features, drops irrelevant features, renames some features, imputes missing values and finally encoding on columns.
* Then there is a training.py file, which does model training and evaluate the model.
* Finally, App.py is the base class that then takes the model building ahead through data pre processing, and training.
* Utils.py file predicts the car price by taking the inputs from the users.

**Team Members :** Tanushka, Tejas, Suchita, Anand

**Team Mentor :** Shriti Datta