**Advanced Machine Learning**

**Group project summary**

**Predicting selling price of used cars using Machine Learning**

**Group 13**

**Roshan Pampari (29)**

**Jayanth (26)**

**Preetham Thatikonda (46)**

**Machine Learning pipeline :**

We have implemented our project in 6 pipelines as follows

1. **Problem Formulation**
2. **Data Collection and Data Labeling**
3. **Data Evaluation and Feature Engineering**
4. **Splitting Data and Standardization**
5. **Model Training and Model Evaluation**
6. **Model Deployment**

**1 Problem Formulation:**

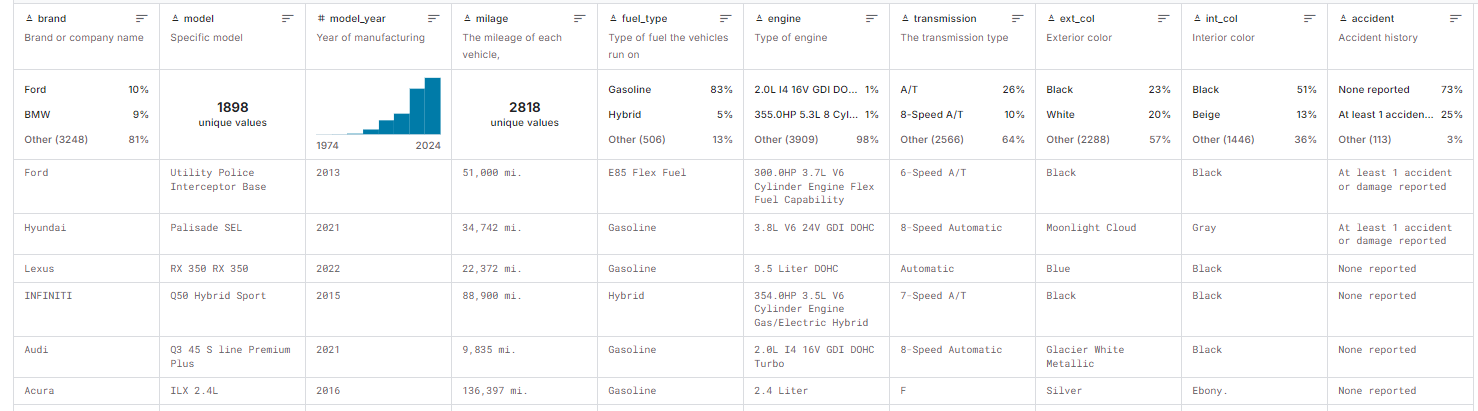
Let us take a general scenario as a students if we want to buy a car we usually don’t go with a Brand new car right ?

We usually go with a used car with good condition and few other features

We decided to create a project focused on predicting the selling price of used cars. With the increase in car buying and selling, forecasting the prices of used cars has gained significant attention. The goal of this project is to estimate used car prices by utilizing attributes that are strongly associated with price. To accomplish this, machine learning techniques have been employed.

**About dataset**

* The Used Car Price Prediction Dataset is a detailed compilation of automotive data sourced from the popular marketplace website [cars.com](https://www.cars.com/). It contains 4,009 entries, each representing a unique vehicle listing, with twelve distinct features offering valuable insights into the automotive market
* **Brand & Model:** Identifies the manufacturer and specific model of each vehicle, providing insight into its make and design.
* **Model Year:** Indicates the year the vehicle was produced, important for understanding its age and technological features.
* **Mileage:** Retrieve the mileage for each vehicle, which serves as a crucial indicator of its usage, wear and tear, and potential future maintenance needs.
* **Fuel Type:** Specifies the type of fuel used, such as gasoline, diesel, electric, or hybrid, affecting efficiency and costs.
* **Engine Type:** Describes the engine specifications, offering details on performance and fuel efficiency.
* **Transmission**: Defines the transmission system, such as automatic, manual, or other types, influencing driving experience.
* **Exterior & Interior Colors:** Highlights the vehicle’s color options, showcasing its style and aesthetic appeal.
* **Accident History:** Indicates whether the vehicle has been involved in accidents, a critical aspect of its condition.
* **Clean Title**: Confirms if the vehicle has a clean ownership record, affecting its resale value and legality.
* **Price:** Displays the vehicle’s listed price, helping in comparing options and planning budgets.

**Having a glance at dataset**

**2 Data collection and Data labeling :**

We have obtained our dataset from kaggle as follows

<https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset/data>

And our dataset is labeled with price being target, hence we need not to perform any methods for finding out target Feature

**3 Data Evaluation and Feature Engineering :**

We have three Features with significant null values as follows

**fuel\_type** : 170

**accident**  : 113

**clean\_title** : 596

In **fuel\_type** feature we have replaced all null values and character “-“ with not supported

In **clean\_title** Feature, all clean titles are with values yes and remaninng are left blank hence these are no-clean title so populated them with no

In **accident** Feature we have replaced all null values with non-reported

In **milage** Feature we have removed all unnecessary characters(“mi.”,”,”)

In **Price** Feature we have removed all unnecessary characters(“$”,”,”)

Adding new Feature **Age** and populated with values of age of car using **model\_year**

Here we have analyzed Feature **Transmission** and transformed all values to three unqiue values – “**automatic”, “manual”, “other”.**

We have analyzed Feature **engine** and have extracted data in it to two separate columns – **“horsepower”** and  **“engine dispacement”**

We have replaced all null values with mean for data in **horsepower** and removed all unnecessary characters in **engine dispacement**

We have updated all colors to standard colors

We found five Features with outliers – **“price”, “horsepower”, ”milage”, “age”** and **“engine\_displacement”**

We have replaced extreme values which are not in boundary region with lower boundaries and higher boundaries respectively( i.e values below lower boundary are replaced with lower boundary and values above higher boundary with higher boundary )

Also Using one hot encoding we have converted values in features **“transmission”**, **“accident”**, **“clean\_title”** into **True** or **False** values and then to binary **0** and **1**

We went with binary encoding for Feature – **“brand”**

And then finally label encoding to Features – **“fuel\_type”, “ext\_col”, “int\_col”**

**4 Splitting Data and Standardization** :

Here we split our dataset into **“train”, “validation” and “test”** in ratios of **70:15:15**

Once datasets are made we went with standardization, Standardization can improve a model's ability to generalize to new, unseen data.

By removing the scale differences between features, the model can focus on learning the underlying patterns and relationships, rather than being biased towards features with larger scales.

**5 Model training and model evaluation:**

Initially, we trained our model using the XGBoost Regressor, followed by the Random Forest Regressor. However, the Random Forest Regressor did not perform as well as the XGBoost Regressor. The performance comparison showed Random Forest Regressor at 83% and XGBoost Regressor at 86%.

To improve the models, we fine-tuned them using GridSearchCV. Even after tuning, the XGBoost Regressor outperformed the Random Forest Regressor, with R² scores of 86% and 84%, respectively.

Finally, when we tested the XGBoost model on the test dataset, it achieved an R² score of 84.1%, which is a strong result compared to the Random Forest Regressor's 83%.

**6 Model deployement**

We finally deploy our model **Xg Boost Regressor** for predicting price of used cars.