**Background & Context**

There is a huge demand for used cars in the Indian Market today. As sales of new cars have slowed down in the recent past, the pre-owned car market has continued to grow over the past years and is larger than the new car market now. Cars4U is a budding tech start-up that aims to find footholes in this market.

In 2018-19, while new car sales were recorded at 3.6 million units, around 4 million second-hand cars were bought and sold. There is a slowdown in new car sales and that could mean that the demand is shifting towards the pre-owned market. In fact, some car sellers replace their old cars with pre-owned cars instead of buying new ones. Unlike new cars, where price and supply are fairly deterministic and managed by OEMs (**O**riginal **E**quipment **M**anufacturer / except for dealership level discounts which come into play only in the last stage of the customer journey), used cars are very different beasts with huge uncertainty in both pricing and supply. Keeping this in mind, the pricing scheme of these used cars becomes important in order to grow in the market.

As a senior data scientist at Cars4U, you have to come up with a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing. For example, if the business knows the market price, it will never sell anything below it.

**Objective**

1. Explore and visualize the dataset.
2. Build a linear regression model to predict the prices of used cars.
3. Generate a set of insights and recommendations that will help the business.

**Data Dictionary**

1. S.No. : Serial Number
2. Name : Name of the car which includes Brand name and Model name
3. Location : The location in which the car is being sold or is available for purchase Cities
4. Year : Manufacturing year of the car
5. Kilometers\_driven : The total kilometers driven in the car by the previous owner(s) in KM.
6. Fuel\_Type : The type of fuel used by the car. (Petrol, Diesel, Electric, CNG, LPG)
7. Transmission : The type of transmission used by the car. (Automatic / Manual)
8. Owner : Type of ownership
9. Mileage : The standard mileage offered by the car company in kmpl or km/kg
10. Engine : The displacement volume of the engine in CC.
11. Power : The maximum power of the engine in bhp.
12. Seats : The number of seats in the car.
13. New\_Price : The price of a new car of the same model in INR Lakhs.(1 Lakh = 100, 000)
14. Price : The price of the used car in INR Lakhs (1 Lakh = 100, 000)

**Best Practices for Notebook :**

* The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* It is preferable to remove all warnings and errors before submission.

**Submission Guidelines :**

1. The submission should be a well-commented Jupyter notebook [format - .html and .ipynb]
2. Any assignment found copied/ plagiarized with other groups will not be graded and awarded zero marks
3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation
4. Submission will not be evaluated if,
   1. it is submitted post-deadline, or,
   2. more than 2 files are submitted

Happy Learning!!

**Scoring guide (Rubric) - Cars4U Project Rubric**

| **Criteria** | **Points** |
| --- | --- |
| **Define the problem and perform an Exploratory Data Analysis**  - Problem definition, questions to be answered - Data background and contents - Univariate analysis - Bivariate analysis | 10 |
| **Illustrate the insights based on EDA**  Key meaningful observations on the relationship between variables | 5 |
| **Data pre-processing**  Data Preparation for Modelling - Missing value Treatment - Outlier Treatment - Feature Engineering | 15 |
| **Model building - Linear Regression**  - Build the model and comment on the model statistics - Print model coefficients with column names in a dataframe - Identify the key variables that have a strong relationship with dependent variable | 12 |
| **Model performance evaluation**  - Evaluate the model on different performance metrics - RMSE, MAE, Adjusted R-square - Comment on the performance measures and if there is any need to improve the model or not | 6 |
| **Actionable Insights & Recommendations**  - Conclude with the key takeaways for the business - what would your advice be to grow the business? | 6 |
| **Notebook - Overall Quality**  - Structure and flow - Well commented code | 6 |
| Points | 60 |