**What drives the price of a car?**

The main objective of this application is to analyse and identify the factors that impact car prices. By conducting this analysis, we will provide valuable recommendations to our client, a used car dealership, regarding the features and qualities that consumers value in a used car. This will help the dealership make more informed decisions and improve their understanding of customer preferences.

The application uses Vehicle Dataset from Kaggle, which contains around 426K samples of used cars. In this application, we have explored various aspects of the dataset, such as the car's manufacturer, model, year, mileage, condition, and other relevant attributes. It is important to keep in mind the specific needs and objectives of the client to ensure that the final recommendations are relevant, practical, and aligned with their goals.

We used industry standard process, named CRISP-DM framework, to work through a data problem.

1. **Business Understanding**

In the "What Drives the Price of a Car" project, we have analyzed a subset of the Vehicle Dataset from Kaggle, which contains around 426K samples of used cars. The main objective of this analysis is to identify the factors that impact car prices. By conducting this analysis, we will provide valuable recommendations to our client, a used car dealership, regarding the features and qualities that consumers value in a used car. This will help the dealership make more informed decisions and improve their understanding of customer preferences.

Throughout the analysis, we will explore various aspects of the dataset, such as the car's manufacturer, model, year, mileage, condition, and other relevant attributes. It is important to keep in mind the specific needs and objectives of the client to ensure that the final recommendations are relevant, practical, and aligned with their goals.

Thank you Kaggle for allowing us to use your dataset <https://www.kaggle.com/datasets> .

1. **Data Understanding: Exploratory Data Analysis (EDA)**

* Load and Read DataSet
* Understand Features and Datatype
* Analyse Data and Statistics

About Dataset

The vehicles dataset contains total 18 features/variables including target variable named price. Here's brief about each feature:

* id: Unique identifier of each sample
* region: Region where vehicle belongs to
* price: Price of a Vehicle
* year: Year of a vehicle
* manufacturer: Manufactures of a vehicle i.e. Ford, Toyota etc.
* model: Model of the vehicle i.e. Elantra, Camry etc.
* condition : Condition of a Vehicle
* cylinders : No. of Cylinders in a vehicle
* fuel : Fuel type of a vehicle i.e. Gas, Electric etc.
* odometer : No. of Kms/Miles vehicle has driven
* title\_status : Title of a vehicle
* transmission : Transmission type of a vehicle i.e. manual/auto etc.
* VIN : Unique Identifier of a vehicle
* drive : Drive type of a vehicle i.e front wheel drive etc.
* size : Size of a car i.e. full-size, compact etc.
* type : Vehicle type
* paint\_color : Color of a Vehicle i.e. white/black etc.
* state : State of a vehicle

1. **Data Preparation**

* Drop features & duplicates
* Handle Missing Values
* Handle Outliers, correct data types
* Feature Encoding
* Handling inconsistent data (example: price shouldn't be less than 1)
* Visualize cleaned Data

1. **Modeling**

We have analysed the data against various regression models through different hyperparameters and perform cross validation to determine the best suited model for a vehicle dataset. The models used in current applications are:

* Linear Regression Model
* Linear Regression Model with Polynomial Degree
* Linear Regression Model with Feature Selection
* Lasso Regression Model
* Ridge Regression Model

**Modeling Verdict**: We have built following models and evaluate their R2 Score and RMSE (Root Mean Squared Error) to determine the best model

* Linear Regression model with Degree complexity level 3: RMSE = 0.9509, R2 = 0.3415
* Linear Regression model with 4 features selection: RMSE = 1.0919, R2 = 0.1317
* Lasso model with 4 features selection: RMSE = 1.1397, R2 = 0.0540
* Ridge model with best alpha 68.66: RMSE = 1.0919, R2 = 0.1317

Though, Linear Regression model with Degree complexity level 3 performs best among all models (lower RMSE and higher R2 Score), but it is overfitting due to degree complexity 3. Therefore, **best performing model is Ridge model** with **best alpha as 68.66: RMSE = 1.0919, R2 = 0.1317**

1. **Evaluation**

Once we have determined the best suited model (Ridge Regression Model), we are going to perform following:

* Predict the Price against Test Dataset (Actual vs Predicted)
* Model Indicators i.e., MSE, RSME, R2 Score
* Coefficients and Intercept
* Visualize Linear Regression on Actual Vs Predicted
* Visualize Price against each Independent Variables

**Interpret the Coefficients and Y-Intercept**

* Coefficient Value: The value of each coefficient represents the expected change in the target variable for a one-unit change in the corresponding feature, assuming all other features remain constant.
* Intercept: The intercept represents the expected value of the target variable when all features are zero.
* year, fuel\_status, size features have positive coefficients which means, unit increase in these features result in increase of target variance as per features coefficient value.
* Other features have negative coefficients which means, unit increase in these features result in decrease of target variance as per features coefficient value. year: 0.025344
* Intercept: -40.41. When all features are zero, the expected value of the target variable (disease progression) is -40.41 units

1. **Deployment**

* Save the trained model to a file for future use.
* Load the saved model and use it to make predictions on new data.

**Key Highlights**

* Year & Transmission features have strong positive impact on price of a vehicle.
* Fuel feature has strong negative impact on vehicle price.
* Odometer has negative impact on vehicle price means vehicle price decreases with odometer value increases
* Year, Condition, Cylinders, Transmission, Size have positive correlation with Price
* There exist group of 2 features that have positive impact on price of a vehicle and they are Cylinders & Condition, Size & Condition, Drive and Cylinders, Size and Cylinders.
* Odometer, Title, Fuel, Drive has negative correlation with Price
* Maximum available vehicles belong to Gas fuel type category.
* There exist vehicles with 0 value price and duplicate samples which are dataset outliers

**Next steps and Recommendations**

* Year and Odometer plays significant role in Vehicle price. Keep vehicles of recent year and less Odometer value vehicles
* Clean title brings higher value to Vehicle.
* Continue to gather new samples for recent years data, evaluate & tune model to improve price prediction and determines the factors that drives Vehicle Price.