**Business Understanding**

From a business perspective, we are tasked with identifying key drivers for used car prices. In the CRISP-DM overview, we are asked to convert this business framing to a data problem definition. Using a few sentences, reframe the task as a data task with the appropriate technical vocabulary.

The objective is to provide the client(Car Dealership) recommendation on what factors make a car more or less expensive. This in turn will help the client to buy and stock the right inventory for maximizing its sales.

**Data Understanding**

After considering the business understanding, we want to get familiar with our data. Write down some steps that you would take to get to know the dataset and identify any quality issues within. Take time to get to know the dataset and explore what information it contains and how this could be used to inform your business understanding.

For this project, we will use dataset from Kaggle. The original dataset contained information on 3 million used cars. The dataset that will be used for analysis contains information on 426K cars. The dataset is in a .csv file. Price is one of the columns. Other columns include region, year, manufacturer, model, condition, cylinders, fuel, odometer, title\_status, transmission, VIN, drive, size, type, paint\_color and state. We will analyze these columns to determine which one influence the price the most. This will then help in providing a recommendation to stock used cars with the top 5 parameters that influence the car pricing.

Import the relevant libraries and load the input data file. The dataset is provided in .csv format. The dataset has columns for Price, region, manufacturer, model, condition, cylinders, fuel, title\_status, transmission, drive, size, type, paint\_color, state, VIN, year, odometer and id. We will analyze all the available parameters to evaluate what impacts the price of the car the most. Accordingly, we will make the recommendation to prioritize the top 5 factors that influence the pricing of the used cars.

**Data Preparation**

After our initial exploration and fine-tuning of the business understanding, it is time to construct our final dataset prior to modeling. Here, we want to make sure to handle any integrity issues and cleaning, the engineering of new features, any transformations that we believe should happen (scaling, logarithms, normalization, etc.), and general preparation for modeling with sklearn.

Firstly, we will clean the input data file by dropping rows with Null, 0 and duplicate values. Secondly, we will separate the categorical columns from the numerical columns. Plotting the heat map or price versus individual column, we can visualize the inputs that have little to no effect on price. We identify and drop those columns (id, region, manufacturer, model, state, VIN). Finally, in this step, we convert the remaining categorical variables into dummy/indicator variables.

**Modeling**

With your (almost?) final dataset in hand, it is now time to build some models. Here, you should build a number of different regression models with the price as the target. In building your models, you should explore different parameters and be sure to cross-validate your findings.

Firstly, we will separate the data into training and test. x will be our input and y will be our output (the price of the used car). We will allocate 70% of the available data for training. We will set random\_state equal to 0 so that the data is split the same way every time we run our model code.

A Pipeline is a way to chain multiple data processing steps (transformers) and a final estimator (model) together. It simplifies the workflow and ensures that the same transformations are applied to both training and testing data. This transformer creates polynomial features from the original features in our data. degree=2 specifies that we want to generate polynomial features up to the second degree (e.g., x^2, xy, y^2). include\_bias=False means that a constant term (bias) won't be added to the features.

LinearRegression is the final estimator in the pipeline, which is a linear regression model. In summary, our pipeline does the following: **Polynomial Feature Expansion:** It takes the original features in our dataset and creates new features by raising them to different powers (up to degree 2 in this case). This allows the model to capture non-linear relationships between the features and the target variable. **Linear Regression:** It fits a linear regression model to the transformed data, which now includes the polynomial features.

**Evaluation**

With some modeling accomplished, we aim to reflect on what we identify as a high-quality model and what we are able to learn from this. We should review our business objective and explore how well we can provide meaningful insight into drivers of used car prices. Your goal now is to distill your findings and determine whether the earlier phases need revisitation and adjustment or if you have information of value to bring back to your client.

We have created a supervised ML model. The pipeline applies all the transformations defined in the pipeline to the training data (x\_train). The final estimator (model) in the pipeline is trained using the transformed training data (x\_train) and corresponding target values (y\_train). We use predict on the pipeline object, which applies all the pre-processing steps defined in the pipeline to the x\_train data, and then uses the final estimator (the machine learning model) to make predictions on the transformed data. Finally, we calculate the Mean Squared Error (MSE) between the true target values (y\_train) and the predicted values (y\_train\_pred) for our training data. We repeat the same step on our test data.

We use permutation\_importance function is used to calculate the importance of features in our ML Model. We evaluate the performance of our trained model (pipe\_model) on the test data (x\_test, y\_test). We calculate the importance score for each feature based on the average decrease in performance across multiple permutations.

**Deployment**

Now that we've settled on our models and findings, it is time to deliver the information to the client. You should organize your work as a basic report that details your primary findings. Keep in mind that your audience is a group of used car dealers interested in fine-tuning their inventory.

We recommend using the top 5 features to consider when deciding to stock used cars. Auto transmission, Manual transmission, Rear Wheel Drive, 4-wheel drive and Forward Wheel Drive are the top 5 features to consider.