Case Study: Used Car Pricing

Data Understanding *(max. 1 Page)* – Please follow along the file 'CaseStudy_Task1.ipynb'

a) Give a short summary of problems or (logical) inconsistencies in the data like outliers, missing values etc. which should be considered or tackled in order to utilize the data.

It is essential to analyze and tackle all the below problems

Problems with data:

1. Features with only one value (Ex. Channel feature)
2. Rows that have value 'NaN' for all features (Ex. Last row of the dataset)
3. Formatting issues for the values in features (Ex. Brand, List price, and Sales price features)
4. Removing duplicate records
5. Representing categorical and date features as numerical features efficiently (Ex. Brand, Region, Sales date, etc.)
6. Missing values for features
7. Outliers in the data

Logical inconsistencies with data:

1. Removing logical ambiguities i.e., records with Sales date earlier than the first registration date.
2. Analyze the cases for which the Sales price is greater than the listed price. However, such cases need not be tackled as it might be a potential case for vintage cars.

b) Provide a brief overview of the data (descriptive analytics). Please select appropriate visualization methods and visualize the aspects that seem to be most important to you.

***Descriptive Analytics*:** The provided dataset has 11 features of which only two attributes modelyear and mileage are numerical attributes without any formatting. The statistical summary of numerical attributes, possible values, and the frequencies of categorical attributes is discussed in the code.

Visualizations:

1. Visualization of categorical attributes: In order to have a clear understanding of the possible values and the frequencies of categorical attributes i.e., brand, model, fuel, region features, we visualize it through bar charts each with a distinguishable color.

2. Visualization of trends in Sales and List Price: It is very important to understand the trends in the sales price and list price from 2017 to 2019, so we visualize it through line charts. We aggregate the sum, mean, standard deviation of the sales price and list price over months and plot it against sales date from 2017 to 2019.

3. Visualization of Correlations among features: Correlation not only identifies the association or statistical dependence among features but also is one of the techniques to identify outliers in the data. The points that are away from the dense regions can be considered as outliers. So, we plot the correlation of each attribute with every other attribute.

Model Development *(1-2 Pages)* – Please follow along the file 'CaseStudy_Task1.ipynb' for data preprocessing, and 'CaseStudy_Task2.ipynb' for remaining questions.

a) Do the data preprocessing and find solutions for possible data problems that you have found in task 1.

Data Preprocessing:

1. Features with only one value (Ex. Channel feature) – Remove the attribute Channel

2. Rows that have value 'NaN' for all features (Ex. Last row of the dataset) – Remove such rows.

3. Formatting issues for the values in features (Ex. Brand, List price, and Sales price features) – Replace VOLKSWAGEN with Volkswagen for the Brand attribute and reformatted the List price and Sale price features by removing $ and commas.

4. Removing logical ambiguities - Records with Sales date earlier than the first registration date are removed using methods in pandas dataframe.

5. Removing duplicate records – Remove 1 duplicate row found using methods in pandas dataframe.

6. Missing values for features - Missing values are identified for first_registration, region features. As replacing missing values for first_registration with a frequent value or computing value based on neighborhood samples does not make sense logically, we drop such a small number of rows. For handling missing values for region feature, we can impute them using SimpleImputer of sklearn with the most frequent strategy.

7. Representing categorical and date features (Ex. Brand, Region, Sales date, etc.) – For each date attribute we create a column for day, month, year. Also, dayofweek, quarter of the year information is also added as these attribute values affect the sales price. The autocorrelation plot of the sales price with the previous 10 lags is observed and the previous day's sales is found to have an impact on the next day. So, the first lag of sales price is added as an additional attribute.

   For categorical attributes, the one-hot encoding increases the dimensionality and label encoding induces a sense of order based on assigned numerical value to each category. So, we use Target Encoder as an alternative. Here the value of the attribute is replaced with the probability of the target given a particular categorical value and is proven to give a better result.

8. Outliers in the data: Outliers are detected based on z-scores which encodes the deviation of an attribute from its mean. The z-score > 3 is considered an outlier and is removed.

9. Feature Scaling: The features are scaled before it is fed to the model. We used different scaling approaches like MinMaxScaler, StandardScaler, and RobustScaler of sklearn library. Based on the performance of the model the scaling mechanism is tuned. RobustScaler is found to be the best for our dataset.

b) Train a model that predicts the sales prices in April 2019. Use only the data of sales up to and including March 2019 for the model training. You should not implement more than one prediction model. Which algorithm did you choose and why?

Model chosen: Long Short-Term Memory (LSTM)
Reason: The dataset given has a huge number of rows with sales information from 2017 to 2019. For predicting the sales in April 2019, the chosen model should not forget the past data seen in 2017 or 2018 i.e., it should learn long-term dependencies. The 3 gates in LSTM encodes these details and make a prediction of the sales price in the future time step based on the learned information from the past. So, LSTM **might** help in providing better predictions than traditional algorithms.

c) Which other algorithms could be used alternatively? Please name up to 3 other approaches and briefly discuss their pros / cons.

Alternative Models to LSTM:

MLP: If there is a meaningful correlation between the inputs and outputs, the Multilayer Perceptrons (MLP) can perform well. Like LSTMs, it can handle noise, approximate nonlinear functions, and accept multivariate inputs. However, the optimal function that maps inputs and outputs have to be fixed.

CNN: CNN can be used to extract features of time series data and has all the pros of MLP. The main limitation is that it cannot capture the temporal dependencies.

Models that might not be used:

ARIMA: Even though there are cases (like stock value forecasting) where ARIMA performed better than the deep learning models. In the case of prediction of linear time series and if there is less change in the overall trend, ARIMA performs the best. However, as ARIMA is a univariate model, it cannot exploit the explanatory variables and long-term dependencies. So, the impact of the input variables for our dataset like the model, fuel, region, brand, etc. on sales price cannot be captured.

d) Provide and explain appropriate model performance evaluation metrics that could be presented to the business stakeholders to help them understand the model.

Mean Absolute Error (MAE): It is the average of the sum of absolute differences between the expected and the predicted forecasted values*

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Root Mean Squared Error (RMSE): It is the root of the Mean Squared Error (MSE), where MSE is the average of the sum of squared differences between the expected and the predicted forecasted values*

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}$$

* Image Reference - Wikipedia

e) During the model development you might have wanted to have more features available. If you could ask the business department for some additional parameters, which ones would you like to get and why?

The sales price of the used car is characterized based on many other parameters apart from the model, brand, mileage, etc. The exterior and interior condition of the car along with any added customizations can greatly affect the sales price. Optionally, features that describe the performance of the parts can be added to the dataset.