Summary of the lead Score prediction project:

The requirement is very clear that we need to identify the correct features to improve the conversion rate. The data provided has many features including the customer activities in the websites and the information about the path which led them to create a lead. Now, let's quickly jump into the data and check what are the patterns available to classify the lead if they will get converted or not.

Data Overview:

The data contains 9240 records or 9240 unique leads and the details about the lead. There are 36 columns which explains the lead. There were more categorical columns than numerical columns which is challenging as no categorical columns should be there to train the model or form a model. The real problem is identifying if the categories are nominal or ordinal and encode them accordingly.

Skewed Data:

There could be columns where most of the shares will be filled by the same value (almost 90%). These columns are useless as they contribute less in predicting the pattern.

Cleaning the data:

- There are values "Select" which are the default values from drop down then replaced with Nan.
- Categorical variables have missing values, in order to impute categorical values we should go with mode, but if the values are evenly distributed between the categories it is better to create a new category "others".
- Drop duplicate records if any
- Drop Nan for columns if it is lesser than 1%

Outliers:

We had 3 numerical columns which had outliers and that it is addressed by the usual way with a formula

- percentile (10%) - (1.5 * IQR) > VALUES > percentile (80%) + (1.5 * IQR)

(10% and 80% are chosen out of purpose as we focused on higher outliers)

Handled Junk Values:

There were 2 features "closed by Horizonn" and "lost to EINS" which means we lost the lead to competitor but the "converted column" value has 1(converted). So the records with this values are dropped.

Encoding and Scaling:

The categorical values are encoded (one hot encoding) and scaled using min max scaler which is the proper approach as the data will be scaled between 0-1.

Feature Selection:

Once we do One Hot Encoding there will be numerous features in the data. We need to find the right way in order to choose the minimum and best features that would contribute to the model.

We know the p-value from statsmodel, RFE method and VIF helps efficiently in choosing the right features for the logistic regression model.

Metrics:

The right metric we need to choose is RECALL/SENSITIVITY/TruePositiveRate where we get around 86% in both training and testing data having the default cut off as 0.5 which is the expected score.

Conclusion: This model helps the intern or sales team to pick the hot leads, check their behaviors and enrich the data with other information like profiling the leads, valuing them, checking their need from their educational or working background and provide them the right offers or guiding them the through the right path to convert them.

This scores not just helps the sales team to convert the leads but also analyze the cold lead on why they show low interest to us? what can be improved from our sides to convert the cold leads? Which are the leads sold to competitors and what is the reason? Are they giving better offers? Finally, we need to get the feedback on the converted leads and try to give the best work to retain them.

This sheet is a spoon feed for the sales team, but the scores may not work in the same way as it is, We need to request the sales team to punch in the feedback about the leads if they get converted or not. We need to focus mainly on the FALSEPOSITIVES and improve the model as the model won't be accurate always. A proper feedback loop helps to improve the model better.