Summary

1. Approach

1. Importing Data
2. Inspecting the Data frame
3. Data Preparation (Encoding Categorical Variables, Handling Null Values)
4. EDA (univariate analysis, outlier detection, checking data imbalance)
5. Dummy Variable Creation
6. Test-Train Split
7. Feature Scaling
8. Looking at Correlations
9. Model Building (Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-vales)
10. Build final model
11. Model evaluation with different metrics Sensitivity, Specificity

2. Analysis Results:

1. The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.

2. Here, the logistic regression model is used to predict the probability of conversion of a customer. Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert).

3. Our final Logistic Regression Model is built with 14 features.
   Features used in final model are
   ['Lead Origin_Lead Add Form', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'Tags_Busy', 'Tags_Closed by Horizzon','Tags_Lost to EINS', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_in touch with EINS', 'Tags_switched off', 'Lead Quality_Not Confirm', 'Lead Quality_Worst', 'Last Notable Activity_Modified']

4. The top three categorical/dummy variables in the final model are “Tags_Closed by Horizzon”,“Tags_Lost to EINS”, “Tags_Will revert after reading the email” with respect to the absolute value of their coefficient factors.

5. “Tags_Closed by Horizzon”,“Tags_Lost to EINS”, “Tags_Will revert after reading the email” are obtained by encoding the categorical variable ‘Tags’.

6. The final model has Sensitivity of 0.881, this means the model is able to predict 88% customers out of all the converted customers, (Positive conversion) correctly.

7. The final model has Precision of 0.848, this means 84.8% of predicted hot leads are True Hot Leads.

8. We have also built an reusable code block which will predict Convert value and Lead Score given training, test data and a cut-off. Different cutoffs can be used depending on the use-cases (for e.g. when high sensitivity is required, when model have optimum precision score etc.)

3. Insights:

- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' Lead Origins and also increasing the number of leads from 'Lead Add Form'

- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'

- Websites can be made more appealing so as to increase the time of the Users on websites

- We should focus on increasing the conversion rate of those having last activity as “Email Opened” by making a call to those leads and also try to increase the count of the ones having last activity as “SMS sent”.

- To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads

- We also observed that there are multiple columns which contain data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis