Summary Report

1. To Identify hot leads and increase conversion, we've built a predictive model using Logistic Regression.
2. Originally data had 9,240 rows and 37 columns, of which 9,240 rows and 12 columns were retained after thoroughly analysing each feature.
3. Data was gathered, studied, cleaned, imputed categorical and numerical features as required, Outliers were taken care of.
4. Checked there was no correlation between variables with help of correlation matrix.
5. For data preparation for Modelling purpose, Dummy variables were created, Train-test data split, scaling performed on Train set.
6. Top 20 Relevant features were picked out using RFE and various models were built. Some features got dropped due to high P-value (>0.5) others due to high VIF (>5).
7. Final model with significant P-value a VIF were used for prediction and further evaluation.
8. Various Evaluation Metrics were deployed to understand model performance such as Confusion metrics, Accuracy, Precision, Recall, Specificity, False Positive Rate (FPR), Area Under the ROC Curve (AUC-ROC). After carefully examining performance based on Accuracy, Sensitivity and Specificity v/s Precision- Recall Trade off, Accuracy, Sensitivity and Specificity was chosen as it was better aligned with our goal of identifying the maximum number of leads that can convert. It gives a slightly higher recall (89.17%), meaning we'll identify more potential customers.
9. With a cut- off of 0.33 we achieved Accuracy (89.79%), Sensitivity (90.22%), Specificity (89.50%), Precision (84.87%), Recall (90.22%) and AUC-ROC (96%).
10. Twelve features were selected as the most significant in predicting the conversion, features having positive impact on conversion probability in decreasing order of important basis model are 1. 'Tags_Closed by Horizzon', 2. 'Tags_Will revert after reading the email', 3. 'Tags_others'. But features having positive impact on conversion probability in decreasing order of impact basis business intuition are: 1. 'Tags_Closed by Horizzon', 2. 'Tags_Will revert after reading the email', 3. 'Last Activity_SMS Sent'
11. More Focus on Last Activity Email opened, Total time spent on website, Lead source Olark chat have great potential and with increased and interactive engagement we can have more conversions.
12. Occupation Others, Specialization Travel and Tourism and Specialization International Business have negative impact on conversions, maybe looking into this profile in the end or putting in least effort is recommended here.

Learnings

1. As part of learnings, Data cleaning and handling irregularities, EDA, Data visualization, Data Preparation for model building and finally building a model was a good refresher for all the concepts and was also very interesting.
2. Alongside, correlating how industry problem can be converted into a Logistic regression problem was fascinating and the exposure of experiencing felt like a real time scenario, found it very engaging.
3. It was also a good brush up on concept learnt during Logistic Regression, few concepts which were not very clear in the first time looked much easier with this industry application case study.
4. Finally, collaboration and Team efforts.