SUMMARY

The procedure adopted to build the logistic regression model is outlined below-

- Importing Data and necessary libraries, Understanding the data using info, shape, describe functions.
- Data Preparation – finding and handling Null/missing values and encoding yes/no values to 1/0, checked value counts of categorical variables, imputing with mode we can replace Nan with Unknown; dropped unnecessary columns/variables
- EDA- Univariate Analysis of categorical variables , some categorical variables with less values are neglected or combined ;Checked Outliers and fixed them using capping and flooring;Data Imbalance in target variable; bivariate analysis between numeric variables and target variable.
- Dummy Variable creation
- Train-test split
- Feature scaling
- Looked at Correlation using heatmap and correlation matrix and dropped variables with high correlation
- Built First Model using RFE and later finetuned by P values and VIF . Rerun model 4 times to reach the final model.
- Found Optimum cut off using ROC and specificity and sensitivity
- Final Model
- Model evaluation
- Conclusion and Insights

Learning-

- API and landing pade submission has lower conversion rate but has considerable counts. However, Lead add form has good conversion rate.
- Google and direct traffic has less conversion rate but reference and weligkit website has good conversion rate
- Country- India has highest count of leads.
- Finance and Marketing specialization leads counts are significant.
- Unemployeed and unknown categroy has more counts but less conversion rate, at other hand working professional CR is high
- Better career prospect is the main reason of most of the applicants and CR is high.
- Even though employee has marked low relevance but lead quality has highest conversion rate
- City- City like Mumbai, Thane has more counts and CR
- The conversion rate-37.86%
- Logistic Regression Model predicts the probability of lead conversion. Cut off value of Probability is used to classify the target variable, here in this case we have cut off value 0.35.
- Here the logistic regression model is used to predict the probabilty of conversion of a customer.
- Optimum cut off is choosen to be 0.35. It means any lead which has greater than 35% probability we can consider that as "Hot lead" and probability less than 35% can be considered as "Cold Lead".
- Final Model is build with Total 12 Features.

- Fetaures Used in Models are ['Do Not Email', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'What is your current occupation_Unknown', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_switched off', 'Lead Quality_Worst', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversat
- Top Three Features which has high Positive coefficient are a) Tags_Lost to EINS --7.47 b) Tags_Closed by Horizzon--6.82 c) Lead Source_Welingak Website--4.12
- Final Model has 86% Sensitivity. We can predict the 86% customer correctly which are truely positive.
- Final Model has 89% precision . Hence 89% of Hot leads are True Hot leads among the predicted Hot leads. ion']