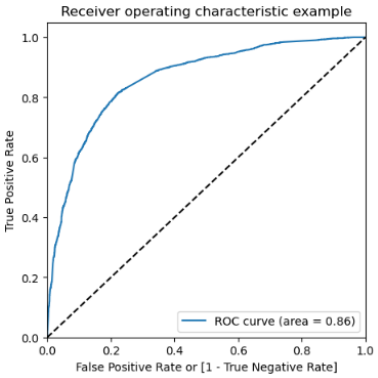
Summary Report on Lead Score Case Study

Problem Statement

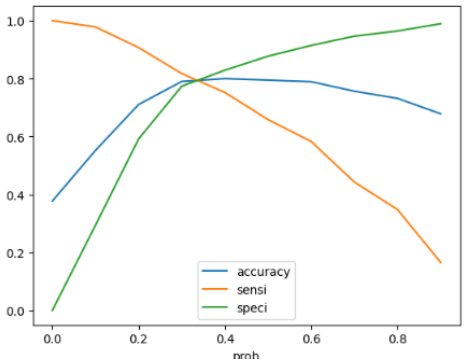
1. To identify the hot leads, so that sales team can focus on the hot leads.
2. Company requires to build a model where we need to assign a lead score to each of the lead so that customer which having high score have high conversion chances. The score should be between 0 and 100.
3. To get to know the which columns are most important and which is not important.

Steps to be Followed.

1. ***Dataset Reading***
2. ***Data Understanding***: This involves examining the structure of the data frame, the data types assigned to each column, the count of null values in each column, the percentage of null values, and the distinct values present within each column.
3. ***Exploratory data analysis***: These tasks encompass discarding columns with a null value percentage exceeding 20 %, performing null value imputation, and conducting both univariate and bivariate analyses.
4. ***Data Preparation***: This involves generating dummy variables and converting columns containing "Yes" and "No" values into corresponding 1s and 0s.
5. ***Splitting the data in Test and Train***: Splitting the dataset in train test in the ratio of 70 and 30
6. ***Feature Scaling***: For numerical columns, an assessment for potential outliers was conducted, revealing no significant outliers. Consequently, the StandardScaler was employed to scale the numerical attributes.
7. ***Model Building***: Constructed a Logistic Regression model utilizing a set of 15 features. All of these features exhibited p-values below 0.05, indicating statistical significance. Additionally, their Variance Inflation Factor (VIF) values remained below 5, ensuring minimal multicollinearity.
8. ***AUR-ROC Curve****:* Area Under the Curve is 0.86, which is good.



9. ***Accuracy sensitivity and specificity trade off value****:* The cut off value is around 0.333.



10. ***Confusion Matrix****:* Train Data
    Accuracy - 79.27%
    Specificity - 80.61%
    Sensitivity - 78.45%
11. ***Predication on Test Set***
    Accuracy: 78.89%
    Sensitivity: 83.25%
    Specificity: 76.20%

Conclusion

Hot Leads Selection

Segmenting leads according to their lead scores provides valuable insights into their likelihood of conversion. Notably, leads with scores surpassing distinct thresholds such as 90, 80, 70, 60 and 50 cluster prominently, signifying different levels of conversion potential. This understanding enables the implementation of focused engagement tactics aligned with individual lead scores, thus enhancing conversion rates by strategically enhancing interactions.

Lead Count for Different Lead Score

1. Number of rows with Lead_score > 90: 228 counts
2. Number of rows with Lead_score > 80: 429 counts
3. Number of rows with Lead_score > 70: 578 counts
4. Number of rows with Lead_score > 60: 813 counts
5. Number of rows with Lead_score > 50: 957 counts

Feature Importance

These listed features are the most important feature.

1. Lead Source_Welingak Website: 5.598763
2. Lead Source_Reference : 4.148916
3. Last Notable Activity_Had a Phone Conversation: 3.070394