Summary

We had to build a logistic regression model for identification of Leads that are most likely to convert into paying customers

Objective:

- Categorize the leads and assign a lead score to each of the leads such that the customers with higher lead score will become hot leads.
- Target lead conversion rate to be 80%.
- For the above, build a logistic regression model

In order to build a good model, we went ahead with the following approach:

- Importing Data
- Dataframe Inspections
- Data Preparation (Encoding Categorical Variables, Handling Null Values)
- EDA (univariate analysis, outlier detection, checking data imbalance)
- Dummy Variable Creation
- Test-Train Split
- Feature Scaling
- Looking at Correlations
- Model Building (Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-vales)
- Build final model
- Model evaluation with different metrics Sensitivity, Specificity.

Conclusion

1. 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'

2.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'

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

4. 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

5. 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

6. 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