SUMMARY REPORT

DATA GATHERING, CLEANSING, MANUAL ANALYSIS

- The key requirement for the customer was to identify hot leads as a focused marketing and conversion strategy, used by the learning portal.
- We choose logical regression model, since we want to predict a categorical variable (binary classification).
- We start with importing the gathered input data to analysis.
- Here while inspecting the imported data, it could be observed that there were a huge number of categorical variables, many of them were redundant like “Last Activity” and “Last Notable Activity” or “Magazine” that had only one value. Such features could be removed. Also, we could remove unique identifiers like “Prospect ID”
- Remove any duplicates and impute missing values.
- Now understand the data by visualization, using EDA – Univariate and bivariate analysis for categorical and numerical variables. This helps understand spread, outliers, need for treating them, relationship with output variable, finally a heatmap that quantifies correlation. Here a positive correlation could be observed between “Total Time Spent on Website” and “Conversion”. This could be a factor to investigate.
- Dummifying categorical variables, so they can be used for model building.
- Repeat EDA step for categorical variables, now check correlation amongst specific values. Also repeat univariate, bivariate analysis. For example, “Tags_Closed by Horizzon” has all conversions.

MODEL BUILDING, TRAINING, TESTING

- Now, before model building, splitting training and test data. Also scaling numerical features using standard scaler.
- Building initial model with around 50 features and an assumed cut-off probability of 0.5, then using RFE to pick top 20 features.
- Top 20 had features like “Tags_Busy”, “Total Time Spent On Website”, etc. This also eliminated low factoring features, like “Page Views per visit”, “Total Visits”
- Check VIF and p-value of features. VIF was < 5, around 2.6. While p-value was as high as 1 for around 5 features.
- Dropping one feature at a time, based on VIF and p-value, and re-building model, to check accuracy.
- After around 8 iterations, the p-values came to < .05 and VIF was maintained, while accuracy was also 87.5%
- Calculating other metrics like sensitivity (we are interested in maximizing this), specificity, positive predictive value, etc.
- Plotting ROC and area under curve. The latter was about .936, which is excellent.
- Identifying optimal cut-off that balances metrics, which comes to .37.
- Using precision, recall as alternative, here cut-off comes to .48.
- Calculating F score, that gives accuracy of 83%
- Predict using test data, accuracy is still 86.39%
- Generate classification report
- Check ROC and AUC. Here also AUC is .936.
- Calculate lead score using probability predicted for train and test data.
- Check for null values, which are none.

RECOMMENDATION

- Take model coefficients excluding intercept to determine importance of a feature
- Sorting and collect top 3 features, and positive coefficient features, negative coefficient features.

	index	0
2	Tags_Closed by Horizzon	100.00
3	Tags_Lost to EINS	92.76
5	Tags_Will revert after reading the email	41.15

After trying several models, we finally chose a model with the following characteristics:

All variables have p-value < `0.05` . All the features have very low VIF values, meaning, there is `hardly any muliticollinearity` among the features. This is also evident from the heat map. The overall accuracy of `0.8776` at a probability threshold of 0.37 on the test dataset is also very acceptable.

The conversion probability of a lead increases with increase in values of the following features in descending order:

Features with Positive Coefficient Values
Tags_Closed by Horizzon
Tags_Lost to EINS
Tags_Will revert after reading the email
Lead Origin_Lead Add Form
Tags_Busy
Tags_in touch with EINS
Last Activity_SMS Sent
Last Activity_Email Opened
Last Activity_Others
Total Time Spent on Website

The conversion probability of a lead increases with decrease in values of the following features in descending order:

Features with Negative Coefficient Values
Tags_Ringing
Last Activity_Email Bounced
Tags_invalid number