**Summary**

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

1. EDA:

Looking at the dataset we observed that following things need to be done:

* Encoding categorical variables
* Missing value handling
* Convert the Select into the NaN
* Dropping columns having more than 40% null values

**Visualisation:**

**OBSERVATION: 1**

* API and Landing Page Submission has less conversion rate but counts of the leads from them are considerable
* The count of leads from the Lead Add Form is pretty low but the conversion rate is very high
* Lead Import has very less count as well as conversion rate and hence can be ignored

**OBSERVATION: 2**

* The count of leads from the Google and Direct Traffic is maximum
* The conversion rate of the leads from Reference and Welingak Website is maximum

**OBSERVATION: 3**

* The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information
* Users spending more time on the website are more likely to get converted

**OBSERVATION: 4**

* The count of last activity as "Email Opened" is max
* The conversion rate of SMS sent as last activity is maximum

**OBSERVATION: 5**

* Looking at above plot, no particular inference can be made for Specialization
* Looking at above plot, we can say that working professionals have high conversion rate
* Number of Unemployed leads are more than any other category

**OBSERVATION: 6**

* We can clearly see that most values in the all the columns are 0 i.e. "No" and hence no inference can be made using these columns

**OBSERVATION: 7**

* 'Will revert after reading the email' and 'Closed by Horizzon' have high conversion rate

2. Train-Test split & Scaling :

· The split was done at 80% and 20% for train and test data respectively.

· We will do standard scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

3. Model Building

· RFE was used for feature selection.

· Then RFE was done to attain the top 15 relevant variables.

· Later the rest of the variables were removed manually depending on the p-value.

· A confusion matrix was created, and overall accuracy was checked which came out to be 92%.

4. Model Evaluation

After Model evaluation we will get on Training Data

Accuracy 92%

Sensitivity 85%

Specificity 96%.

5.Prediction

* The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly.
* Our final Logistic Regression Model is built with 14 features.
* Features used in final model are ['Tags\_Already a student’, 'Lead Origin\_Lead Add Form', 'Lead Source\_Welingak Website', 'Last Activity\_SMS Sent', 'Tags\_Busy', 'Tags\_Closed by Horizzon', 'Tags\_Lost to EINS', 'Tags\_Ringing', 'Tags\_Will revert after reading the email', 'Tags\_switched off', 'Lead Quality\_Not Sure', 'Lead Quality\_Worst', 'Last Notable Activity\_Modified', 'Last Notable Activity\_Olark Chat Conversation']
* The top three categorical/dummy variables in the final model are ‘Tags\_Lost to EINS’, ‘Tags\_Closed by Horizzon’, ‘Lead Quality\_Worst’ with respect to the absolute value of their coefficient factors.
* ‘Tags\_Lost to EINS’, ‘Tags\_Closed by Horizzon’ are obtained by encoding original categorical variable ‘Tags’. ‘Lead Quality\_Worst’ is obtained by encoding the categorical variable ‘Lead Quality’.
* Tags\_Lost to EINS (Coefficient factor = 9.2313)
* Tags\_Closed by Horizzon (Coefficient factor = 8.8325)
* Lead Quality\_Worst (Coefficient factor = -3.9397)The final model has Sensitivity of 0.928, this means the model is able to predict 92% customers out of all the converted customers, (Positive conversion) correctly.

6.Percision -

The final model has Precision of 0.93, this means 93% of predicted hot leads are True Hot Leads.

7.Conclusion -

1. Top variable contributing to conversion: -

* lead source:

-Total Time Spent on Website

* Lead Origin:

-Lead Add Form

* Last Activity:

-Olark chat conversation

2. The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.

3. Observational Conculsions:

* 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'
* 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'
* Websites can be made more appealing so as to increase the time of the Users on websites
* 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
* 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
* We also observed that there are multiple columns which contains data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis.