Objective

- Problem Statement
- To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.

Methodology

- Logistic Regression model to predict the Lead Conversion probabilities for each lead.
- Decide on a probality threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.
- Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

CheckPoints to approach the objectives of the projects.

- Understanding the Data Set & Data Preparation
- Applying RFE to identify the best performing subset of features for building the model.
    - Building the model with features selected by RFE.
    - Eliminate all features with high p-values and VIF values and finalize the model
    - Use the model for prediction on the test dataset and perform model evaluation for the test set.
- Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data
- Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.

Evaluating the model on Train Dataset

- The train dataset final model is used to make predictions for the test dataset
- The train data set was scaled using the scaler.transform function.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of **0.33**, the leads from the test dataset were predicted if they will convert or not.

Lead Score Calculation

- Concatenate train and test dataset to get the entire list of leads available.
- The *Conversion Probability * 100* to obtain the Lead Score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted and vice versa, **0.33** is our final Probability threshold for deciding if a lead will convert or not.

Determining Feature Importance

- 16 features have been used by our model to successfully predict if a lead will get converted or not.
- The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
- Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted. Similarly, features with high negative beta values contribute the least.

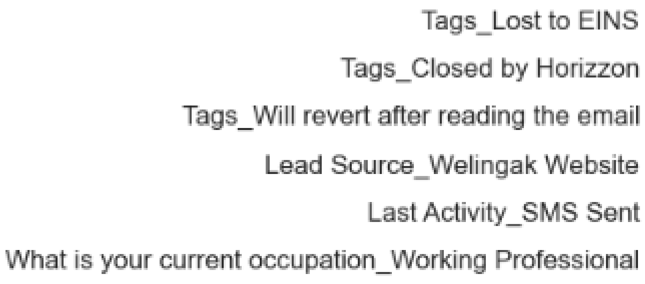
Conclusion

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 multicollinearity among the features. This is also evident from the heat map.
- The overall accuracy of **0.9056** at a probability threshold of **0.33** on the test dataset is also very acceptable..

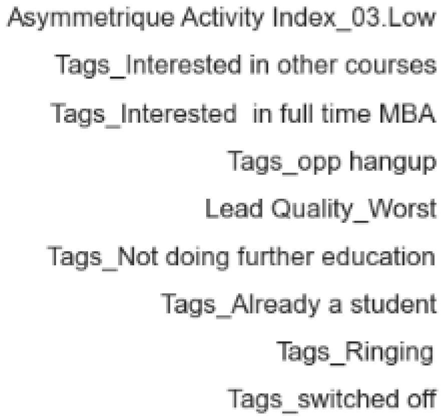
Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

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



Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email
Lead Source_Welingak Website
Last Activity_SMS Sent
What is your current occupation_Working Professional

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



Asymmetrique Activity Index_03.Low

Tags_Interested in other courses

Tags_Interested in full time MBA

Tags_opp hangup

Lead Quality_Worst

Tags_Not doing further education

Tags_Already a student

Tags_Ringing

Tags_switched off