Summary report

On commencing the report, the initial approach was to understand the business, and objective of the project, which was to build a logistic regression model to assign a lead score between 0 and 100, to each of the leads, which can be used by the company to target potential leads. A higher score, would mean it's a Hot Lead, which indicates that there are higher chances of being converted to a customer, and vice versa.

Once the data was understood, we first started by cleaning the data, and then gained a comprehensive understanding of the data provided, through EDA methods. After handling outliers and obtaining dummy variables, we proceeded to build and train the model. We started by splitting the data into train and test and also performed feature scaling for all the numerical variables, where a value between 0 and 1 is assigned to each value, for ease of comparison.

Out of the 80+ variables, we first used RFE to select 15 variables. StatsModel was then used to assess the model. Variables with high p-values were dropped. We then used VIF to determine which variables have to be further dropped. We first dropped the variable with the highest VIF, and then assessed the model again. All the factors now had a low p-value, as well as VIF factor of below 5.

With the final model of 11 variables, we used the confusion matrix to assess the model along with ROC curve and PR trade off curve.

We further identified the optimal cut off point to be 0.2 probability. Which means, anything above a 0.2 probability, will be considered as a hot lead (high chances of being coverted to a customer), and anything below, will be a cold lead. We checked the confusion matrix measures again, to ensure there were no major changes.

Since, there were not much changes in the metrics, we finalized on the model, and proceeded to use it on the test set, with 0.2 as the probability. On doing so, we observed that the accuracy - 92%, sensitivity – 88%, specificity – 94%.

We observed that "Tags_Closed by Horizzon", "Tags_Lost to EINS" and "Tags_Will revert after readin g the email" were the most influential variables.

X Educations' marketing team must focus on those leads with higher lead scores, as the rate of conve rsion for such leads is much higher.

Some of our key learnings from the case study has been that high accuracy is not the only measure of a good model. Specificity and sensitivity must also be determined. Also, that the ROC curve and PR trade off curve, could suggest 2 different optimal probabilities, and both must be tested to obtain the best model. Further, while it is normal for the test set to not perform as well the train set, the metrics still have to be high for the model to be successful.