# Summary Report

Following is a documentary about how our team solved the case study and the learnings we derived from it. The model is built for an education company to improve the conversion rate of non-paying customers into paying ones. The company wanted us to assign a score to each lead such that the customers with a high lead score have a higher conversion chance and a low lead score have a lower conversion chance. The X education company provided dataset with variables. There were around 37 variables and 9000 data points in the dataset. The target variable is "converted" which represents the lead to be successfully converted (1) or not (0) into a paying customer. The Goal we had set for ourselves was to develop solutions that would result in business growth and cost savings for our client.

We explain our journey in following points:

1. The data provided by the company is about leads and the variables associated with them. We started with reading variables and their description. At that moment, we got a vague idea about which variables would be crucial to solve the problem.
2. We imported the data and necessary libraries.
3. We checked and calculated the null values in each column and dropped the columns with more than 20% null values that seemed less important for the analysis.
4. We performed the EDA. We did Univariate and bivariate analysis on relevant columns. We found interesting insights, like, the majority of hot leads came from Google or the time spent on the company's website was significantly higher for the people who converted into paying customers than those who didn't. We also found out that the customer segment of the company mainly comprised people who identified as unemployed and lived in Mumbai. My personal favourite is that almost 90% of people who classified with the Tag "Will revert after reading the mail" got converted into paying customers. We learnt that data speaks and can tell its own stories.
5. We checked the unique values in column and created dummy variables for them.
6. We divided the variable in Train and test dataset and scaled it.
7. We performed RFE-analysis for automatic feature selection. We also manually selected features based on P-value and VIF score.
8. We used Logistic Regression to build the models. We finalized our 3rd model for predictions.
9. We drew ROC curve and found the optimal cutoff probability/score for optimum results.
10. We calculated precision, accuracy, sensitivity and specificity for different probability cutoffs.
11. We predicted results using test data with the cutoff score set to 70 and 30.

**Conclusion & Findings –**

Since the CEO has requested that the lead conversion rate be 80%, we recommend that the sales team target leads with a score of 70 and above. We expect the team to close almost 80% of such

leads. But by this strategy, the team will only be able to target 50% of the hot-leads available in the funnel. We recommend an alternative strategy to target people with a score above 30. This way, the sales team will be able to close almost 65% of cases and also target 90% of the hot leads present in the funnel. Both these strategies are useful in different scenarios. We also tried to predict the monetary impact of our model. Since the conversion rate is 2x using our model, one salesperson can close the same number of cases as two people. So, if the salary of one salesperson is 30,000 and the company has a team of 100 people. Then each month, the company can save around 15 Lakhs in salary by using our model, and we believe it is a good impact.