Summary:

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
- We have performed Data cleaning, EDA and Data Preparation on provided data set.
- We understood that this is a Logistic Regression problem as the target variable "Converted" is a Binary/categorical output.
- We then performed Logistic Regression Modelling and come up with Probability of Hot Leads.
- We found out the lead score of all the Leads.
- **Few categorical feature have very less value of a particular kind compared to other value. Which we then grouped them appropriately** for example Country had India in large volume, other country has very less % we grouped them all as "Other"
- **We first created dummy variable for categorical features**
- **Created X and y Independent and Target variable datframes**
- **Split the data into train and test in 70:30 keeping random_state and stratify**
- **Performed Scaling of numerical feature using StandardScaling**
- **Created 1st Logistic Regression Model using all the features**
- After 1st model, we optimised the model. We used **Recursive Feature Elimination (RFE)** to select top 15 features.

- Then we also performed **Variance Inflation Factor (VIF)** on these selected 15 features and found VIF score is in-acceptable range.
- We chose optimal cut-off point as 0.35 for our final model.
- Which give us approx. 80% Recall and 70% precision