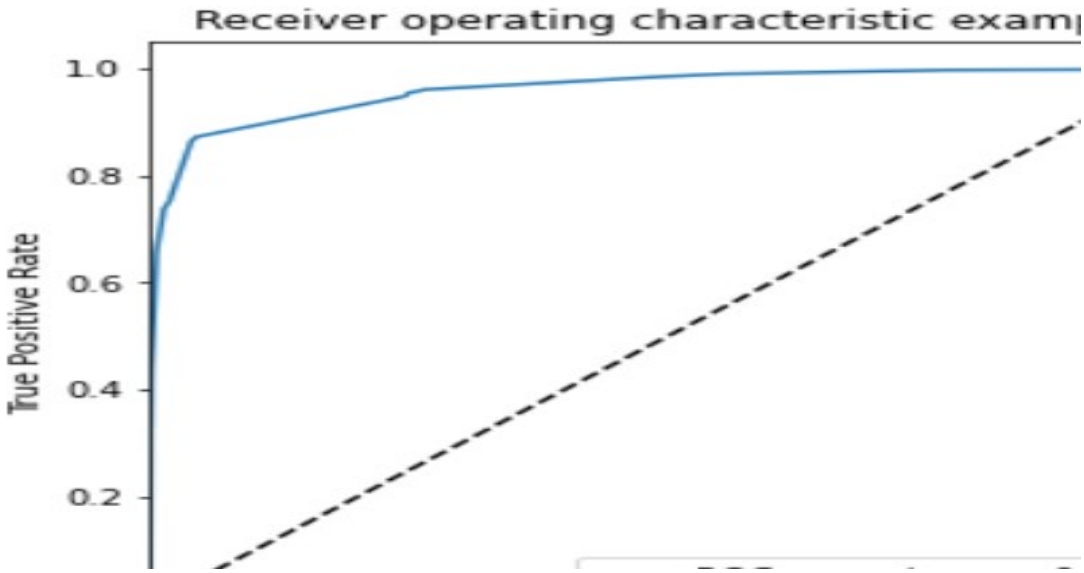
Summary Report

The problem statement in hand is to identify the leads which have high conversion rate above 80% so as to save time and efforts of the company. The given dataset of previous converted customers has 9240 entries with 36 variables. The target variable is 'Converted'.

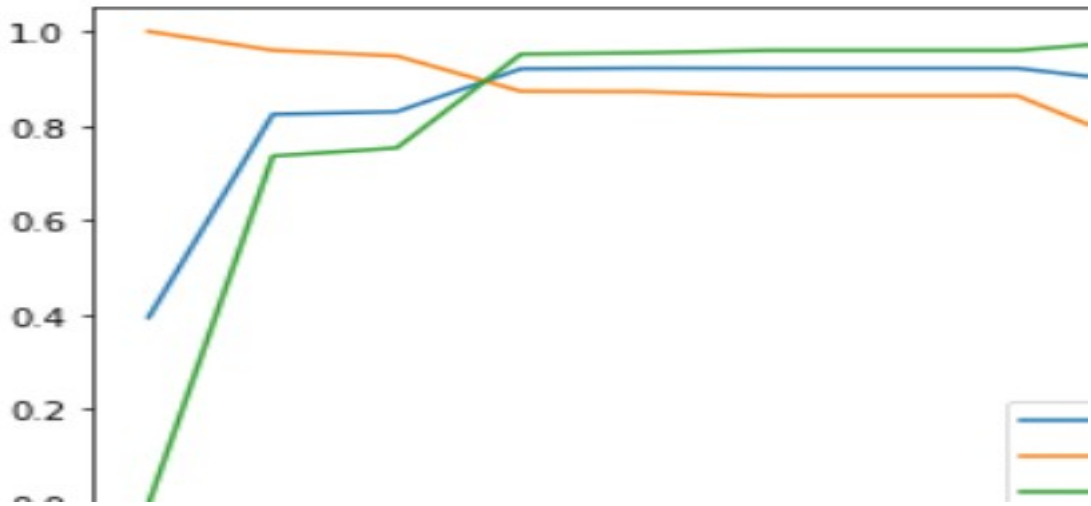
Understanding the data

1. Handling Select Values: The cell values with 'Select' are assumed to have been result of not selecting the value by the customer are taken as Null.
2. High Null Values: Variables with very high 40% to 50% Null values are removed from the dataset. These are 'high did you hear about X Education', 'Lead Quality', 'Lead Profile' and variables with tag Asymmetrique in their name.
3. Repeated Variables: The repeated variables are also removed. These are 'Prospect ID' and 'Last Activity'.
4. Low Frequency with Multiple Categorical Values: Some categorical variables have many multiple values. Therefore, values with low frequency are clubbed together into 'other' category so as to reduce the number of variables to work with. These variables are 'Tags', 'Lead Source' and 'Last Notable Activity'.
5. Mostly Same Value: The variables with almost all as same values are also removed. These are Country, Newspaper Article, Newspaper, Search, Magazine, X Education Forums, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque.
6. Binary Variables: The variables with binary Yes and No are converted to 1 and 0. These are Do Not Email, Do Not Call and A free copy of Mastering The Interview.
7. Categorical Variable with Multiple Values: The categorical variables with multiple values are converted to dummy variables. These are Lead Origin, Lead Source, Last Notable Activity, Specialization and Tags.
8. Removing Outliers: There are outliers identified in the numerical variables namely TotalVisits and Page Views Per Visit and are removed accordingly.

After cleaning the data more than 99% of the original entries are intact. The data is divided into train and test dataset with 30 % as test data. The numerical variables are scaled using standard scaler. At this movement we have 65 variables. Therefore, we use RFE method to identify the important 15 variables affecting the target variable 'Converted'. The RFE model gives top 15 factors with 92% model accuracy. The factors with high p values are removed from the model one by one. These are 'Number not provided', 'wrong number given' and 'invalid number'. The remaining variables have p value lower than .05 and VIFs are all below 5 as desired. The model shows 92% accuracy, sensitivity at 86% and specificity at 96%. The ROC curve shows trade-off between Sensitivity and Specificity as desired as can be seen below.



The optimal cut-off point is identified at .28 probability as can be seen below:



The factors identified are closed to Horizzon, lost to EINS or will revert after reading the email, SMS sent and Welingak website. The gives high accuracy (~90%), high senstivity (~90%) and specificity (~90%).