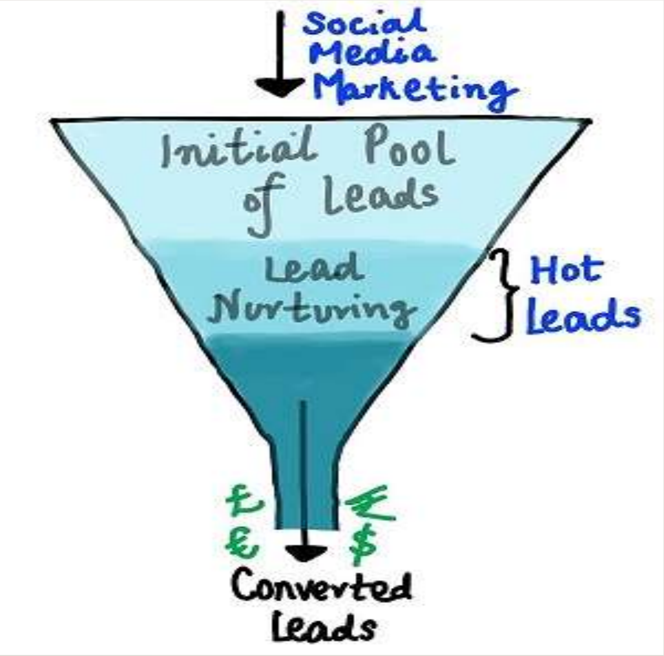
# LEAD SCORING LOGISTIC REGRESSION ASSIGNMENT

SUBMITTED BY: ROHIT BHANDARI AND ANSHU GOYAL

EMAIL-ID: ROHSINGHIT@GMAIL.COM, ANSHUGOYAL.UBS@GMAIL.COM

# PROBLEM STATEMENT / BUSINESS PROBLEM

- The education company X has a very low conversion rate (Conversion of potential leads into customers) of around 30% and thus a lot of time and money resources are wasted in chasing after poor leads.
- The company wants to identify hot leads from these potential leads so that time and money resources could be used more efficiently. The company wants to increase its lead conversion rate to around 80% from present 30%.



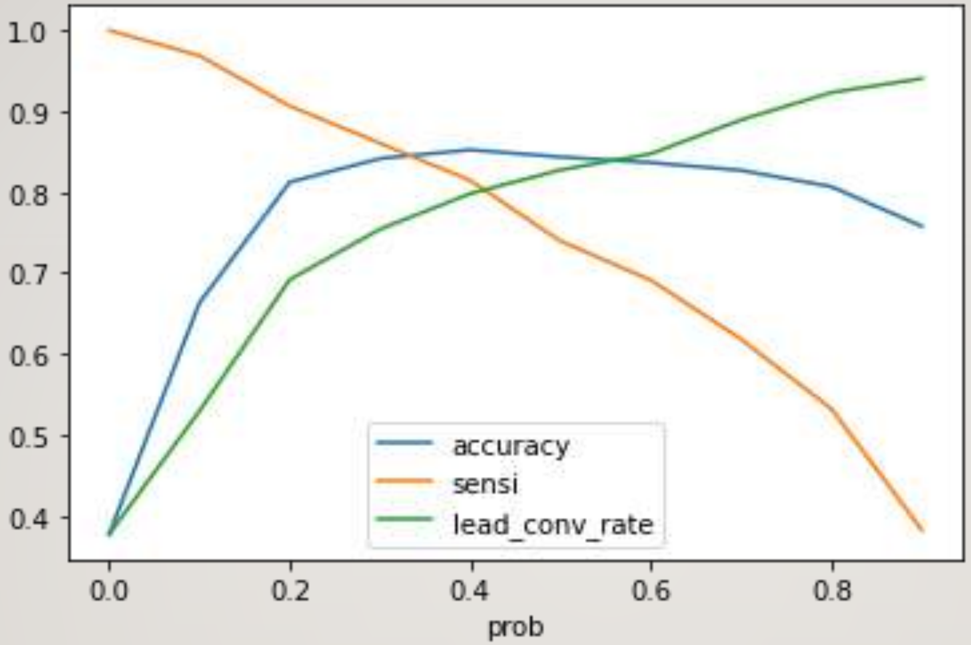
# Data Preparation

- **Some columns (like sources from where lead had seen advertisement) had one single value for almost all records. These columns have been dropped**
- **Some columns (like Country, City, Tags) which have significant number of records with null values and which are of no or little use in model building have been dropped. Other columns having significant number of null records but are of use in model building, have been kept and the missing values are imputed with the value "Data not available"**
- **Some columns (like Total Visits and Page Views Per Visit) have miniscule number of rows with missing values. These records/rows have been dropped.**
- **Many categorical columns (like Lead Source and Last Activity) have large number of categories. The categories with low number of records have been merged into "Others category"**
- **There were some outliers (7 in number) for the columns "Total Visits" and "Page views per visit". These records are removed from the data.**
- **Columns with numeric data have been scaled using Standardized Scaler.**

# ANALYSIS APPROACH

- **Train-Test Split on 70:30 ratio**
- **Model Building on train data using all features**
- **Feature Selection Using RFE. Tested RFE value of 5, 10, 15 and 20 to chose the best model**
- **Removing features with p-value greater than 5%**
- **Checking VIFs and removing features where VIF is around or greater than 5.**
- **Plotting ROC Curve to check for model efficiency**
- **Finding optimal value of the cut off probability so as to simultaneously maximize all of the metrics, particularly Lead Conversion Rate.**
- **Checking metrics on the test data**

# RESULTS



Final Model with values of metrics varying with different value of probability

- All the metrics i.e. Accuracy, Sensitivity and most importantly Lead Conversion rate all have values greater than 80% for cut-off probability in between the range of 0.4 and 0.5
- Cut-off probability of 0.45 is chosen as it leads to Lead Conversion Rate being greater than 81% while values of sensitivity gets to 79%, specificity to 88.35% and accuracy to 85%. (on train data)
- On test data, Sensitivity is 81%, Specificity is 89%, Lead Conversion Rate is 82.5% and Accuracy is 86%.

# RECOMMENDATIONS

- Start with leads who have higher score in terms of probability of conversion and avoid leads who have low score in terms of probability of conversion
- Chose leads who have lead_quality as one of “High in relevance”, “Low in relevance” or “might be”
- Chose leads whose origin is in “Others category”
- Focus on leads whose last notable activity is “SMS_sent”
- Focus on working professionals
- Focus on leads whose source is Olark chat
- Avoid leads whose lead quality is “worst”
- Avoid leads whose asymmetric activity index is low
- Avoid leads who have not opted for emails about the course