Lead Scoring Case Study

Group Members :

- Bhavesh Patel
- Arpit Goyal
- Nupur Singh

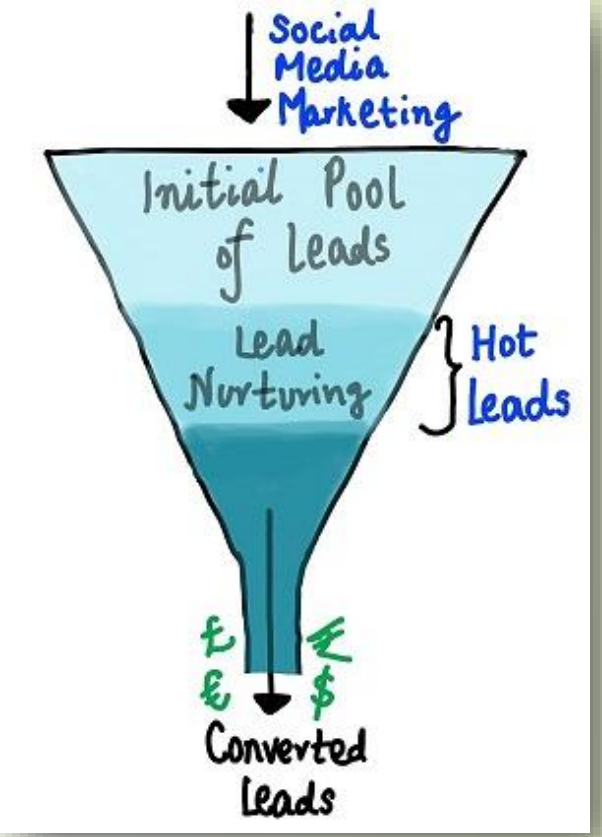
Problem Statement

- An Edu-tech company called X Education provides online technology courses to industry professionals to increase/upgrade their skill set.
- Leads to this particular company comes from various sources:
  - Advertisements on several websites.
  - Search engines like Google.
  - Social media websites like Facebook, twitter, etc.
  - Past referrals.
- When these people sign up, they will have to provide their email address and phone number and this information from the customer is collected and regarded as a potential lead. Once this information is captured, sales team starts making them calls and write emails to persuade them to become customers.
- The typical lead conversion rate right now is 30%
- To improve the lead conversion rate, company wants to identify the potential / converted leads, also known as Hot Leads.

Goals of the Case Study

- 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 that the lead is hot, i.e is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future



Solution Implementation

- Reading and cleaning the data
  - Read the data.
  - Handle missing (NA, ‘Select’) values
  - Drop insignificant columns (with NA >60%),imbalanced (No’s more)
  - Handling outliers
- Exploratory Data Analysis
  - Uni-variate and Bivariate Analysis
- Creating dummy variables, feature scaling and split the data into train and test.
- Algorithm: Logistic Regression
- Model training on train set and validation on test set.
- Summary and recommendation

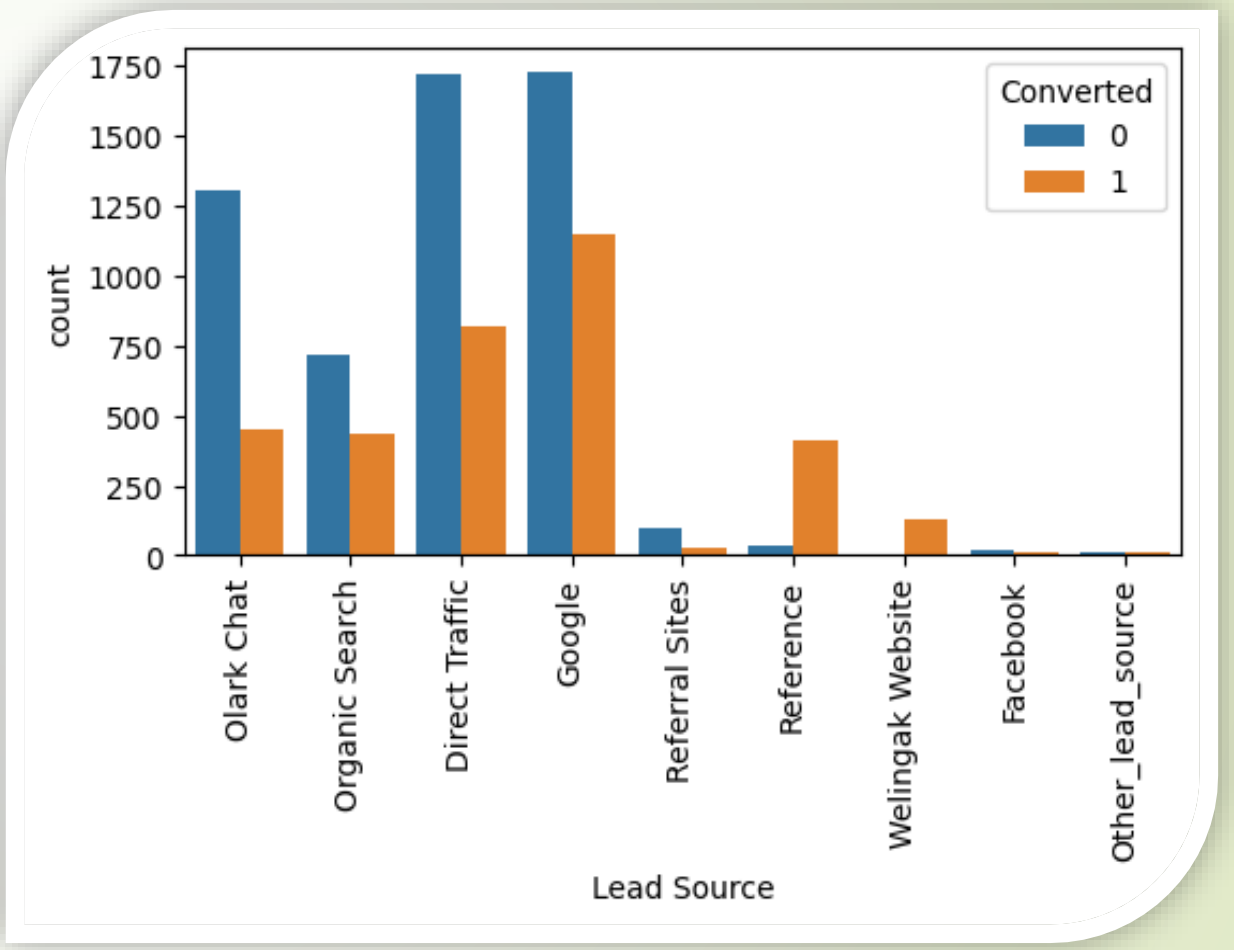
Data Cleaning

- Data shape:(9240, 37)
- Columns with null values more than 60 %are removed from the dataset (How did you hear about X Education, Asymmetrique Profile Score, Asymmetrique Activity Score, Asymmetrique Activity Index, Asymmetrique Profile Index)
- Rest of the columns with null values are imputed with mode (City, Tags, What is your occupation, Country)
- Column Specialization is imputed with ‘Others’ for NA values.
- Rows are deleted for the lesser null value percentage columns. (Total Visits, Page Views Per Visit, Last Activity, Lead Source
- Column ‘What matters to you most in choosing a course’ is completely imbalanced so we removed the column.
- Categorical columns with value as ‘Select’ as value are replaced with NA as they might not have been chosen by the lead.
- Outliers are removed from Total Visits and Page Views Per Visit (data upto 95th percentile is considered)
- Data imbalance:
  - Converted rate –37.9%, Un-converted rate –62.1%

Exploratory Data Analysis (EDA)

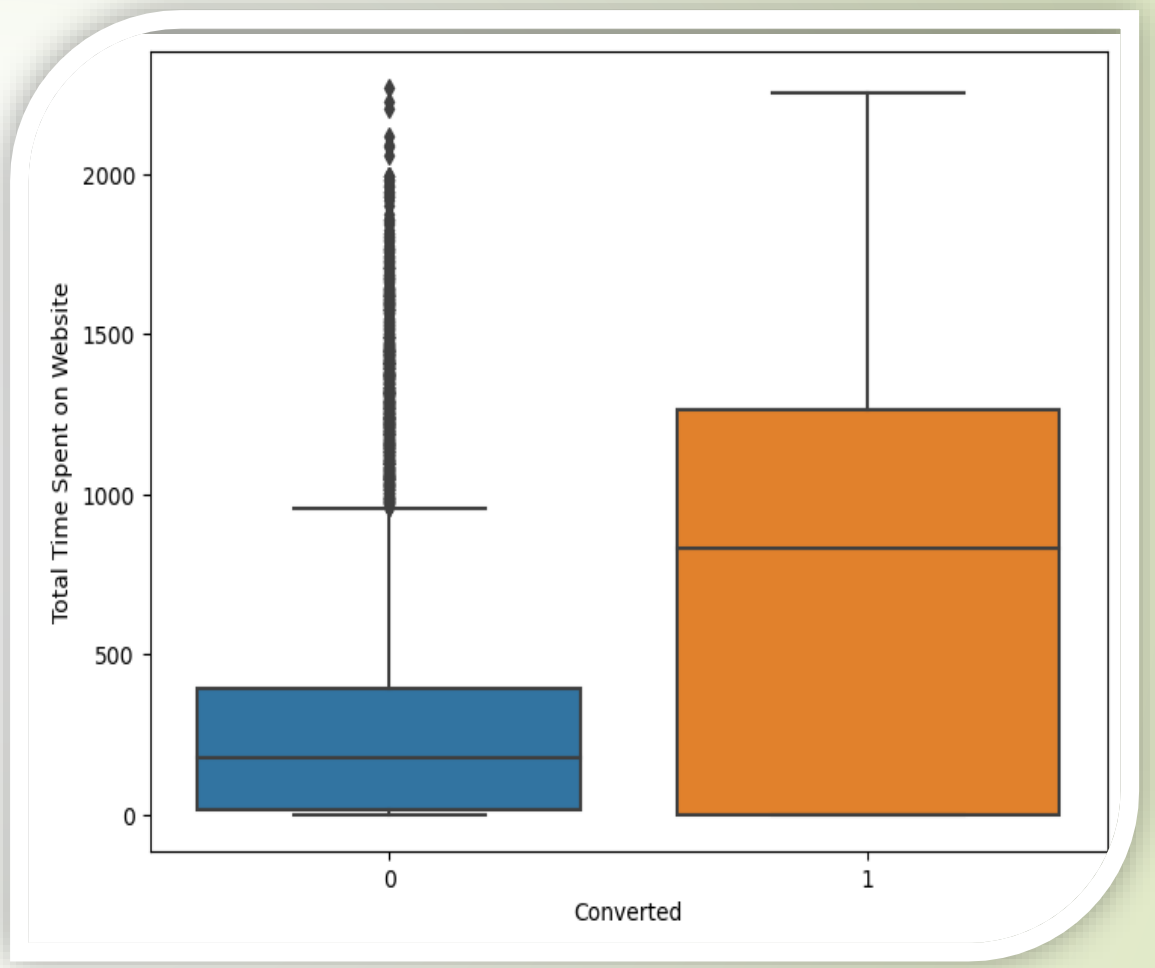
Observation:

- Google and direct traffic generate the most leads.
- The conversion rate of reference leads and leads generated by the Welingak website is very high.
- To increase overall lead conversion rate, focus on improving lead conversion of Olark chat, Organic Search, Direct Traffic, and Google, as well as generating more leads from Referral Sites and Welingak Website



Observation:

- Leads who spend more time on the site are more likely to convert.
- To encourage leads to spend more time on the website, it should be made more engaging



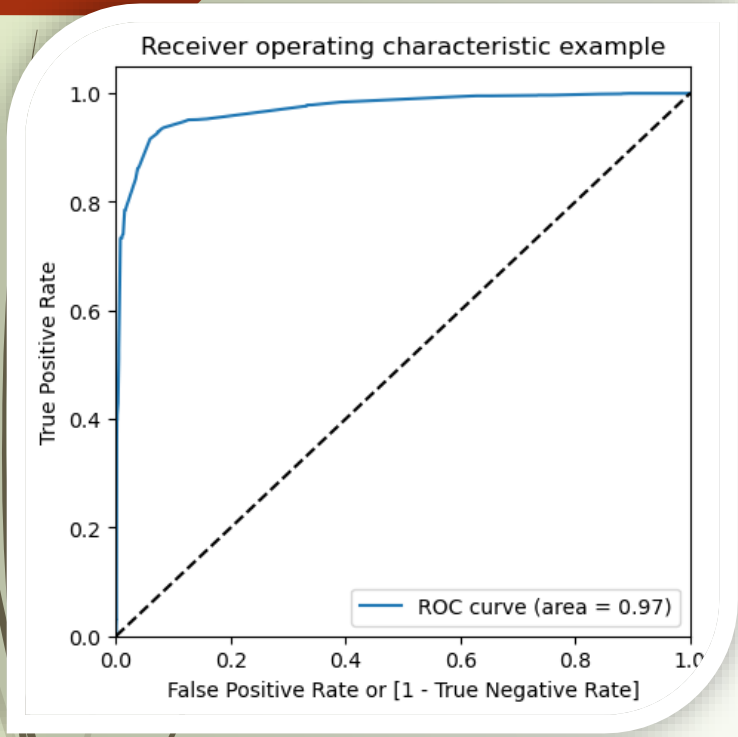
Data Manipulation

- Numerical Variables are scaled using Standard Scaler (TotalVisits, Total Time Spent on Website, Page Views Per Visit).
- Dummy Variables are created for categorical variables (Lead Origin, Lead Source, Last Activity, Specialization,What is your current occupation, City, Last Notable Activity).
- Split the data into 70 %training set and 30 %test data set.

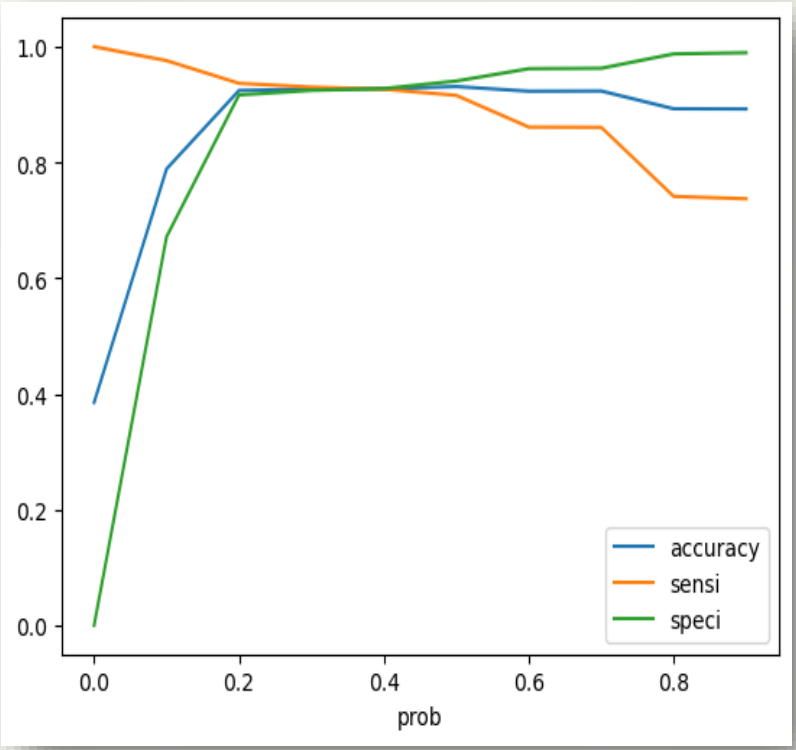
Model Building and Evaluation

- Using RFE to select 15 features
- Fit the model with variables selected by RFE.
- Eliminating the variables whose p-value is greater than 0.05.
- Eliminating the variables whose VIF is higher.
- Repeat the above two steps until all the variables have p-value less than 0.05 and VIF is below 5.
- Finalize the model and make the predictions on test data set.
- Evaluate the model by deriving metrics.
- Accuracy on train and test data respectively: 93.10 % and 91.59 %
- Sensitivity on train and test data respectively: 91.57 % and 90.89 %
- Specificity on train and test data respectively: 94.05 % and 91.98%

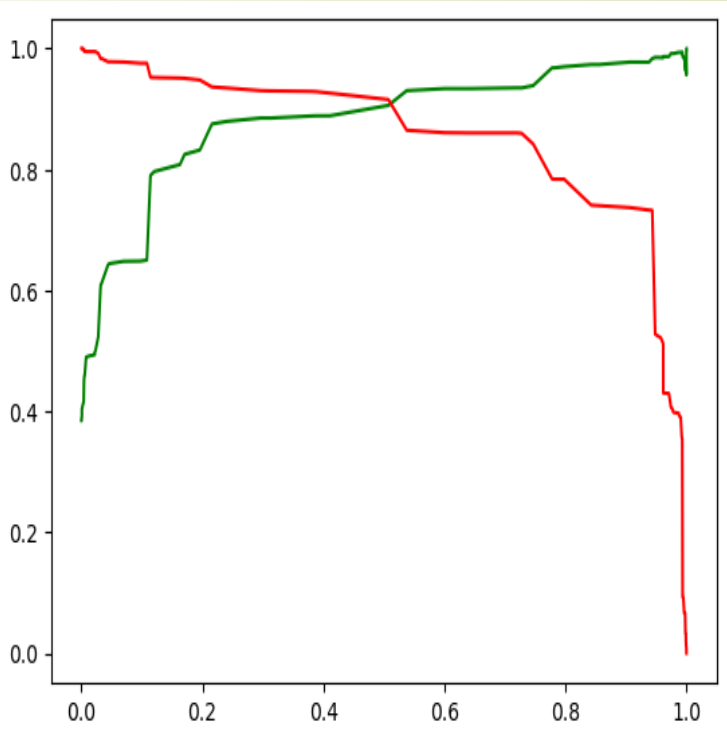
ROC Curve



The area under the ROC curve is 0.97, which is an excellent value.



Plot for accuracy, sensitivity and specificity for various probabilities. It is clear that the optimal cut off is 0.40



Plot a trade-off curve between precision and recall

Summary and Recommendations

- We decided on the optimal cut off based as 0.40 on Sensitivity and Specificity for calculating the final prediction. The test set's metrics such as accuracy, sensitivity, and specificity values are around 90%, respectively, which are roughly close to the respective values calculated using the trained set.

- **The top three variables that contribute to lead conversion are:**
  - Lead Origin
  - Last Activity
  - What is your current occupation_Working Professional
- **By approaching people with:**
  - Lead Origin
  - Last Activity_SMS sent
  - Lead Occupation – Working Professionals
- **Avoid calling people with:**
  - Lead origin as Landing Page Submission
  - Specialization as Others
  - Do Not Email option selected as yes.
  - By making calls, the sales team can convert these lead sources into leads.