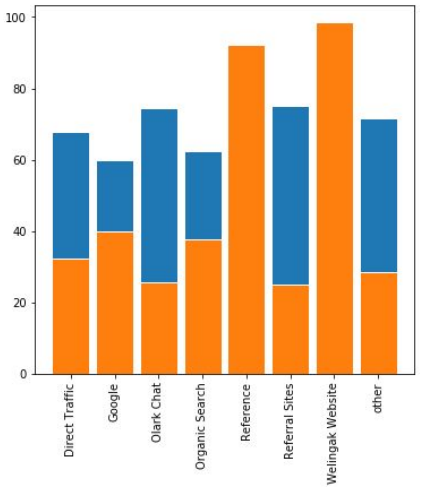
Summary Report

Objective of the case study – Help X-Education to identify Hot Leads from initial pool of leads obtained by the company from different sources. The intent is to help company run a focused marketing and sales program to maximize the output of investment of time and money made by the marketing team.

Solution Proposed – The problem is classification problem. We are expected to predict the leads most likely to be converted into paid customers. However, we also need to assign the weight/probability to each lead. This will be helpful to team in case of varying resource availability (having interns or target achieved). Therefore, the problem is identified as a use case of logistic regression.

Solution Approach – We have 36 different variables. We start with data analysis and treatment. The columns with over 50% null values are dropped. Manually generated columns like Tags and computed columns like AsymmetriqueProfileIndex etc. are also dropped. We find that columns like specialization, city and country have good conversion rates against missing values. We decide to impute these as a separate category – 'Not Specified'. We also group categories with insignificant counts to 'others' in columns like country, lead source etc. Highly skewed columns are dropped.

Exploratory Data Analysis: Key insights obtained from exploratory data analysis.



1. Old References, Welingak Website are the most effective sources of leads.
2. Time Spent on the website is a key factor for conversion.
3. Working professionals are the key customers. Unemployed people have high chances of joining.
4. SMS marketing complements telemarketing effectively.

Model Building:

For logistic regression, we convert are categorical variables to dummies and scale the numerical variables using standard scaler. Then we remove highly correlated variables. We use RFE from sklearn library for identifying 15 most relevant features.

We then use Stats Model for manual feature reduction. We go on eliminating the features having high p-values and VIF factors.

Final list of features retained -

- WhatIsYourCurrentOccupation_NotSpecified
- LeadSource_Olark Chat
- LastNotableActivity_Modified
- TotalTimeSpentOnWebsite
- LastActivity_Email Opened,LastActivity_SMS Sent
- LeadSource_Reference
- LastNotableActivity_Olark Chat Conversation
- LeadSource_Welingak Website
- LastActivity_other.

We plot the Sensitivity, Specificity and Accuracy Score graph. Optimal cut-off lies between 3 and 4. We have taken 0.35 as cut-off. Key metrics for the model on training data.

Metric	Value
Accuracy	78.7%
Sensitivity/Recall	82.2%
Specificity	76.56%
Precision	68.21%
F1-Score	74.5

Then we make the predictions on test data. Here are the key Metrics on test Data

Metric	Value
Accuracy	79.48%
Sensitivity/Recall	83.94%
Specificity	76.76%
Precision	68.77%
F1-Score	75.6

Recommendations made from above study:

Highest contributing factors for lead conversion are LeadSource_Welingak Website, LeadSource_Reference, LastActivity_SMS Sent. Team should focus more on people coming through these sources. People coming through search engine results should be the next priority for the marketing team. Also, combination of SMS and telemarketing is more effective than individual methods. Company can reduce efforts on Olark Chat as the conversion rate is not very high. Website content helps in converting customers. Company should consistently focus on keeping content up to date.