LEAD SCORING CASE STUDY

Vikas Purohit

LEAD SCORING CASE STUDY

- X-Education is generating leads through multiple sources for the online courses they sell.
- Their current lead conversion ratio is around 30%, which is very low and the management is looking to use machine learning to identify leads that have more probability to be converted.
- A dataset of around 9000 rows and 37 columns with a Target variable ”Converted” has been provided.
- Ask is to build a logistic regression model that can predict the potential lead conversion with about 80% accuracy.

EXPLORATORY DATA ANALYSIS

- As initial step, EDA has been performed on the dataset, in order to clean the data by removing outliers, correcting spelling mistakes and also treating outliers.

- There were columns that had no variance, for example – most of the columns capturing whether the customer has seen ads published across different media were "No" and hence were dropped.

- As next step, Univariate and Bivariate analysis was performed on the cleansed data.

UNIVARIATE ANALYSIS

- Major findings from the Univariate Analysis are:

✓ Most leads are generated from India, predominantly from Mumbai.

✓ Most leads are generated from people who are unemployed and are looking for better career options.

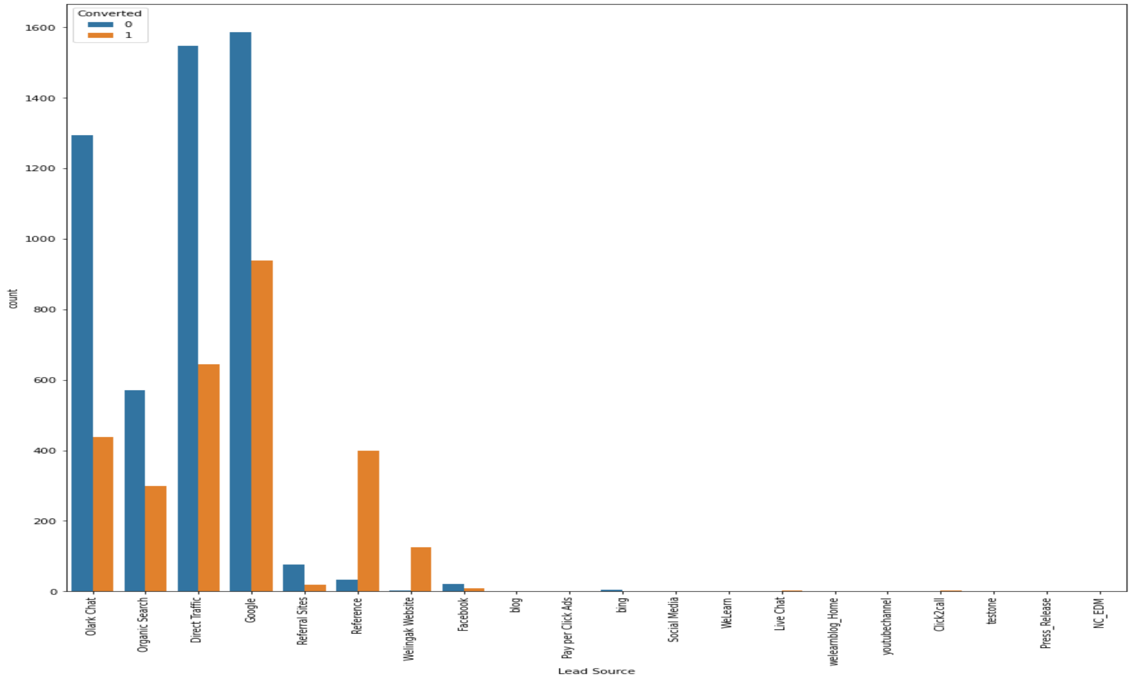
✓ Majority of leads are generated directly from the website landing page submission and are routed from Google.

✓ Large number of leads have opened the e-mail sent to them, but very few have clicked on the link provided in the email.

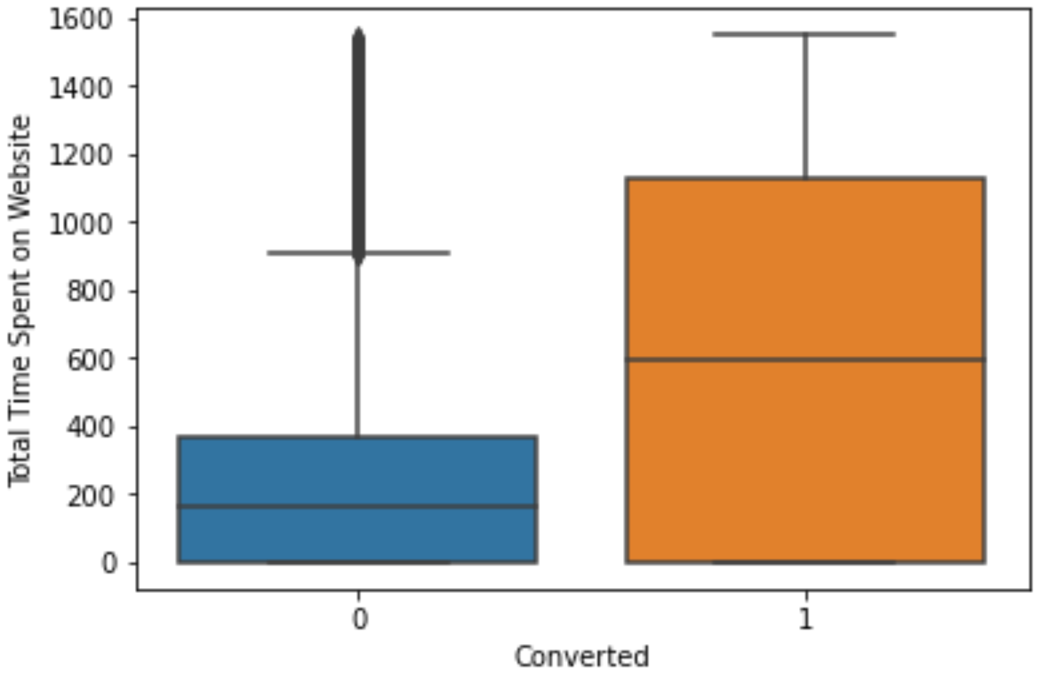
✓ Despite majority of leads being from unemployed people, but majority have not opted to receive free guide on mastering interviews.

BI-VARIATE ANALYSIS

- Bivariate Analysis was done mainly to understand how the target variable "Converted" is behaving with other variables.

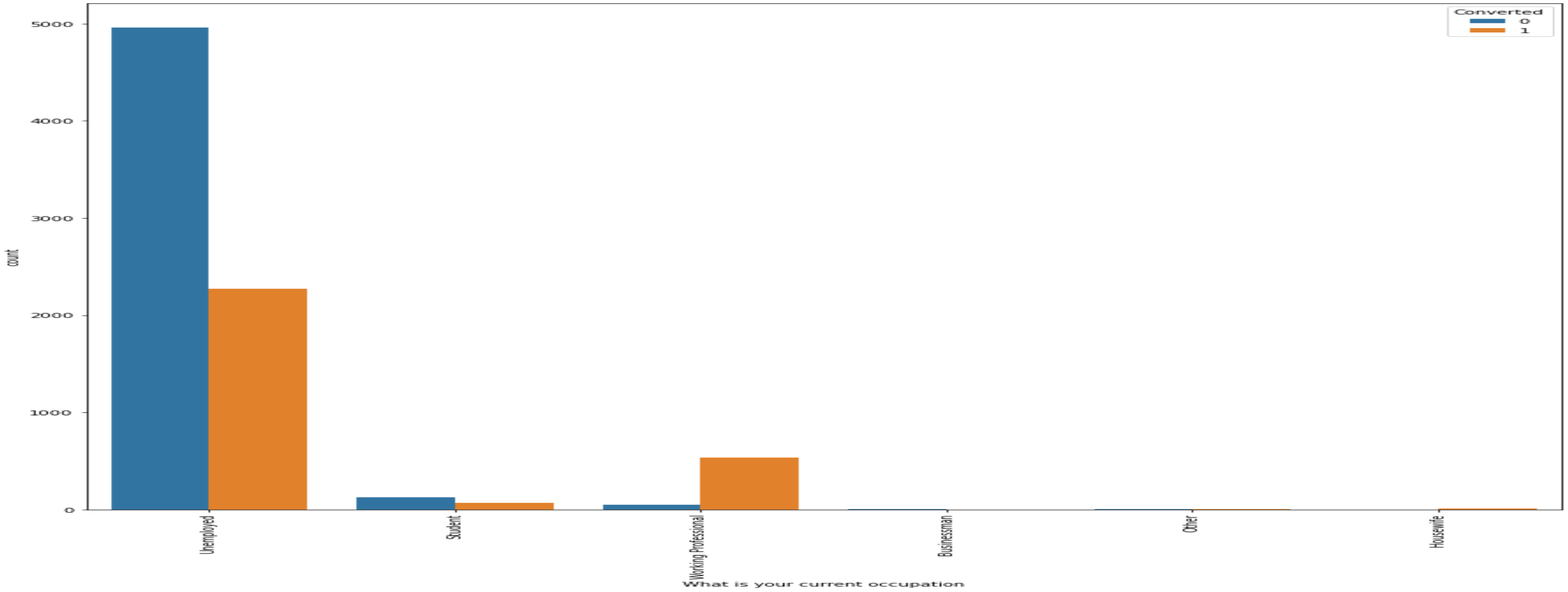


Conversion Rate is much higher when the lead comes from Reference.



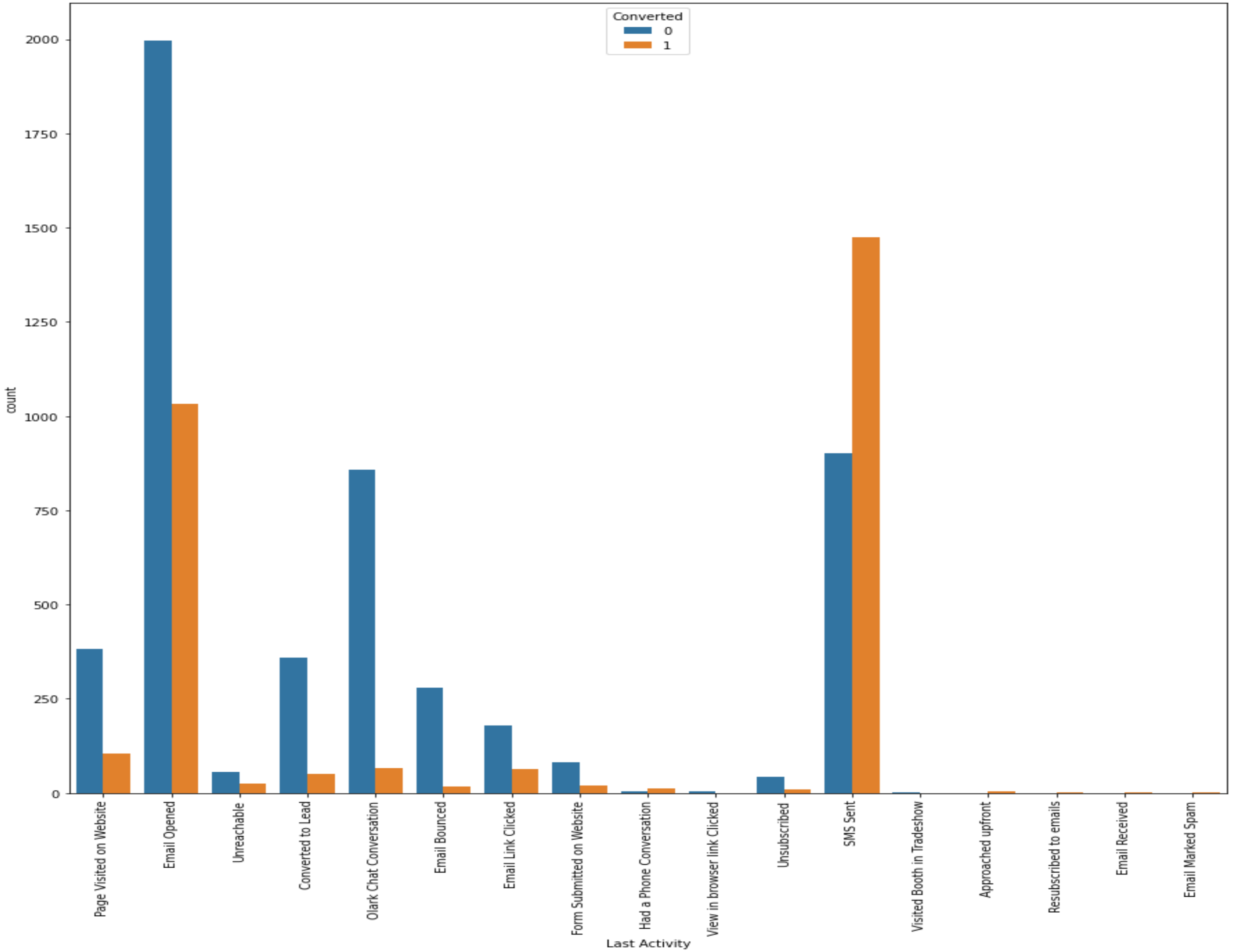
Conversion rate increases with time spent On the website

BI-VARIATE ANALYSIS



Working Professionals have higher conversion rate

BI-VARIATE ANALYSIS



Leads that are made by sending SMS have High conversion rate.

MODEL BUILDING – DATA PREPARATION

- As part of Data Preparation, following steps were taken:

✓ Binary categorical variables were converted to 0s and 1s

✓ Dummy variable creation for categorical variables with multiple hierarchies.

✓Numerical variables were scaled using Scaler function.

✓ RFE was used for feature elimination. Top 20 features were selected.

MODEL BUILDING – TOP FEATURES

- As part of the model building, Top 20 features were selected using RFE
- Further optimization of model was done by manual feature elimination based upon p-values and VIF.
- A final model with 15 features and 81% accuracy was arrived at.
- Top features impacting the Conversion rate were identified as:

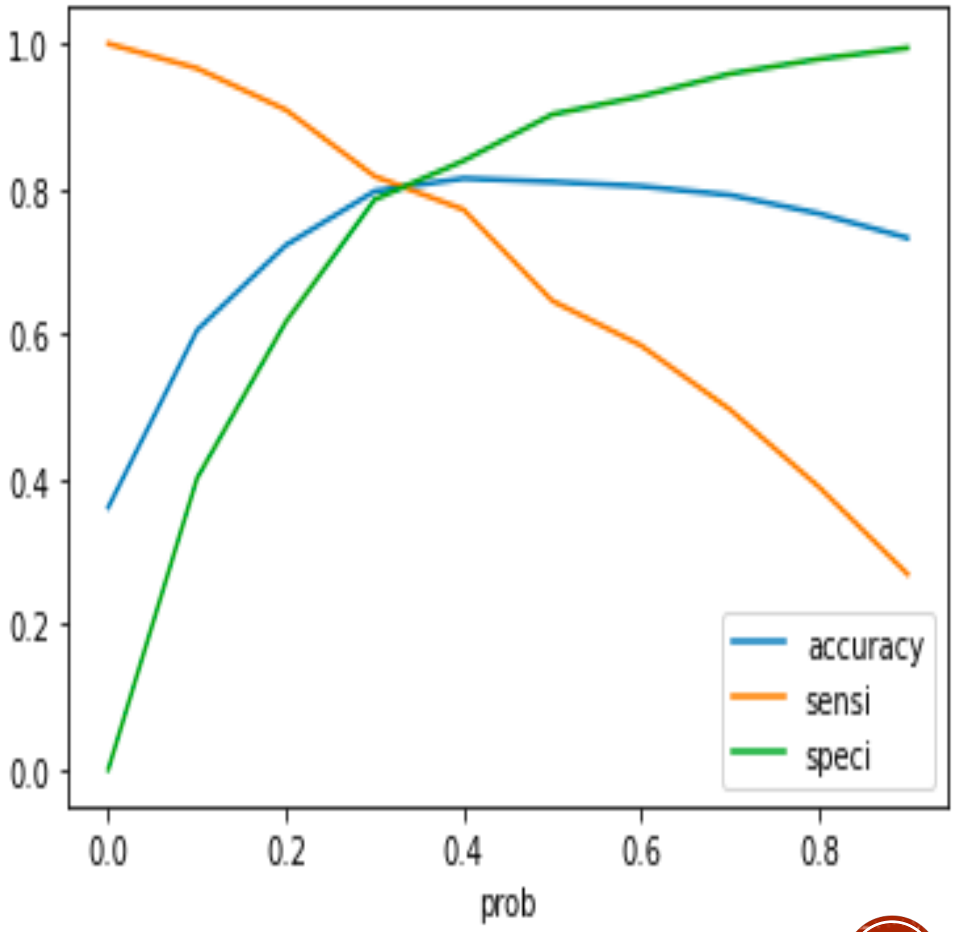
✓ Time Spent on Website – Conversion rate increases with more time spent

✓ Current Occupation – Working professionals have higher conversion rate.

✓ Lead Origin – Highest conversion rate is when lead originates from Landing Page Submission

MODEL BUILDING – GO BEYOND ACCURACY

- In order to improve the lead conversion, it is crucial to understand that out of total converted, how many are predicted.
- Actual Converted = True_Positive + False_Negative = Sensitivity
- Model only had a Sensitivity of 64.5% with a cut-off of 0.5 (Conversion probability of 0.5 and above considered to be Yes for Conversion)
- ROC curve was plotted and Accuracy, Sensitivity and Specificity were determined and plotted for different cutoffs ranging from 0.1 to 0.9
- Optimum Cut-off of 0.35 obtained



RECALIBRATING THE MODEL

- Model Accuracy, Sensitivity and Specificity were re-calculated using 0.35 as cut-off
- Final Accuracy : 81.1%
- Final Sensitivity: 79.8%
- Final Specificity: 83.7%

MODEL PERFORMANCE ON TEST DATA

- Accuracy of the model on Test Data: 81.07%
- Final Sensitivity on Test Data: 79.76%
- Final Specificity on Test Data: 83.66%
- Model performance on Test Data is similar to Train Data, indicating that model is sufficiently generalized and not overfit for Train data.

THANK YOU