Lead Scoring Case Study: Summary

Data size

- The data set contains 9240 rows & 37 columns (features)

Data value "select" under many columns

- Many columns have the value "select". This is the default value while filling the data. This means the data was not provided.
- So, all the "select" values of all columns replaced with Null value

Missing values

- Columns Lead Profile, City, Lead Quality, How did you hear about X Education, Asymmetrique Activity Index, Asymmetrique Profile Index , and Asymmetrique Activity Score have more than 40% missing values ,hence these columns dropped from analysis
- Country, Specialization, What is your current occupation, What matters most to you in choosing a course, Tags had 27-36% missing values, these values are imputed with most appeared value of the corresponding column.

Columns with very few missing values

- Columns TotalVisits, Page Views Per Visit, Last Activity" had 1% missing values, missing values dropped instead imputing

Columns with highly skewed data

- "What matters most to you in choosing a course", "News Paper" ,"News paper article", "Do Not Call", "Country", "Search", "X Education Forums", "Digital Advertisement", "Through Recommendations", and "A free copy of Mastering The Interview"
- These columns dropped

Columns with Most unique value

- Columns Magazine, 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 have 100% unique values. So, these columns dropped from analysis

Columns not necessary for the analysis

- 'Prospect ID', 'Lead Number'

Final chosen features for analysis

- **Categorical:** Lead Origin, Lead Source, Do Not Email, Last Activity, Specialization, What is your current occupation, Tags, Last Notable Activity
- **Numerical:** TotalVisits, Total Time Spent on Website and Page Views Per Visit

Outliers handling for numerical features

- Columns “TotalVisits” and “Page Views Per Visit” have lot of outliers.
- Data capped to 95% percentile, which seems

Data Imbalance

- “Converted” data is 38% compared to “non converted”, after data cleaning and preparation which seems reasonable i.e. not highly imbalanced

Exploratory data analysis

- **Lead Origin:** value 'Lead Add form' has highest conversion rate
- **Lead Source:** values 'Reference' and 'Wellingak Website' have high conversion rate followed by “Google”
- **Last Activity:** "Head Phone Conversation" and 'SMS sent’ have high conversion rate
- **Specialization:** No value has any significant higher conversion rate than others

- **What is your current occupation:** “Working Professionals” have high conversion rate
- **TotalVisits:** A general trend of increasing conversion rate as TotalVisits increased

Logistic regression model building

- **Dummy creation binary categorical features:** “0/1” value conversion done to data having two level categories
- **Dummy creation multilevel categorical features:** One hot encoding done to data having more than two category levels
- **Independent and response features:** From final data set “Converted” assigned to “y” & all other features assigned to “X”
- **Train-Test split:** Data split to Train and Test set in 70-30% ratio
- **Feature scaling**: Feature scaling done on numerical features ('Total Time Spent on Website', 'TotalVisits','Page Views Per Visit’)
- **RFE:** top 20 most significant features selected using RFE method
- **Feature elimination**: further insignificant feature elimination carried based on p values from statsmodes logistic regression fit models summary and VIF values. Features with more than 0.05 “p” value and more than 5 VIF value are dropped
- **Prediction on train set:** prediction on train set carried out using defaults threshold probability value of 0.5
- **Train set performance metrics:** Optimal probability cut off: final model created by changing the threshold probability cut off value to 0.38, which is obtained by the intersection point of “Accuracy”, “Sensitivity” and “Specificity” of the previous model
- **Train set performance metrics from final fine tuned model**

  Accuracy Sensitivity Specificity

80.9	0.78	0.82

- **Prediction on the test set**

Accuracy	Sensitivity	Specificity
80.7	0.86	0.77

- **Assigning lead score to test set:** lead scores of 0 to 100 assigned to test data based on their probability values
- **Predicting hot leads:** hot leads, which have lead score of greater than or equal to 85 as defined by the business were predicted

Model summary

- Accuracy on train set & test set is 80.9 and 80.7 indicates a good model
- Sensitivity on train and test is 0.78 and 0.86 which reasonably good performance indicator