LEAD SCORING CASE STUDY

PROBLEM STATEMENT

This case study involves helping X Education, an education company, improve its lead conversion rate by building a logistic regression model to assign lead scores. The aim is to identify potential leads with the highest chances of converting to paying customers and handling future problems to achieve a target conversion rate of 80%.

GOALS

- 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 our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

During this case study the following steps are done:

READING AND UNDERSTANDING DATA

- The company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- We have to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

TREATING OUTLIERS

Outliers can be caused by measurement errors or may indicate a certain phenomenon. It is important to identify and treat outliers before building a model. We observed outliers with two numerical columns Total visits and Page Views per Visit.

EXPLORATORY DATA ANALYSIS

- Majority of the people who visit the site are unemployed.
- The last activity for majority of the leads is Email opened.
- majority leads have landing pager submissions as lead origin.
- Total time spent on website has a bit high positive correlation with converted whereas Pages per visit is low comparatively related.

MODEL BUILDING

DATA PREPARATION

Created dummy variables for features with multiple levels using pd.get_dummies.

FEATURE SCALING

Since the range of values of raw data varies widely, we have used Recursive Feature Elimination(RFE) for selecting the features in the training dataset that help to predict the target variable.

MODEL EVALUATION

- Model Evaluation is an important step and we have used this to assess the performance of our model
- Reached to the following observations:

Recall - 0.7011447260834015

Accuracy - 0.8141755416819684

Precision - 0.7932469935245143

Sensitivity - 0.7542972699696663

Specificity - 0.8483275663206459

FINDINGS & SUGGESTIONS

1. The final model works with

Accuracy = 0.80

Precision = 0.75

Recall = 0.70

2. The model is good, simple and easy to use

3. It can help to identify the correct leads which can be converted.

4. Focusing on the unemployed people help to generate more leads.

5. Similarly, the sales team should spend more time on websites and try to contact through sms or phone calls which can easily create more hot leads and has higher chances of conversion.