Lead Scoring Case Study Summary

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Our goal is to identify factors that influence conversion to a professional training course offered online. Understanding these factors will allow us to target effective marketing strategies and improve enrollment rates.

SOLUTION SUMMARY

Lead Scoring Logistic Regression Case Study is divided into below sections:

- Import Python Packages
- Reading and Understanding the dataset
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Train-Test split
- Feature Scaling
- Model Building
- Model Evaluation
- Making Predictions on test dataset
- Model Parameters
- Conclusion
- Recommendations

IMPORT PYTHON PACKAGES:

As a first step, we import the python packages such as numpy, pandas, matplotlip, seaborn, sklearn, statsmodel etc.

READING AND UNDERSTANDING THE DATASET:

- Read the dataset as a pandas dataframe and start analysing the dataset
- Perform some basic checks on the dataset like get the head details, describe the dataset using percentiles, info, shape of the dataset, columns
- After the basic check it was seen that the columns 'TotalVisits' and 'Page Views Per Visit' might have some outliers

DATA CLEANING:

- **Treatment for 'Select' values**
    - Many of the categorical variables have a level called 'Select' which needs to be handled. The columns having ‘Select’ were imputed with NaN
- **Delete redundant columns i.e 'Prospect ID' and 'Lead Number'**
- **Treatment for columns with unique values**
- **Treatment for null values**
    - Delete the columns where null value percentage is > 40%
    - Impute the null values for the columns where null value percentage > 15% with mode
    - Treat columns where null value percentage ~1%
- **Outlier Treatment**
    - There were outliers for column 'TotalVisits' and 'Page Views Per Visit'.
    - The outliers are present only in the upper range so the outliers were treated by capping the upper range to 99%
- **General Analysis for remaining columns**
    - Lead origin:
        - 'API' and 'Landing Page Submission' both have high numbers or leads and conversion rate
        - 'Lead Add Form' has a very high conversion rate as compared to the number of leads they have
        - 'Leads Import' has very few leads
    - Do not email:
        - 92% of the people don't want to be emailed about the course
    - Do not call:

        - We can see from the value count data that 'No' is having more than 99.9% of the data. We can safely drop this column, as this will not add much to the analysis
    - A free copy of Mastering The Interview:
        - The number of leads who do not ask for the free copy are high. This group can be focused for conversion
    - Last notable activity:
        - 'Modified' and 'Email Opened' have a high number of leads. This section can be targeted to increase the conversion rate
        - SMS sent have a high conversion rate

EXPLORATORY DATA ANALYSIS:

- Conversion rate is 37.92%, meaning only 37.92% of the people have converted to leads
- Time spent on the website shows a positive impact on lead conversion
- Performed Univariate and Bivariate Analysis for categorical and numerical columns using countplot, pairplot and correlation matrix
    - 'TotalVisits' and 'Page Views Per Visit' have a high correlation index of 0.72
    - 'Total Time Spent on Website' has a correlation index of 0.35 with target variable 'Converted'
    - All other numerical columns have low correlation with target variable 'Converted'

DATA PREPARATION

- Converting Binary Variables to 0/1 for 'Do Not Email' and 'A free copy of Mastering The Interview' features
- Created Dummy features for categorical variables

TRAIN-TEST SPLIT

- Splitting Train and Test Sets: 70:30 ratio

FEATURE SCALING

- Scaled the numerical features using StandardScaler
- Insights after scaling the features:

    - 'Lead Source_Facebook' and 'Lead Origin_Lead Import' have a high correlation of 0.98
    - 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' have a high correlation of 0.85
    - 'TotalVisits' and 'Page Views Per Visit' have a correlation of 0.72
    - 'Lead Origin_Lead Add Form' ,'Lead Source_Welingak Website', 'Last Activity_SMS Sent' and 'What is your current Occupation_Working Professionals' have positive correlation with the target variable 'Converted'

MODEL BUILDING

- Used Recursive Feature Elimination(RFE) to reduce features from 49 to 15
- Manual Feature Elimination process was used to build models by dropping variables with p-value > 0.05 and VIF > 5
- Total 2 models were built before reaching final Model 3 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5
- The final model had 13 variables, we used it for making predictions on the train and test set

MODEL EVALUATION

- Used Confusion Matrix, Accuracy, Sensitivity, Specificity, Threshold determination using ROC & Finding Optimal cutoff point, Precision and Recall evaluation metrics for evaluating the model
- Confusion matrix was made and cut off point of 0.286 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 90%
- Lead score was assigned to train data using 0.286 as cut off

MAKING PREDICTIONS ON TEST DATASET

- Making Predictions on Test: Scaling and predicting using the final model
- Evaluation metrics for train and test are very close to around 91%
- Lead score was assigned
- The sensitivity value for test data is 92.86% while for train data is also 91.35%.
- The accuracy value is 92.70%. This indicates that the model is performing well for the test data set also

MODEL PARAMETERS

- The final logistic regression model has 13 features
- **Equation of line:**
    - **Converted** = -0.765079 - 0.367659 X Lead Origin_Landing Page Submission + 0.403844 X Lead Origin_Lead Add Form + 0.274769 X Lead Source_Olark Chat + 0.468541 X Lead Source_Welingak Website + 0.547402 X Last Activity_SMS Sent + 1.291309 X Tags_Closed by Horizzon + 0.840598 X Tags_Lost to EINS - 1.010874 X Tags_Others - 1.170860 X Tags_Ringing + 1.851406 X Tags_Will revert after reading the email + 0.626698 X Last Notable Activity_Email Opened + 0.978982 X Last Notable Activity_SMS Sent

CONCLUSION

- The model achieved a sensitivity of 91.35% in the train set and 92.86% in the test set, using a cut-off value of 0.286
- The model also achieved an accuracy of ~91%
- The Optimal cutoff probability point is 0.286. Converted probability greater than 0.286 will be predicted as Converted lead and probability smaller than 0.286 will be predicted as not Converted lead
- **Top three features** that contribute positively to predict hot leads are:
    - Tags_Will revert after reading the email
    - Total Time Spent on Website
    - Last Notable Activity_SMS Sent

RECOMMENDATIONS

To improve the potential lead conversion rate X-Education should focus on the top important features:

- **Tags_Will revert after reading the email:** As the leads with tags, will revert after reading the email is high, so the company should focus more on email marketing
- **Total Time Spent on Website:** Leads spending more time on the website can be our potential lead
- **Last Notable Activity_SMS Sent** Lead whose last activity is sms sent can be potential lead for the company
- **Tags_Closed by Horizzon:** Tags closed by Horizzon have a good conversion rate
- Focus on features with **positive coefficients**
- **Working professionals** to be targeted as they have a high conversion rate

- Develop strategies to attract high-quality leads from **performing lead sources**