UpGrad Lead Scoring Case Study

Name -Dhrudeep, Kalpana and Soumalya
Project - Lead Scoring case study

Problem Statement

- X Education, an online education company, has a low lead conversion rate and wants to identify the most promising leads, or "Hot Leads," to improve its conversion rate.
- The company wants to develop a model that assigns a lead score to each lead, with higher scores indicating a higher likelihood of conversion. The CEO has set a target lead conversion rate of 80%.

Approach and Steps

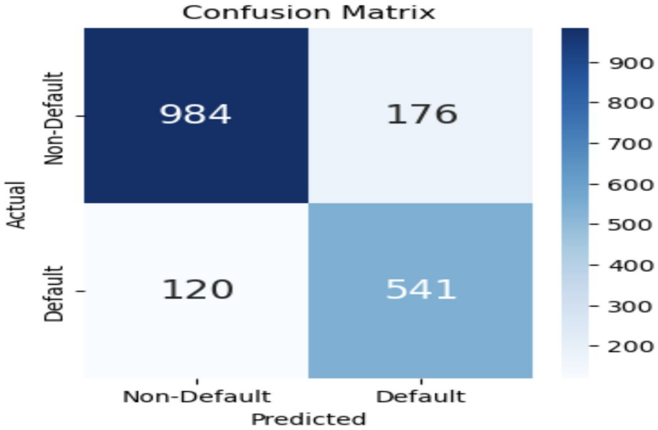
1. The approach involves developing a predictive model that evaluates various features associated with leads, such as browsing behavior, form submissions, and referral sources. This model will assign a lead score based on the probability of conversion.
2. By implementing this lead scoring system, X Education anticipates optimizing its sales team's efficiency, allowing them to focus on leads with the greatest potential for becoming paying customers and, consequently, increasing the overall lead conversion rate.

Exploratory Data Analysis

- Bivariate analysis gives lot of understanding
- Key Insights:
    - Current occupation of Housewife, Businessman and Working professionals have very high lead conversion
    - Lead source of Welingak,NC_EDM, Reference and Click2Call have very high lead conversion
    - Lead Origin -> Lead Add form has highest conversion
    - Total time spent on website -> for converted people it is 728 minutes v/s 329 for non converted.

Model Building

- We started with RFE to select top 30 features to fit on the logit model.
- Post selected top 30 features, we checked VIF of the variables and removed variables with VIF>6
- We had final model which gave training accuracy of 0.83 and test accuracy of 0.84
-



Confusion Matrix

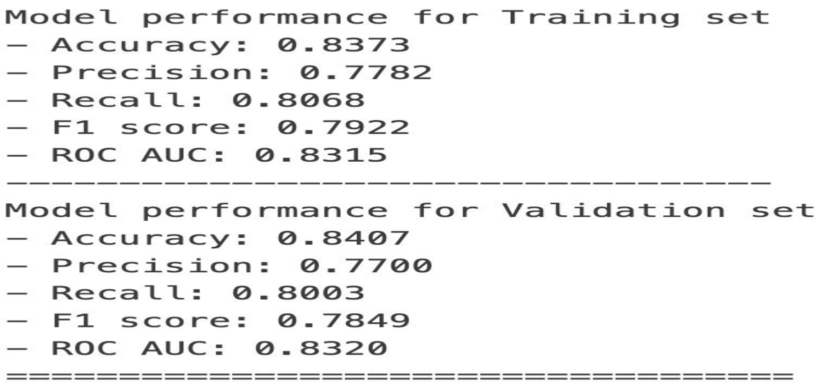
Actual \ Predicted	Non-Default	Default
Non-Default	984	176
Default	120	541

Model Tuning

- We used different models including random forest, decision trees, xg boost and logistic regression
- We trained the model with different combination of hyper parameters to optimize for recall using Randomized search cv and cross validation
- We arrived at final model of Decision tree which gives decently good recall of 0.8 and ROC AUC of 0.83

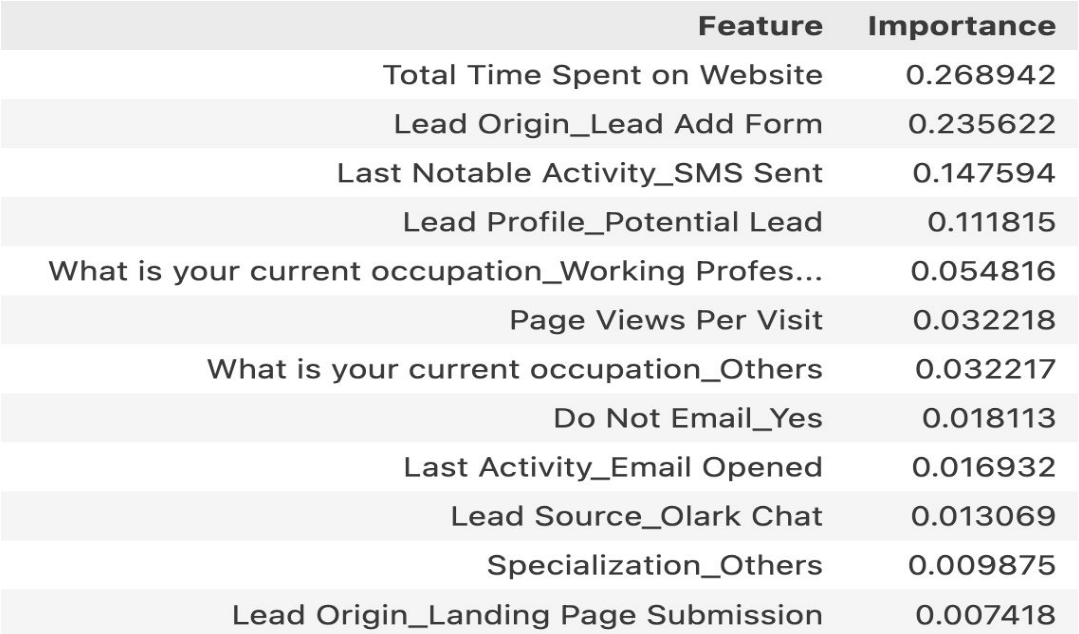
Decision tree model

- We arrived at final model of Decision tree which gives decently good recall of 0.8 and ROC AUC of 0.83
- Best Parameters: {'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': 7, 'criterion': 'entropy'}



```
Model performance for Training set
- Accuracy: 0.8373
- Precision: 0.7782
- Recall: 0.8068
- F1 score: 0.7922
- ROC AUC: 0.8315
----------------------------------
Model performance for Validation set
- Accuracy: 0.8407
- Precision: 0.7700
- Recall: 0.8003
- F1 score: 0.7849
- ROC AUC: 0.8320
===================================
```

Feature Importance Decision Tree



Feature	Importance
Total Time Spent on Website	0.268942
Lead Origin_Lead Add Form	0.235622
Last Notable Activity_SMS Sent	0.147594
Lead Profile_Potential Lead	0.111815
What is your current occupation_Working Profes...	0.054816
Page Views Per Visit	0.032218
What is your current occupation_Others	0.032217
Do Not Email_Yes	0.018113
Last Activity_Email Opened	0.016932
Lead Source_Olark Chat	0.013069
Specialization_Others	0.009875
Lead Origin_Landing Page Submission	0.007418

Summary

The key steps followed in the process:

- Identifying and treating missing values
- EDA to understand top features that could possibly impact the conversion
- Scaling for numerical features
- Dummy variable creation for categorical features
- Creating based model using RFE and VIF for feature selection
- Fine tuned the model using hyper param tuning to arrive at model with 0.8 recall and 0.84 accuracy
- Convert output probability to lead score between 0 and 100
- We can use this model to target new leads to improve conversion

THANK YOU