LEAD SCORING CASE STUDY

Expectations

The following case study is primarily intended to facilitate a discussion on modeling choices and to have a tangible problem to discuss. As a guideline, you should expect to spend around 4 hours to complete this exercise. The assignment does not have to be completed all at once.

Once completed, submit your deliverables to HR team coordinator

Problem Description

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Additionally, the company also gets leads through referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X Education is around 30%.

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, the company nurtures the potential leads (i.e. educating the leads about the product, constantly communicating, etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads for follow-up, i.e. the leads that are most likely to convert into paying customers.

The Task

Build a model to assign a score between 0 and 100 to leads which can be used by the company to target potential leads. A higher score would indicate 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.

Data Supplied:

Leads Data Dictionary.xlsx:
Leads.csv

Deliverables:

- The code that you wrote to solve the problem. (R or Python is preferred)
- Anything that helps a reader understand how well the model will work (if needed)

• Answers to the questions below.

Questions: Please answer the following questions.

1. How long did it take you to solve the problem?
   Around 6-8 hours

2. What software language and libraries did you use to solve the problem?
   Software language -Python
   Libraries- sklearn, matplolib, seaborn, pandas, pickle, gc, numpy, xgboost

3. What steps did you take to prepare the data for the project? Was any cleaning necessary?
   - Missing value imputation-
     A) using mean, mode or a creating new category 'other' in some cases where variable is categorical
     B)missing value imputation of 'TotalVisits' and 'Page Views Per Visit' by creating 4 segments (<q1, q1-q2, q2-q3,>q3) using quartiles of 'Total Time Spent on Website' and finding mean of the variable which has to be imputed in these segments and then replacing missing values of variable to be imputed with these mean values
   - Removing categorical variables with only one category
   - Label encoding categorical variables having yes and no or having rank ordered categories
   - One hot encoding other categorical variables
   - Removing correlated features
   - Adding features or feature engineering (added 'how long' feature to measure the time difference between lead id creation for leads)
   - Visualization to understand how target variable is distributed w.r.t other variables
   - Feature selection of important numeric features using extratree classifier and select kbest library
   - Feature selection of important categorical features using cramer's V
   - Standardization using standard scaler

4. What algorithmic method did you apply? Why? What other methods did you consider?

   Tried XGBoost, Gradient Boost, MLP classifier, naïve bayes, decision tree, random forest, k nearest neigbours,logistic regression. Also tried voting classifier using mixed models
   XGBoost performed the best on test set.
   Used gridsearch to select best parameters of XGBClassifier

5. What features did you use? Why?

   Features used- Numerical['how_long',
   'Total_Time_Spent_on_Website',
   'Asymmetrique_Activity_Score',
   'TotalVisits',
   'Page_Views_Per_Visit',
   'Asymmetrique_Profile_Score',
   'Do_Not_Email',
   'A_free_copy_of_Mastering_The_Interview',
   'Search',
   'Through_Recommendations']

   Categorical ['Tags',
   'Lead_Quality',
   'Lead_Profile',
   'Last_Activity',
   'Last_Notable_Activity',
   'Lead_Source',
   'Lead_Origin',

'What_is_your_current_occupation',
'Specialization',
'City']

The numerical features were selected using extratree classifier's variable importance and the categorical features were selected using cramer's v score. The correlated features were dropped before selecting the final features for the model.

6. How did you train your model? During training, what issues concerned you?

I trained the model using cross validation with 5 folds using XGBoost, Gradient Boost, MLP classifier, naïve bayes, decision tree, random forest, k nearest neigbours,logistic regression. XGBoost performed the best on cross validation as well as test set (0.936 F1_score)
Also trained the model using voting classifier which enables use of mixed models. I selected XGB, Random forest, svm and Logistic regression. Though the model performed well on cross validation set (F1 score 0.937), it did not work so well on test set.
Hence went ahead with XGB and performed Grid search to identify best parameters.

7. How did you assess the accuracy of your predictions? Why did you choose that method? Would you consider any alternative approaches for assessing accuracy?

Used F1_score metrics to assess the accuracy. As it is important to identify hot leads with high precision and recall since a high recall ensures capturing all the actual Converted and a high precision ensures the right set of customers are being targeted (very few FP), hence F1-score seemed to be the right metrics to go ahead with.
ROC AUC score can also be used as an alternative as it indicates how good the model is fitting and how well it is able to predict the class of the lead.

8.Which features had the greatest impact? How did you identify these to be most significant?

Tags- Will revert after reading the email (0/1) One hot encoded
is the most significant variable. I identified this using feature_importances_ of final XGBoost model after training it.