Lead Scoring Case Study Summary

Summary

The objective of this case study was to help the sales team of X education identify & distinguish hot leads from the ones who might not get converted. They also would need to be given the top variables and their values to look out for while trying to assess if the lead will get converted or not.

The following process was performed in order to arrive at the list of leads & impacting features:

- Data Processing: data checked for outliers, duplicates & missing values. Few of the categorical variables with a very high number of levels were recoded. Two of the few numerical variables did exhibit very few cases of outliers which were handled by capping these variables to a particular upper limit as it would be incorrect to exclude these students from the study.The numerical features were scaled & dummy variables were created for categorical variables.
- EDA: A thourough EDA was performed for both the categorical & numerical variables to understand the distribution of conversion ratio across different levels of these variables. Variables which did not depict any difference in the trend of conversion ratio were discarded from the study.
- Correlations: This test was performed to understand the pairs of variables which were highly correlated and also removed from the set of features.
- Lead Scoring Model: The model chosen in this case was a logistic regression model which depicts the probability of conversion.

  This model was built in the following manner:
  - Build the model with the current set of variables
  - Check the VIF / p values & eliminate insignificant variables
  - Repeat above steps until all the features are significant!

  Once all the features utilized were significant, we can optimize the threshold value to arrive at the optimal precision & accuracy which suits the scenario.

  In this problem it is more important to have a higher precision than recall as it is more important to reduce number of false positives than false negatives as if you have a high number of false positives, that is, the model predicted leads who actually would not be potentially converted as a hot lead, then the counselors are actually wasting their time in concentrating on leads who are not beneficial and leaving out the actual hot leads. This might cause a huge loss in the overall conversion.

  As we plot the precision & recall across threshold, a threshold of 0.6 was found to give the best precision and a good recall value. The model has an overall accuracy of 88% on both the train and test data and we can thus say that the model is not overfitting to the train data and is performing well. As it was also required to understand the most impactful variables as well, we were able to identify the degree & direction of importance of the features

and arrived at the conclusion that the below variables and their levels are most impactful and the sales team has to look out for these signs:

- Tag
  - Tags_Already a student
  - Tags_Closed by Horizzon
  - Tags_Lost to EINS
  - Tags_Not doing further education
- What matters most to you in choosing a course
  - Better Career Prospects