Summary

X Education has a lead conversion rate of just 30%. The company has asked us to build them a model that has better conversion rate. We need to build a model that give lead score , higher lead score have higher conversion rate, lower lead score has lower conversion rate.

Steps taken:

Step-1: Loading and Understanding the data: The data has 9240 rows and 37 columns in the beginning.

Step-2: Cleaning the data: After cleaning the data we are left with 9240 rows and 21 columns. (we dropped columns that have so many null values which can effect analysis).

Step-3:Exploratory Data Analysis: It helps us to understand the trends in the data by visualizing it . We can clearly see the difference or relation in the variables in this step. Below are the graphs or plots of the analysis that I have performed. It will help us to identify the required and unrequired parts quickly .

Step-4:Dummy Variables: If there are many types in single columns like lead source: Google, direct,etc. We seperate them using dummies. It helps for easier analysis.

Step-5: Test-Train split and Building model: Here, We use a part of data to train the model. Then we use the rest of the part for evaluating the model.

Step-6: Model Evaluation: We use several method to check the efficiency of the model. Here, we can also identify important variables and keep only them to imcrease the accuracy of the model.

Step-7: Making prediction: Here, we ask the model to make prediction to see how well it will align with the actual data to test the model further.

The graph will tell you how the final prediction is aligning with the actual data.

If you ask me it works pretty well. Im not saying this because I made it though.

The Observations are made from the model:

1. accuracy of train set is 79.6%, test set is 75.9% difference is around 0.4
2. Sensitivity train set is 78.6%, test set is 78.9% increased by 0.3%
3. Specificity(TN) of train: 82.2%, test: 83.6%
4. False Positive(FP) of train: 17.7%, test: 16.3%
5. True negative(TN) of train: 86.2%, test: 85.9% decreased by 0.3%
6. 6.precision of train: 79.6%, test: 75.9%, decreased by 3.7%
7. recall: train: 69.8%, test: 78.9%, increased by 9.1%t

Overall the high recall and precision values indicates that it identifies positive cases while maintaining good number false

positives.High specificity and low false positive rates futher validate the model's reliability. Overall its a good model

Conclusion:

1)Important variables that contribute to the increase in probability:

1. Total time spent on Website
2. Total Visits
3. lead source with google

2) 3 top categories/ dummy variables :

1. Lead source with Google
2. Lead source with direct trafic
3. Lead source with organic search

3) Call Leads if:

1. If they spend a lot of time on website. We can see people who spend a lot of time on website are mostly converted successfully
2. . If Repeatedly comes back to website. There is a high possibility of them converting into successful leads.
3. If their last activity is through E-mail or SMS or Olark chat conversation.
4. If they are Working professionals or unemployed.

Since most people don't like to receive calls as well there are other ways to reach. We can send them E-mails, as most of strategies successfully working.

Suggestion:

1. It would be better if the budget is spent on developing the website rather than using it on advertisement such as newspaper articles,etc.
2. Giving out discounts for customers for spreading word of mouth would be a better advertisement.
3. Use Google SEO to make the website more visible.
4. Target the working professional as they have high conversion rate.