Logistic Regression Assignment Submission

Lead Scoring Case Study

Problem Statement and Goal

Problem Statement :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these 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. Moreover, the company also gets leads through past 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%.

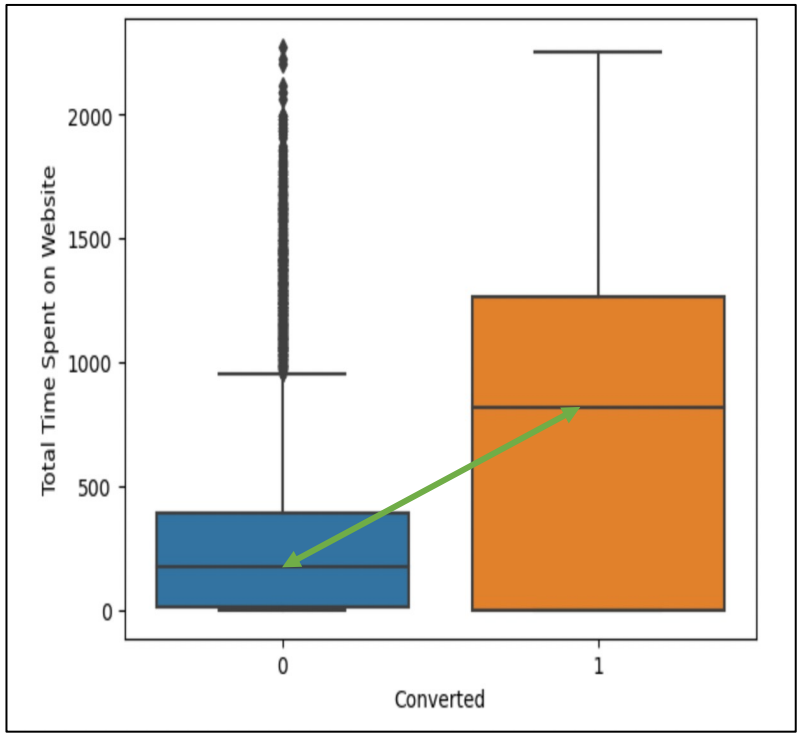
Business Goal :

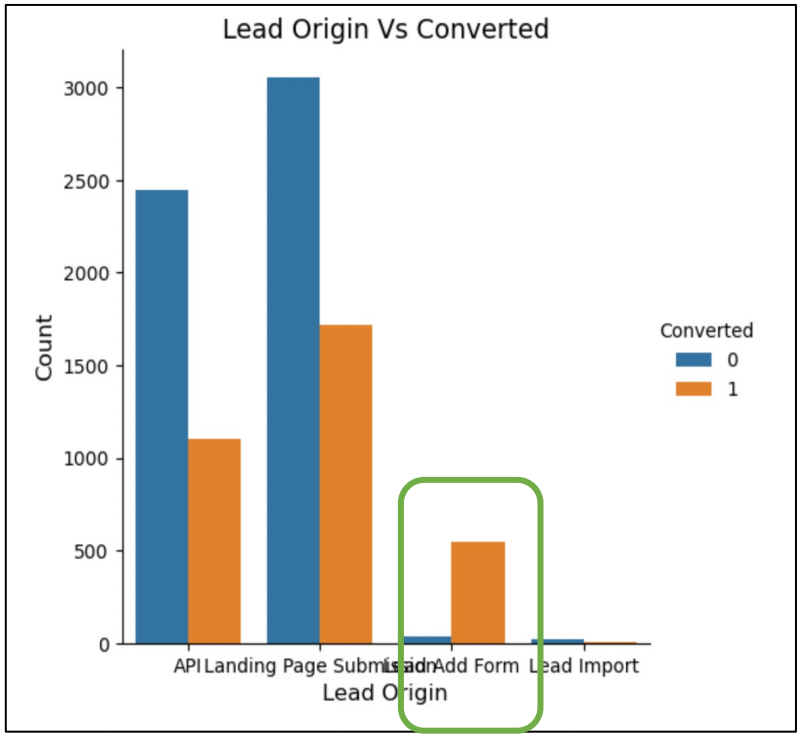
X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the 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%

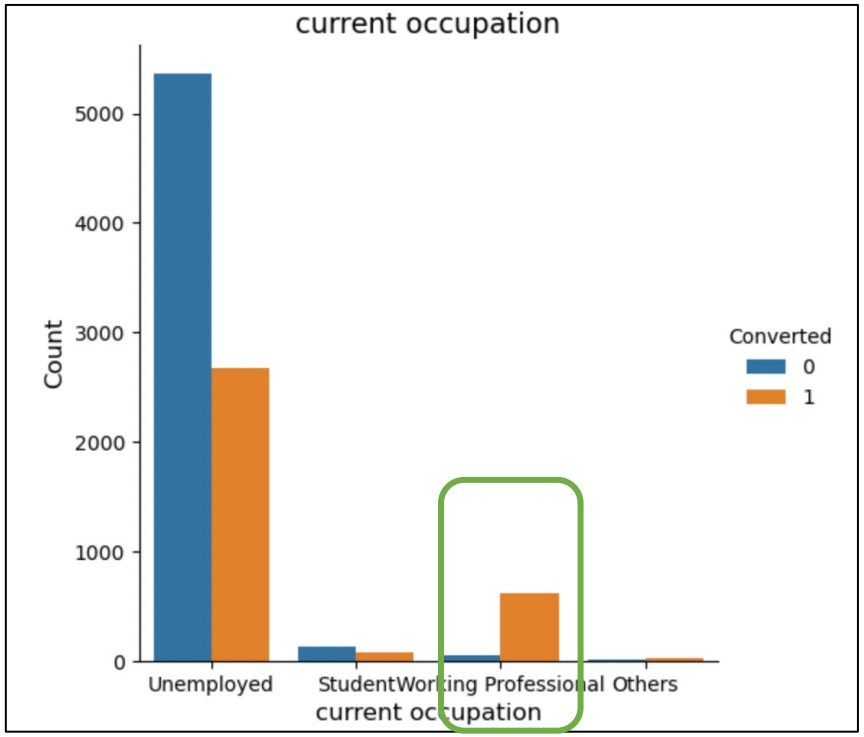
Solution Steps

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis
- Feature Scaling
- Splitting the data into Test and Train dataset
- Building a logistic Regression model and calculate Lead Score
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Bivariate checks



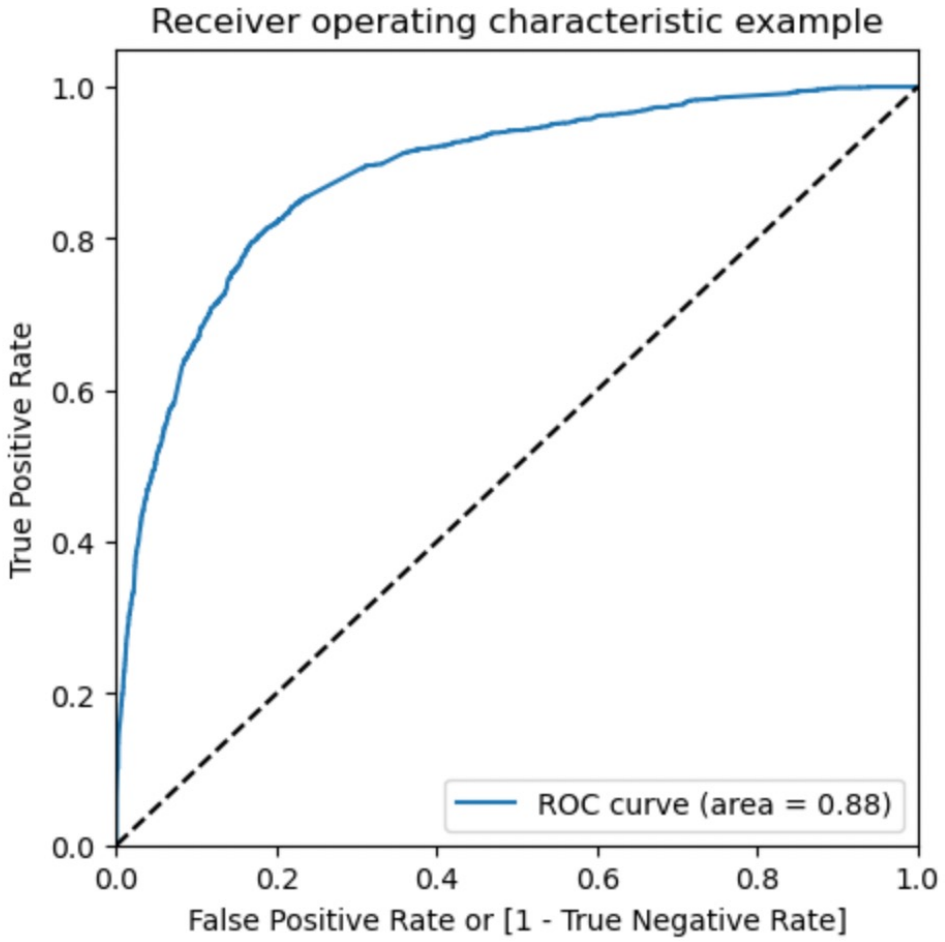


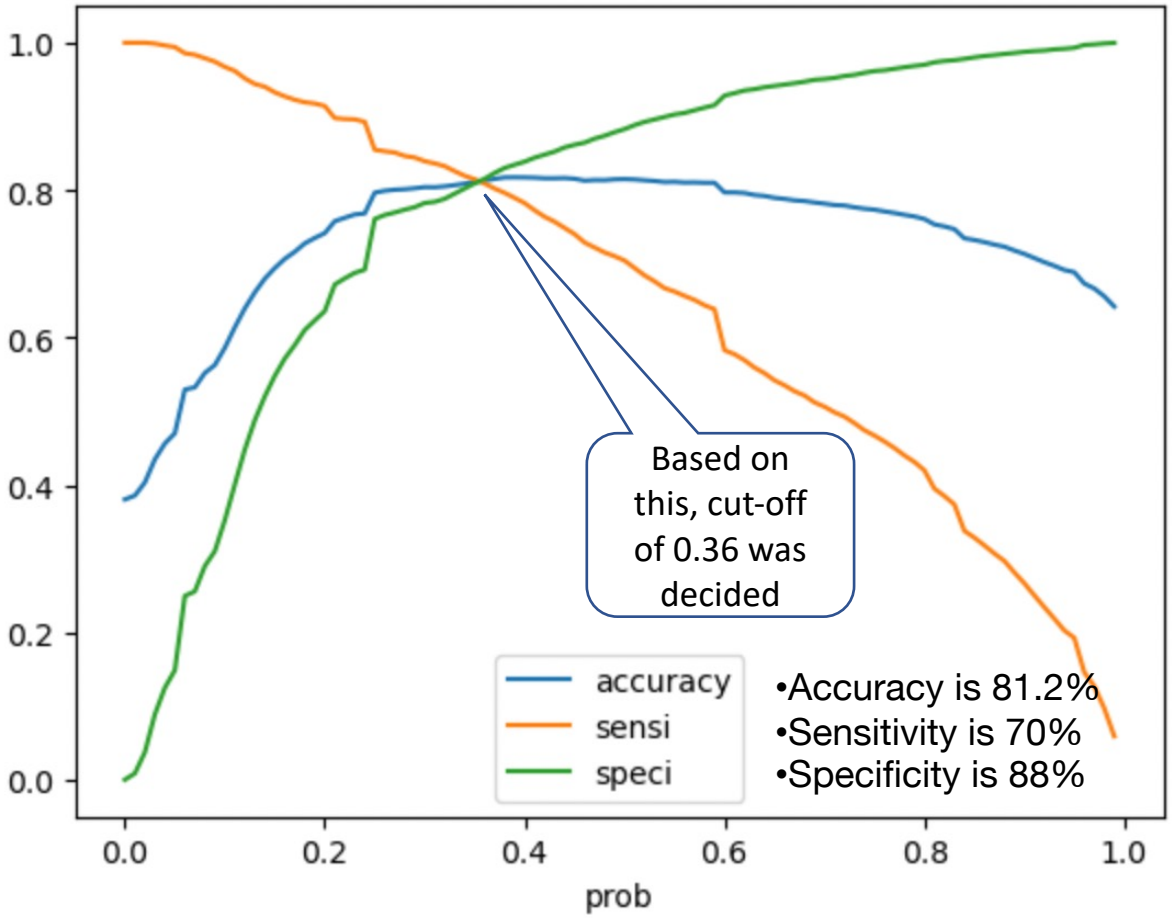


Final list of variable after 7 regressions (removal of variables)

	coef
const	-0.8958
Do Not Email	-1.3639
Total Time Spent on Website	1.1213
Lead Origin_Lead Add Form	3.9107
Lead Origin_Lead Import	1.3961
What is your current occupation_Unemployed	-0.4107
What is your current occupation_Working Professional	2.3205
Last Activity_Converted to Lead	-1.1256
Last Activity_Email Bounced	-0.9049
Last Activity_Had a Phone Conversation	3.0742
Last Activity_Olark Chat Conversation	-1.4173
Lead Source_Olark Chat	1.1569
Lead Source_Welingak Website	1.8619
Last Notable Activity_Modified	-0.2299
Last Notable Activity_SMS Sent	1.5231
Last Notable Activity_Unreachable	1.7727

Train Set: ROC and Sensitivity, Specificity and Accuracy





- Accuracy is 81.2%
- Sensitivity is 70%
- Specificity is 88%

Test Set: Sensitivity, Specificity and Accuracy

Confusion Matrix

```
[[1290,  394],
 [ 173,  821]]
```

Accuracy	Sensitivity	Specificity
79%	83%	77%

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 80%
- The top 3 variables that contribute for lead getting converted in the model are
  - Total time spent on website
  - Lead Add Form from Lead Origin
  - Had a Phone Conversation from Last Notable Activity
- Hence overall this model seems to be good.