Lead Scoring Case Study

Submitted by:
Ishan Agrawal
Shyam Sundar
Akanksha Gupta

Problem Statement

An education company named 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%

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Solution Approach

This is clearly a Classification problem. We have to find the leads that are having high probability of conversion/buying the course. High probability data points can be given high Lead Score hence we can find out all the possible ‘HOT LEADS’.

Here, we will use Logistic Regression and solve the problem.

Metrics- More attention will be given to SENSITIVITY according to the business needs.

Let's have a lot further!!

Steps in project -Part 1

1. We started with simple dataset, information, description.
2. Did dataset cleaning like treating missing rows, some columns have SELECT value, replaced according to needs.
3. After data cleaning, plotted visualisations to see how different columns relate to output label- converted/not-converted.
4. Now, model building process starts.
5. Used RFE for automatic feature selection(15 features)
6. Afterwards, built model, looked at accuracy.
7. Used VIF & p-value for feature elimination leading to better model.

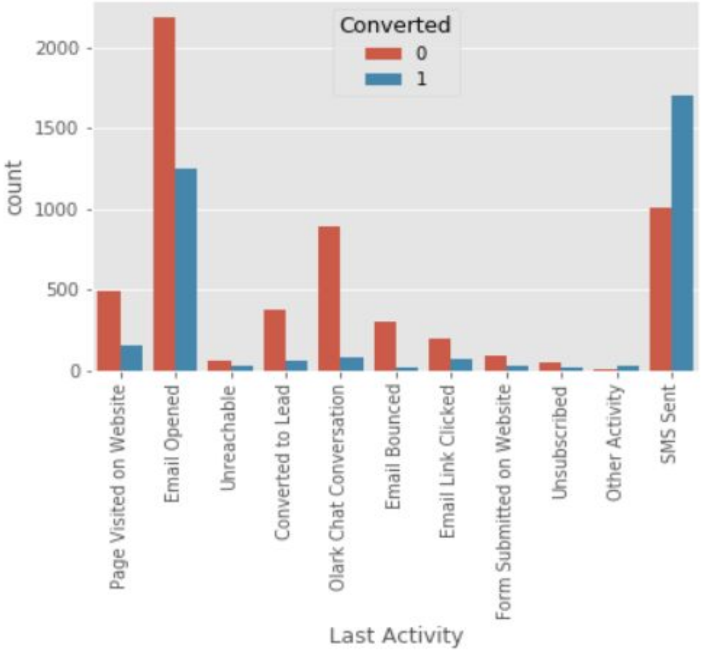
Steps in project -Part 2

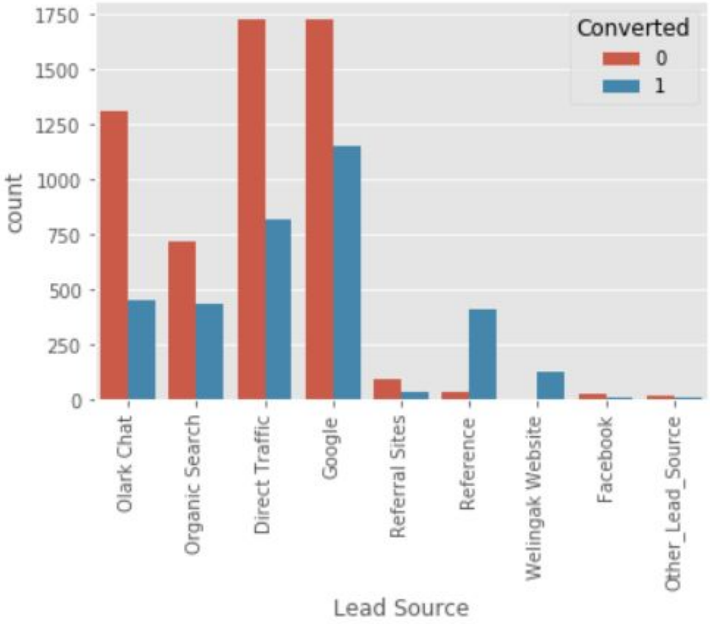
8. After proper feature selection, continued with finding the right cut-off for best sensitivity.

9. Also saw ROC curves for trade-off between sensitivity and specificity.

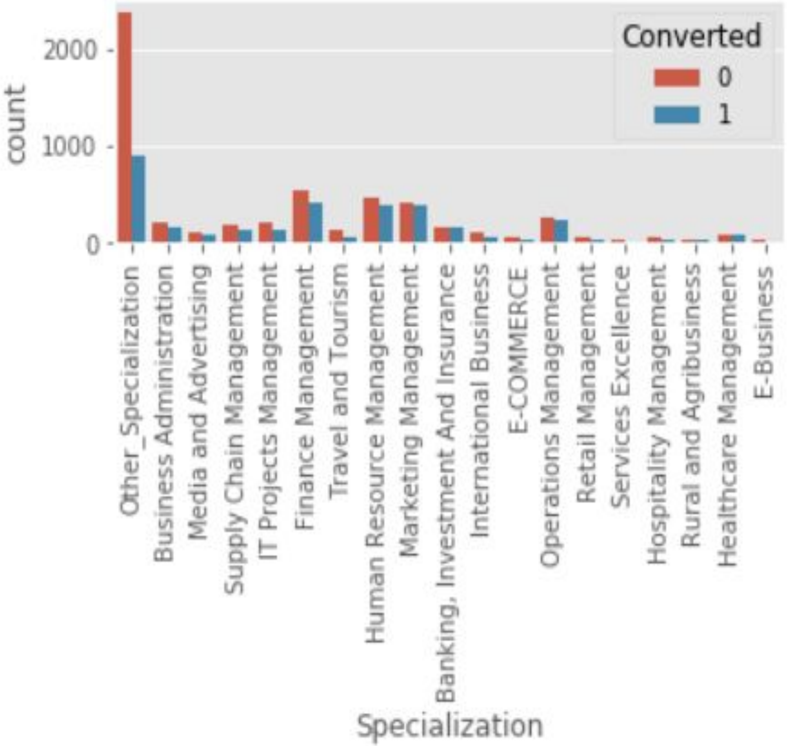
10. Also last got good model.

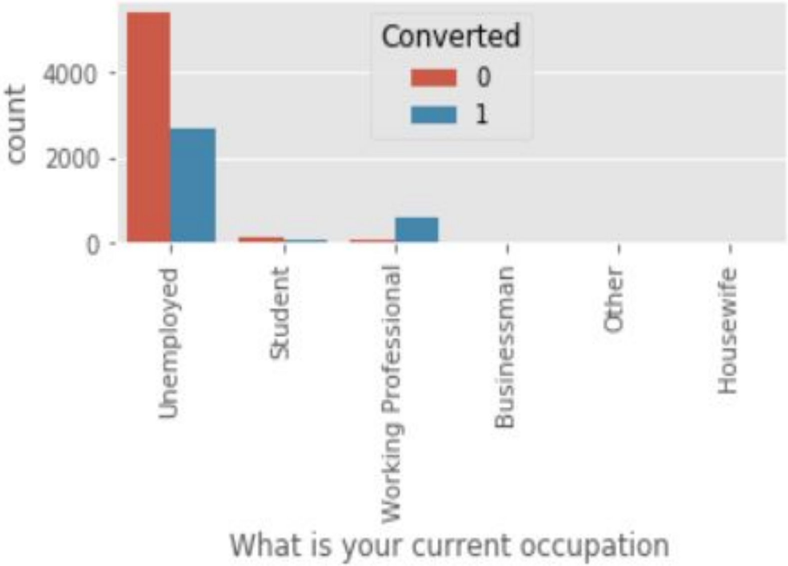
Lead Activity/Source relation to conversion



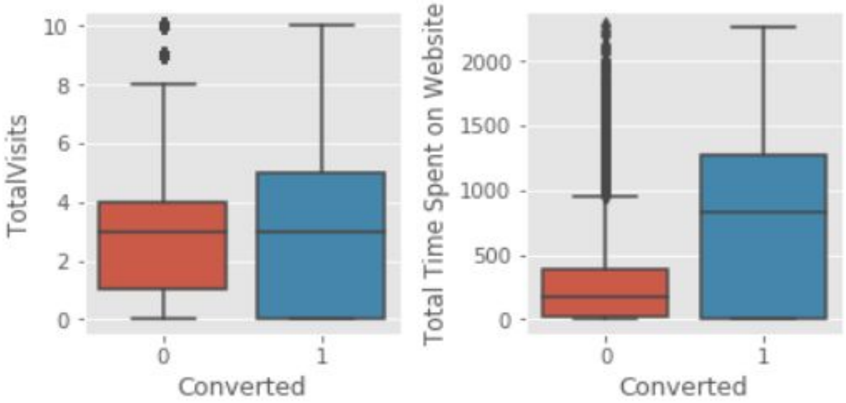


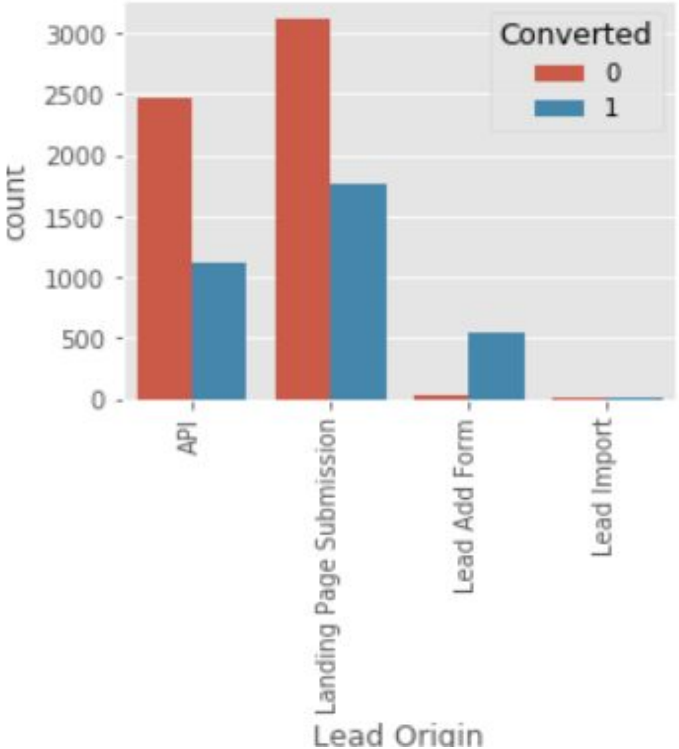
Customer’s Occupation and Course Specialization relation



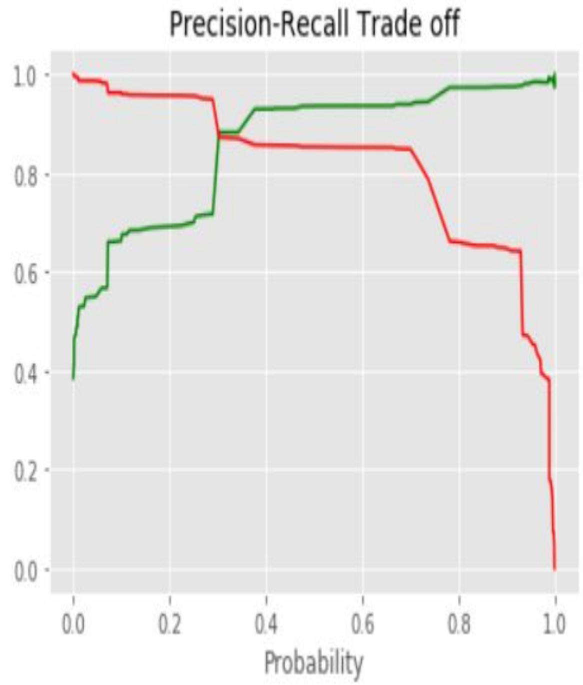


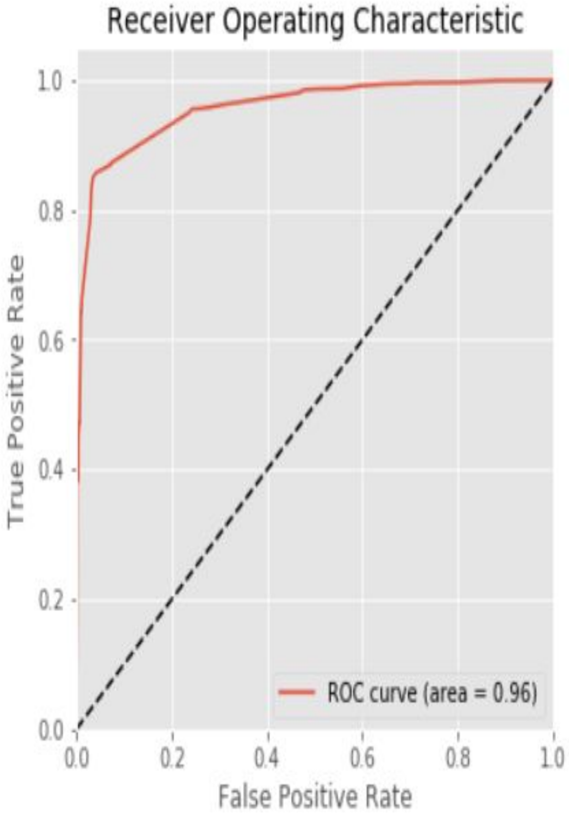
Total Visit, Total time spent on website, Lead Origin

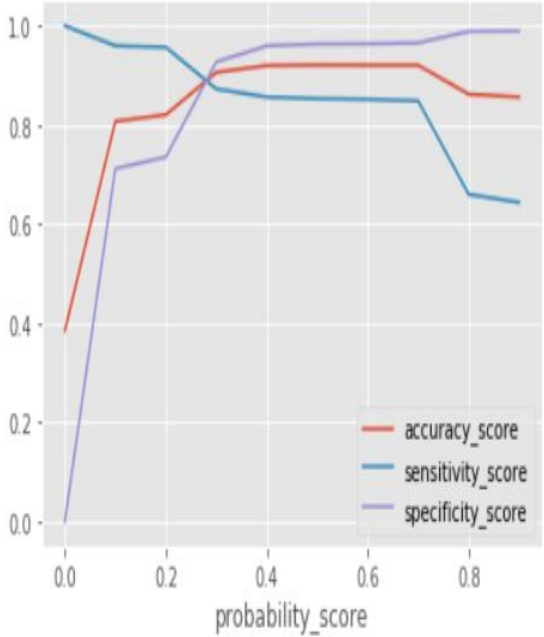




Trade-off graphs







Conclusion

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
- Here, the logistic regression model is used to predict the probability of conversion of a customer.
- Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert)
- Our final Logistic Regression Model is built with 14 features.

Wishing best for X-education!!!

THANK YOU!!!!