Lead Score Case Study

Submitted by :
Sakshi Munde

Lead score case study for X Education

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%.

Strategy

1. Problem Statement
2. Methodology
3. Exploratory data analysis
4. Model Evaluation
5. Identifying the most promising leads
6. Conclusion
7. Recommendation

Methodology

1. Data cleaning and preparation

• The data was cleaned by dropping unnecessary columns and imputing missing values with mode or 'unknown'. Outliers were removed from numerical columns using boxplots.

• Exploratory data analysis was performed to identify patterns and trends in the data.

• Data was prepared by binary conversion of data and creating of dummy variables for categorical columns.

2. Feature scaling and splitting data into train & test sets

- Feature Scaling is done on numeric columns. Data is splitted into train and test sets.

3. Model Building

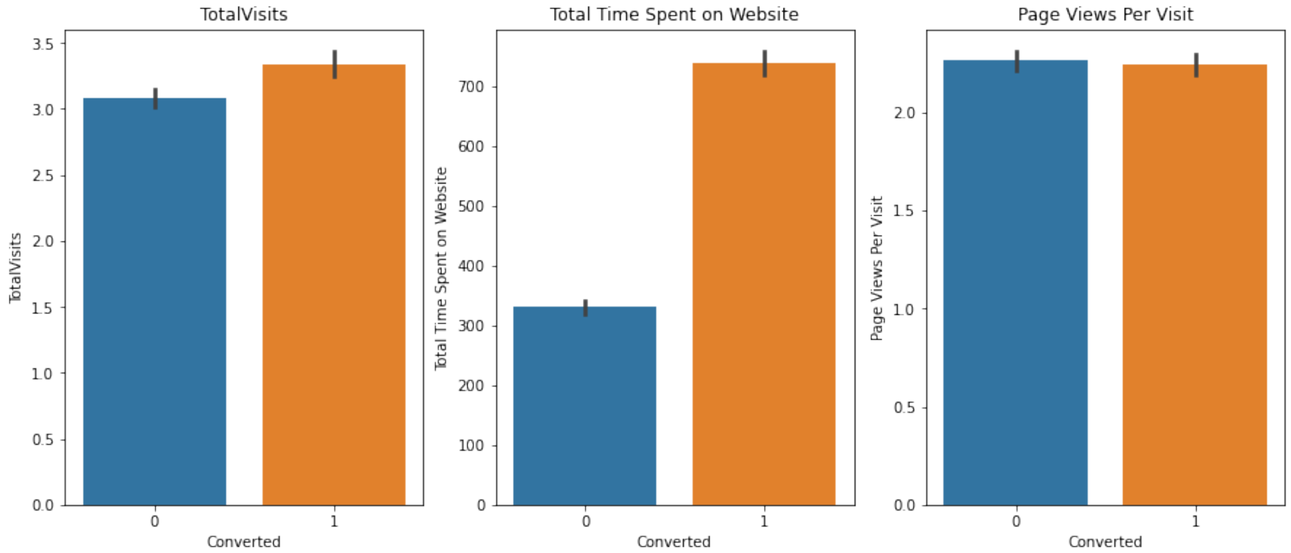
- Feature selection was performed using RFE and manual selection based on p-values and VIF values.
- The model was evaluated on the training dataset using various performance metrics such as accuracy, sensitivity, specificity, and confusion matrix. The optimal cut-off point was determined using sensitivity.
- ROC curve was used to evaluate model performance with an AUC of 0.89.
- Predictions were made on the test dataset after applying the same preprocessing steps as the training dataset.

4. Result

- The accuracy of the model is 81%, which means that it correctly predicted the target variable 81% of the time.
- Sensitivity is 80%, which is a good indication that the model is able to identify the majority of actual positive cases.
- Specificity is 81%, which indicates model has correctly predicted actual negative cases.

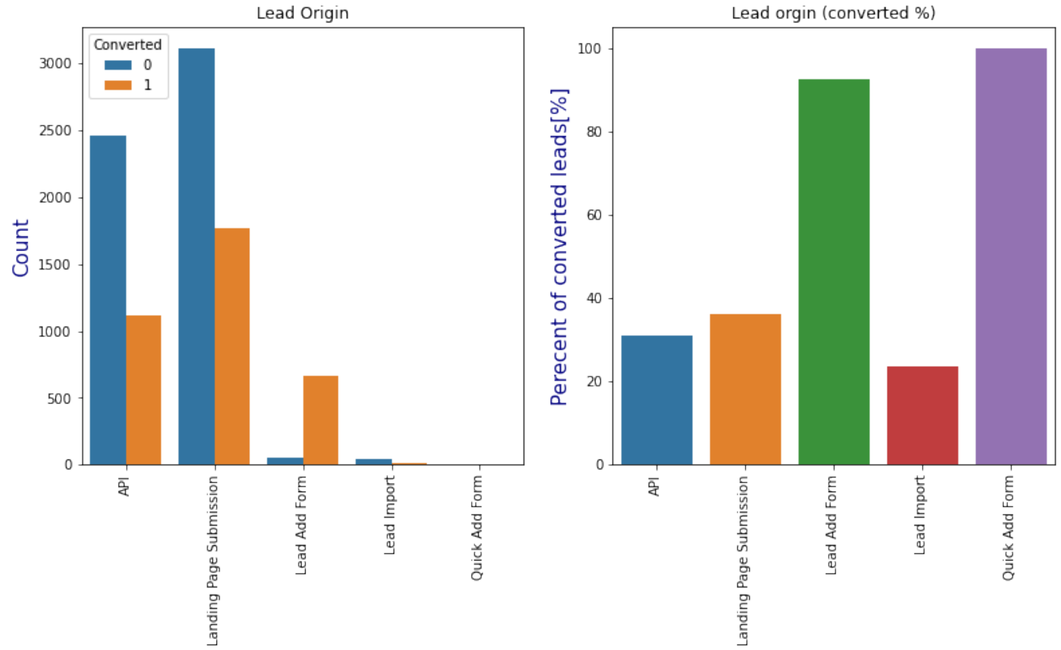
Exploratory data analysis

Numerical variables



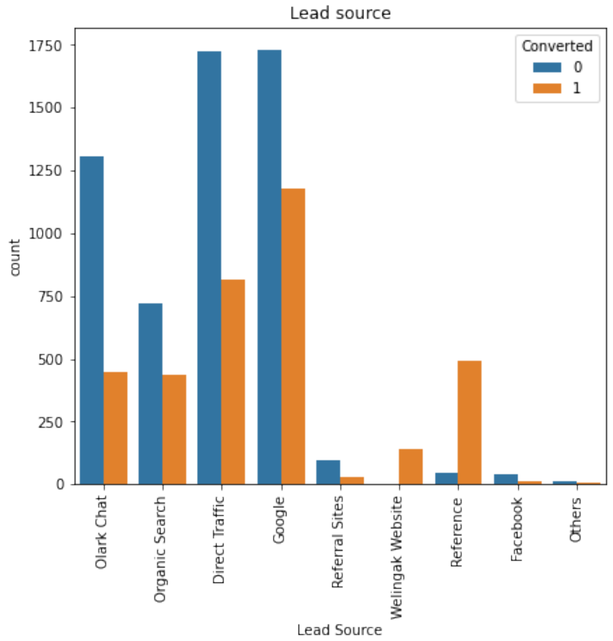
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.

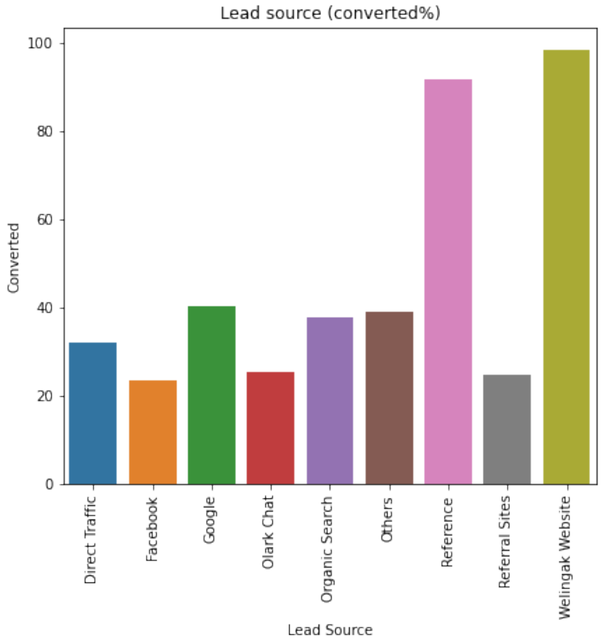
Categorical variables : Lead origin



Most of the leads are originated from "landing page submission" and “API”. The "Lead add form" has a very high conversion rate of 92%.

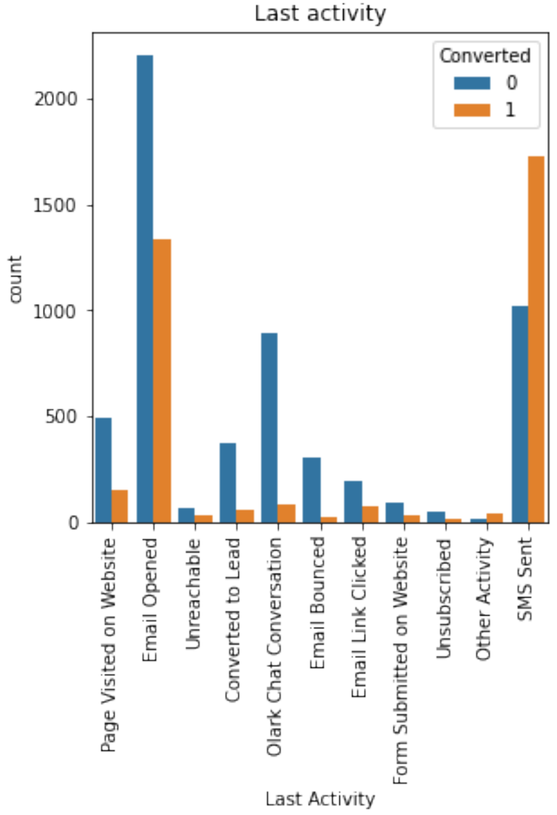
Lead source

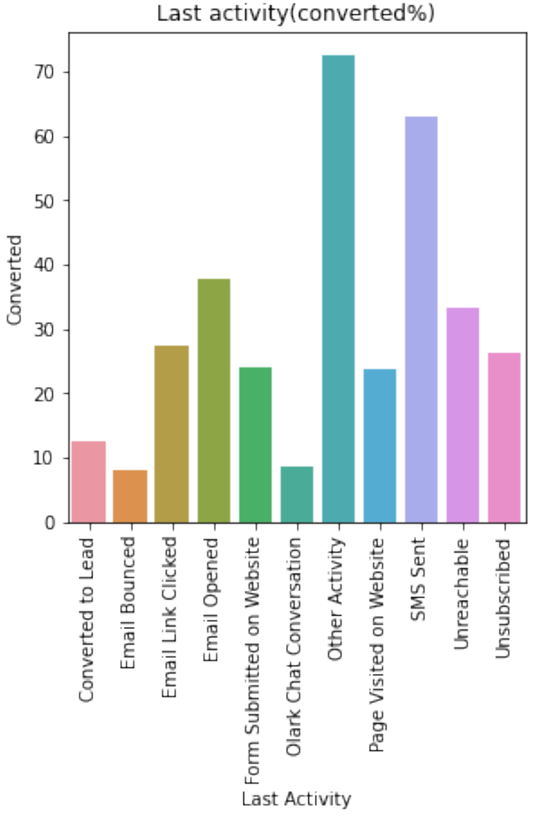




Focus on improving the overall conversion rate of leads from sources like google, olark chat, direct traffic, organic search as they generate significant number of leads.

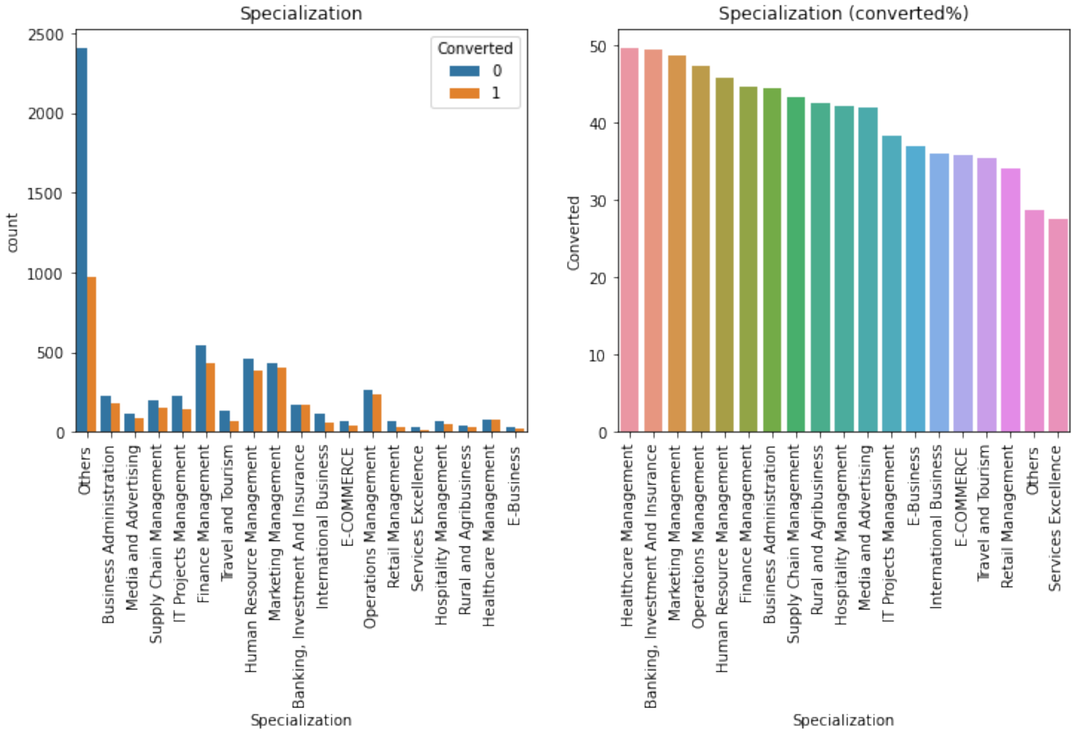
Last activity





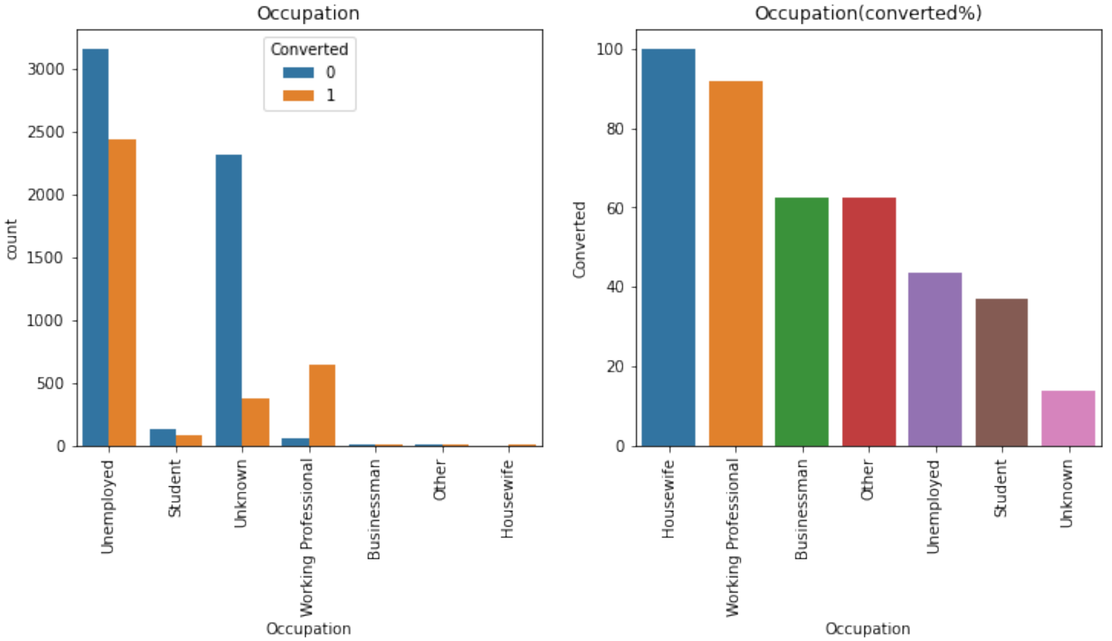
Conversion rate of leads with last activity as 'SMS sent' is almost 60%.

Specialisation



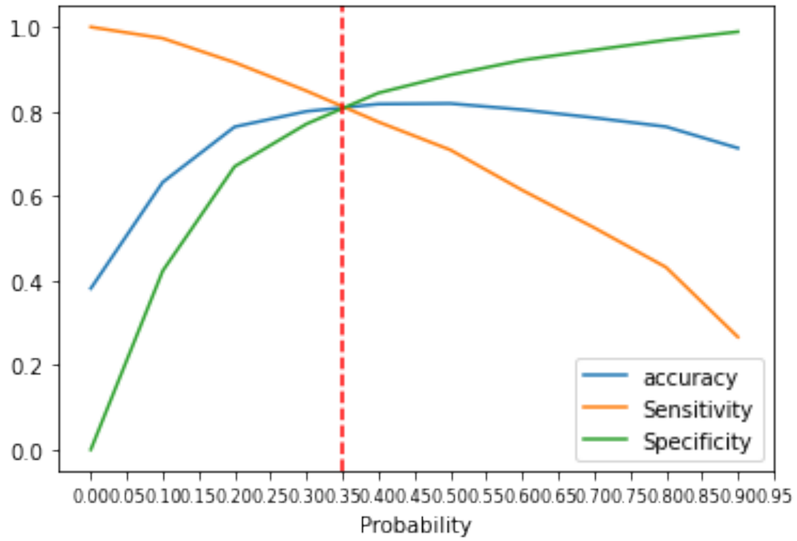
- Leads with Finance Management ,Human Resource Management & Marketing Management has high and their conversion rate is around 45%.

Occupation

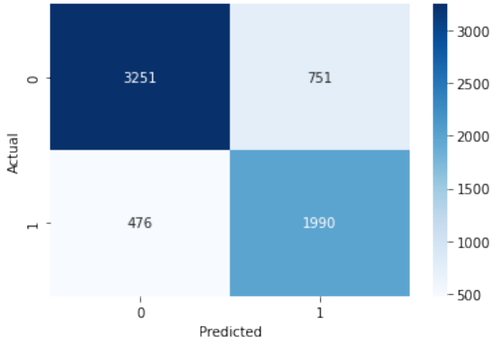


- Most of the leads are from Unemployed category but their conversion rate is very low so we should focus more on unemployed leads for conversion.
- "Working professionals ,businessman and Other" occupation leads are having high conversion rate.

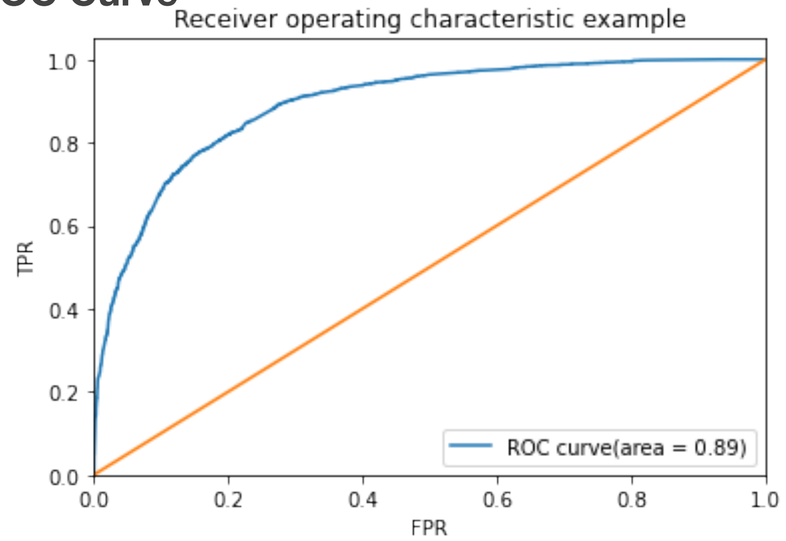
Model Evaluation - Sensitivity and Specificity on Train dataset



Confusion Matrix



ROC Curve



- 0.35 is optimal cut-off point.
- The accuracy of the model is 81%, which means that it correctly predicted the target variable 81% of the time.
- Sensitivity is 80%, which is a good indication that the model is able to identify the majority of actual positive cases.
- Specificity is 81%, which indicates model has correctly predicted actual negative cases.
- The precision shows that out of all the predicted positive cases, only 72% were actually positive.

Identifying the most promising leads

- Do Not Email,
- Total Time Spent on Website
- Lead Origin Landing Page Submission
- Lead Origin Lead Add Form
- Lead Source Olark Chat,
- Lead Source Welingak Website,
- Last Activity Email Opened ,
- Last Activity Olark Chat Conversation,
- Last Activity Other Activity.

This are the features that contribute for lead getting converted.

- Total time spent on website, lead origin landing page submission, lead origin lead add form, lead source Clark chat etc, has higher chance of conversion.
- Leads who spend more time on website are more likely to convert and interested in the courses offered.Similarly, leads who submit a landing page form or a lead add form are indicating a stronger interest in the courses compared to leads who only browse the website. Leads who come from olark chat or welingak website are likely to have a more personalised experience and therefore more likely to convert.

Conclusion

- In conclusion, X Education's challenge of selecting the most promising leads for conversion into paying customers can be addressed using the lead scoring model developed in this project. The methodology involved data cleaning, exploratory data analysis, feature scaling, and model building feature selection using Recursive Feature Elimination and manual selection based on p-values and VIF values.

- The performance of the model was evaluated on the training set with an accuracy of 81%, sensitivity of 80%, and specificity of 81%. The Roc curve was also used to evaluate the model performance, which resulted in an AUC of 0.89. The predictions made on the test dataset after applying the same preprocessing steps as the training dataset also demonstrated a high accuracy.

Recommendation

- Prioritise leads generated through the 'landing page submission' and 'API' channels as they have a higher conversion rate. Allocate more resources towards this channels to generate more leads.

- Optimise the 'don not email' feature to ensure that only interested leads are contacted. This can be achieved by sensing targeted and personalised emails instead of generic ones.

- Improve the overall user experience of the website to increase the 'total time spent on the website'. This can be done by providing easy navigation and relevant content.

- Focus on leads generated through 'reference' and 'welingak website' as they have a higher chance of converting. Leverage customer testimonial and success stories to encourage more referrals.

- Prioritise lead generated from' working professionals' as they have a higher propensity to convert. Offer tailored courses and flexible payment options to appeal to this segment.

- These recommendations, if implemented effectively, can help X Education increase the conversion rate of leads into paying customers and accede its business goal.