Lead Scoring Case Study

Submitted by:
Abhishek Verma
Amit Kumar Gupta

Problem Statement

Help X Education in selecting the most promising leads (that are most likely to convert into paying customers) by building a model which will assign a lead score to each lead such that customers with higher lead score have a higher conversion rate and the one with lower lead score have a lower conversion rate, with target lead conversion rate around 80%.

Methodology

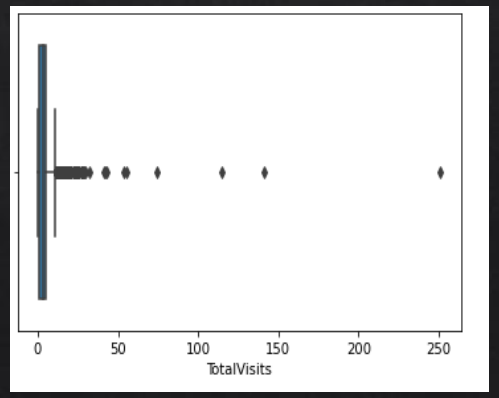
Based on the provided leads dataset from the past having datapoints related to attributes like Lead Source, Lead Origin, Total Time Spent on Website, Last Activity, etc. we have built a Logistic Regression model which will help us in defining the important attributes in deciding whether the lead will be converted or not. The target variable in this case is the column "Converted" which tells us whether the past lead was converted or not.

Assumptions: In the dataset many of the categorical attributes (City, How did you hear about X Education, Specialization, etc.) had a level called 'Select' which have been treated as a null value, based on the fact that due to some reason the customer didn't fill those fields and by default the value was stored as 'Select'.

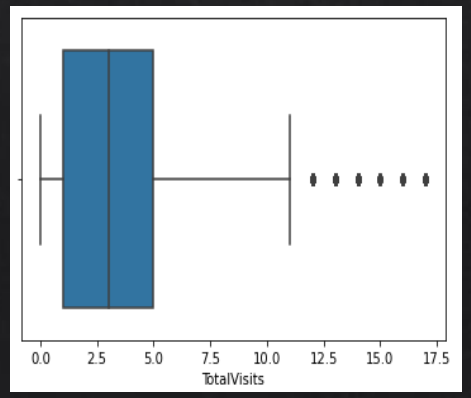
Data Sanity Check

- The attribute ‘Lead Source’ had various levels out of which 4 of them (Google, Direct Traffic, Olark Chat and Organic Search) were contributing towards more than 80% of data points. Therefore, rest of the levels have been clubbed as ‘Others’ so that they are not overshadowed by the dominant ones.
- Attributes like Country, Specialization, How did you hear about X Education, What is your current occupation, What matters most to you in choosing a course and City were not considered for model building as 40% or more data points were missing, along with them attributes which indicate whether the customer has seen the ad or not were also dropped as very few customers have seen the ad.
- Columns added by the sales team (Tags, Lead Quality, Lead Profile, etc.) were also dropped as they are totally based on knowledge and intuition of sales team and are of no help in model building.
- Numerical columns having missing values were imputed with their median values as the data was skewed. The attributes Page Views Per Visit and Total Visits were soft capped to 99 percentile as they have really extreme values as outliers

```
Google               31.428571
Direct Traffic       27.521645
Olark Chat           18.993506
Organic Search       12.489177
Others                9.567100
Name: Lead Source, dtype: float64
```



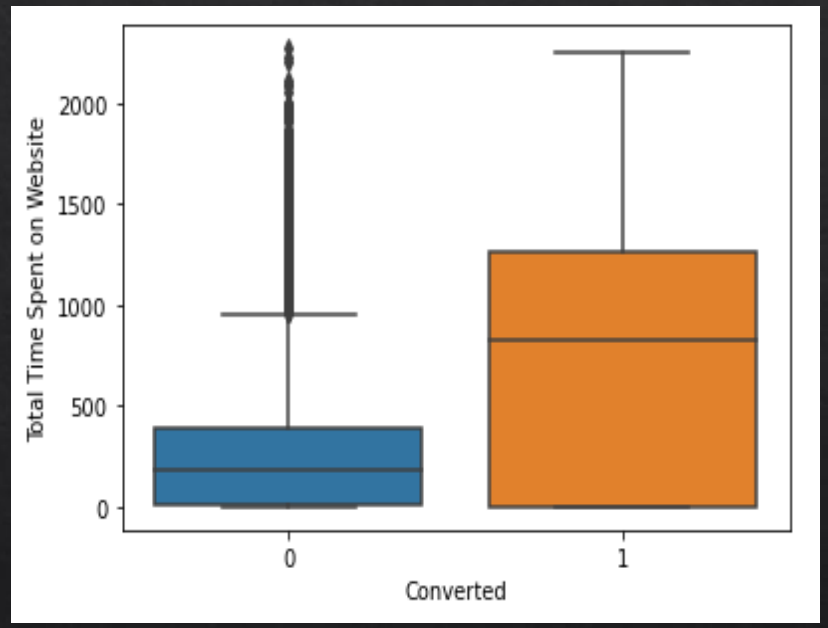
Outlier treatment



Model Results

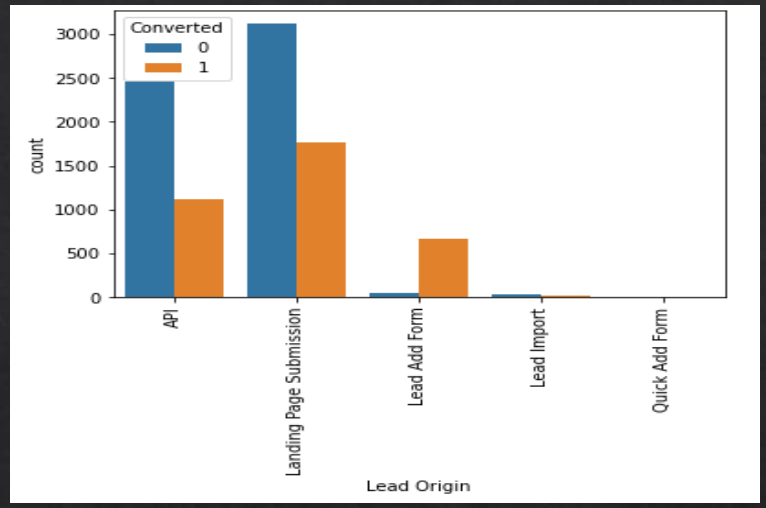
The model shows that the people who spend most time on the website have higher probability of getting converted.

As enrolling for a course is like an investment for the students so before making a call, they spend a significant amount of time on the website, probably to gather more insights regarding the course they want to enrol for.

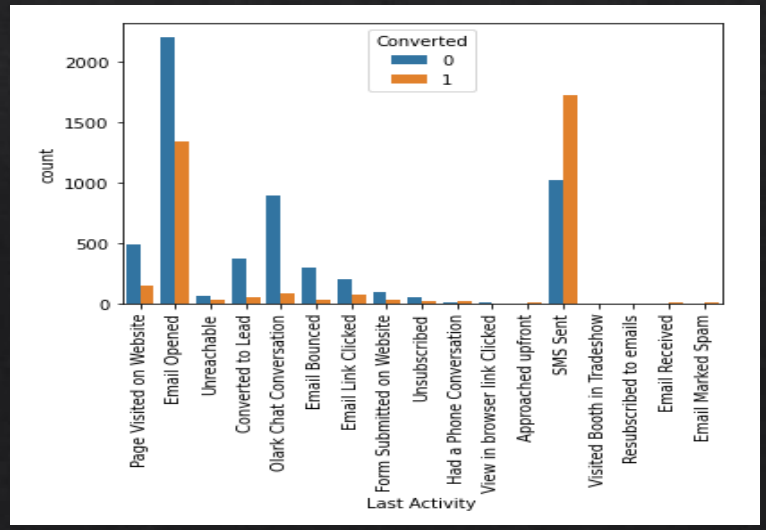


Model Results

The leads being generate from the “Lead Add Form” also have a higher probability of getting converted. Conversion rate from 'Lead add Form' is very high. Since we want to get the hot leads. This seems to be a good candidate for that. However, we see that the percentage of leads through 'Lead add Form' is low.



Based on the last activity performed by the customer we can say that “SMS Sent” resulted in more conversions. So, when the last activity is SMS-sent, user is very likely to convert.

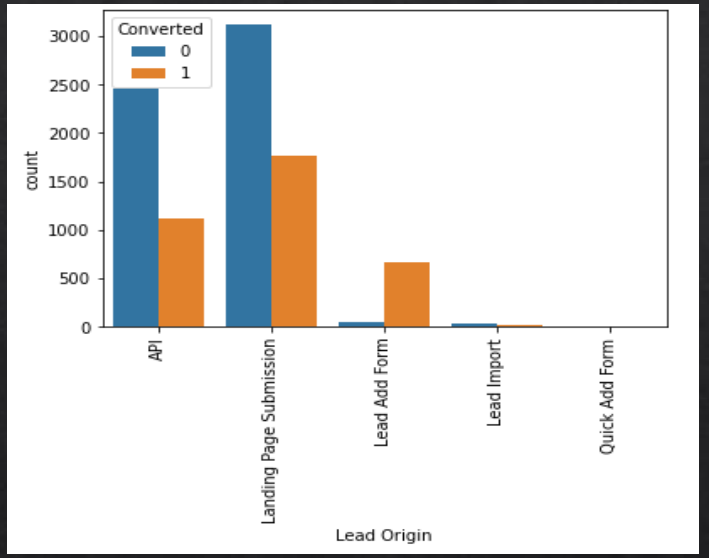


Recommendations

Conversion rate from 'Lead add Form' is very high. Since we want to get the hot leads. This seems to be a good candidate for that. However, we see that the percentage of leads through 'Lead add Form' is low.

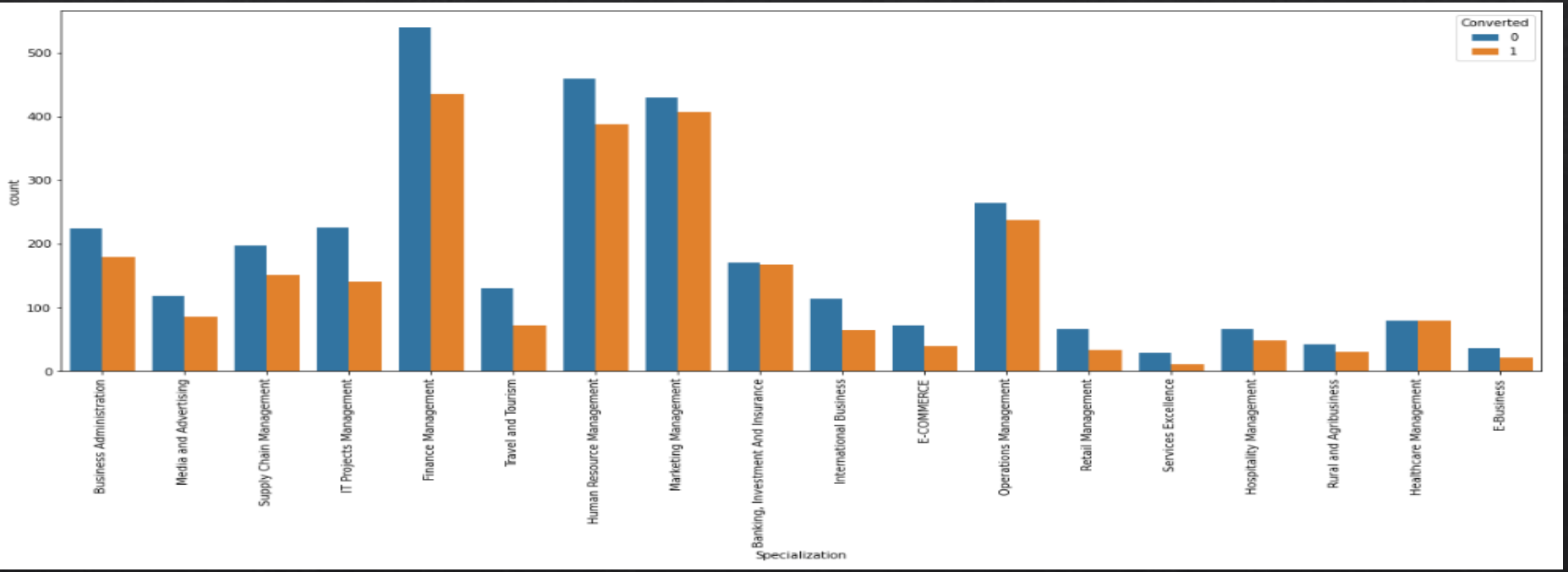
Therefore, during the two months times when the company hires interns, and they wish to make the lead conversion more aggressive the company can take these actions based on these results:

1. Use marketing to increase leads through 'Lead add Form'
2. API and Landing Page Submission has very high percentage of leads, but the conversion is low. Company may try to improve the lead conversion from these 2 Lead origins.



Recommendations

Most of the specialization have similar conversion ratio. However few specializations have better conversion like Healthcare management and Marketing Management. More focus should be given on these specializations. Although percentage of leads with Healthcare management is very low. Therefore, during the time when company reaches the targets for a quarter and company wants the sales teams to focus on some new work then they can build a strategy on how to get more leads from this sector by circulating surveys and analysing their needs.



Conclusion

The model built was able to predict more than 80% of leads which were converted. Also, with the help of this model we can see that the attributes like ‘Total Time Spent on Website’, ‘Lead Origin’ and ‘Last Activity’ were among the top 3 to contribute towards the probability of lead conversion.

The company can still work on generating more leads from ‘Lead origin’ and by targeting promising specializations like Healthcare management and Marketing Management.