Lead Scoring Case Study

By Siddhartha Puttuguntala

Technical Approach For Solving Business Problem

- Business Understanding
- Data Sourcing
- Data Cleaning & Manipulation
- EDA
- Data Vizualisation
- Data Preparation
- Model Building
- Model Evaluation
- Conclusions, Observations & Recommendations

Data Cleaning

Handling 'Select' variable

- "Select" variable indicates that the user has not selected any option.
- We impute the same with null values.

Dropping Score and Activity variables

- Score and Activity variables : This is the data that is obtained after contact with the lead. So we need to remove them.
- Score variables: Tags, Lead Quality, Lead Profile, Activity Index, Activity Score and Profile Score.
- Activity variables: Last Notable Activit

Treating Categorical data

- High Data Imbalance – Columns having high data imbalance must be removed. For e.g. : Category A has 98% , and Category B has 2% - This data is irrelevant to our analysis as one category is overpowering the other.
- In other categorical columns where there are columns with small percentages should be removed

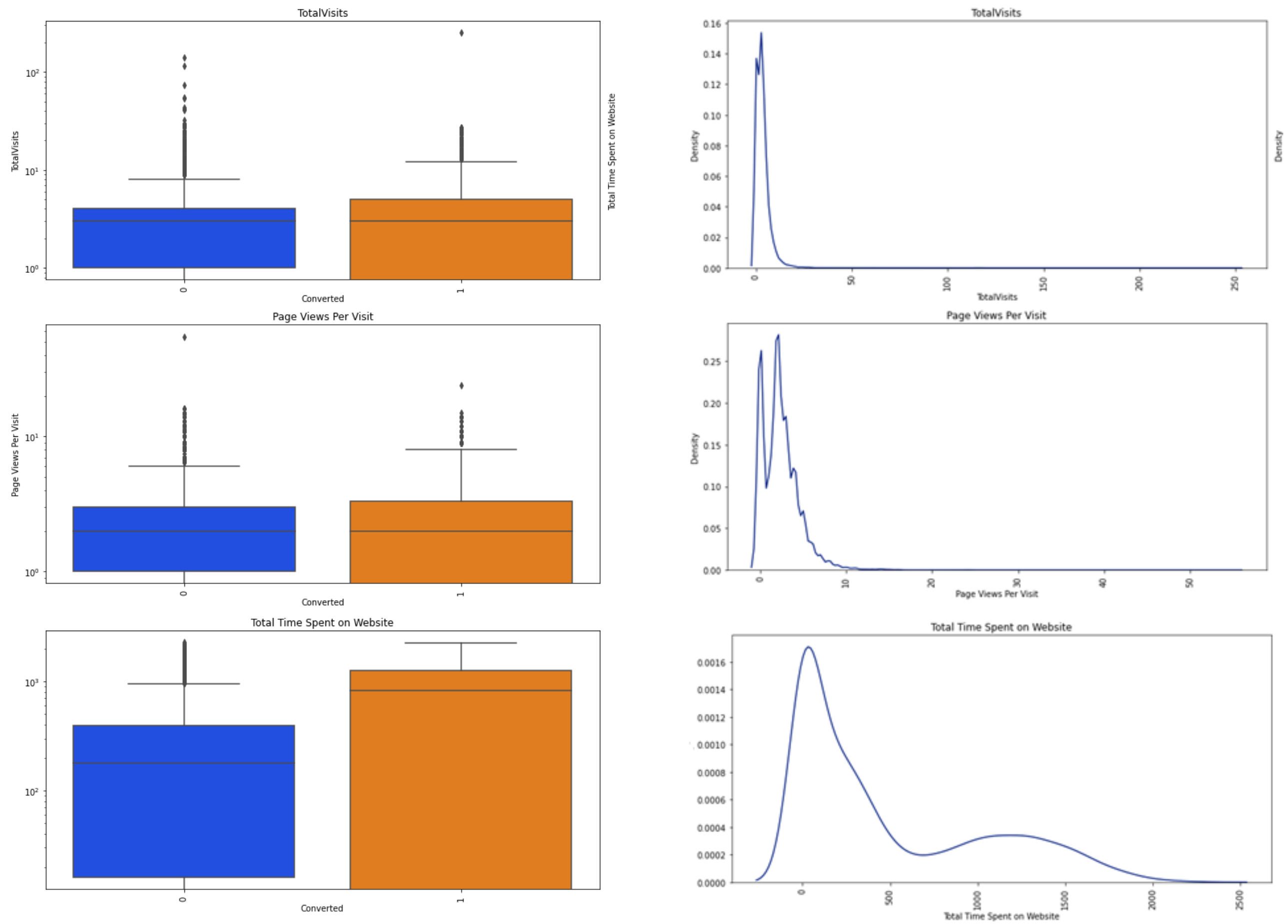
Dropping column with high null values

- Columns having null values greater than 40% does not have meaning to the data, hence we drop these columns
- For Specialization, we consider the column where people have not selected any value into one more column known as Not Specified and we use this for model building.

Final list of features

- Lead number
- Lead Origin
- Lead Source
- Total Visits
- Total Time Spent
- Page Views per visit
- Last Activity
- Specialization
- Current Occupation
- Free Copy of Book

EDA : Numerical Data



Total Visits

- The max probability for Total Visits is found to be around 15-20. It increases initially but decreases further.
- The average total visits for both converted and non converted people is found to be the same.

Total Time Spent On The Website

- The probability of time spent is found to be high for time between 0-300 seconds and decreases further.
- The mean is found to be higher in case of Converted people rather than nonconverted people.

Page Views Per Visit

- The max probability for Page Views Per Visit is found to be around to be 3-5.
- The average page views for both converted and non converted is found to be the same.

EDA : Categorical Data

- Number of features after scaling and dummy variable creation : 35
- Target Variable : Converted o Libraries used: StandardScaler()
- Columns that are not considered : Lead Number and Prospect ID (these variables do not help in model building)
- The steps are as follows:

Outlier Treatment:

- Total Visits and Page Views Per Visit had some outliers.
- We perform capping using Soft Capping (Checking for 99th percentile) and complete the outlier treatment process before we continue to the next step.

Binary Mapping:

- “A free copy of mastering the interview” contains values in terms of Yes/No , we convert these to 1/0 so it converts into numerical values and helps in model building

Dummy Variable Creation:

- We need to create dummy variables for all the categorical columns as they enable us to use a regression equation on multiple groups.

Test Train Split:

- Division of data into test data and train data to check the stability of the model.
- We have randomly sampled 70% of the data as the test data and 30% of the data as test data.
- Random State = 100

Scaling:

- Division of Train Data into X and Y where X has all the features and Y has the target variable – Converted.
- We perform scaling to normalize the data within a particular range
- Technique : Standard Scaler

Model Building

Model – I and II: Basic Model

We build a basic model using 35 features. Since it is not efficient we perform RFE to obtain a model with Top – 20 features. There are so many variables with high p-values and VIF value, we need to remove them

Model – III and IV : Removing variable with p-values > 50%

- Two columns having p –values > 50% : Lead Source_Others and Lead Source Origin.
- Since p –value > 50% , it does not seem to significant at all

Model V, VI and VII : Removing variables having p-value > 10%

- Since we have a cut off for significance value > 10 % does not improve our model.
- Hence, we remove these variables which are : Current Occupation Student , Specialization International Business and LastActivityEmail.

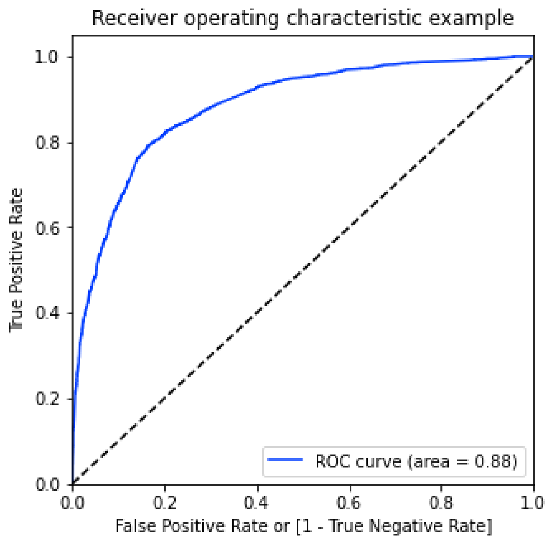
Model VIII : Removing variables having high VIF

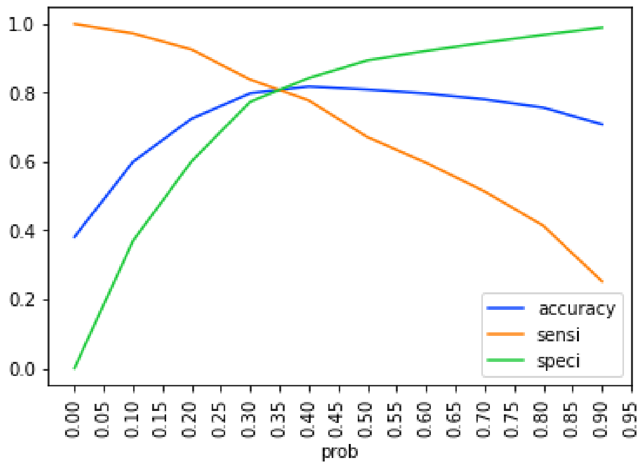
- After model –VII , all p-values < 5%, hence we need to check VIF.
- VIF for Current Occupation_Unemployed = 12.20 which is > 5% .
- Hence we drop this variable from our analysis.

Model – VIII : The Final Model

- All p-values < 5% - Hence they are highly significant .
- All VIF values are < 5. Hence the dependency of variable with another is tolerable.
- Final model has 14 features in total.

ROC Curve And Optical Cut-Off Probability





- ROC Curve represents how much the model is able to distinguish between the classes.
- AUC – Area under the curve represents that it is distinguishing the 1's and 0's correctly.
- On plotting the ROC curve for our data we see that, AUC is around 0.88 which means at around 88% of the times, the model is able to distinguish the 1's as 1's and 0's as 0's.
- AUC of 0.88 is found to be very stable model.
- When we plot the sensitivity, accuracy and specificity of the model together, the optimal cut off point is found to be at 0.35. This means that at 35% probability, the sensitivity and specificity are found to be balanced.
- With probability = 0.35 , we predict y-values with X-Train, in such a way that, any conversion prob > 35% is said to be converted to a lead

Model Performance Test

Train Set

- ACCURACY - 81.19%
- SENSITIVITY - 80.45%
- SPECIFICITY - 81.7%

Test Set

- ACCURACY - 80.08%
- SENSITIVITY - 80.0%
- SPECIFICITY - 80.3%

The sensitivity value after model building process is found to be greater than 80% as required.
When the model is evaluated for Test Set, the model evaluation parameters remains to be the same. Hence the model is highly stable.

Lead Score And Conversion Rate

- Conversion Rate is the number of customers who are converted to leads and interested in the course.
- Before model building the Conversion Rate was found to be 38.53% .
- After model building, the conversion rate is increased to 72.87%
- Hence we can conclude that our final model has served to the business purpose

Lead Score And Conversion Rate

Steps taken to assign a lead score variable for all customers.

1) Train the data with the model.

- Run the model on the entire Leads dataset.
- Do not divide into Test and Train and run the obtained LR model on the entire data frame

2) Predict the Conversion Probability using Cutoff

- Predict the Conversion probability for all the customers using the cutoff value = 0.35.
- Create a new data frame and store Conversion_Probability and actual converted values in this.

3)Adding Lead Score for all variables

- Create a new column called Lead Score.
- Convert the probability score into Lead Score by multiplying by 100 and store it in this column.

4) Calculate the conversion rate

- Once we obtain the complete model result on the data, we filter only the leads as predicted by the model.
- Calculate the Conversion Rate using this filtered result..

Hot Leads

- Hot leads are people who have a high probability to be converted as a Lead and thus needs to be identified. They have a higher conversion rate.
- The leads whose lead score is greater than 35% are considered as potential leads. The conversion rate is around 73%. When we increase this threshold from 35% to 95% we get Hot Leads.
- Conversion Rate for hot leads is increases from 73% to 96%. This means they have a 96% probability of getting converted to a lead.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

Conclusion

From our model, we can conclude following points :

- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is SMS sent or Email opened.
- It’s always good to focus on customers, who have spent significant time on our website.
- It’s better to focus least on customers to whom the sent mail is bounced back.
- If the lead source is referral, he/she may not be the potential lead.
- If the lead didn’t fill specialization, he/she may not know what to study and are not right people to target. So, it’s better to focus less on such cases.

Recommendations

- It’s good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.
- Along with phone calls, it’s good to mail the leads also to keep them reminding as email is as powerful as cold calling.
- Reducing the number of call attempts to 2-4 and increasing the frequency of usage of other media like advertisements in Google, or via emails to keep in touch with the lead will save a lot of time.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

Thank you