X Education - Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving conversion rates for X Education

By – Shashiraj Arunachalamath

Table of Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Understanding Dataset
- Technical Approach For Solving Business Problem
- Data Cleaning
- EDA
- Data Preprocessing
- Model Building
- Model Testing
- Recommendations
- Conclusions

❑ Background of X Education Company

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- 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.
- 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%.

❑ Problem Statement & Objective of the Study

Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%.
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads.
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we 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.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

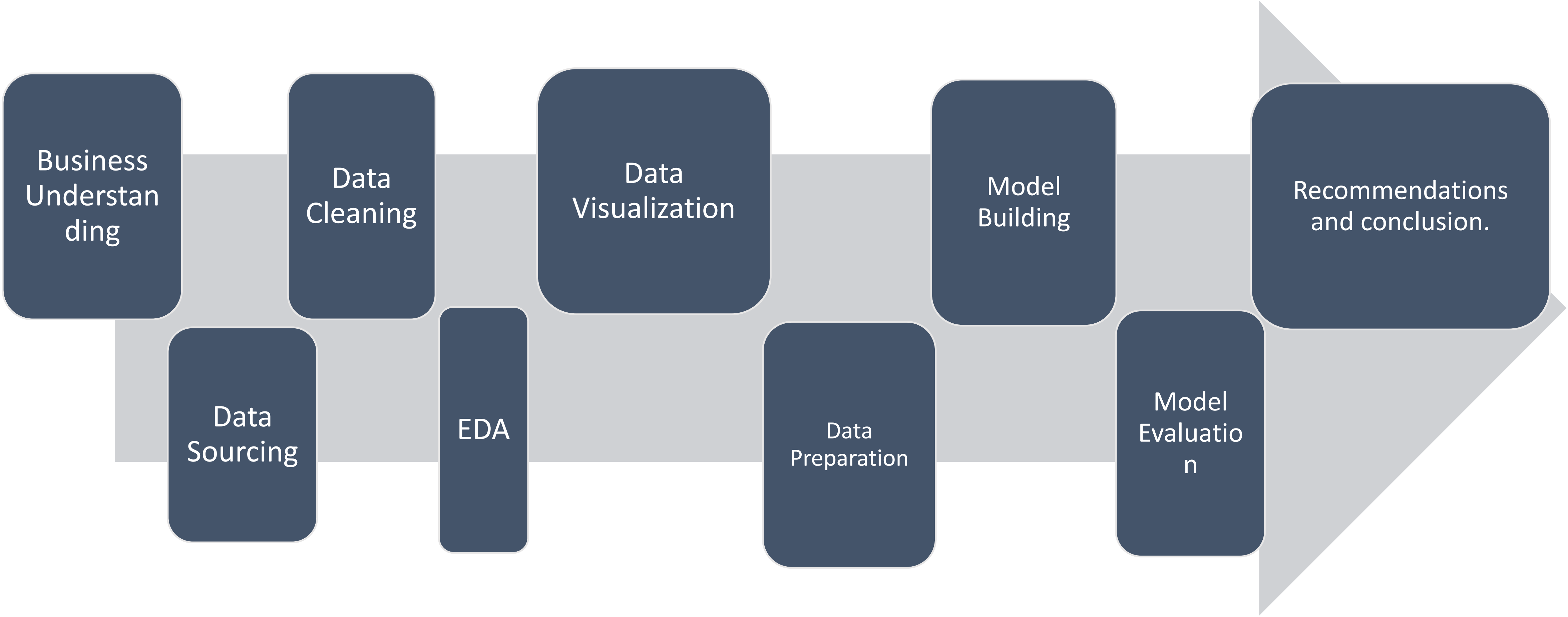
❑ Understanding Dataset

- We got a file named “Leads.csv” provided with a leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- To learn more about the dataset we got the data dictionary.
- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- Another thing that to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

Data Knowledge

- Dataset used : **“Leads.csv “**
- Total number of customers present : **9240**
- Total number of features : **37**
- After initial analysis, we see that there are multiple factors that influence conversion rate.
- The target column in our dataset : “**Converted**” .
- We need to reduce the features to maximize the conversion rate.
- Current Conversion Rate = **38.53%**

Technical Approach For Solving Business Problem



Business Understanding

Data Sourcing

Data Cleaning

EDA

Data Visualization

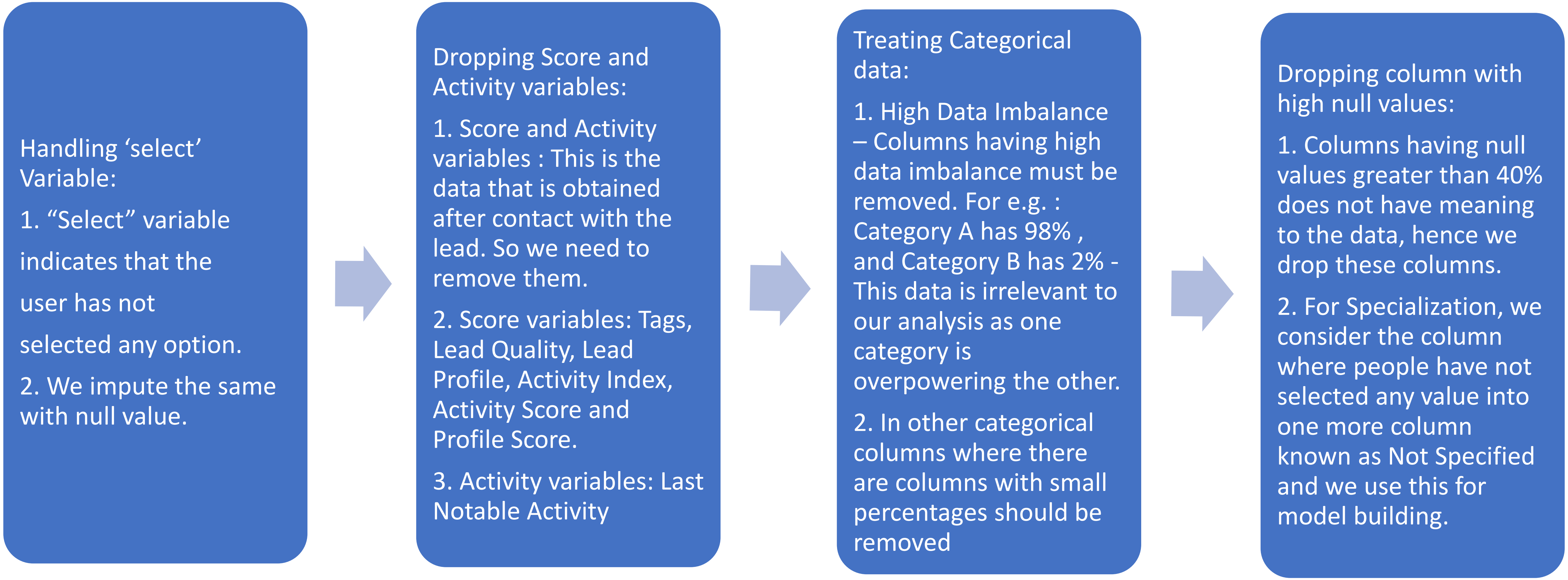
Data Preparation

Model Building

Model Evaluation

Recommendations and conclusion.

Data Cleaning



Handling ‘select’ Variable:

1. “Select” variable indicates that the user has not selected any option.

2. We impute the same with null value.

Dropping Score and Activity variables:

1. Score and Activity variables : This is the data that is obtained after contact with the lead. So we need to remove them.

2. Score variables: Tags, Lead Quality, Lead Profile, Activity Index, Activity Score and Profile Score.

3. Activity variables: Last Notable Activity

Treating Categorical data:

1. 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.

2. In other categorical columns where there are columns with small percentages should be removed

Dropping column with high null values:

1. Columns having null values greater than 40% does not have meaning to the data, hence we drop these columns.

2. 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.

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.

Categorical Data

Lead Origin:

- The percentage of Converted people is found to be greater for Landing Page Submission.
- We can also see that if Lead source is Add Form, the ratio of lead con

Lead Source:

- Google is found to be the important source for Lead Conversion.
- Direct Traffic also proves to be important to secure leads. version is very high (almost not converted is very less).

Last Activity:

- We need to target people via Emails and SMS as it is found that the probability of response in case Converted leads is found to be higher.

❑ Data Preprocessing:

- Binary level categorical columns were already mapped to 1 / 0 in previous steps.
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation.
- Splitting Train & Test Sets ○ 70:30 % ratio was chosen for the split.
- Feature scaling ○ Standardization method was used to scale the features.
- Checking the correlations ○ Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

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

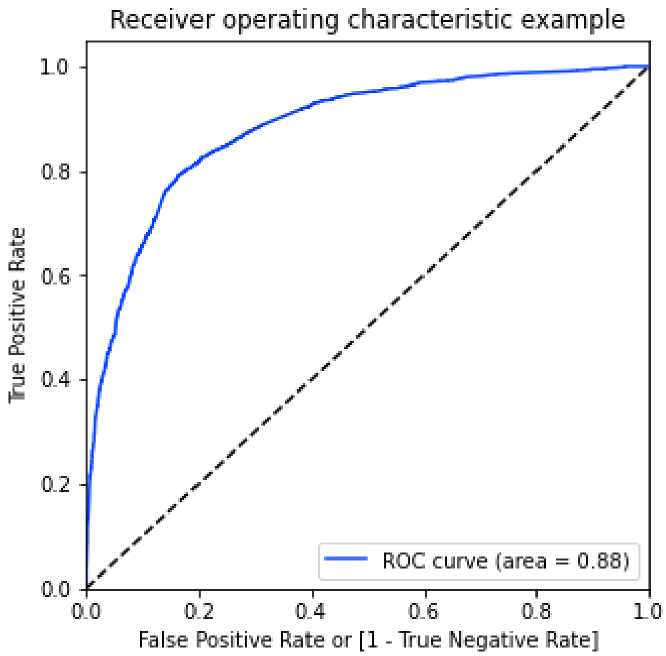
- After model –VII , all p-values < 5%, hence we need to check VIF.
- VIF for Current Occupation_Unemployed = 12.20 which is > 5% .
- Hence we drop this variable from our analysis.

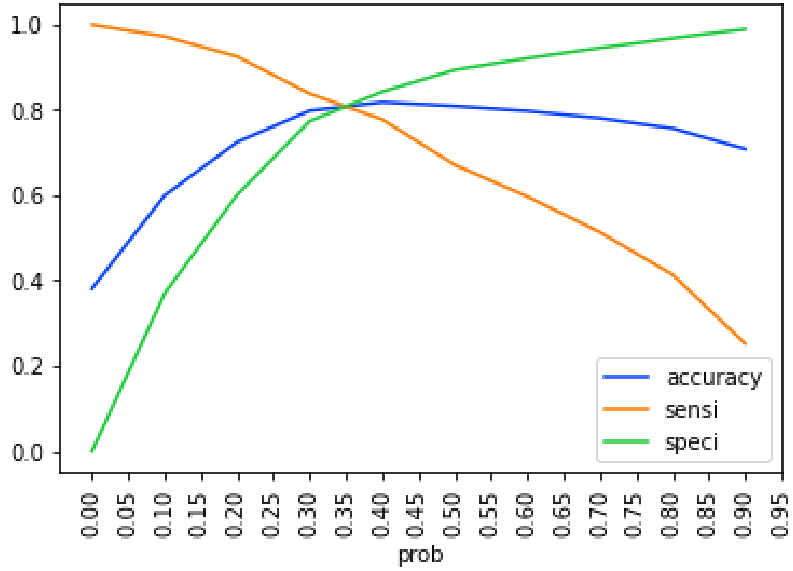
Model – VIII : The Final Model

- 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 XTrain, in such a way that, any conversion prob > 35% is said to be converted to a lead.





Model Testing:

- 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.

Train Set

ACCURACY - 81.19%

SENSITIVITY - 80.45%

SPECIFICITY - 81.7%

VS

Test Set

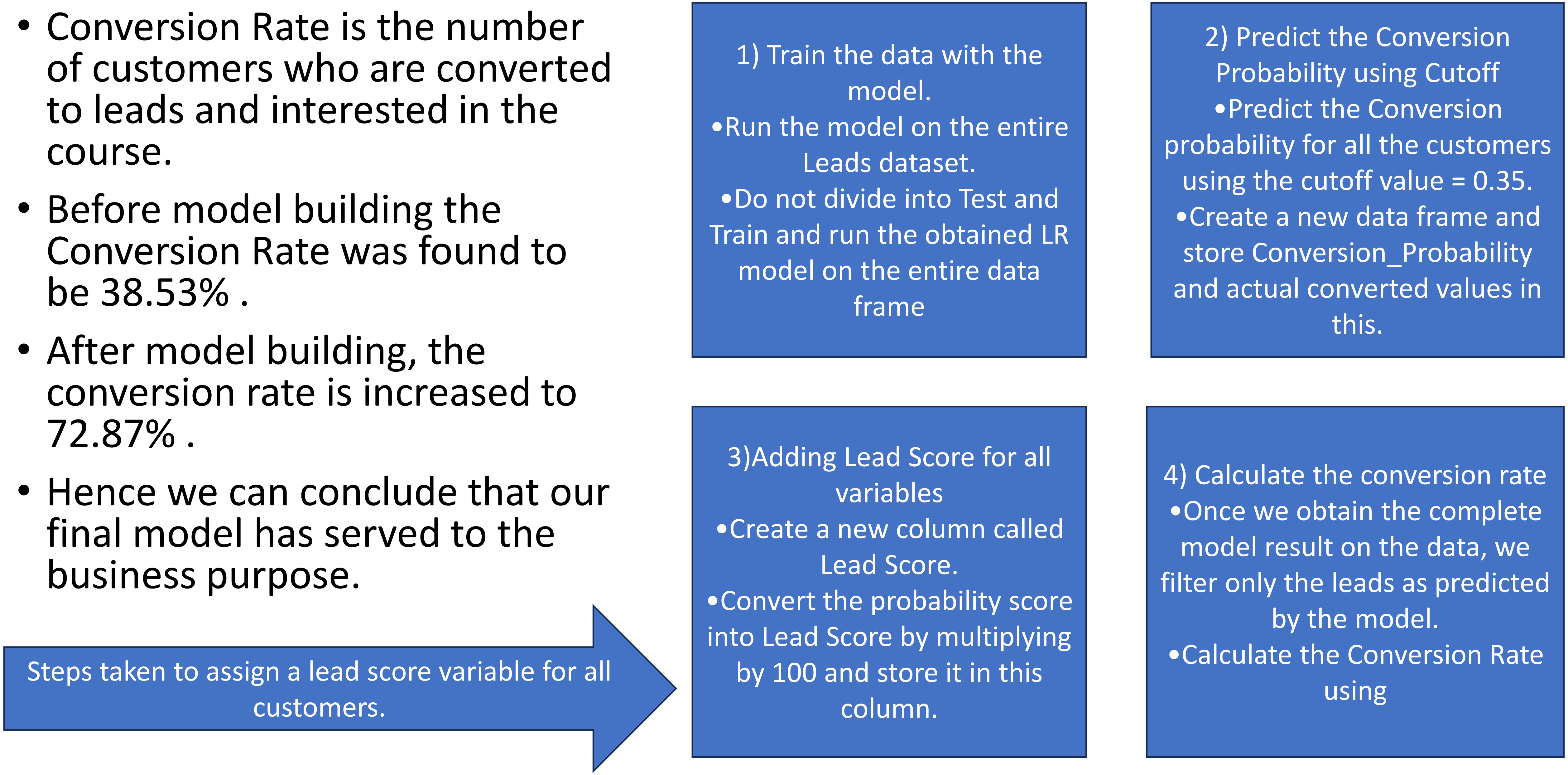
ACCURACY - 80.08%

SENSITIVITY - 80.0%

SPECIFICITY - 80.3%

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.



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

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

- **Equation :-**

-1.0565 * const + 0.1944 * TotalVisits + 1.0574* Time Spent -0.3186 * Free Copy -1.0199 * Lead Origin_Landing Page Submission + 4.4017 * Lead Origin_Lead Add Form + 1.2101 * Lead Source_Olark Chat-1.1764 * Lead Source_Reference -1.1921 * Last Activity_Email Bounced + 0.8166 * Last Activity_Email Opened -0.6859 * Last Activity_Olark Chat Conversation + 0.6463 * Last Activity_Others - 1.9097 * Last Activity_SMS Sent -1.1380 * Specialization_Not Specified + 2.6908 * Current Occupation_Working Professional

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.

Lead Generation

THANK YOU