LEAD SCORING – CASE STUDY

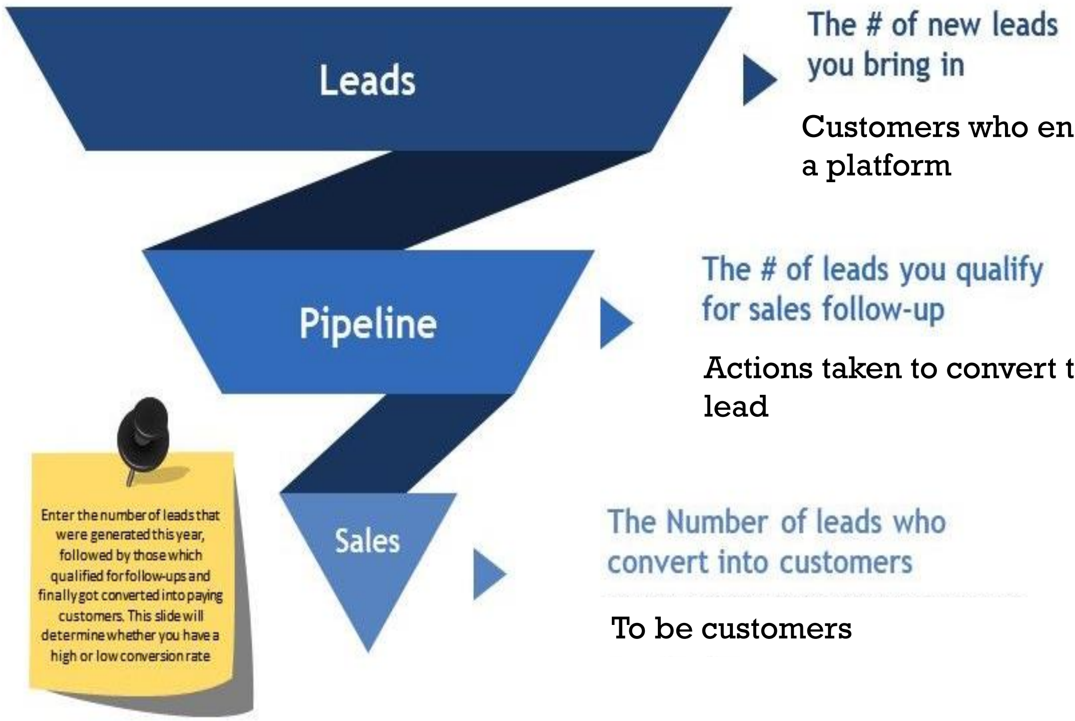
WILL THE LEAD CONVERTS!!!

PRESENTED BY :-

- PRITAM NAIK
- SHREY WADHWA

INTRODUCTION

A lead refers to contact with a potential customer, also known as a “prospect”. For some companies, a “lead” is a contact already determined to be a prospective customer, whereas other companies consider a “lead” to be any sales contact. Companies generate leads from a variety of sources, then follow up with each one to see if the business lead is a good fit for what they sell.


Leads
The # of new leads you bring in
Pipeline
The # of leads you qualify for sales follow-up
Sales
The Number of leads who convert into customers

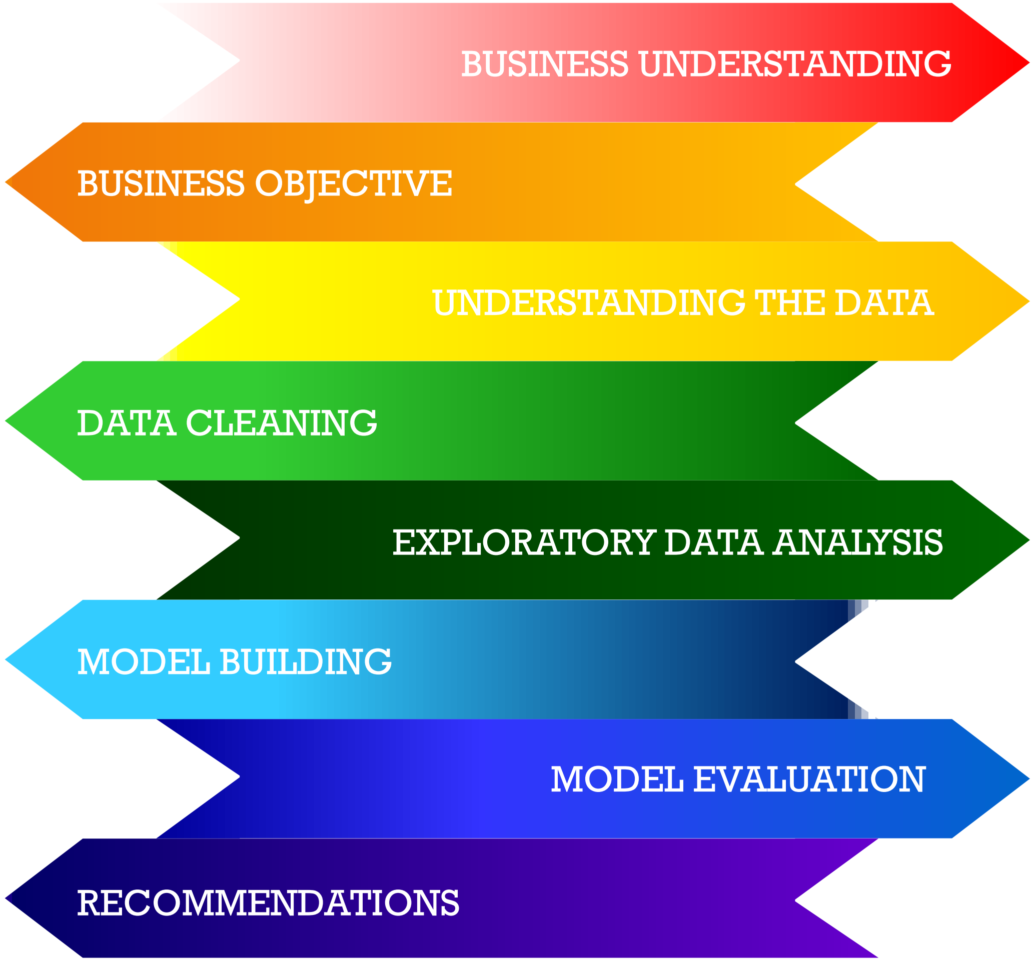
Enter the number of leads that were generated this year, followed by those which qualified for follow-ups and finally got converted into paying customers. This slide will determine whether you have a high or low conversion rate

Customers who ends up on a platform

Actions taken to convert the lead

To be customers

APPROACH TOWARDS THE PROBLEM



BUSINESS UNDERSTANDING

How the relevant business works, and what the driving factors of the business problem.

BUSINESS OBJECTIVE

What is the outcome desired by company.

UNDERSTANDING THE DATA

What is the data that we have. Gathering other relevant data from other sources.

DATA CLEANING

Cleaning the data to the extent that it will be free of null values, noisy points as well as outliers.

EXPLORATORY DATA ANALYSIS

Understanding the data graphically. Many business problems can be solved by EDA alone.

MODEL BUILDING

Building the relevant machine learning model based on business outcome.

MODEL EVALUATION

Evaluating the model on unseen data and analyzing its problem.

RECOMMENDATIONS

Major features contributing towards the business and how to tackle them.

BUSINESS UNDERSTANDING

- A company markets its courses on several websites and search engines. Once the 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 of an X education is around 30%

- Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

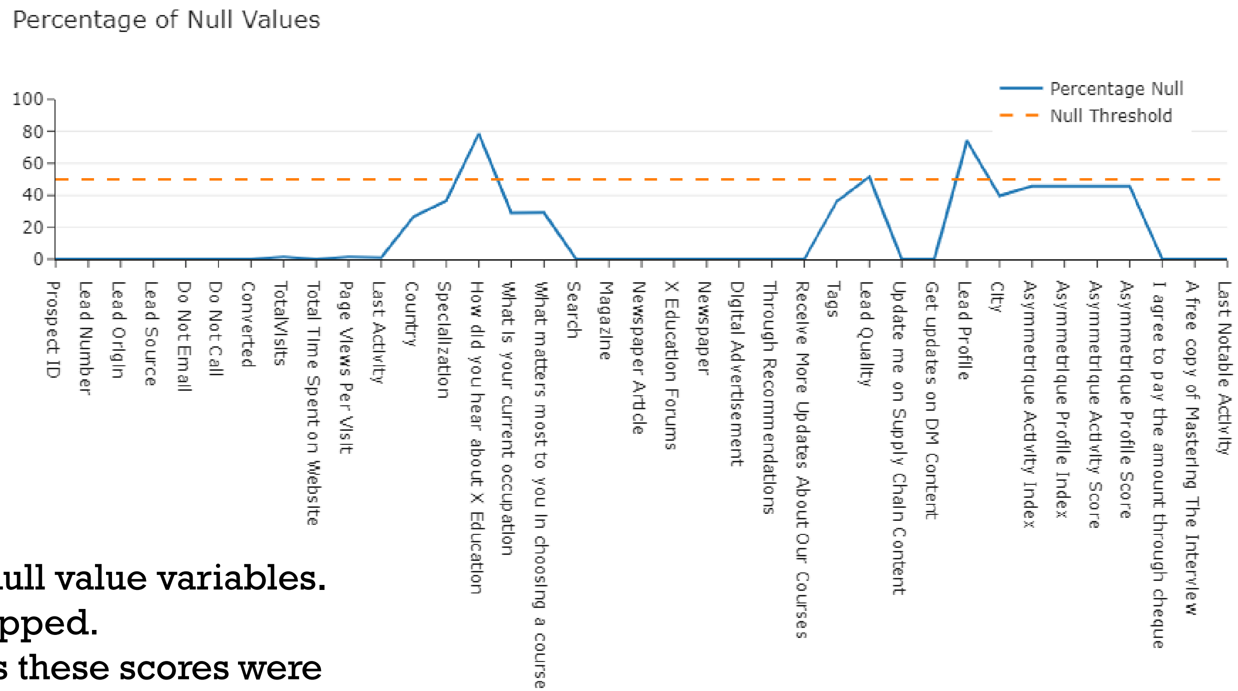
BUSINESS OBJECTIVE

- There are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom. In the middle stage, we have to nurture the potential leads well in order to get a higher lead conversion. The X Education wants to build a model wherein we have to assign a lead score 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. We have to achieve a target lead conversion rate of around 80%.

UNDERSTANDING THE DATA

- Our aim is to Identify Clients who can be a Probable Lead or their probability of conversion into a lead is very high. Based on the Problem Statement variable **"Converted"** will be our **Target variable**.
- Our target variable is **Categorical** in nature having binary outputs. Hence, we will use **Logistic Regression Algorithm** to Classify the leads. We will assign a lead score between **0 and 100** to each of the leads which can be used and easily interpreted by various stakeholders to target the potential leads. A **higher score** would mean that the **lead is hot**, i.e. is most likely to be converted whereas a **lower score** would mean that the **lead is cold** and will mostly not get converted.
- Apart from Determining the lead we will make the model such that if company's requirement changes in future our model will still able to make good predictions.

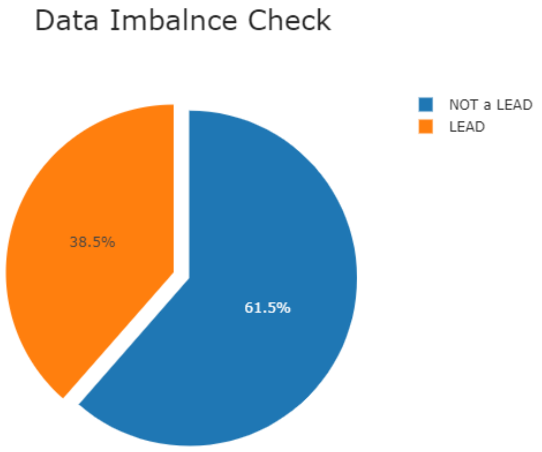
DATA CLEANING

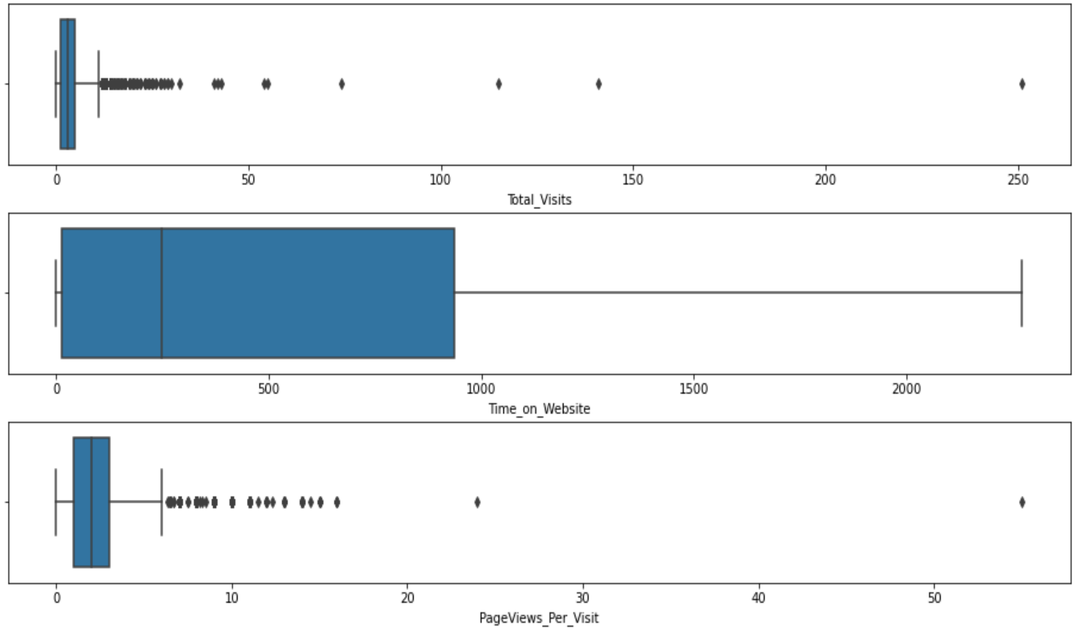


1. Data was full of null values. Some values were missing at random but some were missing due to some reason.
2. Some variables were having select as a value. This values were also equivalent to null as customer filled the online form and they just don’t want to answer that question.
3. We used a thresh hold of 50% for dropping the null value variables. Variables like lead profile, lead quality were dropped.
4. Variables related to Scores were also dropped as these scores were assigned after the client was contacted many times and assigned based on conversion rate.
5. Variables related to Country as well as City were dropped. Country having high null values and was monotonous in nature, similarly city variable was also having high null values and imputing with mode or creating another category will not justify the null values.
6. Other Categorical variables like specialization, occupation, choosing the course were having high null values therefore another category were created for each column.
7. Numerical variables such as total visits, time spent on website, page views per visit were having less number of missing values. To avoid affect of outliers, missing values were imputed with median values.

EXPLORATORY DATA ANALYSIS

- **Data Imbalance Check**
  - It is very important to perform the imbalance check. If data is imbalanced high accuracy can be achieved by predicting the all values as the highest value.
  - In our case data not much imbalanced.

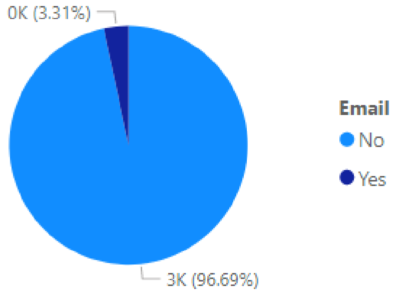


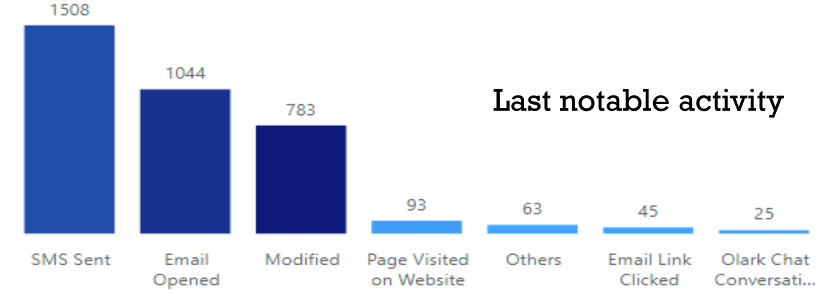


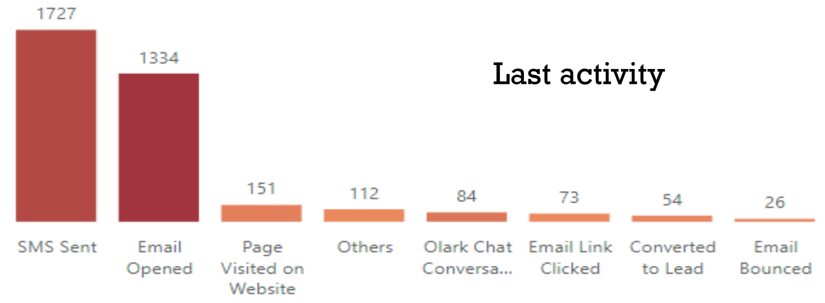
- **Univariate Analysis and Handling of Outliers**
  - It can be seen that numerical variables were influenced by outliers.
  - As our data points are few in number. We will not remove the outliers.
  - Categories were created for these numerical columns. This will help in handling the outliers as well as make the model stable.
  - Numerical variables can vary a lot and makes the mode susceptible to fluctuations with small changes.

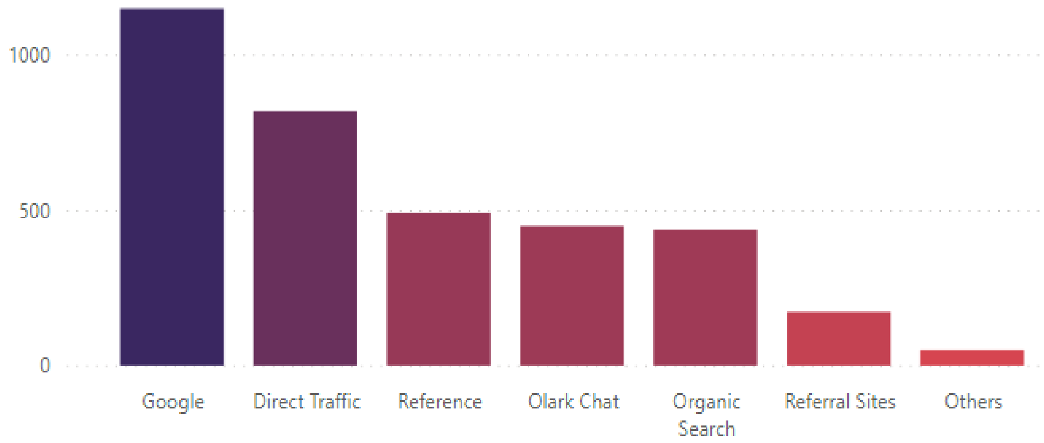
EXPLORATORY DATA ANALYSIS – CONT…

- **Univariate analysis**
  - Most of clients came to platform from Google followed by Direct Traffic.
  - Mostly unemployed Clients wants to join course with X-Education.
  - Most Clients have not chosen to specify their current occupation. It is correct as most of the Clients who are probable leads are unemployed.



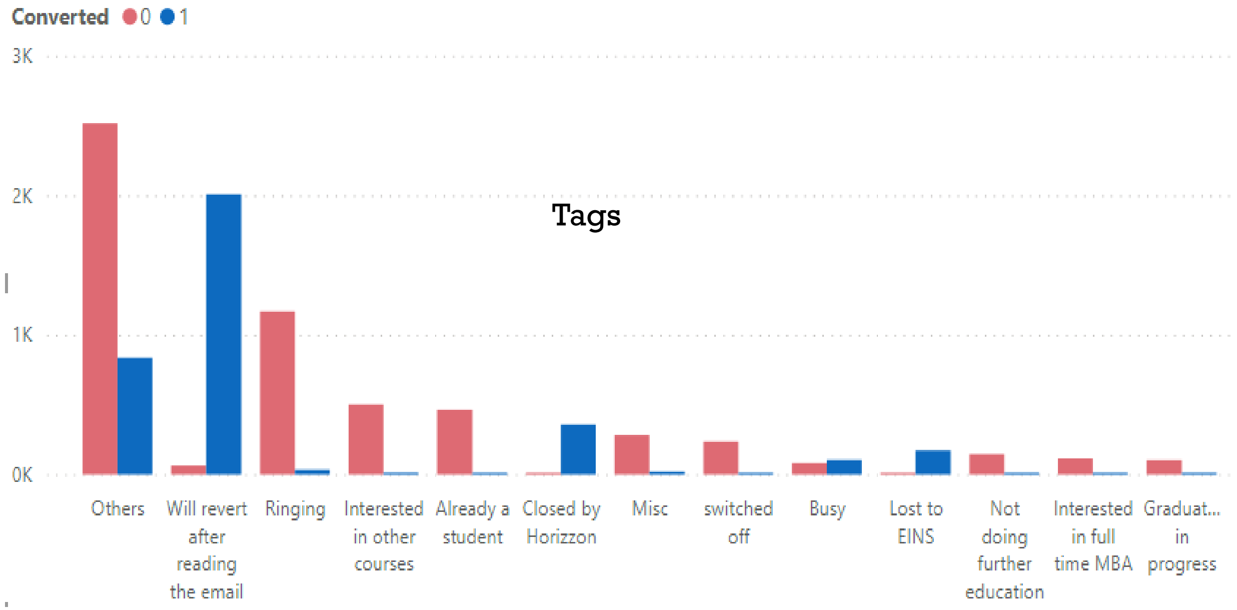

Last notable activity

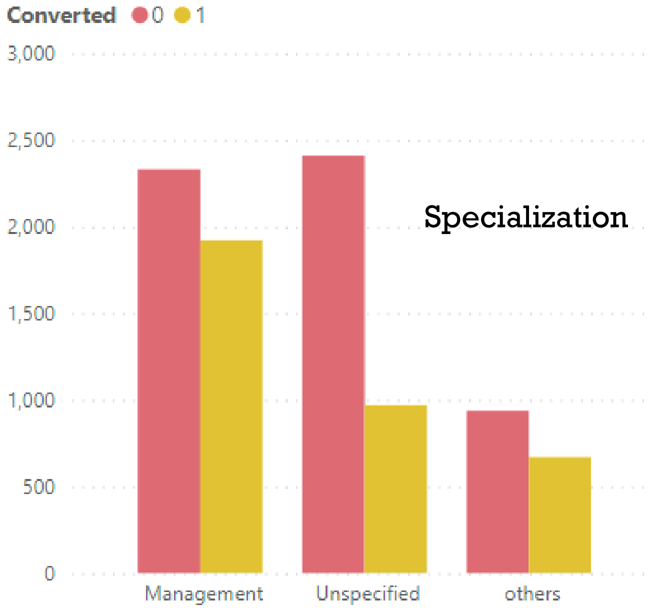

Last activity



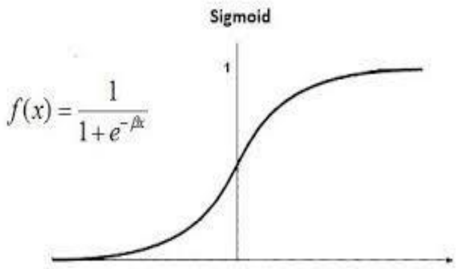
EXPLORATORY DATA ANALYSIS – CONT…

- **Bivariate analysis**
  - Clients belonging to Management have high conversion rates.
  - Although large number of Clients join through Google but clients through reference have highest conversion rate.
  - Similarly, clients to whom sms is sent have highest conversion rates.
  - Clients opted for revert after email also have high conversion rates.





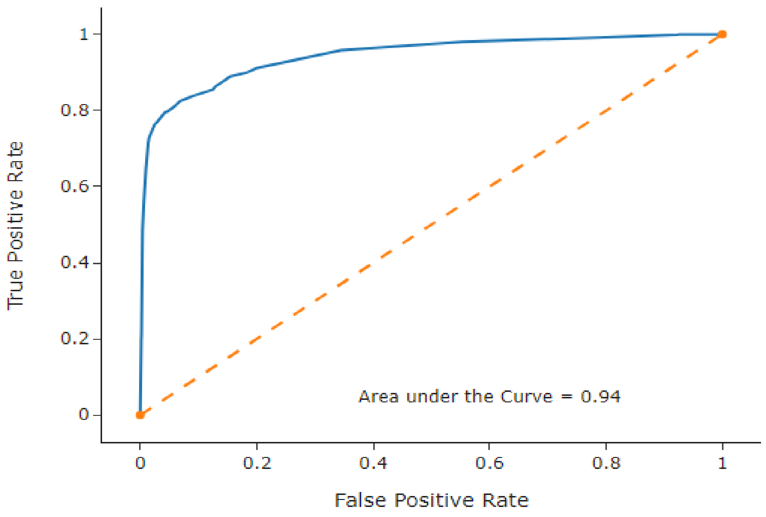
MODEL BUILDING

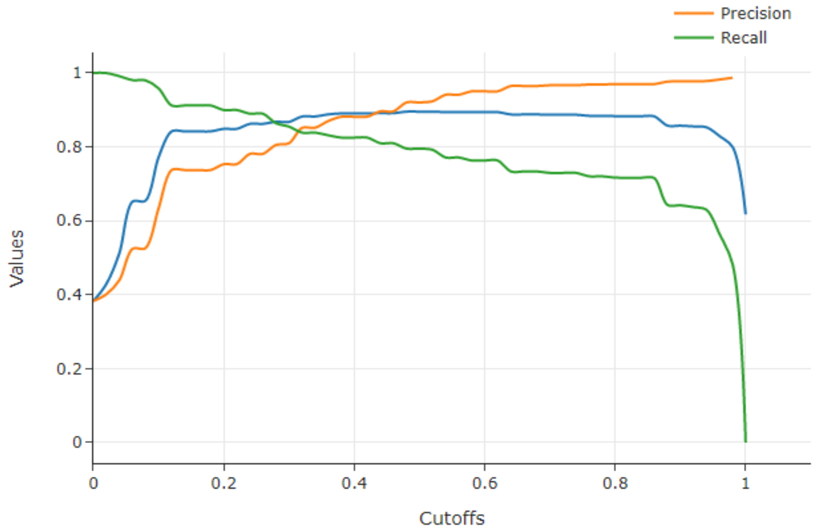


- Finally our data is free of null values and outliers.
- Our Target variable is Categorical in nature having binary outputs. So we have used Logistic Regression to Classify the Data points.
- All the variables finally present in Dataset were categorical in nature. So dummies are created for all variables.
- Using sklearn library top 15 features were extracted using Recursive Feature Elimination. Finally, to make the model leaner are making the variables easier to target, more features were removed based on business knowledge.
- Final Model had variables - **'Lead_Origin_API', 'Lead_Origin_Landing Page Submission', 'Lead_Source_Social Media', 'Occupation_Unspecified', 'Tags_Already a student', 'Tags_Busy', 'Tags_Closed by Horizon', 'Tags_Lost to EINS', 'Tags_Will revert after reading the email', Time_Spent_Range_500-1000', 'Time_Spent_Range_1000-1500', 'Time_Spent_Range_More than 1500‘.**
- The most important variables were contributing positively towards the Lead conversion are Tags-Closed by horizon, Tags-lost to EINS and Tags-will revert after reading the email.
- The Variables negatively contributing towards the lead conversion were Tags-Already a Student, Lead Source-Social Media and Lead origin-landing page submission

MODEL BUILDING – CONT…

1. As per our business objective we have to find the probable leads which high conversion rates.
2. Precision and Recall were the evaluating criteria that were used to evaluate the model.
3. Cutoff of 0.3 was chosen which had high Recall (0.86) as well as Accuracy (0.87) and Precision of 0.81.
4. Model also had High Precision, Recall as well as Accuracy of 0.82, 0.86 and 0.87 respectively on Test Dataset.


Receiver Operating Characterstics Curve

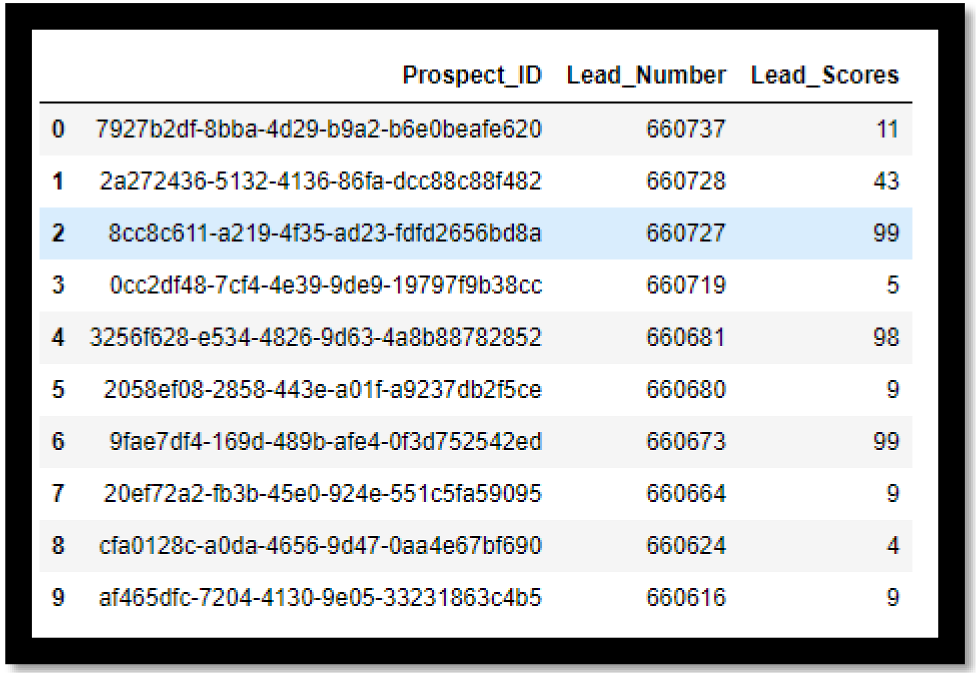

Accuracy vs Precision vs Recall

MODEL EVALUATION

- Train Dataset
  - Accuracy on Train Data - 87%
  - Precision on Train Data - 86%
  - Recall on Train Data - 81%

- Test Dataset
  - Accuracy on Test Data - 87%
  - Precision on Test Data - 82%
  - Recall on Test Data - 86%

Some of the lead Scores allotted to the Prospective LEADS



	Prospect_ID	Lead_Number	Lead_Scores
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	11
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	43
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	99
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	5
4	3256f628-e534-4826-9d63-4a8b88782852	660681	98
5	2058ef08-2858-443e-a01f-a9237db2f5ce	660680	9
6	9fae7df4-169d-489b-afe4-0f3d752542ed	660673	99
7	20ef72a2-fb3b-45e0-924e-551c5fa59095	660664	9
8	cfa0128c-a0da-4656-9d47-0aa4e67bf690	660624	4
9	af465dfc-7204-4130-9e05-33231863c4b5	660616	9

Our Recall is good for Test as well as Train Data Set. It can inferred that the model built can be used for predicting the Customers who are probable leads having a success rate of 86-87%

RECOMMENDATIONS

- The **3** variables which are majorly contributing in determining the Probable leads are **Tags-Closed by Horizon, Tags-Lost to EINS and Tags-Will revert after reading the email.**
  - These variables should be targeted the most, if client is falling in any one these categories, he/she has high chances of being converted.
  - If client is spending more than 500 hours on Platform, he/she will more likely to be prospective candidate.
- The **3** variables which are majorly contributing towards in determining the Person is **NOT** a Probable Lead are Tags-Already a Student, Lead Source-Social Media and Lead origin-landing page submission.
  - This variables should be checked, if client is falling in any one these categories, he/she has high chances of **NOT** being converted.
- Tags are most important characteristics while determining the lead. If the clients are associated with Tags stated above.
- Tags can only be provided after the Clients have been contacted already. Therefore, initial attempt have to made to connect with Client and based on the response or tag provided, Clients should be followed up and resources should be allocated for the Clients to successfully converting them.
- Apart from looking at Tags other variables should also be checked like time spent on page and Lead sources. Clients coming through API, Landing Page, Social Media not specifying their employment are less likely to be converted into a lead

RECOMMENDATIONS – CONT…

- During the Internship program, company gets to train many interns. At this stage Manpower is much more and more number of Clients can be focused.
  - To make the lead conversion more and more aggressive at this stage X-Education can focus on clients with lower lead Scores as well.
  - Earlier a cutoff of 30 was chosen and leads having a lead score beyond 30 were targeted. Now we can reduce the cutoff value so that leads with lead Scores as low as 10 can also be targeted.
  - Already converted leads can be contacted for referrals. Referrals should be focused on as they have high chance of being converted because on influence of already converted lead.
  - Will revert after reading the Email was a major factor contributing towards lead conversion. Hence, attractive emails offering discounts should be sent to prospective leads.
- Sometimes, company reaches the Quarter target before the Deadline. At this point, only necessary calls should be made to avoid unnecessary phone call charges.
  - During this period company can focus on marketing the products much more. Marketing can be done on platforms such as televisions, colleges etc.
  - Providing offers to already converted leads to provide referrals. Provide leads with Emails with attractive offers.
  - X-Education can form collaborations with a Companies related to the courses offered. So that, employees of companies can be trained through X-education only.
  - Free Courses, books, blogs etc. should be provided by X-Education, so that leads spend much more time on their platform. The leads spending more time on the Platform have higher chance of being converted.

Thank you