# Lead Scoring Case Study

FIND HOT LEADS

# Why We are Here-

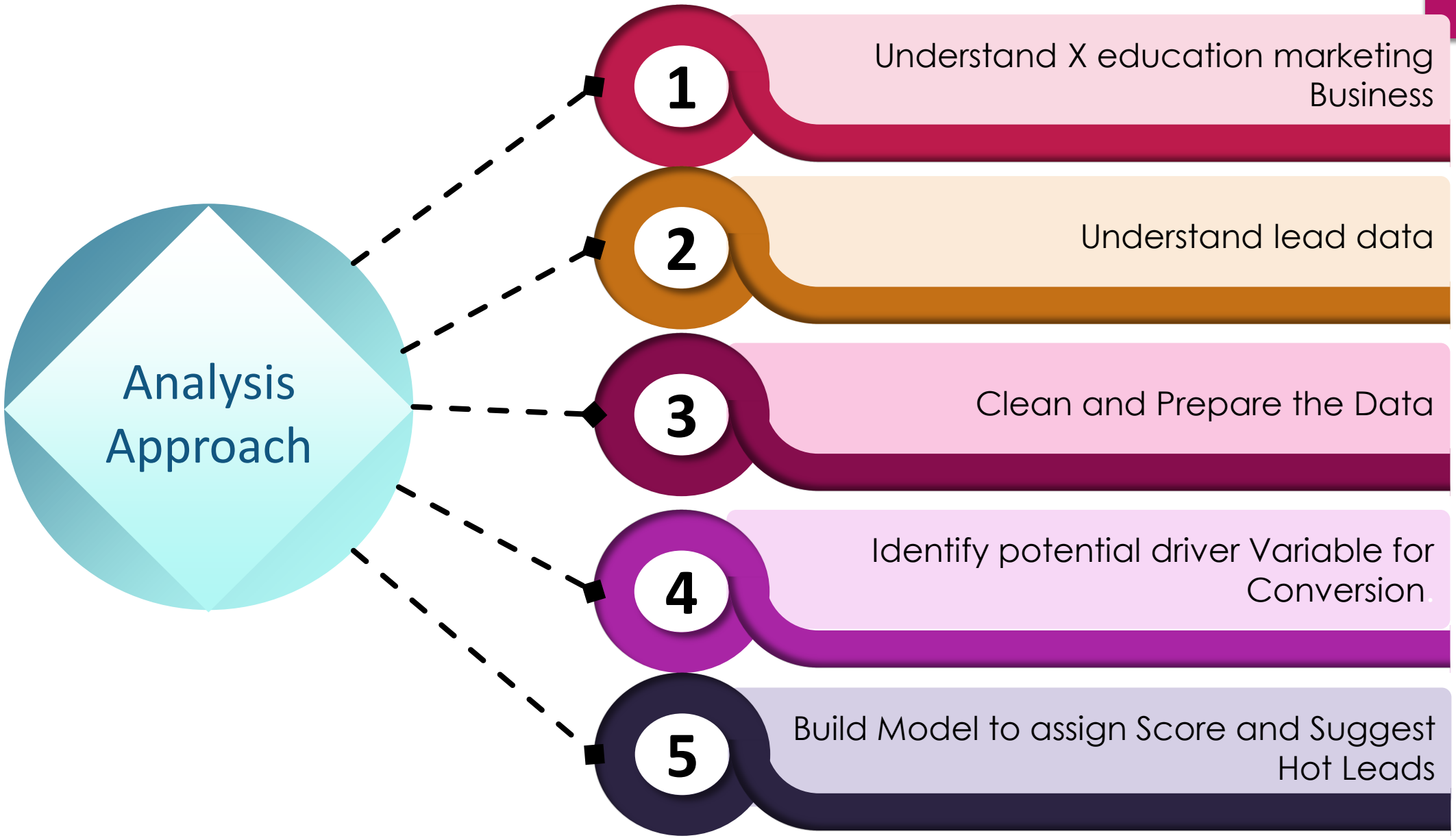
## Problem

X educations sells Online courses to industry professionals. Marketing of courses are done in multiple way like advertise in Search engine like google and in multiple website. When any person see the website and fill a form or browse a course data got captured as a lead. Marketing Team does cold call or write email. Marketing got huge amount of lead but conversion rate is less as lead conversion rate is 30%

**What do we want—**

- To identify the most promising Lead.
- Build a model to assign lead score.
- Score the Lead by Conversion Probability Percentage
- Minimize marketing call to non potential leads

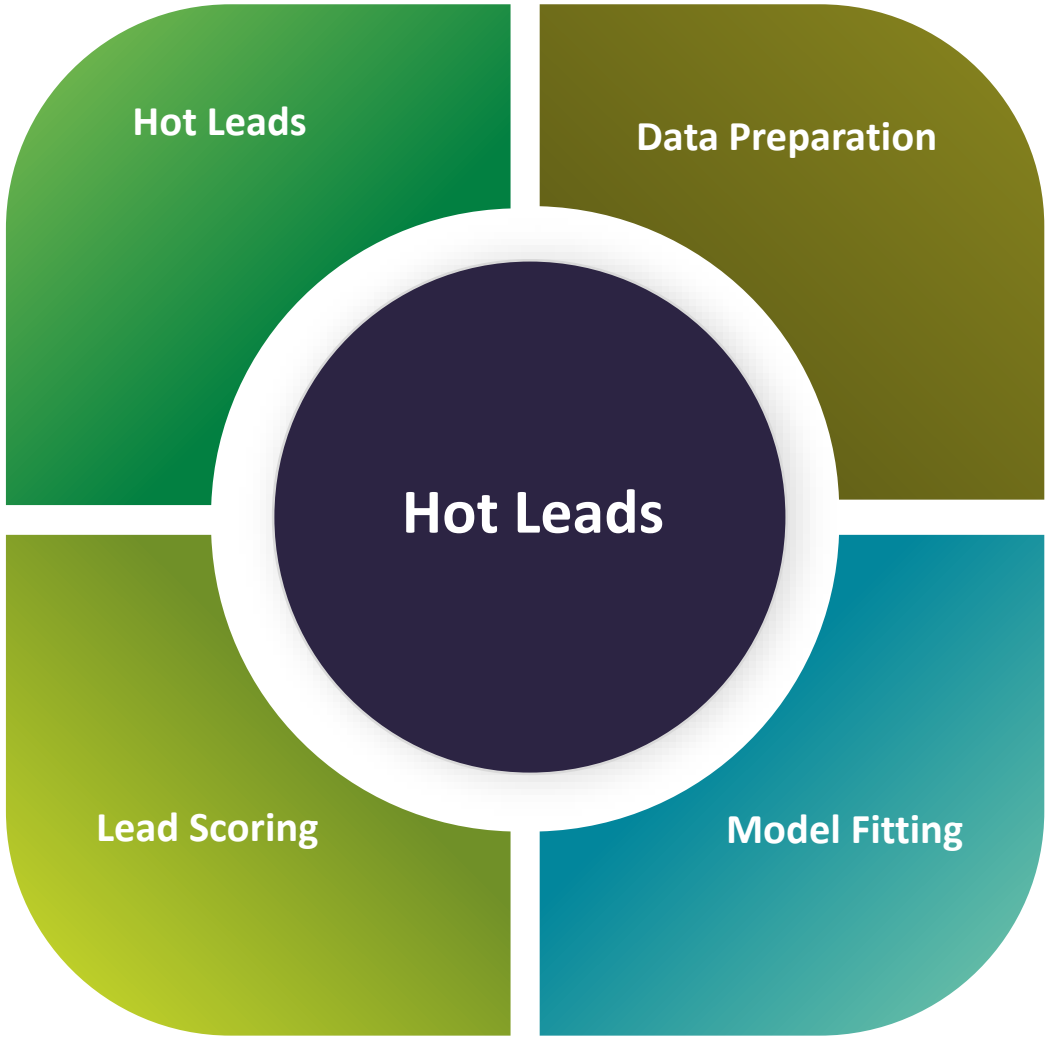
# How to Find the Hot lead– Approach



# Solution Overview

**Main Criteria**
- Find the hot leads .
- Suggest marketing team About Hot Leads
- Describe Hot Leads Criteria
- Provide Lead Score options

**Main Criteria**
- Calculate and Assign score for All Lead data
- Find optimum cut off point for Lead Score




**Overview**
- Inspect lead data and Treat user input.
- Remove Column with missing values more than 30%
- Remove Outliers
- Create dummy categorical variable
- Remove the low variance categorical variable
- Find and Remove highly Correlated Variables
- Scale the data

**Steps**
- Build Logistic Regression model
- Find optimum model with optimum accuracy and specificity
- Predict on Test dataset
- Find Probability Conversion Score on test data

# Solution Methodology-Data Analysis and Cleaning

| |
|---|
| Do Not Call |
| What matters most to you in choosing a course |
| Search |
| Magazine |
| Newspaper Article |
| X Education Forums |
| Newspaper |
| Digital Advertisement |
| Through Recommendations |
| Receive More Updates About Our Courses |
| Update me on Supply Chain Content |
| Get updates on DM Content |
| I agree to pay the amount through cheque |

- Columns showed in left table provides a view of very low variance features
- For most of the columns 99% of the leads has same value
- So these column or feature will not be use full to identify potential leads

| | |
|---|---|
| Prospect ID | 0.0 |
| Lead Number | 0.0 |
| Lead Origin | 0.0 |
| Lead Source | 0.0 |
| Do Not Email | 0.0 |
| Converted | 0.0 |
| TotalVisits | 1.0 |
| Total Time Spent on Website | 0.0 |
| Page Views Per Visit | 1.0 |
| Last Activity | 1.0 |
| Country | 27.0 |
| Specialization | 16.0 |
| How did you hear about X Education | 78.0 |
| What is your current occupation | 29.0 |
| Tags | 36.0 |
| Lead Quality | 52.0 |
| Lead Profile | 74.0 |
| City | 40.0 |
| Asymmetrique Activity Index | 46.0 |
| Asymmetrique Profile Index | 46.0 |
| Asymmetrique Activity Score | 46.0 |
| Asymmetrique Profile Score | 46.0 |
| A free copy of Mastering The Interview | 0.0 |
| Last Notable Activity | 0.0 |

- 9240 rows are the captured in Lead Scoring set .
- For Few columns or feature more than 30% leads done have values.
- Columns showed in left table provides a view of missing Values /rows for each features
- Below columns or features are removed as they have more than 30% missing/null value.
    a) How did you hear about X Education
    b) Lead Quality
    c) Lead Profile
    d) City
    e) Asymmetrique Activity Index
    f) Asymmetrique Profile Index
    g) Asymmetrique Profile Score
    h) Asymmetrique Activity Score
    i) Tags

# Solution Methodology-Data Cleaning and Preparation

## Below are the Assumptions for Data Cleaning:

- Replaced 'Select' value in Specialization, Lead Profile, City, How did you hear about X Education.
- Since there is no Other option for Specialization, replaced 'Select' with Others
- Since Lead Profile has Other Leads and City has Other City, How did you hear about X Education has others imputing 'Select' with nan values
- Replaced /Imputed null values with median for numerical variables /columns.
- Dropped rows with more than 3 columns blank.
- Replaced /Imputed null value with highest occurring value for categorical columns.
- Total visits per page is capped to less than 50, Pages per visit is capped to less than 20

## Below are the Assumptions for Data Preparation:

- Dummy variable has been created for categorical column to make the dataset compatible for Model
- Dropping Low variance column which has near to 95 % data same
- 70 % of the lead data is taken to train the model
- Scaled the train and test data to keep all variable in same range
- Verified the correlation between columns or variable and dropped high correlated variable

# Solution Methodology-Model Selection

Below are the Assumptions for Model Selection:

- Selected the features using RFE. Took top 9 features
- Built 3 model based on RFE suggested features. Selecting Model one for prediction
- Concluded that model 1 is better compared to other two for below considerations
    - Difference between accuracy of train and test being minimal(<2%)
    - The accuracy is above 80% for both

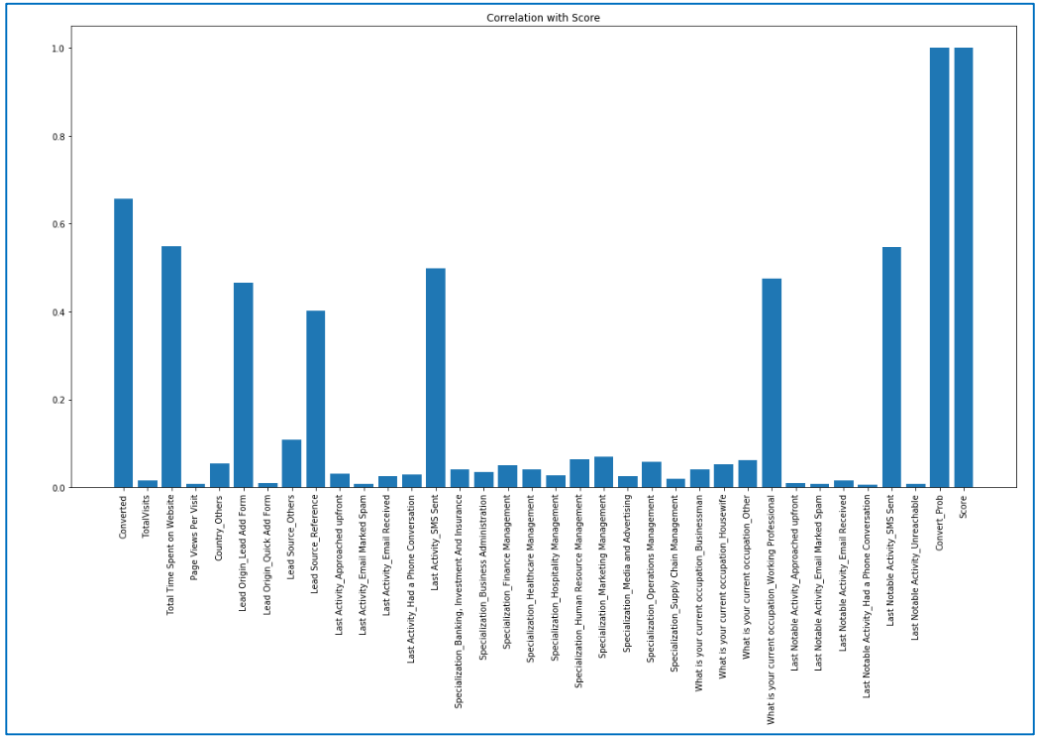
- Below features are suggested by models

| |
|---|
| Do Not Email |
| Last Activity_Email Bounced |
| Last Activity_Page Visited on Website |
| Specialization_IT Projects Management |
| What is your current occupation_Student |
| Total Time Spent on Website |
| Last Activity_Converted to Lead |
| Last Activity_Email Link Clicked |

# Model Result-Findings

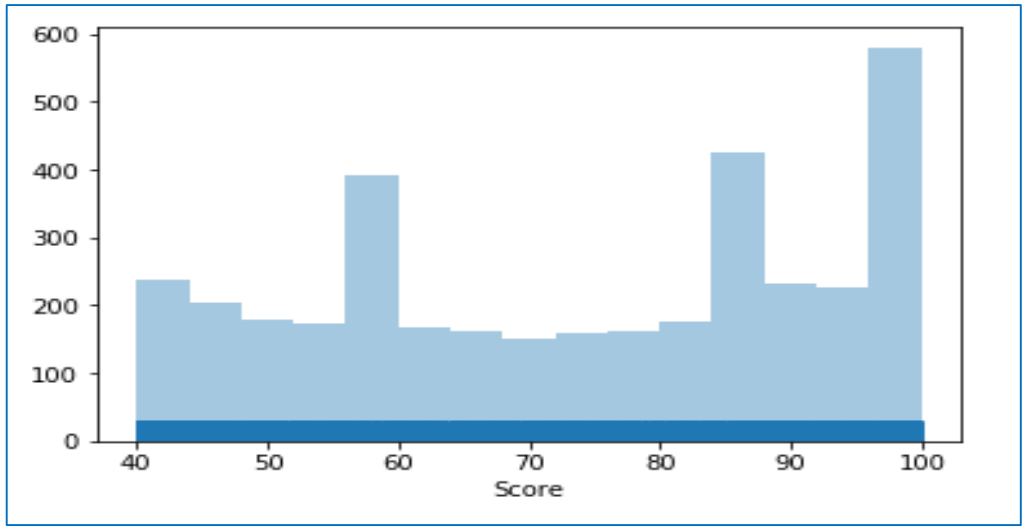
- Model gives a Overview of Conversion probability for each leads
- Scores are assigned to each lead based on Probability of conversion.
- Selected Model can predict the potential Leads in **80%** accuracy
- It is assumed that 18 % of the lead can be marked as potential lead which may not be a potential lead
- Do not email 'No' is favorable for higher score
- Total time Spent on website is more for higher score
- If lead_origin has add form, it is highly favorable for higher score
- If Lead source is Olark Chat it is highly favorable
- If Last Activity is Converted to lead, it is least favorable for higher score
- If Last Activity Olark Chat Conversation is least favored for higher score
- If current occupation is student, it is less desirable for higher core
- If occupation is unemployed, more favorable for higher score
- If Last Notable Activity is SMS Sent, it is favorable for higher score
- Top 3 desirable variables
    - Total time spent on website
    - Lead Origin is Lead Add From
    - Last Notable Activity is SMS Sent
- Observe that for higher score
    - Total time spent on website between 1000-2000 is more favourable
    - Do not email 'Yes' is less favourable
    - Olark Chat, Reference and Others is more for Lead Source
    - For last activity if Page visited on website is less,Email options is less,Email Link clicked is less, Converted to Lead less,Olark chat conversation is less, unreachable less
    - Current occupation is housewife or working professional and not student or unemployed

# Model Result-Analysis




**We notice that Score has Strong positive correlation/dependency with**

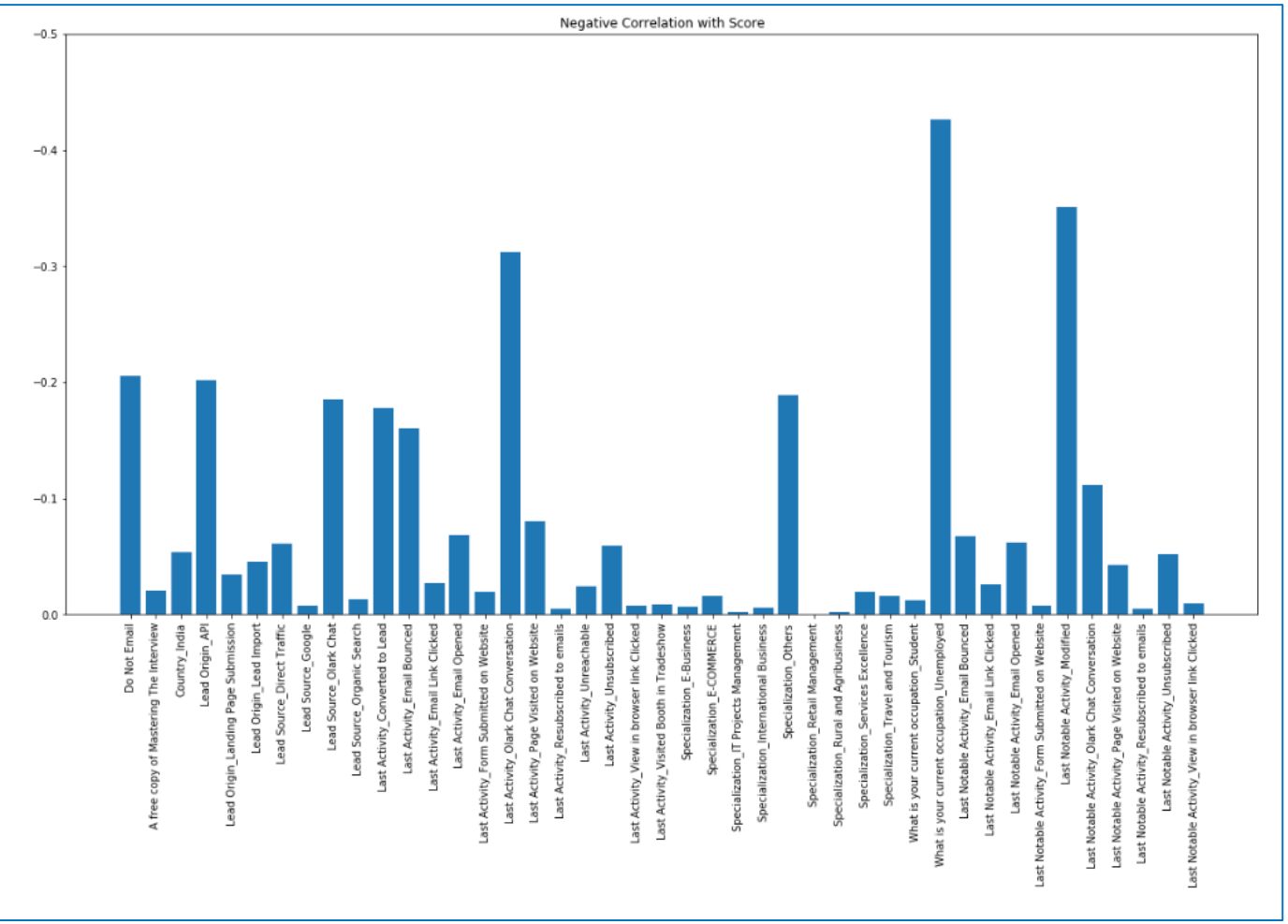
a) Total time spent on website
b) Lead Origin is Lead Add Form
c) Lead Score is Reference
d) Last Activity is SMS Sent
e) Current Occupation is Working Professional
f) Last notable Activity is SMS sent




**Graphs of the left shows the Distribution of Scores by Conversion Count**

a) Lead Conversion is Peaks for Score between 95-100, 85-90, 55-60.
b) Leads Score more than 40 will only convert.

# Model Result-Analysis




**We notice that score has a strong negative correlation with. So These should be avoided to improve the Lead conversion rate .**

- Do not email 'Yes'
- Lead Origin is API
- Lead Source is Olark chat
- Last Activity is Converted to Lead
- Last Activity is Email bounced
- Last Activity is Olark Chat Converation
- Specialization Others
- current Occupation is unemployed
- Last notable Activity is Modified

# Recommendations to X-Education Team

Below are the Suggestion by category/ Priority for Marketing executives

## Peak Call Time Suggestion

- Call first to the leads has scores
    a) Between 95 to 100
    b) Between 85 to 95
    c) Between 55 to 60
- Then Call all leads has score more than 40

## After reaching target Suggestion

- Call only to the leads has scores
    a) Between 95 to 100
    b) Between 85 to 95
    c) Between 55 to 60
- They are suggested to only high potential leads as it's a busy time for sales team.

## Top 3 desirable variables

- Lead Origin with value as Add Lead Form
- Current Occupation is not Student or Unemployed
- Last Notable Activity with value as SMS sent

## Top 3 desirable Categorical variables

- Lead Origin with value as Add Lead Form
- Current Occupation is not Student or Unemployed
- Last Notable Activity with value as SMS sent

## High Score Favorable fact

- Do not email 'No' is favorable for higher score
- Total time Spent on website is more for higher score
- If lead_origin has add form, it is highly favorable for higher score
- If Lead source is Olark Chat it is highly favorable
- If occupation is unemployed, more favorable for higher score
- If Last Notable Activity is SMS Sent,it is favorable for higher score
- Total time spent on website between 1000-2000 is more favourable