



# **Lead Scoring Case Study**

## **Analysis and Inferences**

Submitted by

**Ekta Gupta**

**Jyothi R**

**Mutyala Sridevi**




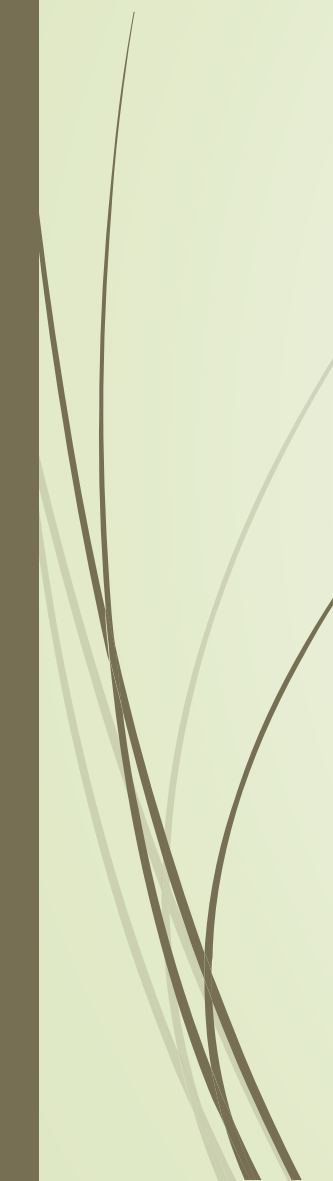
# Problem Statement

- 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.
- 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.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
- 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.

# Goal

- Build model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

# Approach followed

- 
- 
- 1- Data Understanding
    - - Importing Data and Check Statistics
  - 2- Data Cleaning
    - - Check missing values/checking outliers and fix those by checking their statistics
  - 3- Exploratory Analysis
    - - Uni-Variate, Bi-Variate and Correlation or pair plots
  - 4- Data Preparation
    - - Convert in binary column and dummy variables creation
    - - Feature Scaling
  - 5- Build Model
    - - Split the data in train and test, features scaling, check correlation matrix,
    - - Features Selection using RFE and manual
  - 6- Model Evaluation
    - - Confusion Matrix
    - - Accuracy , Sensitivity, Specificity, Precision and recall, ROC Curve
  - 7- Prediction of test Data



# Data Pre-processing

- ▶ Null values have been handled appropriately using the imputation, retaining the missing values as special category and deletion methods
- ▶ Features having same values throughout are dropped as they do not contribute in the prediction and may actually lead to bias
- ▶ Dummy variables are obtained for categorical values
- ▶ Data transformed for binary categorical features
- ▶ Data scaled to prevent the effect of outliers



# Exploratory Data Analysis





# Inferences from uni and bi-variate analysis

- More leads are from landing page submission
- Google is the top most lead source
- Most of the leads are from Mumbai city
- There is only 38.54% lead conversion
- 'Search' followed by 'Recommendations' and 'Digital Advertisement' works out to be better channels for the leads
- All the leads landed with a motive to achieve better career prospects
- There are 22.3% leads with neutral attitude and 13.7% potential leads who actually converted to be positive leads
- Most of the leads are from banking and insurance, and healthcare management domains.
- Most of them are currently working professionals and housewives
- 'Interested in next batch' and 'lateral students' are the tags frequently given
- Dual specialization and lateral student are the top lead profiles




# Inferences from uni and bi-variate analysis

- Most of the positive leads didn't opt for free content
- Approached upfront and resubscribed to emails are some among the frequent last notable activities
- There is no much correlation between the numerical variables 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit'
- Most of the working professionals from management domain were among the positive leads
- Leads who responded positively indicating they shall revert converted to positive leads
- Having phone conversation had a positive impact on lead conversion





# Feature Selection using RFE

- Top 20 features as ranked by RFE mechanism has been considered for model building
- 



# Model Built using Logistic Regression

- ▶ Performance metrics on Training Set with optimal cut-off as 0.5

**Overall Accuracy: 0.84**

**Sensitivity: 0.90**

**Specificity: 0.80**

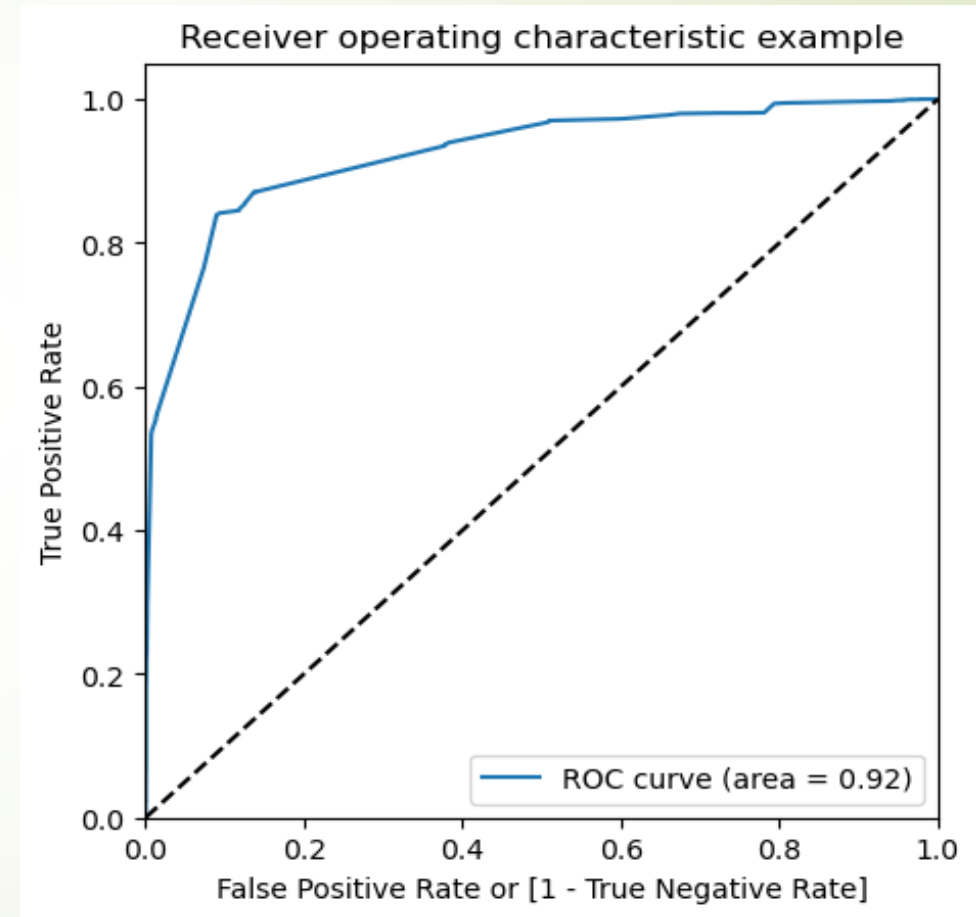
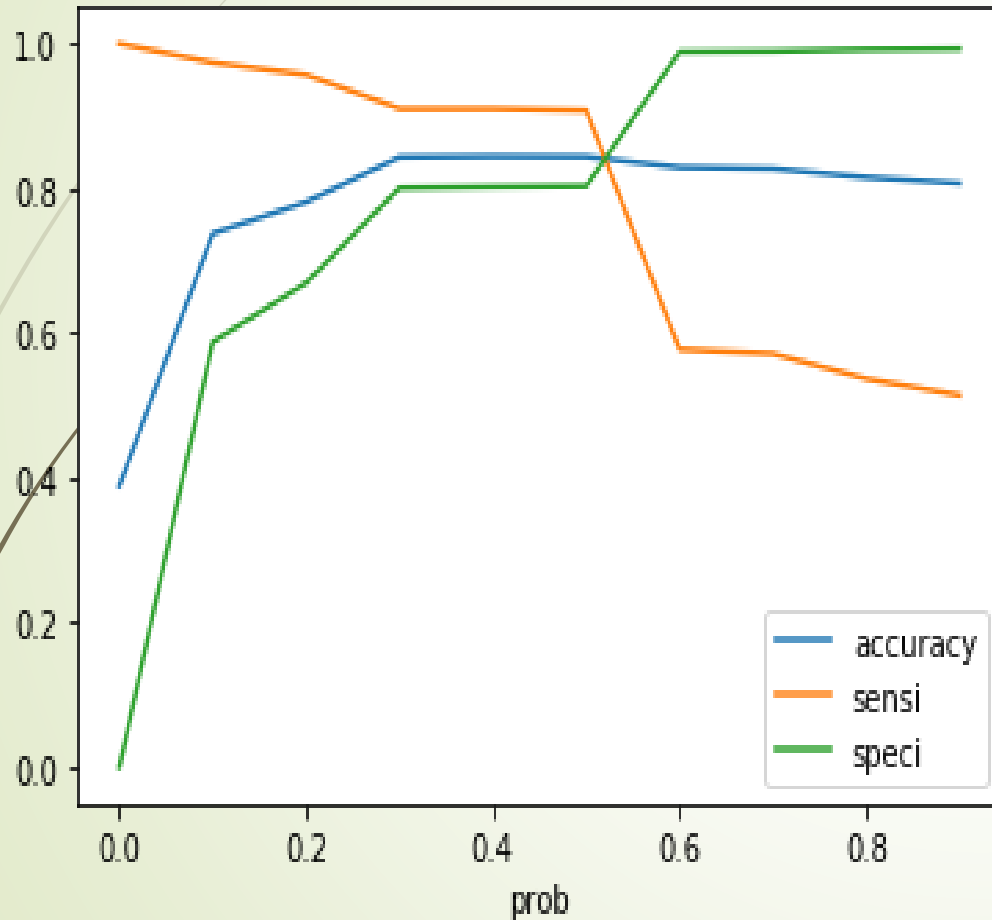
**ROC AUC: 0.93**

**Precision: 0.74**

**Recall: 0.90**

- ▶ Precision tells us how accurate the model was in predicting the positive samples out of all the samples predicted to be positive. Recall tells us how accurately the model was able to identify the positive samples out of all positive samples that were actually present.
- ▶ Here, we should prefer high recall as the positive leads identified correctly need to be approached without wasting time on leads that are wrongly predicted as positive.

# Visualization of Optimal Cut-Off and ROC Curve





# Model Built using Logistic Regression

- ▶ Performance metrics on Test Set

**Overall Accuracy: 0.85**

**Sensitivity: 0.92**

**Specificity: 0.80**

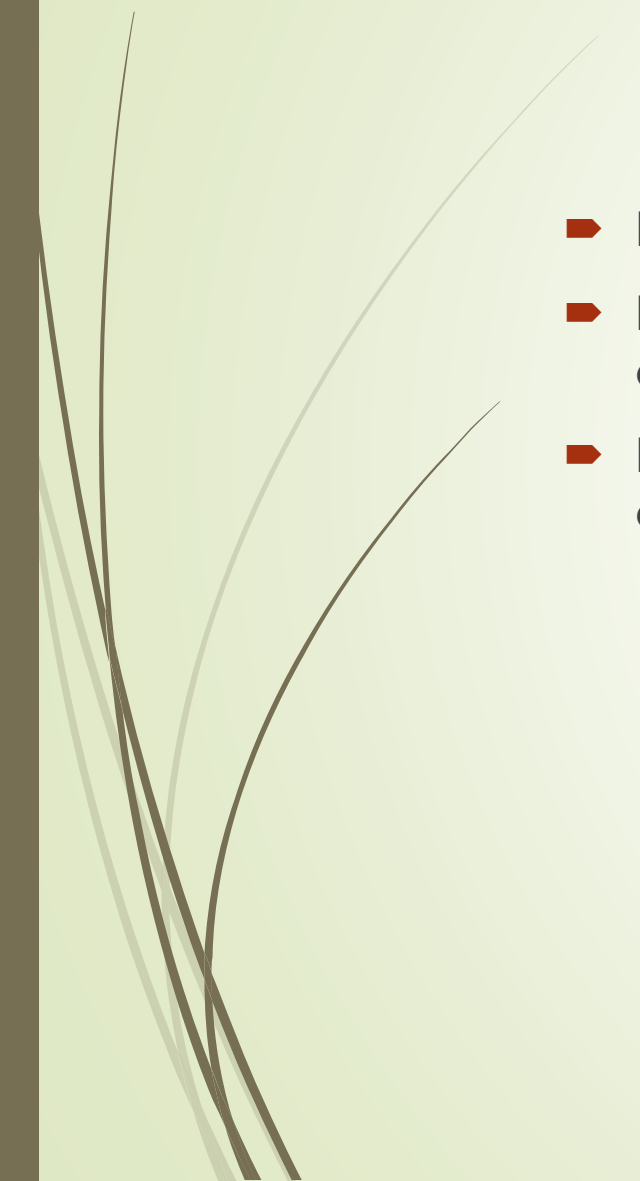
**Precision: 0.74**

**Recall: 0.92**

- ▶ The performance metrics with test data are also in line with the performance metrics of the training data.



# Lead Score

- ▶ Lead score between 0-100 is assigned to each data point in the test set.
  - ▶ High lead score between 50 -100 indicates hot lead i.e., lead having higher chances of conversion. These leads can be approached.
  - ▶ Low lead score between 0-50 indicates cold lead i.e., lead having lower chances of conversion.
- 



# Suggestions to the sales team

- Lead Source, Current occupation of the lead, and Tags assigned to the leads are the top three variables which contribute most towards the probability of a lead getting converted.
- 'Reference', 'Welingak Website', and 'Working Professional' are the top 3 categorical/dummy variables which should be focused the most on, in order to increase the probability of lead conversion.
- The sales team can suggest the interns to approach 'working professionals' as they would want to continue education while working with the flexibility provided by online learning.
- At the same time, leads suggested by 'references' has the potential for higher conversion.
- Offering more referral benefits than usual over a short duration can attract more aspirants in the said duration.
- Even leads through 'Welingak Website' would be hot leads as the traffic through this lead source is significantly high.
- Moreover, the tags assigned to the leads are aptly indicative that provide good support in the prediction of lead conversion.
- Lead profile also would provide relevant information to follow them up for positive response.





# Suggestions to the sales team

- As the utilization of time have to be optimized by the sales team, it is good to focus only on highly potential leads rather than spending time with no outcome.
- 'Lead profile' assigned to the lead would help in this as a first step.
- The 'working professionals' would be among hot leads as they have the capability to pay the fee for the course and would take independent decisions when compared to student community or unemployed group.
- Leads referred by alumni would be more promising as they would have already enquired the alumni about the course/program and have been inspired by them to take up the course for better career prospects.



Thank You