


LEAD SCORING CASE STUDY

Submitted By

1. Chayanika Hazra
2. Bhagyashree Bose
3. Prineet Tiwari



CONTENTS

1. Problem Statement
 2. Goal
 3. Approach
 4. EDA
 5. Model Evaluation
 6. Insights and Suggestions
- 



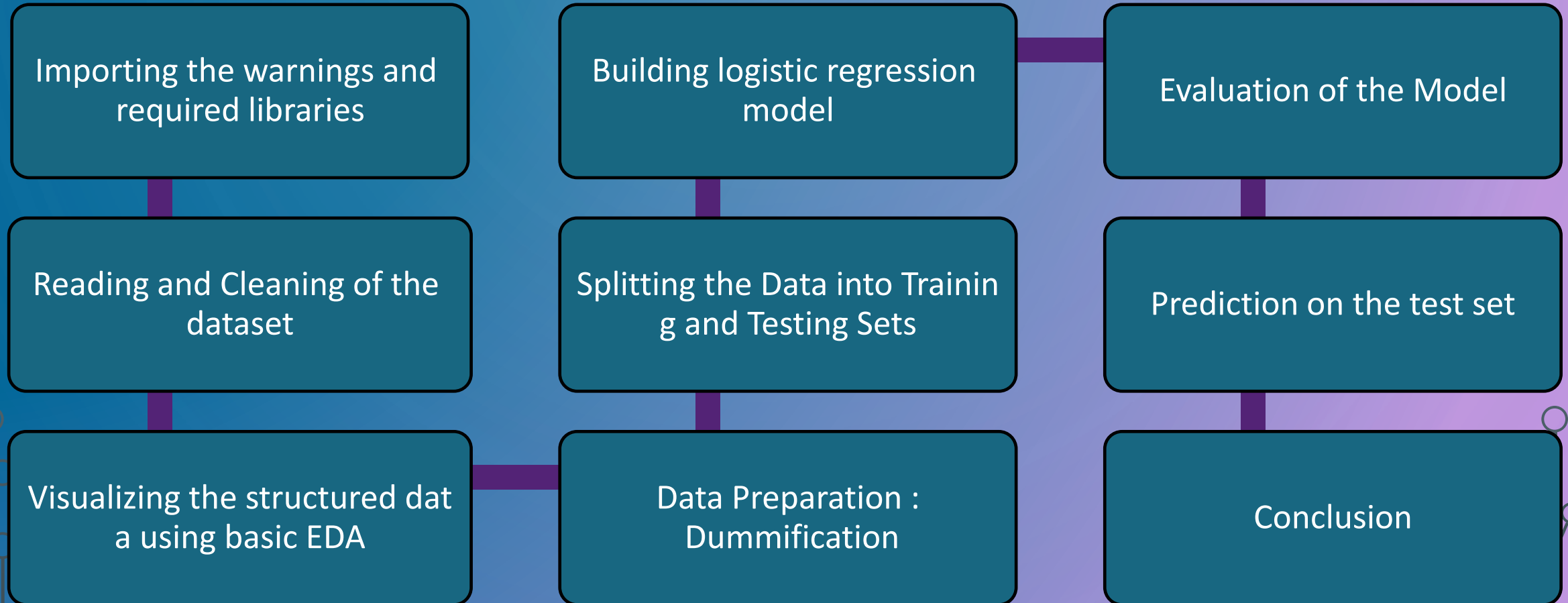
PROBLEM STATEMENT

- 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 through which people land on the website where they might fill up a form for the course in which they provide their email address or phone number. These are classified to be a lead. Additionally, 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 at X education is around 30%.
- Now, 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.

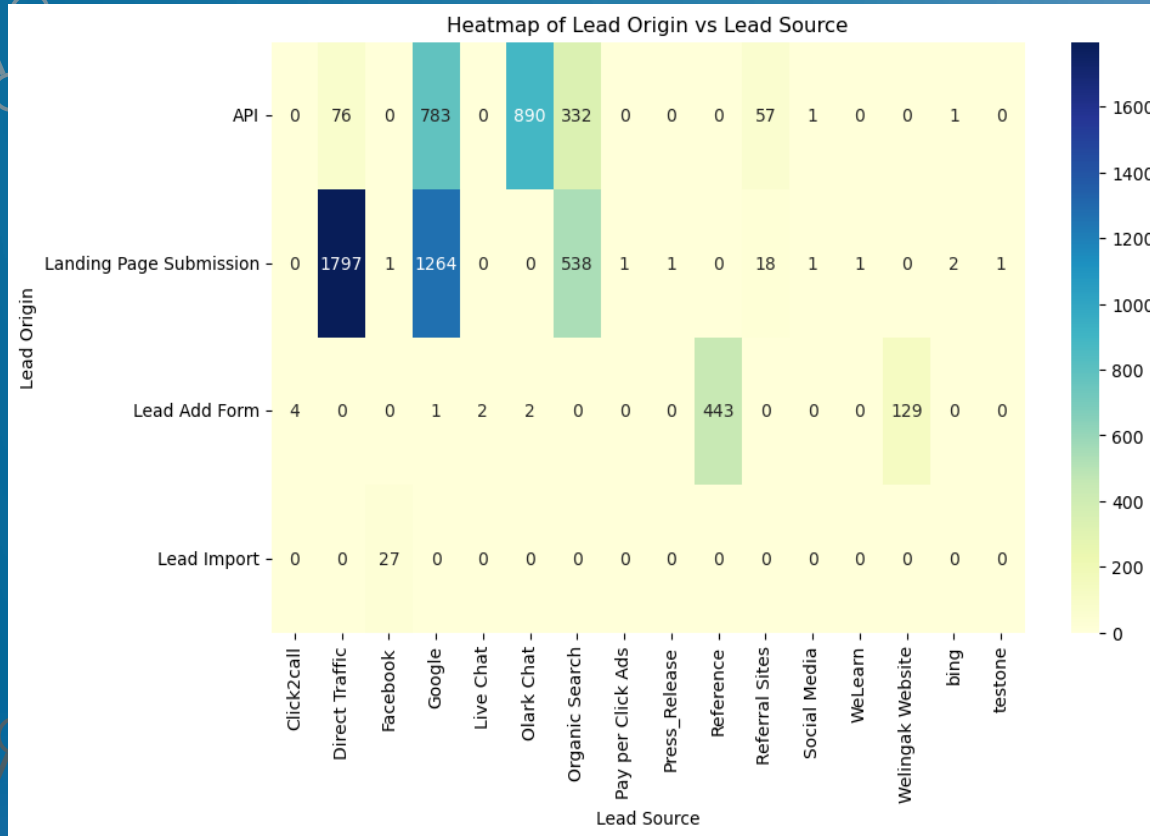
GOAL

- Building a logistic regression 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. %.
- The CEO wishes to achieve a lead conversion rate of 80%.

APPROACH



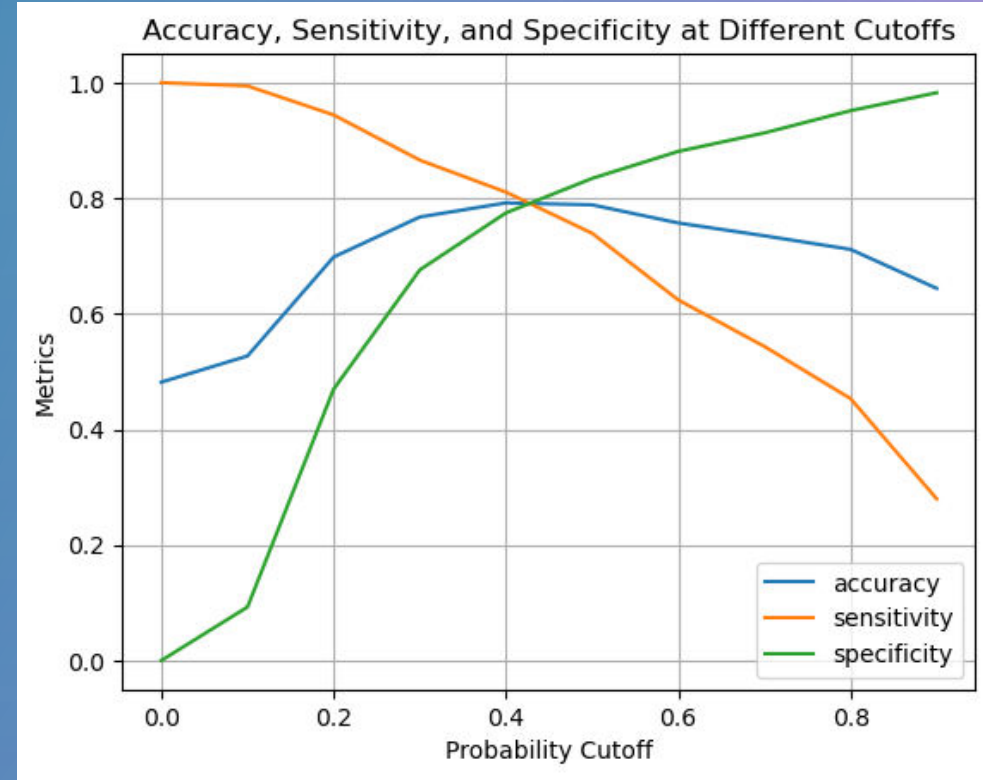
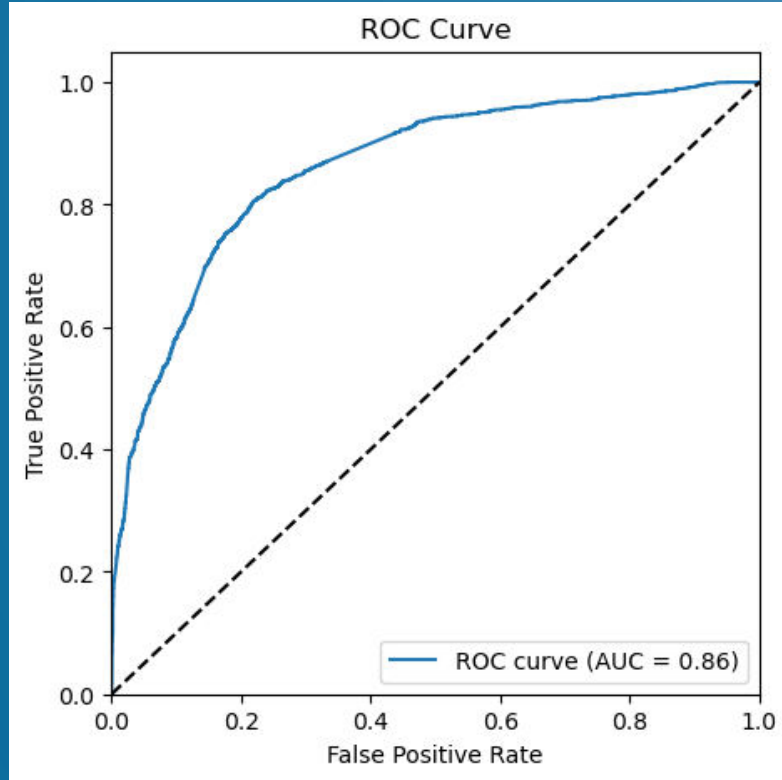
EDA



Insights-

1. Landing Page Submissions are the primary driver of leads across various sources.
2. The Lead Add Form is not generating significant leads, conducting a thorough analysis to identify potential issues with the form's design, placement, or user experience.
3. Google and Direct Traffic are consistently high-performing sources for leads.
4. Referral Sites and Social Media are not contributing significantly to lead generation.
5. While Landing Page Submissions and Google/Direct Traffic are performing well, diversifying lead sources can reduce risk and improve overall lead quality.

MODEL EVALUATION



INSIGHTS

1. The plot above shows the relationship between different probability cutoffs and the associated accuracy, sensitivity, and specificity metrics of the model. As the cutoff increases, sensitivity tends to decrease, indicating that fewer true positives are being captured, while specificity increases, suggesting better identification of true negatives.
2. Accuracy rises initially with increasing cutoffs but eventually plateaus, reflecting the trade-off between capturing more positive cases versus correctly identifying negatives.
3. The final Cut-off was selected as 0.42.

MODEL EVALUATION

1. Overall Accuracy: 78%
2. Precision: 80%
3. Recall: 73%

INSIGHTS

- The top three variables which are contributing to the final model are
 1. Total visits
 2. Views per visit
 3. Total time spent on Website
- The top three categorical/dummy variables on which we should focus the most are
 1. Total visits
 2. Views per visit
 3. Total time spent on Website

SUGGESTIONS

1. Optimizing Precision and Recall:

- The balance between precision and recall is crucial for maximizing conversion opportunities. Given the importance of capturing positive cases, strategies should be considered to enhance recall without significantly impacting precision. A deeper dive into the features and potential adjustments to the classification threshold could yield improvements.

2. Addressing Class Imbalance:

- If the dataset reflects an imbalance between classes, it is critical to employ additional metrics beyond accuracy. Metrics such as precision, recall, and the F1 score will provide a more nuanced understanding of model performance across all classifications.

3. Improvement Initiatives:

- With 249 false negatives identified, it is recommended to explore avenues such as feature engineering, hyperparameter tuning, and experimenting with alternative algorithms to boost recall. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) may also be beneficial in balancing class representation during training.

4. Comprehensive Evaluation:

- It is advisable to utilize additional evaluation metrics, such as F1 score, AUC-ROC (Area Under the Receiver Operating Characteristic Curve), and PR AUC (Area Under the Precision-Recall Curve), to gain a holistic view of the model's performance.

SUGGESTIONS

- The company should make calls to the following leads that-
 1. Spend more time on the website
 2. Originate from Landing Page Submissions
 3. Originate from Google and Direct Traffic