LEAD SCORING CASE STUDY

PRESENTED BY : RICHITA HIRANI | YASH PANDEY | JATIN ARORA

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. 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. 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 at 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.

GOAL OF THE CASE STUDY

Build 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.

AGENDA

The purpose of this case study is to optimize the lead scoring mechanism based on their behavior, demographics, tendency etc. The study aims to implement a lead point system for identifying the probability of lead conversion.

APPROACH

1. Import necessary libraries
2. Sourcing the data for analysis
3. Reading and understanding the dataset (Reading CSV file)
4. Data Cleaning (Missing Values, Outliers, formatting, standardization)
5. Exploratory data analysis (Uni / Bi / Multi Variate analysis)
6. Feature Scaling
7. Splitting the dataset into Test & Train datasets
8. Prepare the data for model building
9. Logistic Regression Model Building
10. Model Evaluation - Specificity | Sensitivity | Precision Recall
11. Predictions on the Test dataset

DATA CLEANING & PREPARATION :

- After replacing the missing & ‘Select’ values with NaN, the columns with more than 40% of missing values were excluded from the dataset as they become insignificant in our analysis.
- The columns with less than 5% of missing values were replaced with mode of their respective columns.
- Further, for columns with missing values ranging between 6% to 39% were analysed individually.
- The country column was excluded as India was the most common occurrence and became insignificant for further analysis.
- Missing values in City & Current Occupation columns were replaced with Mode values
- Missing values in Specialization, Tags, Lead Source columns were tagged as Unknown and the smaller representing variables were consolidated into Others category
- Dropped columns with unique values & single unique entries & Data imbalance (99% same entry) as they are insignificant in our analysis
- Numerical Variables Outliers : Below 5% & above 95% quantiles are excluded in the data.

EXPLORATORY DATA ANALYSIS :

- Conversion Rate is 38%

- **LEAD ORIGIN** : Absolute counts of leads for API (39%) & Landing Page Submission (53%) is very high However, the Lead add form Lead origin has very high conversion rate. To improve the target conversion goal, we must increase lead count for Lead Add Form Lead Origin and we must increase conversion rate for API & Landing Page Submission
- **LEAD SOURCE** : The conversion rate for welingkar website & references is very high Followed by good conversion rate for Google & Organic search Absolute counts for Google and Direct Traffic are high
- **LAST ACITIVITY** : SMS Sent has very high conversion rate as well as absolute count. However, Email Opened has high count but low conversion rate.
- **SPECIALIZATION** : Highest conversion rate is for Investment & Insurance specialization followed by Management Speiclisations, Business administration and Rural Agribusiness.

EXPLORATORY DATA ANALYSIS :

- **WHAT IS YOUR CURRENT OCCUPATION** : Working professionals have very high conversion rate.
- **TAGS**: Will revert after reading the email & closed by Horizzon have extremely high conversion rate.
- **LAST NOTABLE ACTIVITY**: SMS sent has highest conversion rate
- **CORRELATION MATRIX FOR NUMERICAL VARIABLES** :

Strong Correlation (0.77) between Page Views Per Visit & TotalVisits

Good correlation (0.36) between Total time spend on website with Converted variable

DATA PREPARATION FOR MODEL BUILDING:

- Converted Binary variables into 0 and 1
- Created dummy variables for categorical variables.
- Feature scaling of numerical data
- Splitting data into Train & Test datasets

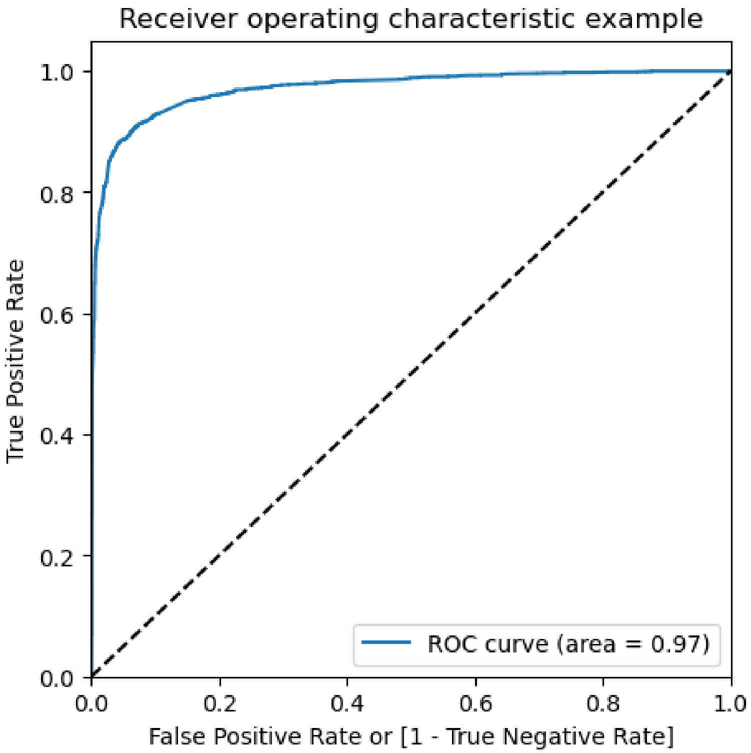
LOGISTIC REGRESSION MODEL BUILDING:

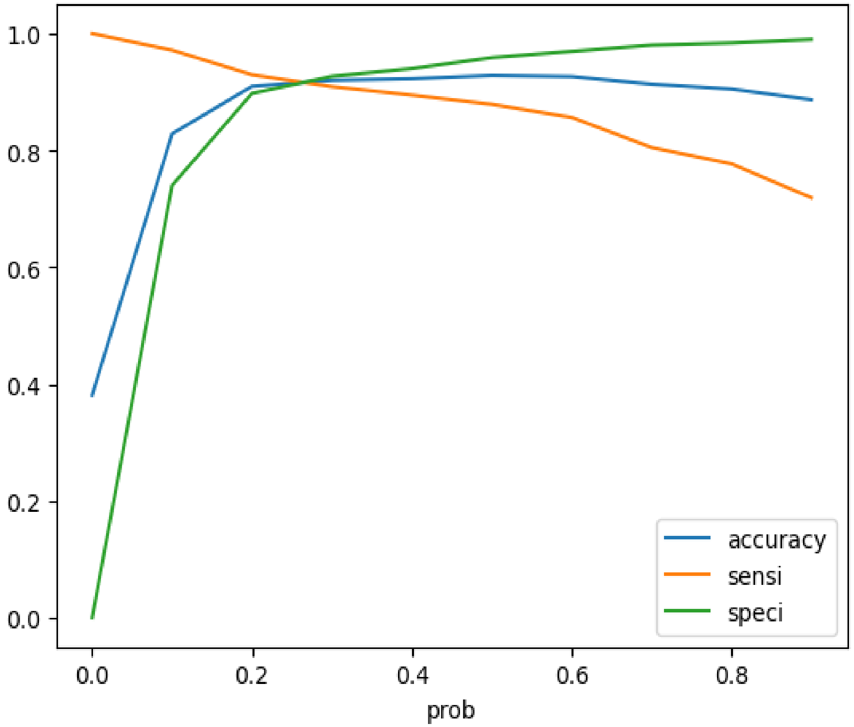
- Feature Selection using RFE
- Determine Optimal model using logistic regression
- Calculate Accuracy, Sensitivity, Specificity, Precision, Recall, Model Evaluation

VARIABLES IMPACTING THE CONVERSION RATE :

- Tags (Closed by Horizzon, Lost to ENIS, Will Revert after reading the Email)
- Lead Source (Welingak Website, Reference)
- Last Activity (SMS sent)
- Total Time Spent on Website
- Total Visits

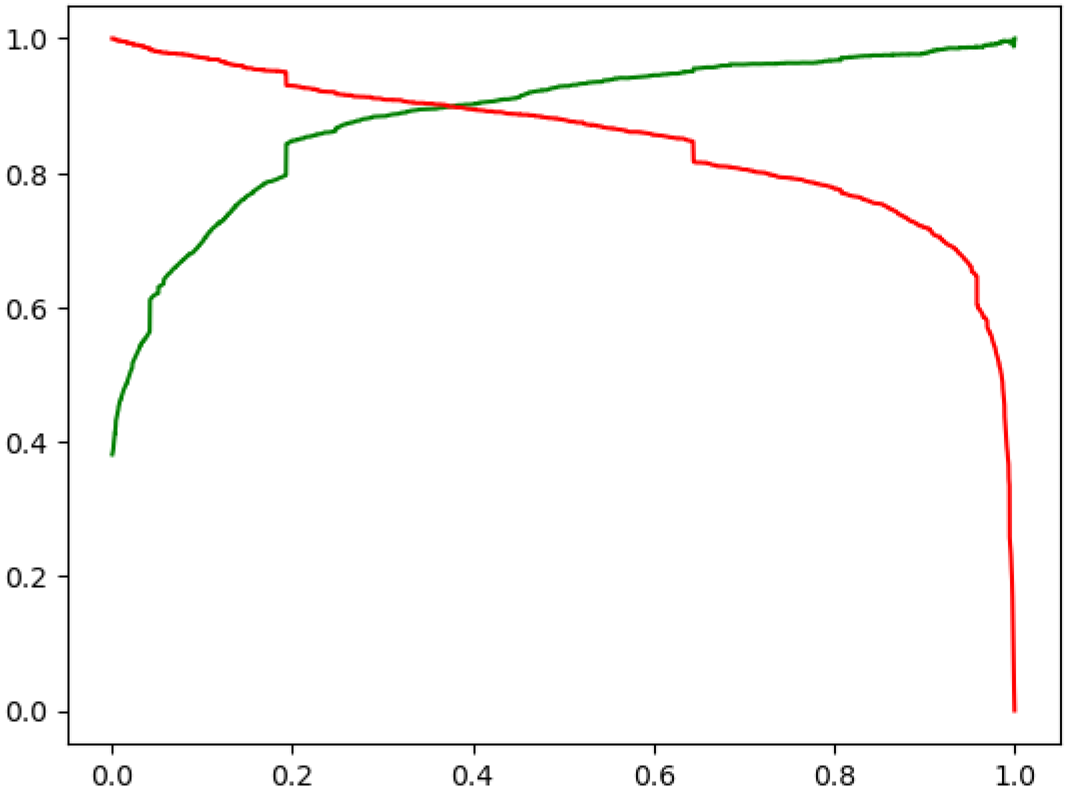
MODEL EVALUATION : TRAIN DATASET :





- Accuracy : 92.01%
- Sensitivity : 90.92%
- Specificity : 92.68%

MODEL EVALUATION : PRECISION & RECALL : TRAIN DATASET :



- Precision : 88.44%
- Recall : 90.92%

MODEL EVALUATION : TEST DATASET :

- Accuracy : 92.39%

- Sensitivity : 92.69%

- Specificity : 92.19%

FINAL OBSERVATIONS :

- **TRAIN DATASET**

- Accuracy : 92.01%
- Sensitivity : 90.92%
- Specificity : 92.68%
- Precision : 88.44%
- Recall : 90.92%

- **TEST DATASET**

- Accuracy : 92.39%
- Sensitivity: 92.69%
- Specificity : 92.19%
- Precision : 88.56%
- Recall : 92.69%

The Model predicts the conversion rate very well and thus can be shared with the business team for decision making.

SUMMARY :

Depending the important variables the target customers must be identified and these customers must be attended with personalized experience such as the information must be kept handy and elaborate about the desired products. Monitor each of these lead carefully so that you can tailor the information you sent to them. A carefully drafter plan which caters the needs to each of these leads will go a long way to improve probability of lead conversion. Hold open discussion forums and informative session for resolution of enquiries and FAQs.