# Lead Score Case Study

Pravat Jena

&

Anusha N

# Agenda

- Problem Statement
- Objective
- Data
- Analysis & Preparation of Data
- Model
- Model Accuracy and Metrics
- Conclusion

# 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%
- The CEO has given a ballpark of the target lead conversion rate to be around 80%

# Goal

- 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
- Find the following based on your model:
  - Top three variables which contribute most towards the probability of a lead getting converted
  - Top 3 categorical/dummy variables which should be focused the most on in order to increase the probability of lead conversion
  - Strategy for phone call to the potential lead candidates
  - Strategy to find whether phone call is necessary or not

# Data - Summary

- Data Source: Leads dataset from the past of X Education
- No. of records (rows): 9240
- No. Attributes (columns): 37
- Attributes: Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- Target variable: Converted. Values: 0 ( Not concerted) & 1(Converted)

# Exploratory Data Analysis and Treatment

- **Columns with missing values**

| Columns | Missing % |
|---|---|
| Lead Quality | 51.59 |
| Asymmetrique Profile Score | 45.65 |
| Asymmetrique Activity Score | 45.65 |
| Asymmetrique Profile Index | 45.65 |
| Asymmetrique Activity Index | 45.65 |
| Tags | 36.29 |
| What matters most to you in choosing a course | 29.32 |
| Lead Profile | 29.32 |
| What is your current occupation | 29.11 |
| Country | 26.63 |
| How did you hear about X Education | 23.89 |
| Specialization | 15.56 |
| City | 15.37 |
| TotalVisits | 1.48 |
| Page Views Per Visit | 1.48 |
| Last Activity | 1.11 |
| Lead Source | 0.39 |

# Exploratory Data Analysis and Treatment

- Columns with values as: "Select" which can treated as missing values:
  - How did you hear about X Education: 5043
  - Lead Profile: 4146
  - City: 2249
  - Specialization: 1942
- Columns with single unique values:
  - Get updates on DM Content
  - I agree to pay the amount through cheque
  - Receive More Updates About Our Courses
  - Magazine
  - Update me on Supply Chain Content

# Exploratory Data Analysis and Treatment

- Columns with almost single unique values
  - Do Not Call: No - 9238 & Yes – 2
  - What matters most to you in choosing a course: Better Career Prospects- 6528 out of 6531
  - Search: No - 9226 out of 9240
  - Newspaper Article: No - 9238 out of 9240
  - X Education Forums: No - 9239 out of 9240
  - Newspaper: No - 9239 out of 9240
  - Digital Advertisement: No - 9236 out 0f 9240
  - Through Recommendations: No - 9233 out of 9240

# Exploratory Data Analysis and Treatment

- Drop columns
  - Missing values > 60 %
  - Univariate Analysis  (and higher missing values)
    - Lead Number
    - 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
    - A free copy of Mastering The Interview
    - Country

# Exploratory Data Analysis and Treatment

- Imputing
  - City : Replace missing values by Mumbai
  - Specialization – Replace the missing values by another value('Others')
  - Tags: Replace missing values by 'Will revert after reading the email'
  - What is your current occupation: Missing values by Unemployed
  - Country: Missing values by India as most occurring in the dataset
  - Lead Source: Check for similar options as well as club options
  - Last Activity: Club options to Other_Activity

# Exploratory Data Analysis and Treatment

- Dummy Features
  - For categorical variables with multiple levels, create dummy features
    - Lead Origin
    - Lead Source
    - Last Activity
    - Specialization
    - City
    - What is your current occupation
    - Tags
    - Lead Quality
    - Last Notable Activity
- Feature Scaling
  - Feature scaling helps us to scale all data to one range so that it is easy to compare the columns within same scale
  - StandardScaler

# Model

- Feature Selection: 15 features selected using Recursive Feature Selection (RFE)
  - Do Not Email
  - Lead Origin_Lead Add Form
  - Lead Source_Welingak Website
  - What is your current occupation_Working Professional
  - Tags_Busy
  - Tags_Closed by Horizzon
  - Tags_Lost to EINS
  - Tags_Ringing
  - Tags_Will revert after reading the email
  - Tags_invalid number
  - Tags_switched off
  - Tags_wrong number given
  - Lead Quality_Not Sure
  - Lead Quality_Worst
  - Last Notable Activity_SMS Sent

# Model
## Summary

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6351 |
| **Model:** | GLM | **Df Residuals:** | 6337 |
| **Model Family:** | Binomial | **Df Model:** | 13 |
| **Link Function:** | logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -1588.8 |
| **Date:** | Sun, 01 Mar 2020 | **Deviance:** | 3177.6 |
| **Time:** | 12:59:26 | **Pearson chi2:** | 3.08e+04 |
| **No. Iterations:** | 8 | **Covariance Type:** | nonrobust |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -2.0888 | 0.216 | -9.654 | 0.000 | -2.513 | -1.665 |
| **Do Not Email** | -1.3012 | 0.212 | -6.134 | 0.000 | -1.717 | -0.885 |
| **Lead Origin_Lead Add Form** | 1.0894 | 0.363 | 3.001 | 0.003 | 0.378 | 1.801 |
| **Lead Source_Welingak Website** | 3.4138 | 0.818 | 4.173 | 0.000 | 1.810 | 5.017 |
| **What is your current occupation_Working Professional** | 1.3403 | 0.291 | 4.602 | 0.000 | 0.769 | 1.911 |
| **Tags_Busy** | 3.8040 | 0.330 | 11.532 | 0.000 | 3.157 | 4.450 |
| **Tags_Closed by Horizzon** | 7.9562 | 0.763 | 10.433 | 0.000 | 6.461 | 9.451 |
| **Tags_Lost to EINS** | 9.1785 | 0.754 | 12.177 | 0.000 | 7.701 | 10.656 |
| **Tags_Ringing** | -1.6947 | 0.337 | -5.036 | 0.000 | -2.354 | -1.035 |
| **Tags_Will revert after reading the email** | 3.9665 | 0.229 | 17.311 | 0.000 | 3.517 | 4.416 |
| **Tags_switched off** | -2.2882 | 0.587 | -3.900 | 0.000 | -3.438 | -1.138 |
| **Lead Quality_Not Sure** | -3.3406 | 0.128 | -26.026 | 0.000 | -3.592 | -3.089 |
| **Lead Quality_Worst** | -3.7624 | 0.850 | -4.426 | 0.000 | -5.428 | -2.096 |
| **Last Notable Activity_SMS Sent** | 2.7406 | 0.120 | 22.847 | 0.000 | 2.506 | 2.976 |

# Model - Summary

# Model
## Features VIF

| | | |
|---|---|---|
| **8** | Tags_Will revert after reading the email | 2.89 |
| **12** | Last Notable Activity_SMS Sent | 2.85 |
| **1** | Lead Origin_Lead Add Form | 1.62 |
| **7** | Tags_Ringing | 1.56 |
| **2** | Lead Source_Welingak Website | 1.36 |
| **3** | What is your current occupation_Working Profes... | 1.26 |
| **5** | Tags_Closed by Horizzon | 1.15 |
| **0** | Do Not Email | 1.11 |
| **4** | Tags_Busy | 1.11 |
| **10** | Lead Quality_Not Sure | 1.11 |
| **6** | Tags_Lost to EINS | 1.05 |
| **9** | Tags_switched off | 1.04 |
| **11** | Lead Quality_Worst | 1.02 |

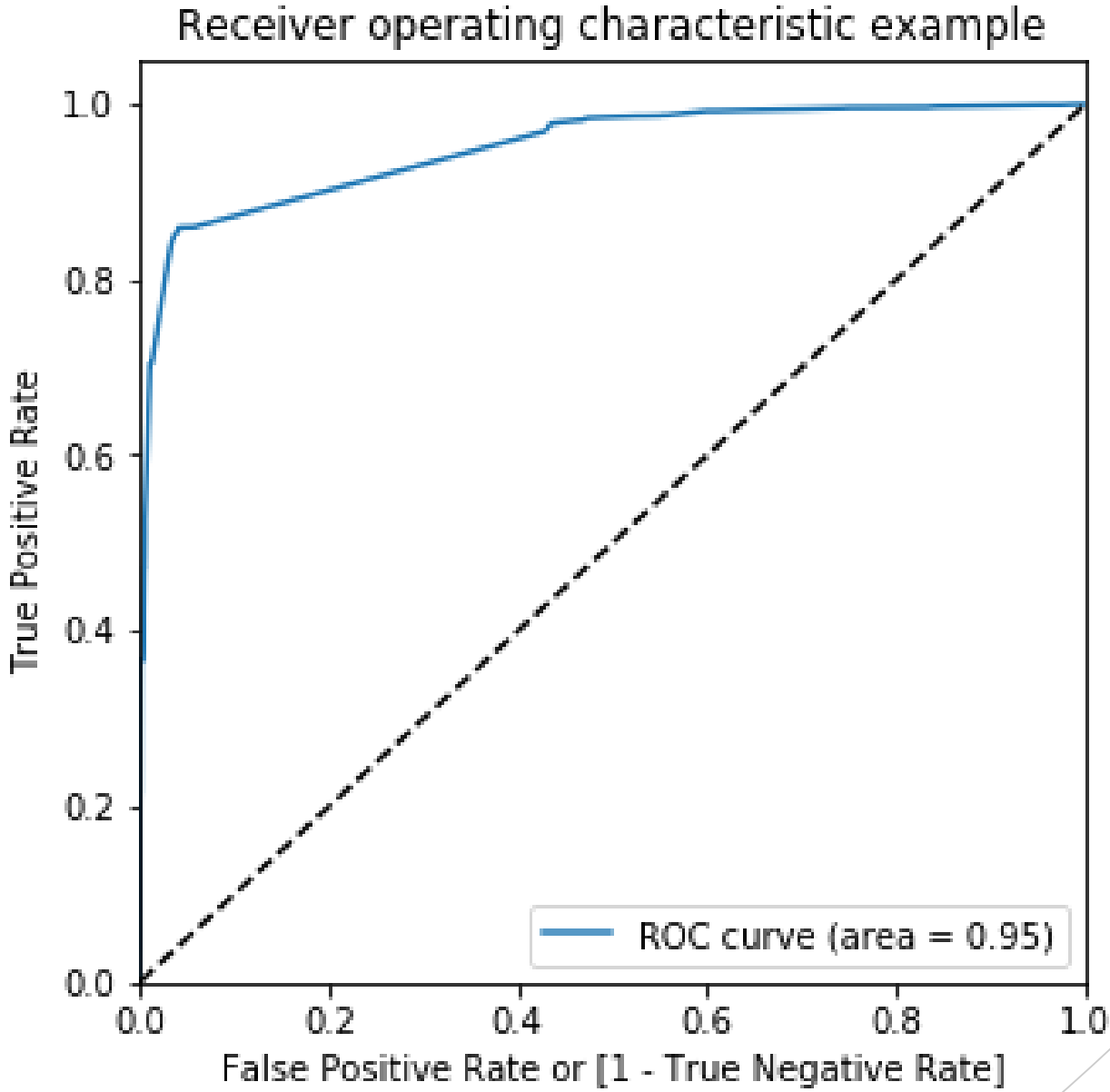
# Model
## Insights

- p-Values and VIF for features within the standard range
  - Max p-Valaues for all the features = 0.003
  - Max VIF for all the features = 2.89
- Assumption:
  - Churn probability > 0.5 is converted
  - Otherwise not converted
- Accuracy: 92%

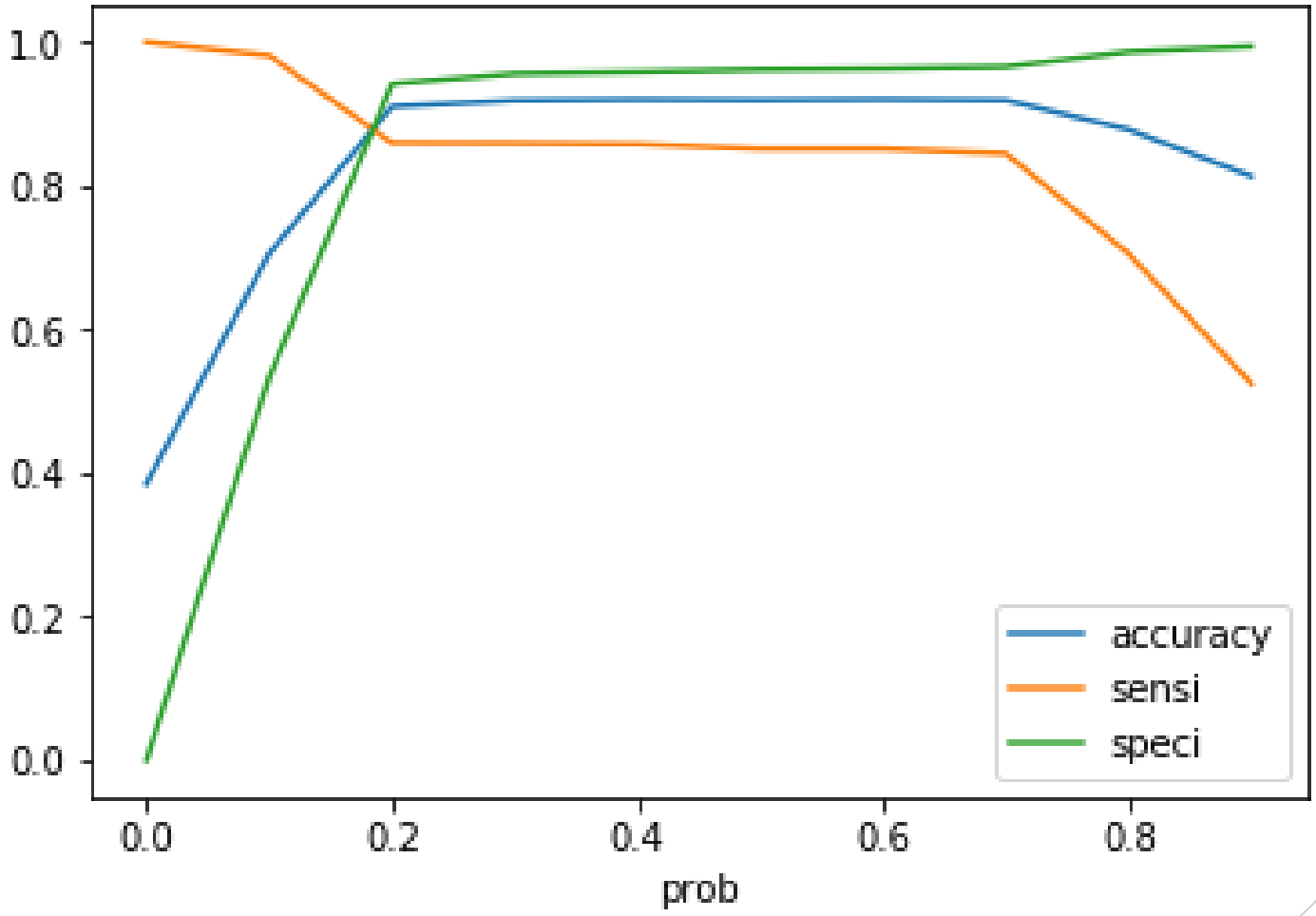
# Model Accuracy & Metrics

| Metrics | Values (%) | Inferences on Model |
|---|---|---|
| Sensitivity | 85.16 | High accuracy of positive prediction to lead conversion |
| Specificity | 96.18 | High accuracy of negative prediction to non lead conversion |
| False Positive Rate | 3.81 | Low rate of non potential leads being predicted as potential |
| Positive Predictive Values | 93.32 | High percentage lead conversion with respect to hot leads |
| Negative Predictive Value | 91.19 | High percentage of lead non conversion with respect to cold leads |

# Model ROC Curve



# Model
## Optimal Probability



# Model Accuracy & Metrics

| Metrics | Values (%) |
|---|---|
| Accuracy | 90.67 |
| Sensitivity | 84.33 |
| Specificity | 94.29 |
| Precision | 93.32 |

*Metrics for test data based on the optimal cut off churning probability of 0.2*

# Conclusion

- The ROC Curve Area = 0.95
- Over all accuracy of the model is more than 90 % which is more than the objective of > 80% conversion
- So, if we pick up dataset with a lead score greater than 20, we will get more than 80 percent conversation rate

# Conclusion

Considerations for better  hot leads

- Continue sourcing leads  from **Welingak Website**
- Target more **working professionals**
- Continue leads closure by **Horizzon**
- **Don't** target leads who opt for **don't email**
- Target leads who says**: will revert back after reading email (Tags)**
- **Don't** target leads if **lead quality** is **worst** or **not sure**
- **Don't** target leads if the **phone is switched off (tags)**