LEAD SCORING CASE STUDY

KANISHKA SINHA

CONTENT SLIDE

- AGENDA
- PROBLEM STATEMENT
- GOAL
- DATA SOURCING & CLEANING
- OUTLIERS
- EXPLORATORY DATA ANALYSIS
- DATA PREPARATION
- MODEL BUILDING
- MODEL EVALUATION
- VARIABLES IMPACTING CONVERSION RATE
- RECOMMENDATIONS
- REFERENCES

AGENDA

The purpose is to optimize the lead scoring mechanism based on their fit, demographics, behaviors, interaction tendency etc. by implementing explicit & Implicit lead scoring modelling with lead point system.

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%.

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.

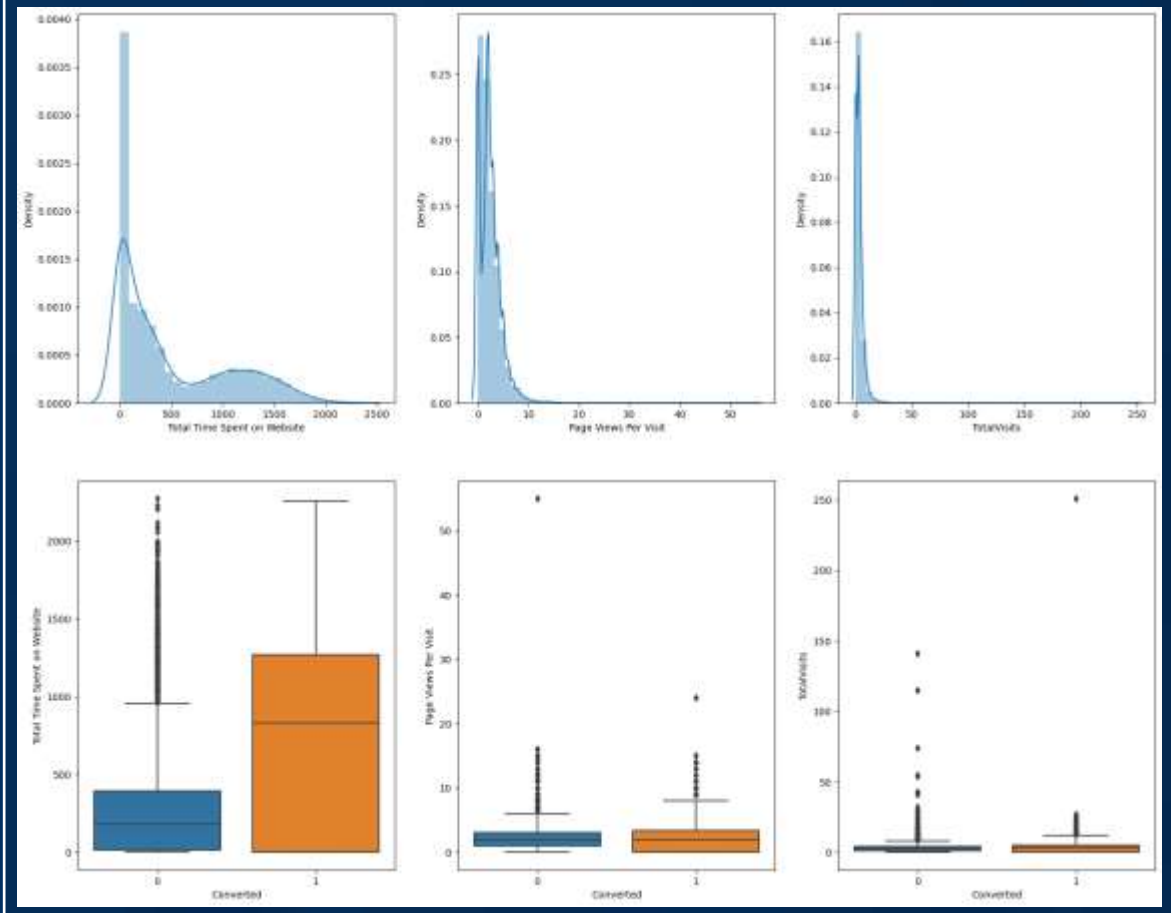
GOAL

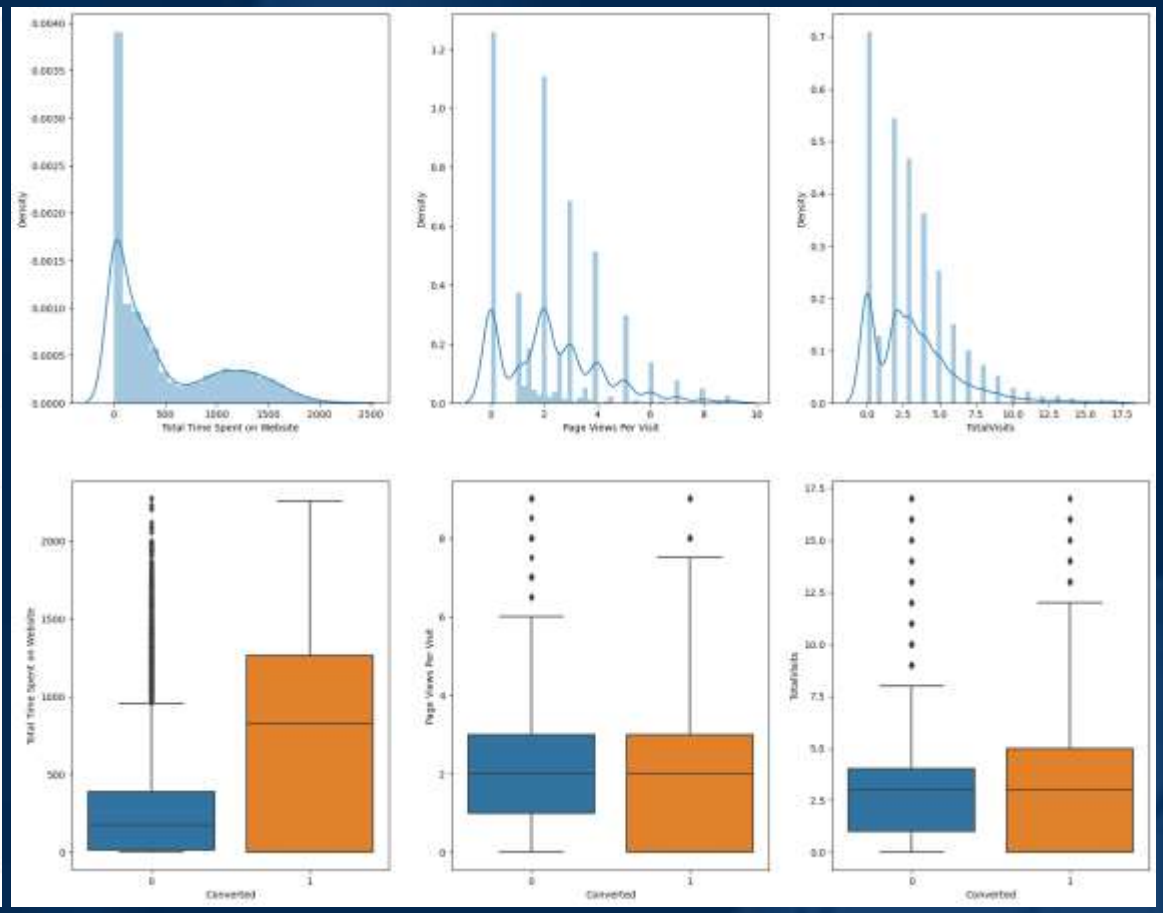
Build a logistic regression model to give each lead a score between 0 and 100 that the business may use to target potential leads. In contrast, a lower number would indicate that the lead is chilly and unlikely to convert, while a higher score would indicate that the lead is hot and most likely to convert. You will also need to deal with some additional issues that the firm has raised and that your model should be prepared to address if the company's requirements change in the future.

DATA SOURCING & CLEANING

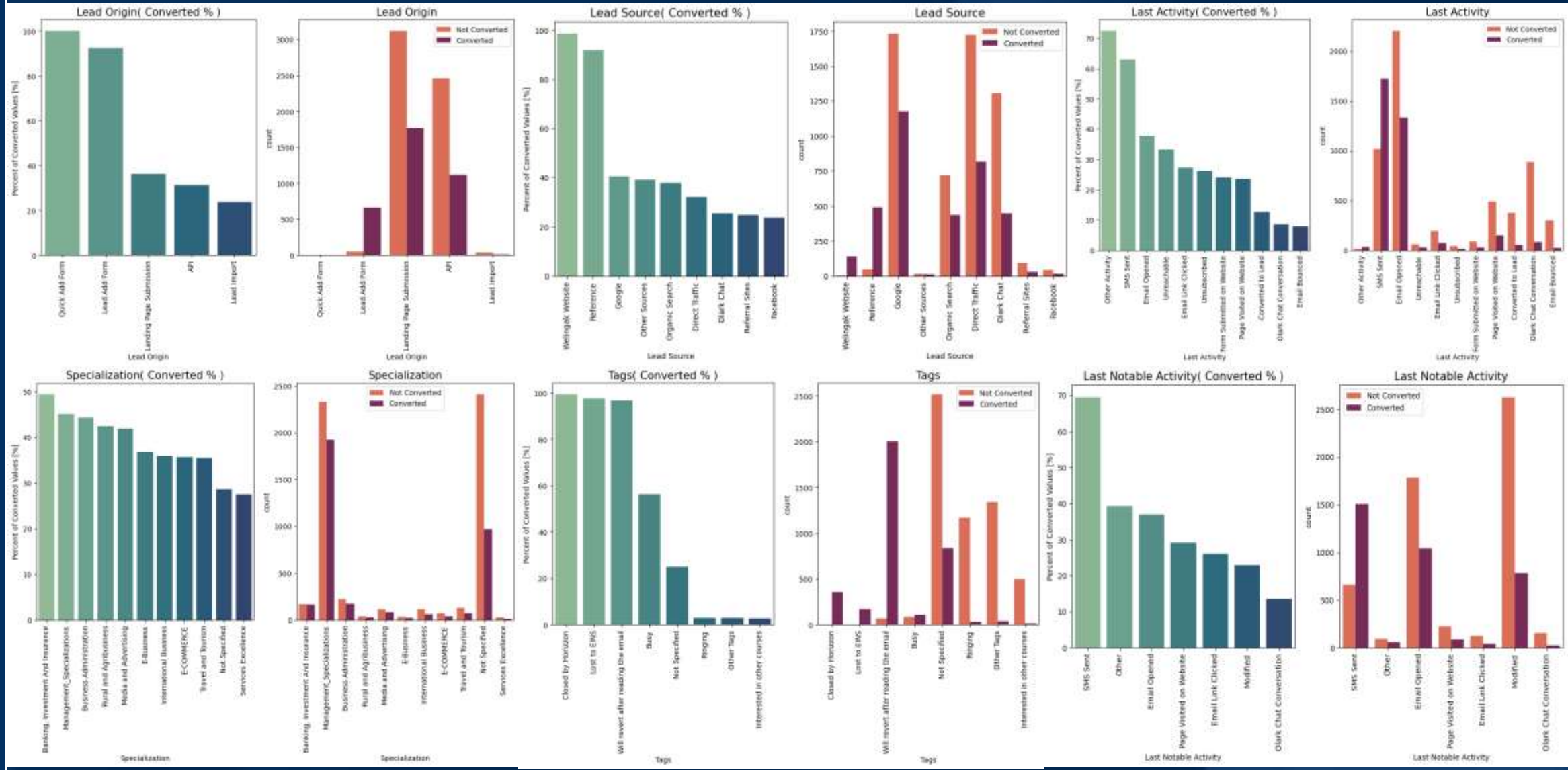
- Read the data from .csv file.
- Changed ‘Select’ to NaN.
- Removed columns consisting of more than 40% missing values.
- Removed columns with only one unique value and index columns.
- Imputed Null Values of some categorical variables with mode, while assigned the rest to a new value.
- Imputed Null Values of numerical variables with median.
- Merged lower frequency values of some categorical variables.
- Handled outliers.

OUTLIERS


Density
Total Time Spent on Website
Page Views Per Visit
TotalVisits
Converted


Density
Total Time Spent on Website
Page Views Per Visit
TotalVisits
Converted

EXPLORATORY DATA ANALYSIS



DATA PREPARATION

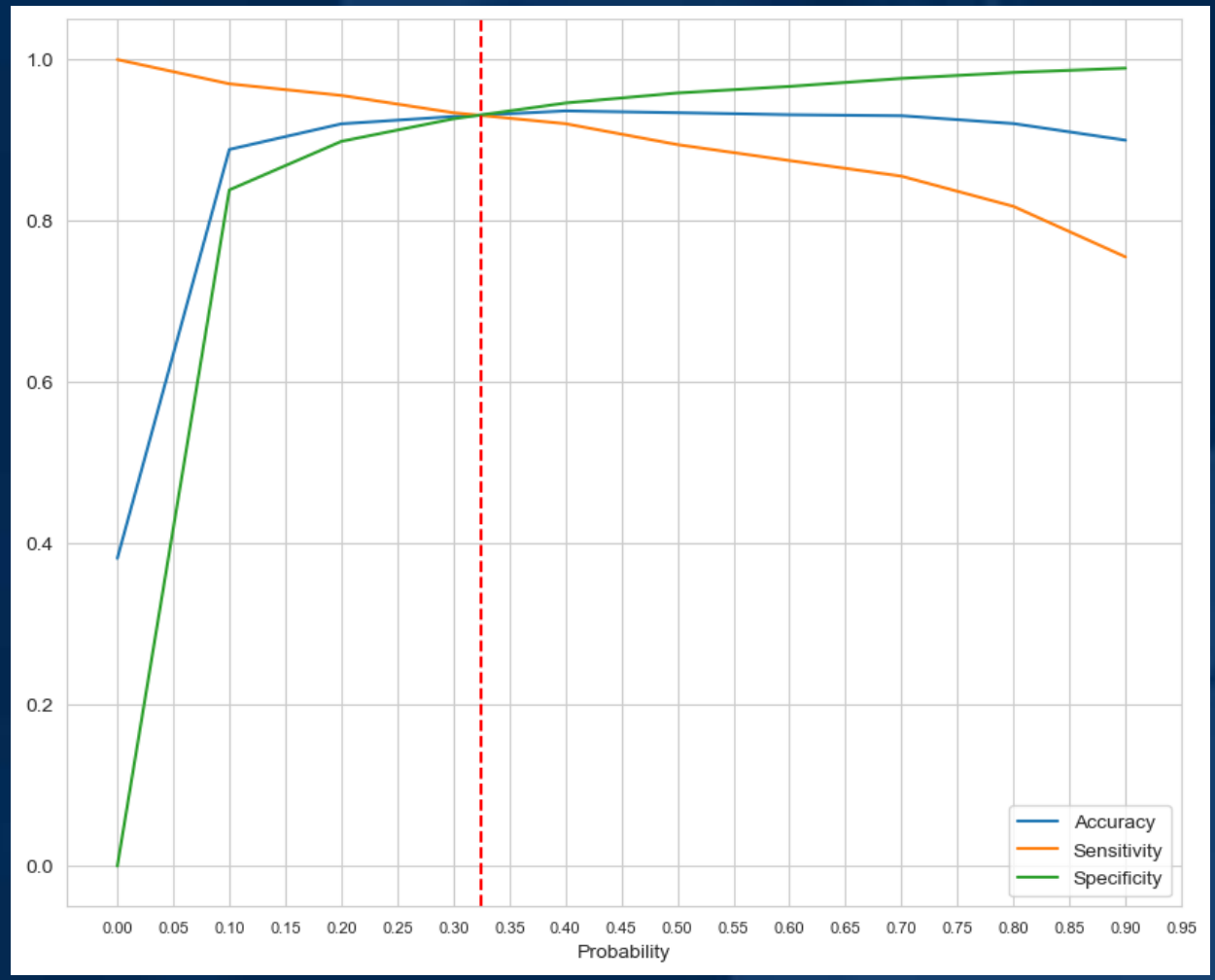
- Converted Binary variables into 0 & 1
- Created dummy variables for categorical variables
- Data was split in 70:30 ratio of train-test set.
- Feature Scaling was done.

MODEL BUILDING

- Feature Selection using RFE
- Determined Optimal Model using Logistic Regression
- Calculated Accuracy ,Sensitivity ,Specificity, Precision, Recall & evaluated model

MODEL EVALUATION

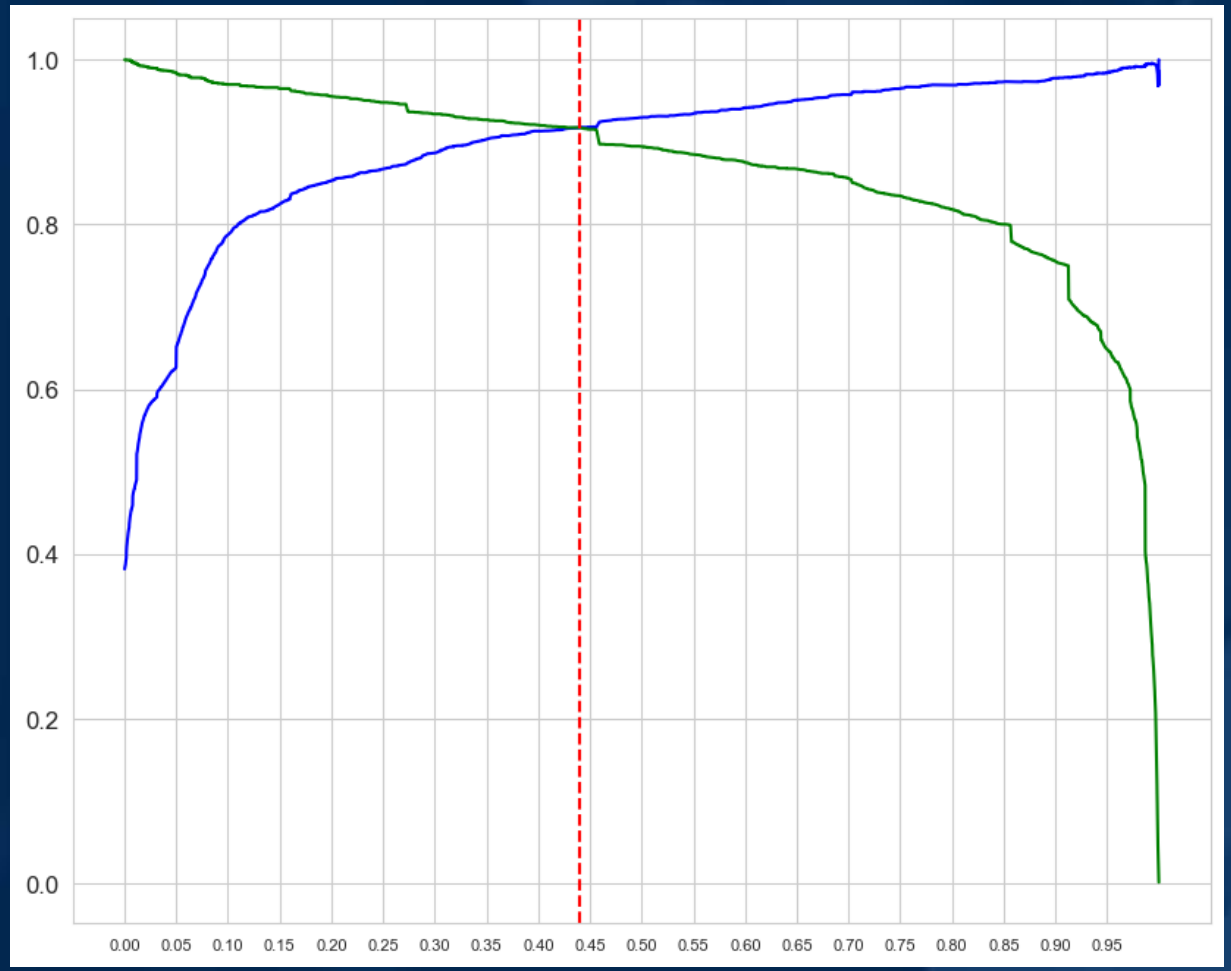
Graph depicts an optimal cutoff of 0.325 based on Accuracy, Sensitivity and Specificity.



- Accuracy: 93.2%
- Sensitivity: 93%
- Specificity: 93.32%

MODEL EVALUATION

Graph depicts an optimal cutoff of 0.44 based Precision and Recall.



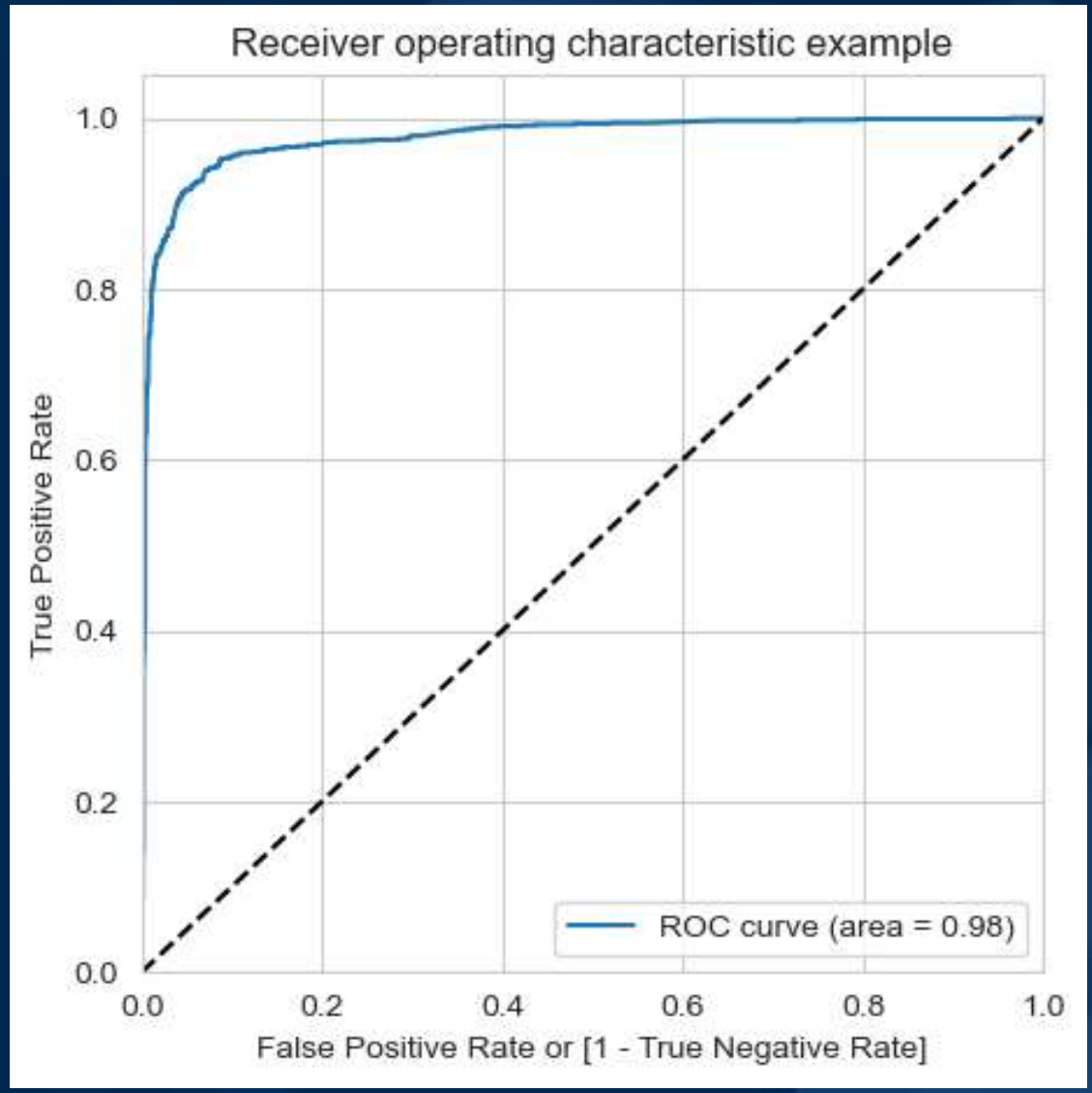
- Precision: 91.8%
- Recall: 91.76%

MODEL EVALUATION

On Test dataset

- Accuracy: 93.47 %
- Sensitivity: 92.13 %
- Specificity: 94.34 %

ROC is 0.98 which is very good



VARIABLES IMPACTING CONVERSION RATE

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Lead Source_Welingak Website
- Tags_Will revert after reading the email
- Last Activity_SMS Sent
- Total Time Spent on Website

RECOMMENDATIONS

- **Optimize Lead Scoring :** Refine lead scoring for better accuracy.
- **Data Quality & Enrichment :** Ensure clean, enriched lead data.
- **Personalized Nurturing :** Tailor emails and content for leads.
- **Segmentation & Targeting :** Segment leads, run targeted campaigns.
- **Sales-Marketing Collaboration :** Foster teamwork, define lead handoff.
- **CRM Utilization :** Maximize CRM for tracking and training.
- **Effective Lead Sourcing :** Analyze sources, allocate resources.
- **Lead Engagement Tracking :** Monitor engagement metric
- **Competitor Awareness** : Stay updated on competitors.
- **Customer Feedback Loop** : Gather insights, drive improvements.

REFERENCES

- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
- https://medium.com/@rithpansanga/logistic-regression-for-feature-selection-selecting-the-right-features-for-your-model-410ca093c5e0
- https://machinelearningmastery.com/rfe-feature-selection-in-python/
- https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/
- https://www.upgrad.com/learn/regression/feature-elimination-using-rfe-in-logistic-regression-5400-32542-192781-593248-3029313/
- https://datascience.stackexchange.com/questions/70337/recursive-feature-elimination-rfe-with-logistic-regression-and-little-correlat


THANK YOU