Lead Score Case Study

Group Members:
Neha Joshi
Soumya Gupta

Problem Statement

- A company X Education sells online courses to industry professionals and markets its courses on several websites and search engines like Google.
- When people fill up a form providing their email address or phone number, they are classified to be a lead. Some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%
- The company wants to you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

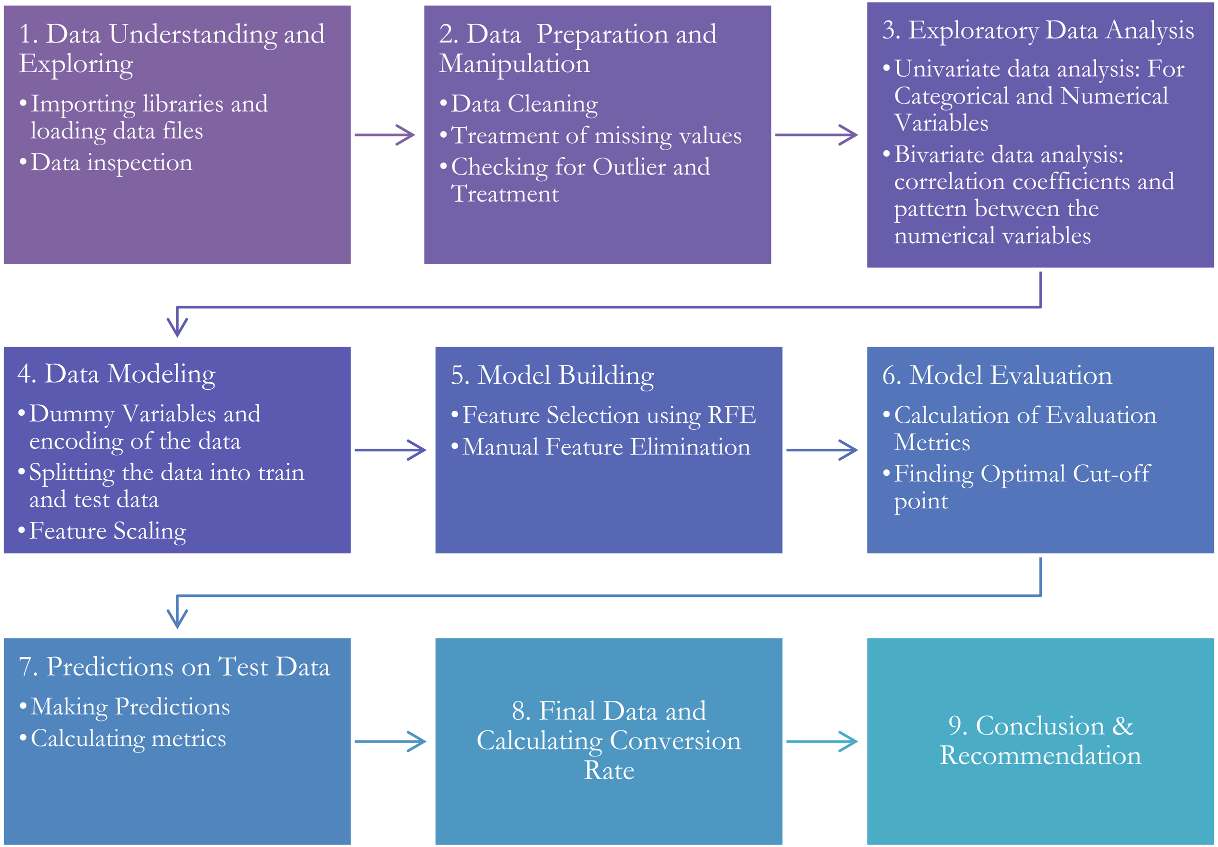
Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ ₹
€ $
Converted Leads

Objective

- To 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.
- The company requires a model that will assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot Leads
£ ₹
€ $
Converted Leads

Analysis Approach



1. Data Understanding and Exploring

- Importing libraries and loading data files
- Data inspection

2. Data Preparation and Manipulation

- Data Cleaning
- Treatment of missing values
- Checking for Outlier and Treatment

3. Exploratory Data Analysis

- Univariate data analysis: For Categorical and Numerical Variables
- Bivariate data analysis: correlation coefficients and pattern between the numerical variables

4. Data Modeling

- Dummy Variables and encoding of the data
- Splitting the data into train and test data
- Feature Scaling

5. Model Building

- Feature Selection using RFE
- Manual Feature Elimination

6. Model Evaluation

- Calculation of Evaluation Metrics
- Finding Optimal Cut-off point

7. Predictions on Test Data

- Making Predictions
- Calculating metrics

8. Final Data and Calculating Conversion Rate

9. Conclusion & Recommendation

Data Understanding and Manipulation Summary

Data contains total 9240 rows and 37 columns

Few categorical columns with ‘Select’ as a value was observed. As, customer may not have selected any option from the list.	Columns with missing values > 45% have been dropped	Few columns with Single value were present in data	Data Imbalance observed in some columns
We have replaced the Select option from categorical variables as null value.	Missing value treatment was done for columns having < 45% missing data	These features have been dropped as they don't serve any purpose for analysis	We have dropped these columns as they won't help with our analysis

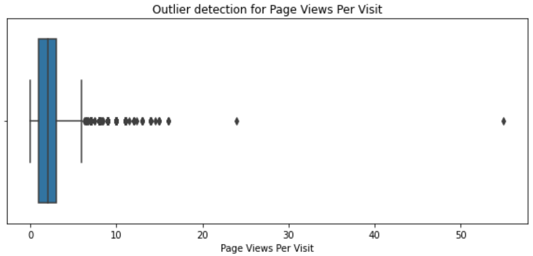
After Data Cleaning Process: Total rows 9240 and columns 13

Outlier Analysis

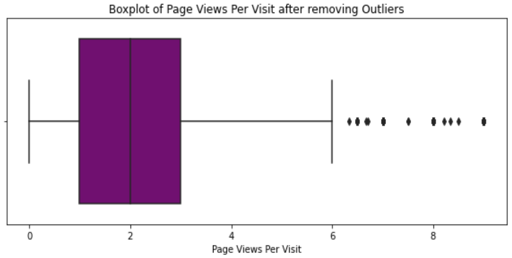
- ❖ Outlier check was done on numerical columns.
- ❖ The features 'TotalVisits', 'Page Views Per Visit' have outliers and they can be capped at 0.01 and 0.99th percentile.
- ❖ No Outliers present in "Total Time Spent on Website "data variable.

Column: Page Views Per Visit

Before Treatment

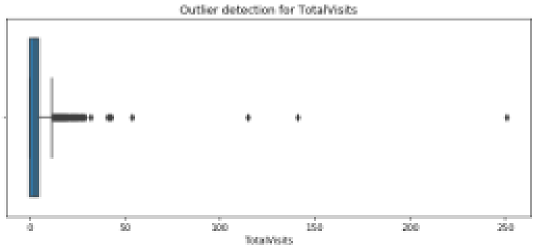


After Outlier Treatment

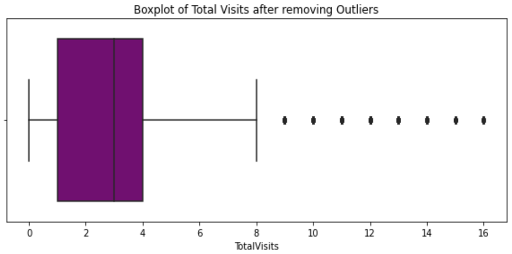


Column: TotalVisits

Before Treatment



After Outlier Treatment



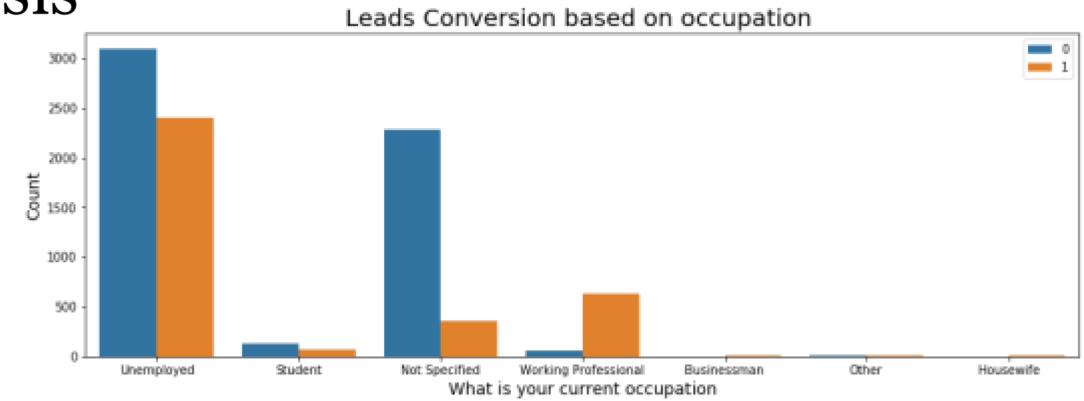
Exploratory Data Analysis

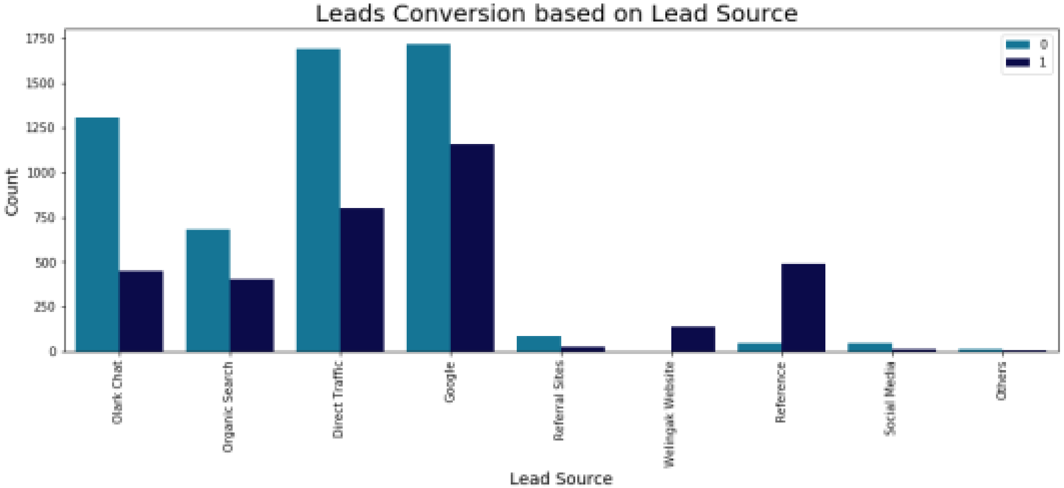
1. Leads Conversion based on occupation

- Working Professionals going for the course have high chances of joining it
- Unemployed leads are the most in numbers but has around 30-35% conversion rate
- Businessman and Housewife are very less in count.

2. Leads Conversion based on Lead Source

- Most of the leads generated are through Google and Direct traffic and the least through Live Chat
- Welingak website has the most conversion rate
- Lead conversion is maximum from Reference and welingak website.

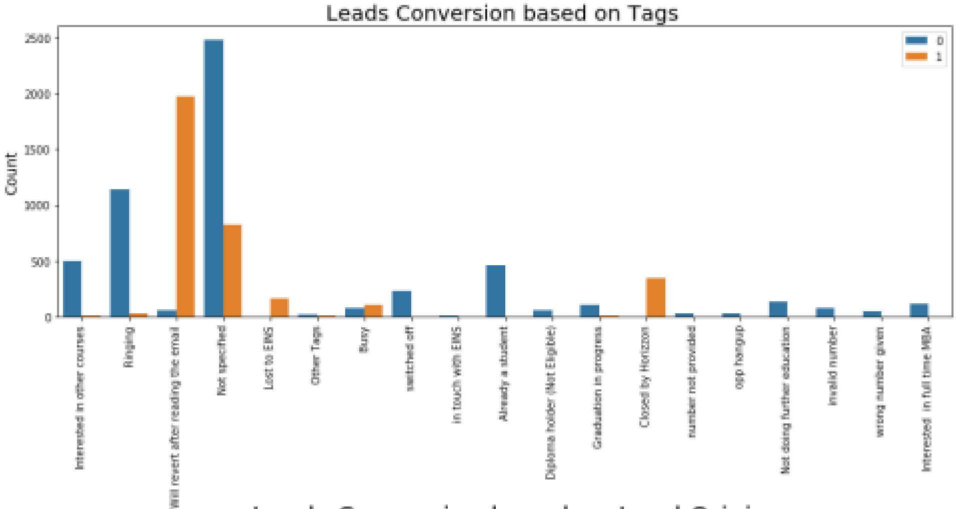




Exploratory Data Analysis

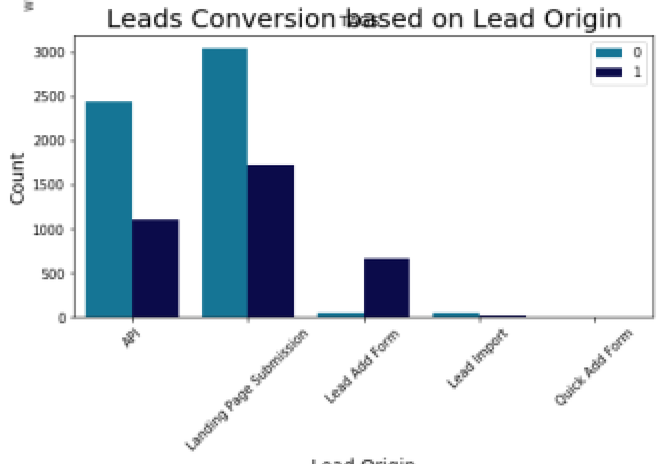
3. Leads Conversion based on Tags

- 'Will revert after reading the email' and 'Closed by horizon' have high conversion rate.
- 'Lost to ENS' also has good conversion rate but count of lead is very low.


Leads Conversion based on Tags

4. Leads Conversion based on Lead Origin

- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.


Leads Conversion based on Lead Origin

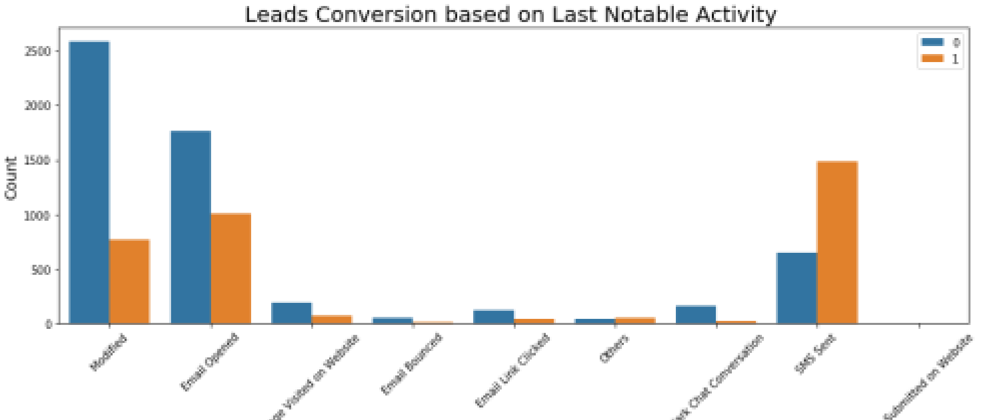
Exploratory Data Analysis

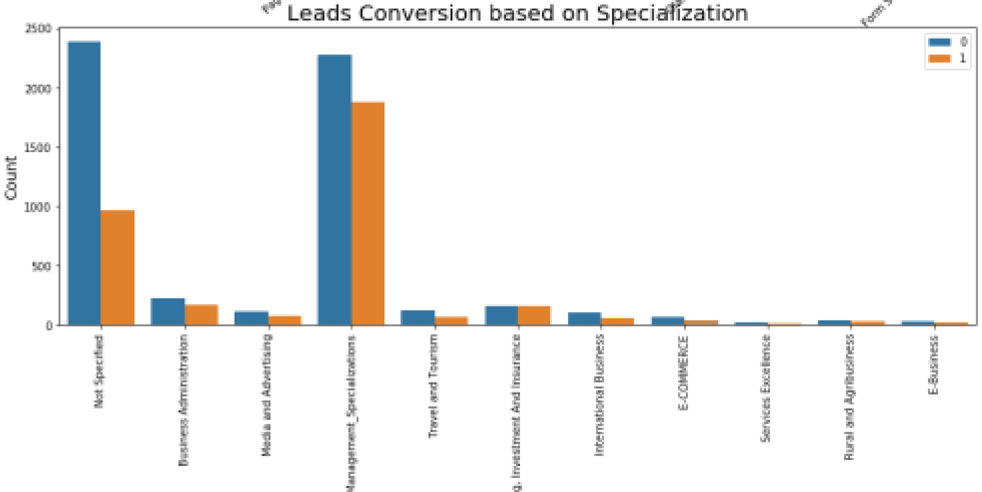
5. Leads Conversion based on Last Notable Activity

- 'Email Opened' has 30-35% conversion rate.
- 'SMS sent' has high number of conversion rate as compared to the count of leads who do not convert.
- 'Modified' has very high number of leads who do not convert.

6. Leads Conversion based on Specialization

- Focus should be more on the Specialization with high conversion rate such as leads with specialization in management field, banking, investment and insurance domain.

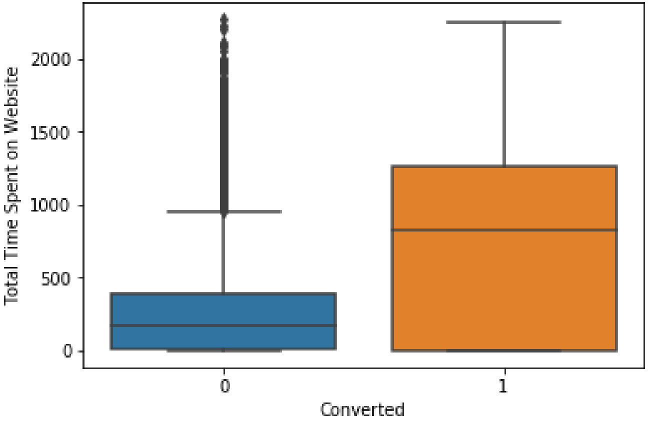

Leads Conversion based on Last Notable Activity


Leads Conversion based on Specialization

Exploratory Data Analysis

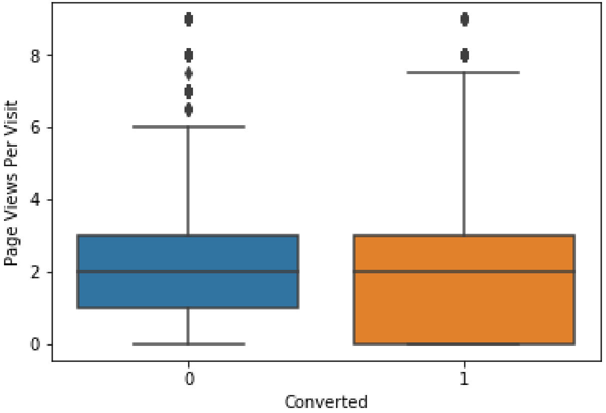
7. Leads Conversion based on Total Time Spent on Website

- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.



8. Leads Conversion based on Page Views Per Visit

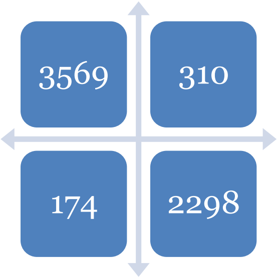
- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit



Data Modeling and Model Building

- Dummy variables were created for categorical variables.
- We have performed a train-test split, we have chosen 70:30 ratio.
- Feature scaling of numerical variables performed.
- Automated Feature selection was performed using RFE with 15 variables as output to build the model.
- Through Manual feature elimination 5 variables were removed based on high P-Value.

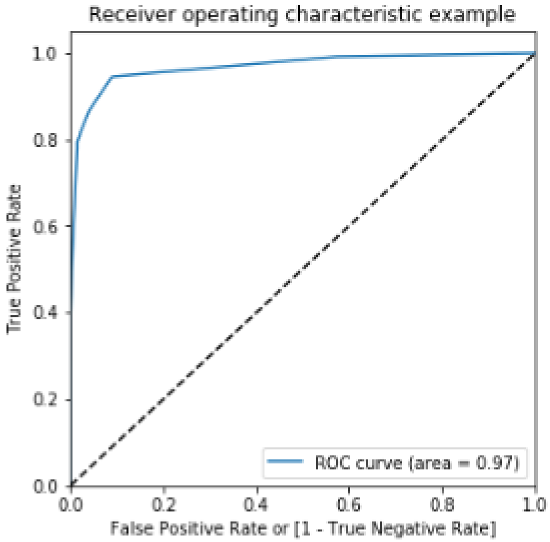
Confusion Matrix



3569	310
174	2298

Model Evaluation (Train Data)

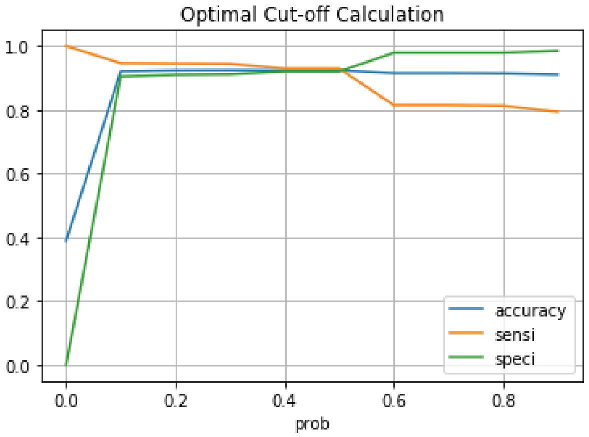
- As per predictions made on Train-data, we achieved accuracy of 92.37%.
- Other evaluation metrics:
- Sensitivity : 92.96%
- Specificity : 92.0%
- To visualize tradeoff between sensitivity and specificity, ROC curve was plot.
- The model seems to be performing well as the ROC curve has a value of 0.97, which is very good.



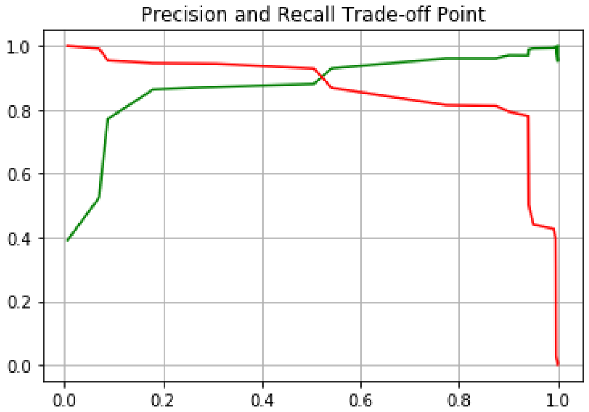
Model Evaluation

- **Optimal Cut off Point** of probabilities where we get balanced sensitivity and specificity.
- From the plot of Probability values vs. Accuracy, Sensitivity and Specificity, we can observe that from 0.4 to 0.5 all the three metrics have similar value around 0.92.
- Hence, we have selected 0.5 as optimal cut-off value for probability to get better conversion prediction.
- We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.5.
- Evaluation Metrics:

- **Accuracy : 92.37%**
- **Sensitivity : 92.96%**
- **Specificity : 92.0%**
- **False Positive Rate : 7.99%**
- **Positive Predictive Value : 88.11%**
- **Negative Predictive Value : 95.35%**
- **Precision : 88.11%**
- **Recall : 92.96%**



Optimal Cut-off Calculation

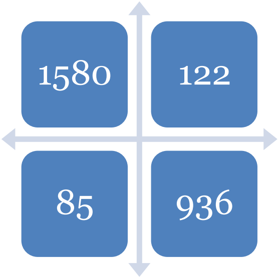


Precision and Recall Trade-off Point

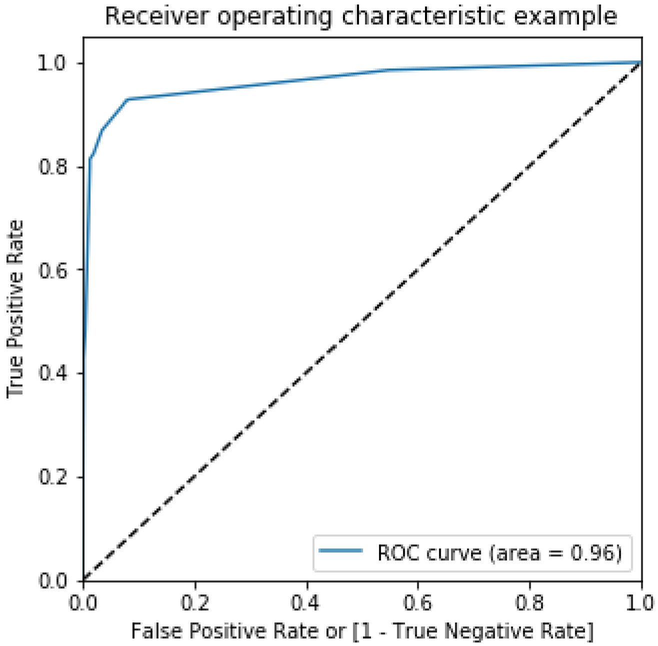
Predictions on Test Data

- ➢ Predictions were made on Test Data.
- ➢ Evaluation metrics on Test- Data
- **Accuracy : 92.39%**
- **Sensitivity : 91.67%**
- **Specificity : 92.83%**
- **Precision : 88.46%**
- **Recall : 91.67%**

Confusion Matrix



1580	122
85	936



Conclusion and Recommendation

- Evaluation Metrics like Sensitivity- Specificity as well as Precision and Recall were calculated on Train and Test Data.
- Optimal Cut-off value of 0.5 was used to make final predictions.
- Accuracy, Sensitivity and Specificity values of Test Data are around 92.39%, 91.67% and 92.83% which are approximately closer to the values of respective metrics calculated on Train Data and in the acceptable region
- As per the lead score calculated on final data shows conversion rate of 88%. Hence, we can say the business requirement of having 80% conversion rate was achieved.
- The top three dummy variables contributing for lead conversion are as below:
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tag_We will revert after reading the email
- Hence, Overall this model seems to be Good.
- To improve lead conversion, Company should target leads which are having specialization in Management fields
- The leads which were having last notable activity as 'SMS Sent', 'Modified' and 'E-mail Opened' should be targeted more as they showed higher conversion rate
- We would like to recommend the company to follow up with the leads frequently and maintain their current state by assigning proper tags and not to leave this field empty as we can see the model gave importance to Tags_Not specified variable
- Lead conversion can be improved by maximizing leads from Reference and welingak website
- Feedbacks from leads can be taken to understand their requirements and expectations.