UPGRAD & IIIT-B

Lead Scoring Case Study

Prepared by:
Prateek Rana & Uday Suri

Table of Contents

Presentation Agenda

01 Problem Statement

02 Analysis Approach

03 EDA

04 Analysis Results

05 Conclusion

PROBLEM STATEMENT

Achieving a target lead conversion rate to be around 80%.

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. 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. The company has a conversion rate of 30% and wishes to make this process efficient by identifying the more potentially convertible leads or 'Hot Leads'.

Problem Statement:

To help them select the most promising leads. The company requires us to build a model which helps identify the leads' conversion rate.

→

ANALYSIS APPROACH

1.Data Understanding and Cleaning

Checking & treating the data-set for null values and inappropriate data types.

2.Exploratory Data Analysis (EDA)

Analyzing the data set to summarize and visualize the columns

3.Outlier Analysis

Checking for Outliers/noise in the columns and treating them accordingly.

4.Scaling

Scaling the numerical columns to avoid weightage bias.

5.Training & Testing Data

Randomly dividing the data-set between train and test in 70:30 ratio.

6.Building the logistic regression model

Using RFE and manual selection to build a logistic regression model that best predicts the dependent variable i.e. Converted column.

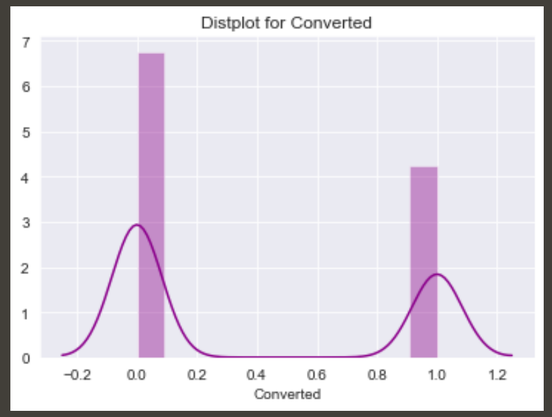
7.Conclusion

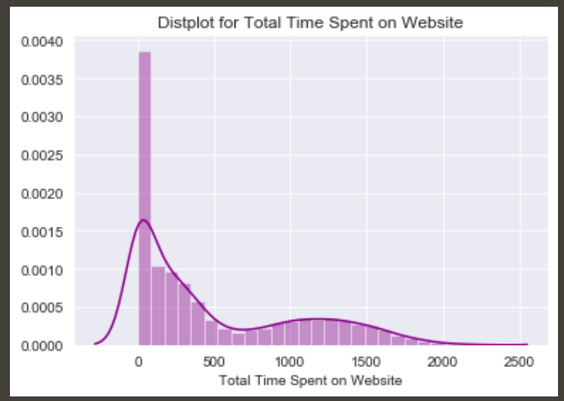
Optimizing the model and summarizing the results.

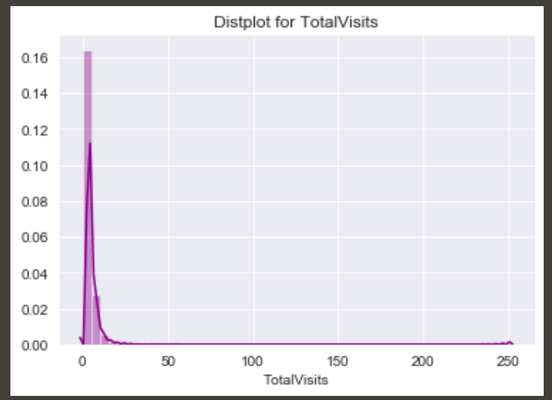
EXPLORATORY DATA ANALYSIS

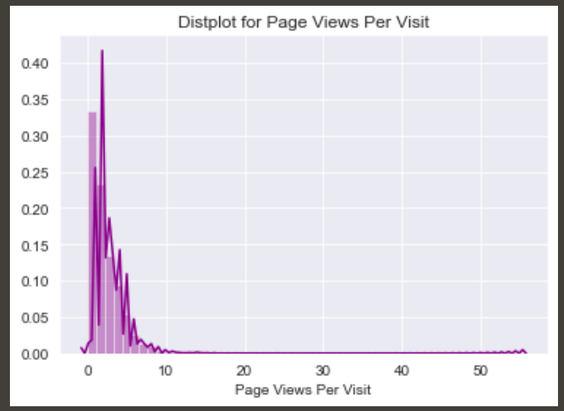
Insights:

- *We already know that converted is the target variable and as we can see it only has two values which tell us that the lead is either converted(1) or not(0).*
- *All the other 3 displots for Total Time Spent, Total Visits & Page Views Per Visit are right-skewed with outliers only in Total Visits.*







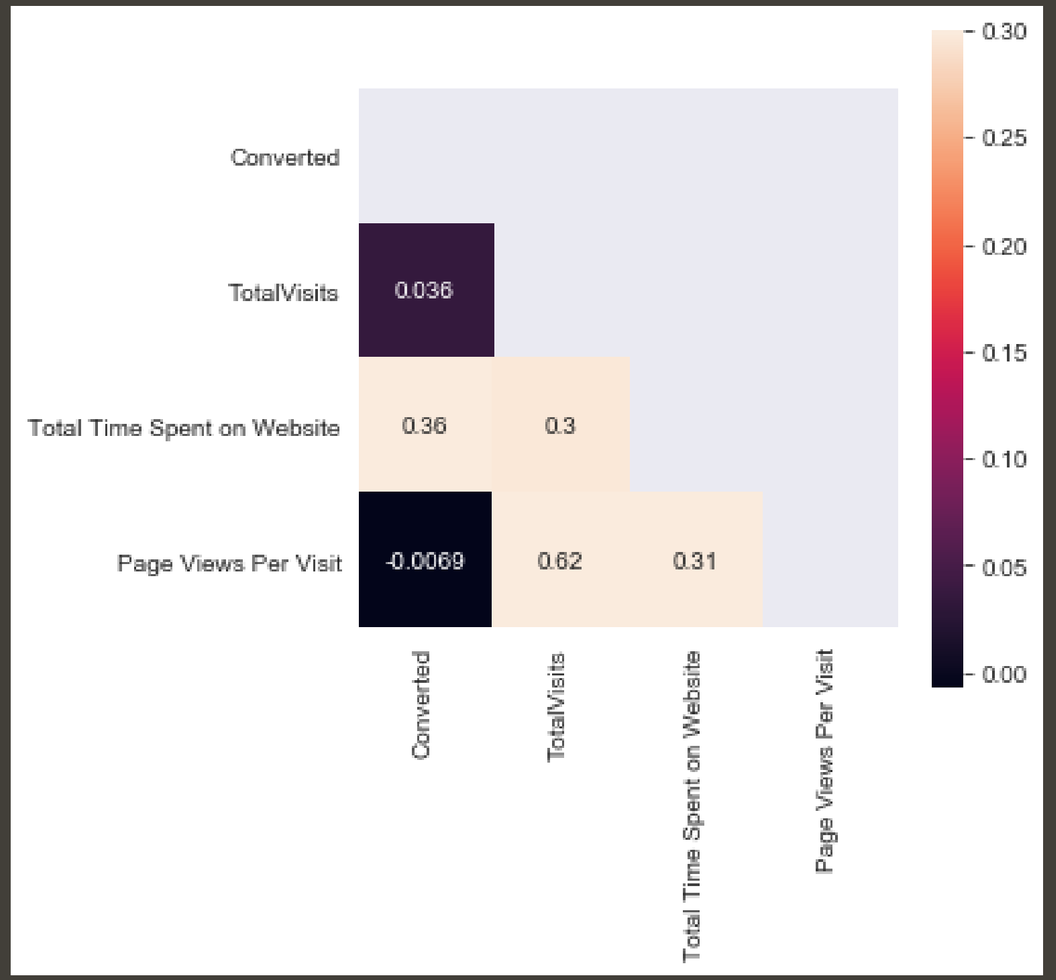


EXPLORATORY DATA ANALYSIS

Correlation Analysis

- A person who visits the website more often is also likely to view more pages per visit.

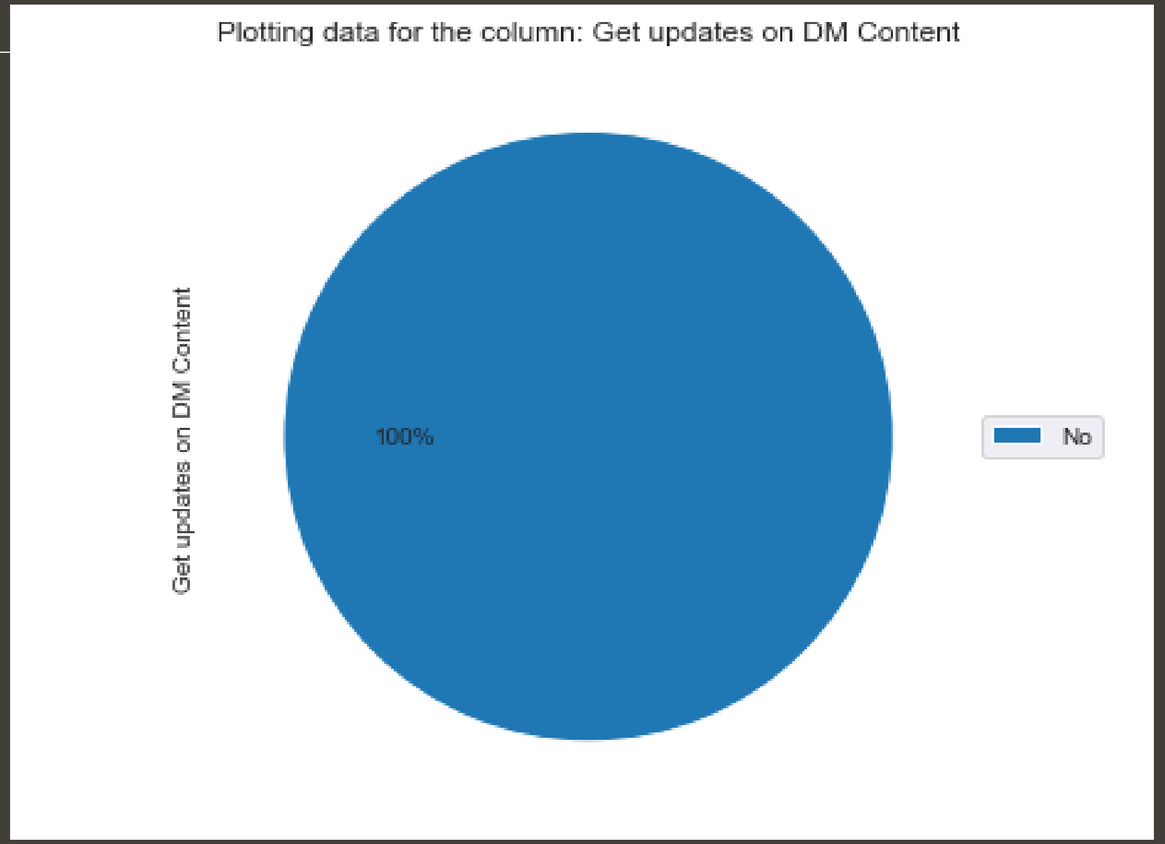
Other numeric columns have a relatively low R coefficient so no such relationship can be drawn for them.



EXPLORATORY DATA ANALYSIS

The following columns have 0 or negligible yes values so it is safe to drop them:

- Do Not Call'
- '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'



Analyzing the Model

BEST-FIT MODEL

We followed the following steps to get the best-fit model (represented by the figure):

- *Dropped the columns mentioned in the previous slide.*
- *Dividing the data-set between test & train.*
- *Fitting and transforming the train data-set.*
- *Selecting Variables using RFE and Manual Selection*

→

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6464
Model:	GLM	**Df Residuals:**	6453
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2896.1
Date:	Mon, 08 Mar 2021	**Deviance:**	5792.2
Time:	15:57:51	**Pearson chi2:**	6.76e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

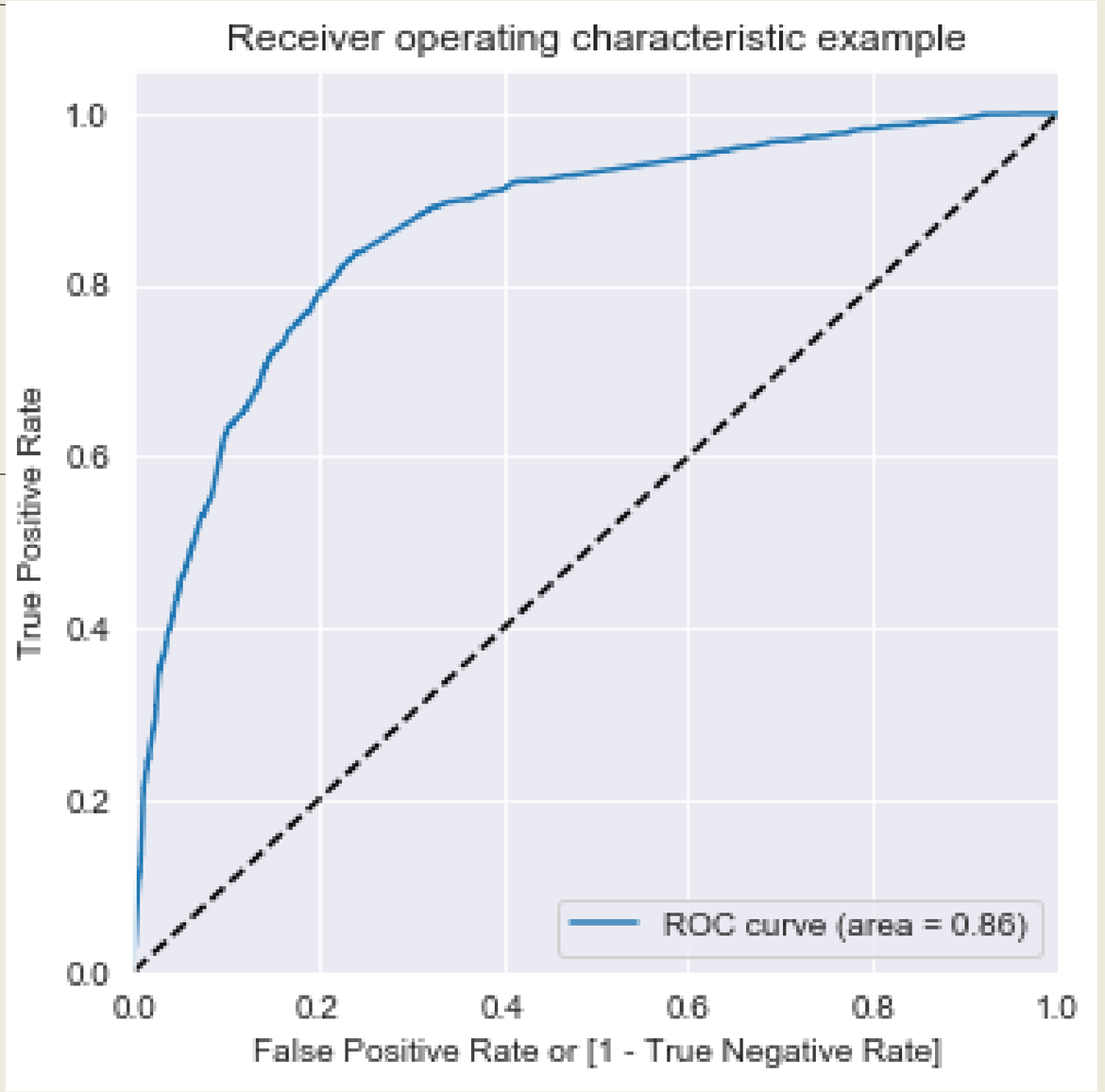
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0316	0.055	-18.621	0.000	-1.140	-0.923
Do Not Email	-1.7652	0.182	-9.686	0.000	-2.122	-1.408
Total Time Spent on Website	1.1207	0.039	29.053	0.000	1.045	1.196
Lead Add Form	4.4326	0.188	23.548	0.000	4.064	4.802
Olark Chat	1.0094	0.095	10.624	0.000	0.823	1.196
Had a Phone Conversation	3.3217	1.091	3.045	0.002	1.183	5.460
Modified	-0.7275	0.080	-9.119	0.000	-0.884	-0.571
Olark Chat Conversation	-1.4666	0.317	-4.622	0.000	-2.088	-0.845
SMS Sent	1.3671	0.082	16.735	0.000	1.207	1.527
Unreachable	1.7662	0.522	3.381	0.001	0.742	2.790
Unsubscribed	1.5538	0.498	3.122	0.002	0.578	2.529

Analyzing the Model

MODEL EVALUATION: ROC CURVE

The ROC curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

- *The ROC curve for this model follows the left-hand border and then the top border of the ROC space, so it is safe to assume that the model is adequately accurate.*

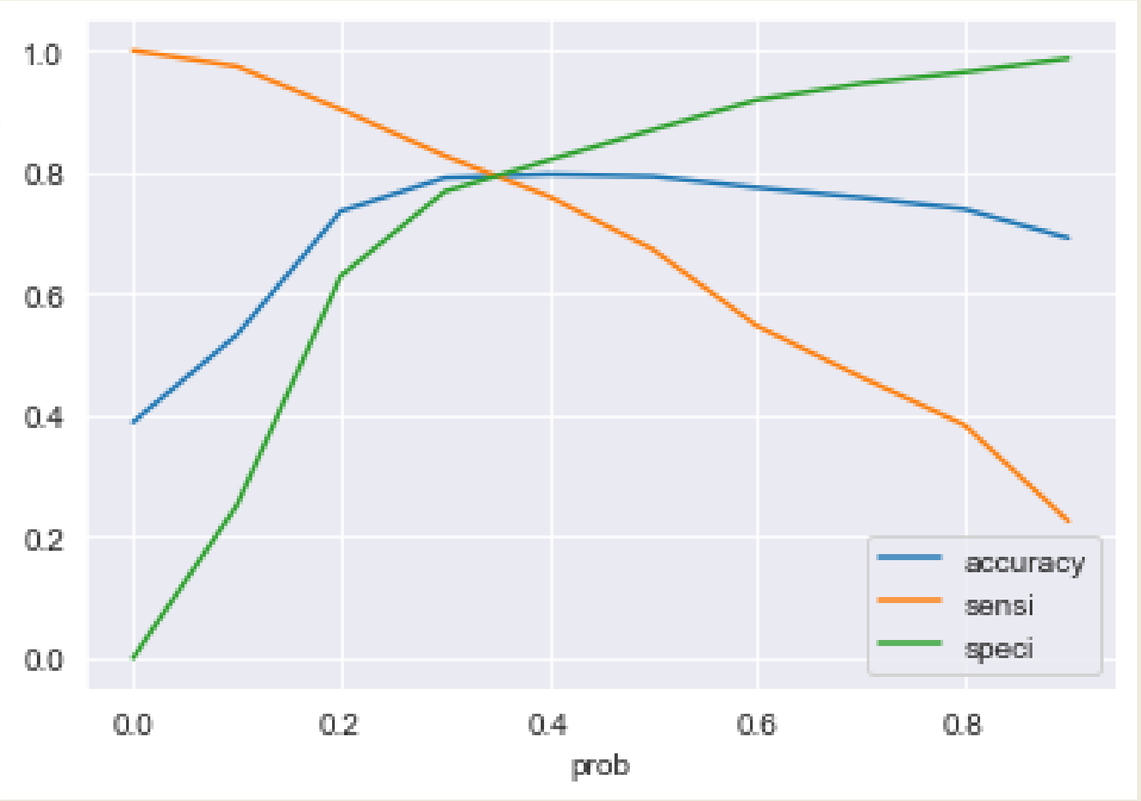


Analyzing the Model

OPTIMIZING

Accuracy, Sensitivity and Specificity

In order to select the best possible cut-off value, we visualize the accuracy, sensitivity and specificity of the model at every cut-off value.
For this model, we require fairly equal values of accuracy, sensitivity, and specificity so we chose the cut-off point at their intersection i.e. at 0.35.



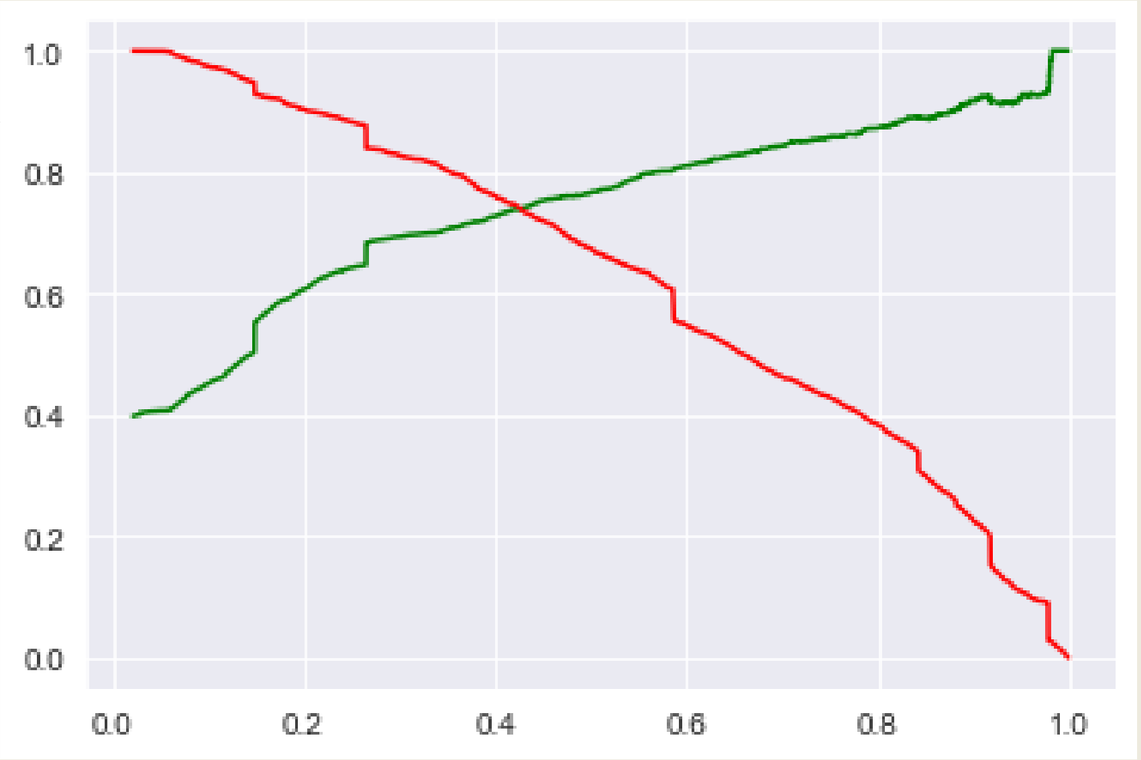
→

Analyzing the Model

OPTIMIZING

Precision and Recall Tradeoff

To further optimize the model, we check the precision-recall tradeoff by visualizing them at different probability/cut-off values.
For this model, the Precision-Recall trade-off is at 0.42, so we go ahead and predict the test data-set using this cut-off value.



→

Analyzing the Model

METRICS FOR THE TEST DATA-SET

ACCURACY, SENSITIVITY, AND SPECIFICITY

ACCURACY	SENSITIVITY	SPECIFICITY
0.8029	0.7416	0.8403

All the metrics are as expected, which means that the model predicts the dependent variable fairly well.

→

CONCLUSION

FINAL RECOMMENDATIONS

THE COMPANY SHOULD FOCUS ON THE FOLLOWING VARIABLES TO IMPROVE THE CONVERSION RATE:

- GET MORE LEADS FROM AD FORMS
- HAVE A TELEPHONIC CONVERSATION WITH THE LEAD INSTEAD OF OLARK CHAT CONVERSATION
- ASK THE CONCERNED EMPLOYEE TO SEND EMAIL REMINDERS.
- IMPROVE THE WEBSITE TO INCREASE THE DWELL TIME.

THANK YOU!

PREPARED BY

PRATEEK RANA & UDAY SURI

FOR

UPGRAD AND IIIT BANGALORE

COURSE:

PGDDS