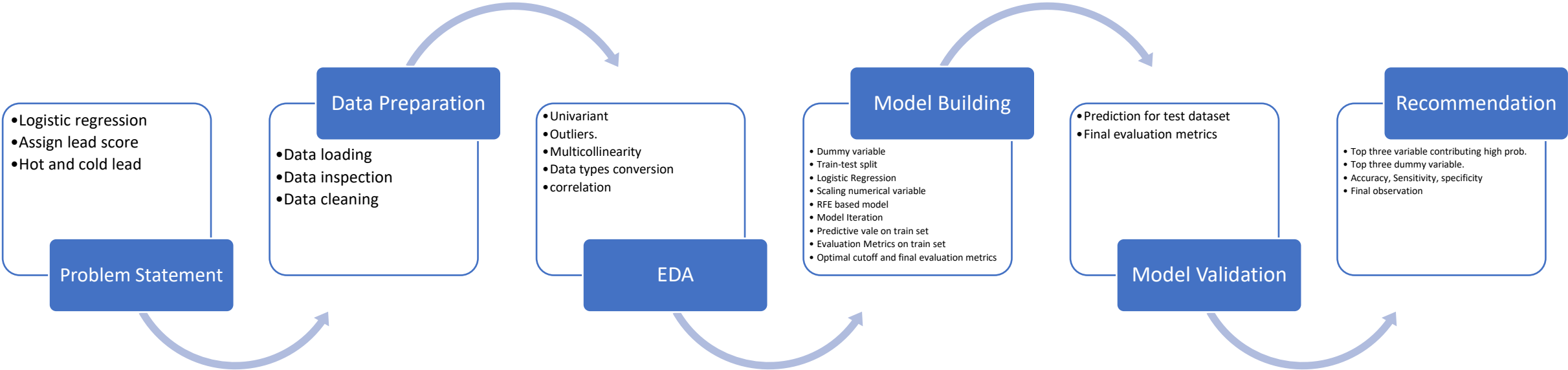# Lead score – Case study

Presented by:-

- Mohammad Shahid Rashid (DCS 27)

- Dharmendra Budha (DCS 27)

Lead score – Case study Life cycle

Problem Statement
- Logistic regression
- Assign lead score
- Hot and cold lead

Data Preparation
- Data loading
- Data inspection
- Data cleaning

EDA
- Univariant
- Outliers.
- Multicollinearity
- Data types conversion
- correlation

Model Building
- Dummy variable
- Train-test split
- Logistic Regression
- Scaling numerical variable
- RFE based model
- Model Iteration
- Predictive vale on train set
- Evaluation Metrics on train set
- Optimal cutoff and final evaluation metrics

Model Validation
- Prediction for test dataset
- Final evaluation metrics

Recommendation
- Top three variable contributing high prob.
- Top three dummy variable.
- Accuracy, Sensitivity, specificity
- Final observation

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.
- Lead :- User fill up online form , their email address or phone number, they are classified to be a lead
- The typical lead conversion rate at X education is around 30%.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Goals 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.

Data Preparation

Data loading	→	Data inspection	→	Data cleaning

```
#importing dataset to csv

leads=pd.read_csv("Leads.csv")
leads.head()
```
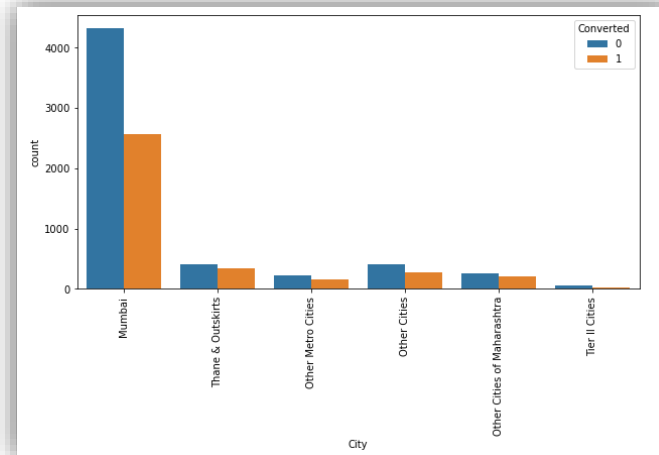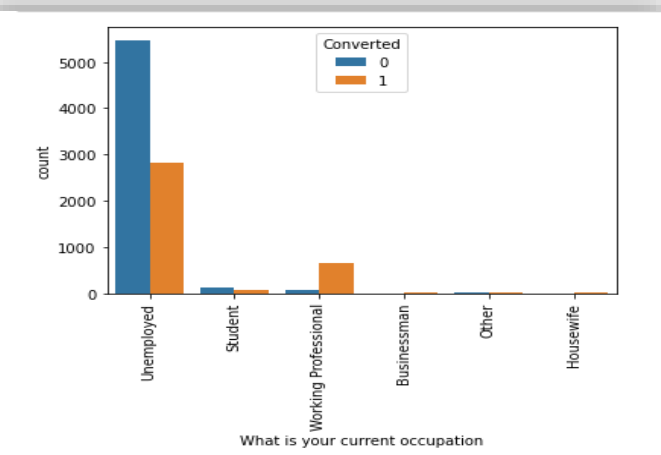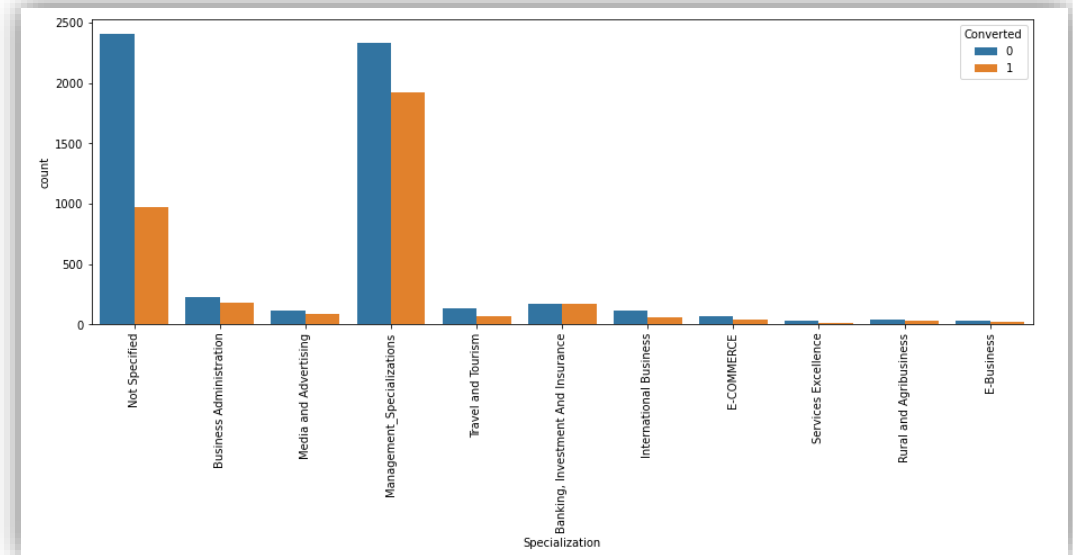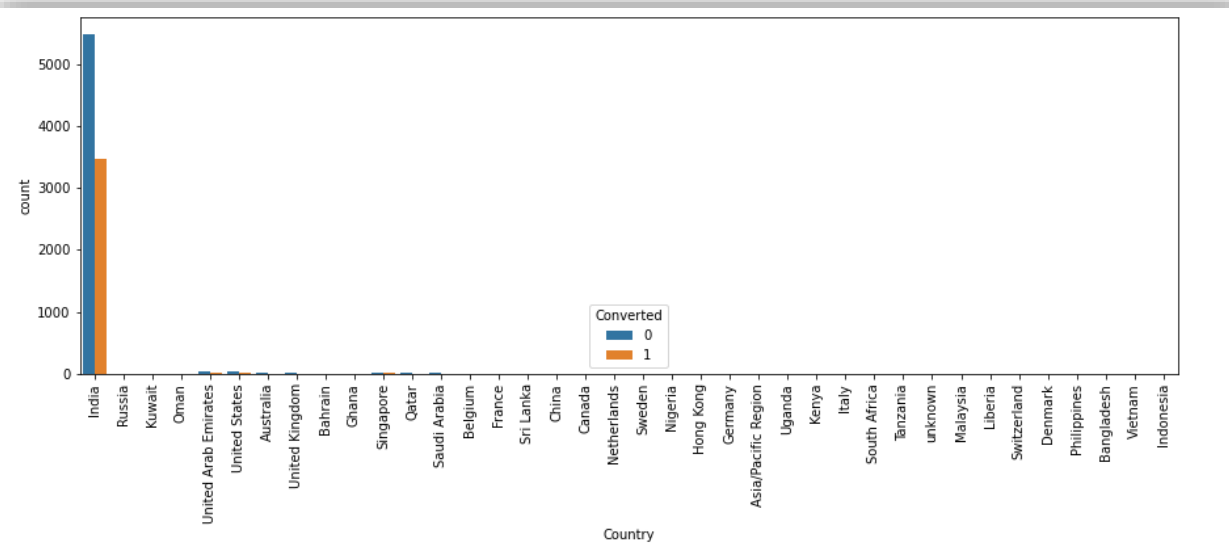
- leads.head()
- leads.shape
- leads.info()
- leads.describe()
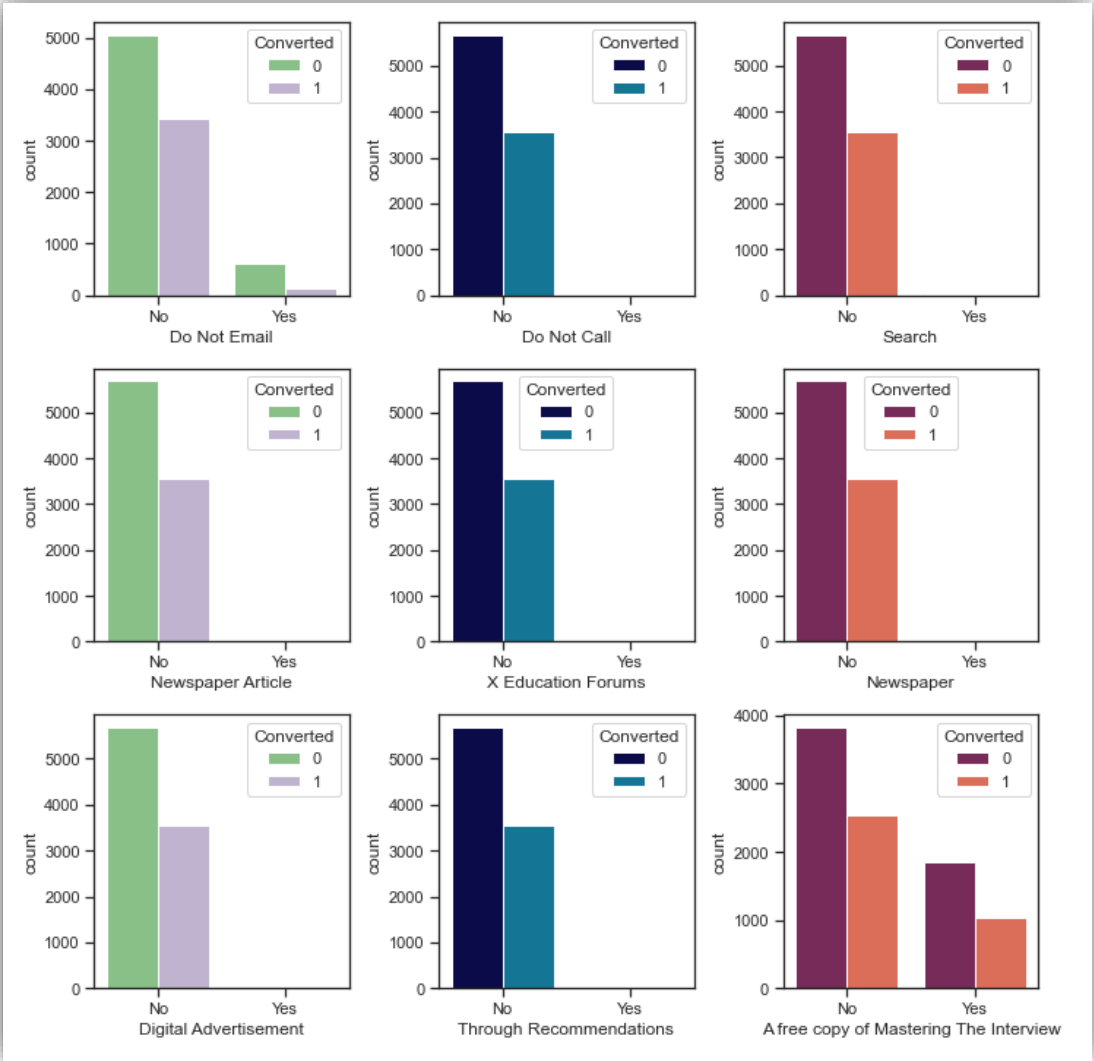- leads.isnull().sum()/len(leads)*100
- leads.duplicated().sum()

- Converting 'Select' values to NaN.
- dropping variables > 45% missing values
- impute all missing values with India
- Dropping Country variable
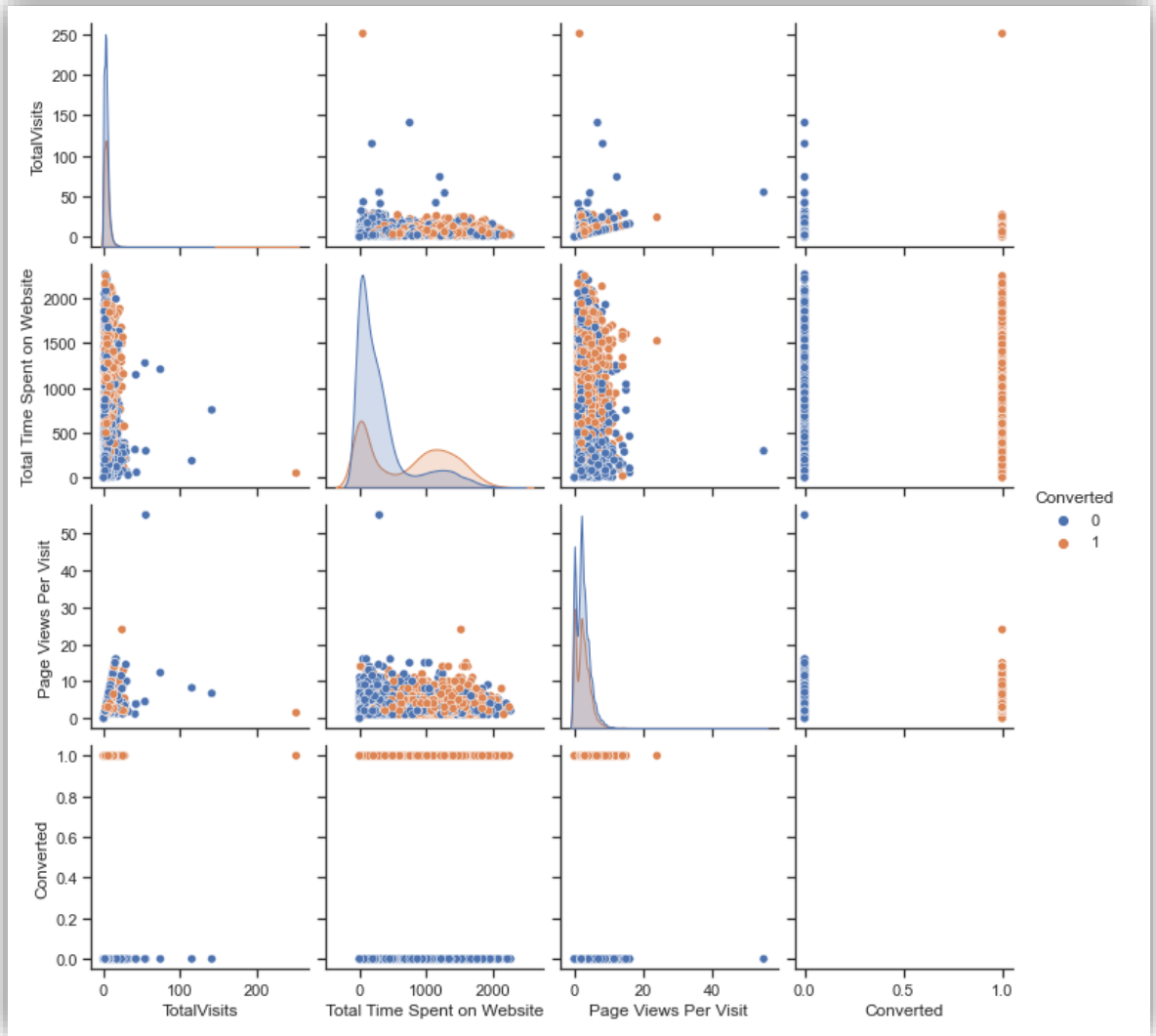- Replacing NaN values to Mumbai in city variable
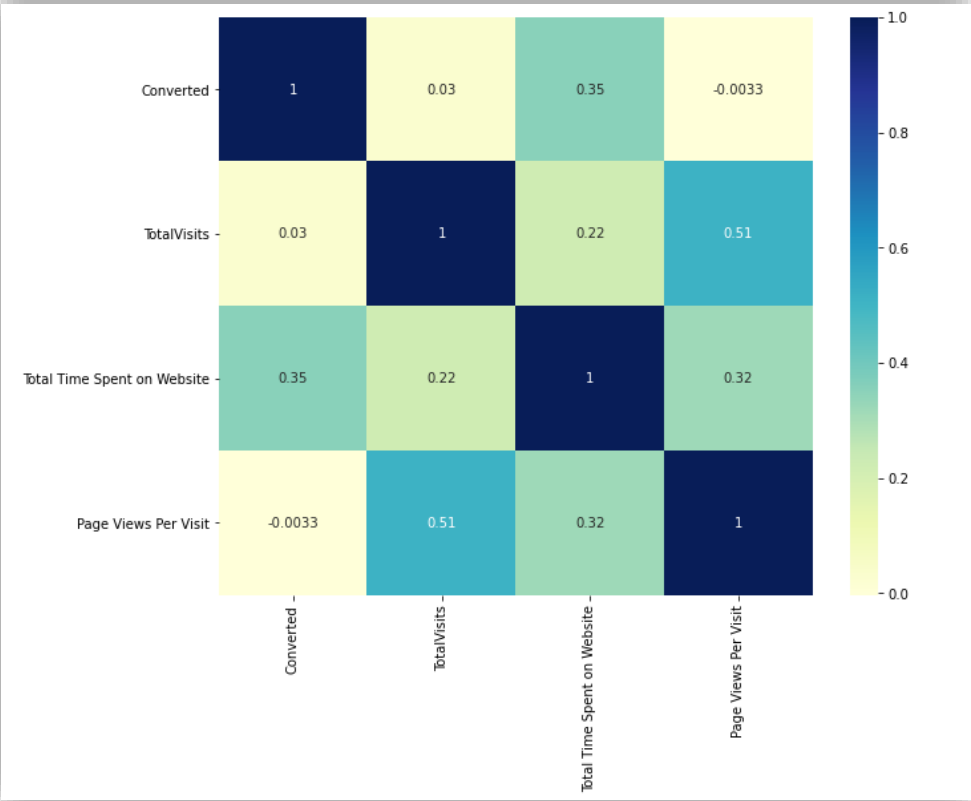
EDA

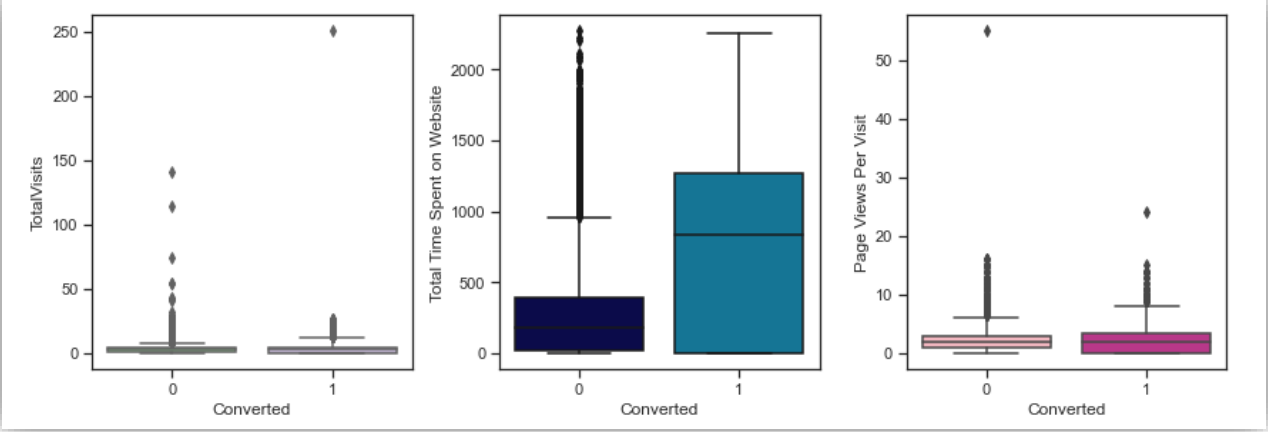## Categorical analysis

Boolean Variable

Numerical Variable

Continue ..

Numerical Analysis

Outliers

Before

After

Model Building

| Create Dummy Variable | Train and test split | Scaling data | Stats Model & RFE | RFE variables |

Dummy variable for datatype "object"

- Lead Origin
- Lead Source
- Last Activity
- Specialization
- What is your current occupation
- Tags
- City
- Last Notable Activity

Train – Test data selection
70-30

Scaling numerical variable.
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit

RFE with **15** variables

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Last Notable Activity_Modified
- Last Notable Activity_Olark Chat Conversation
- Last Notable Activity_SMS Sent
- Tags_Closed by Horizzon
- Tags_Interested in other courses
- Tags_Lost to EINS
- Tags_Other_Tags
- Tags_Ringing
- Tags_Will revert after reading the email

```
import statsmodels.api as sm

from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()

from sklearn.feature_selection import RFE
rfe = RFE(logreg, 15)              # running RFE with 15 variables as output
rfe = rfe.fit(X_train, y_train)
```

Confusion metrics

- Dropped **Lead Source_Referral Sites** due to high p-value
- Dropped **Last Notable Activity_SMS** Sent due to high VIF

```
#BUILDING MODEL #1

X_train_sm = sm.add_constant(X_train[col])
logm1 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm1.fit()
res.summary()
```

```
from sklearn import metrics

# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.Predicted )
print(confusion)
```

Model Validation

ROC curve	→	Prediction (Train Data)	→	Prediction (Test Data)	→	Confusion metrics	→	Observation
• ROC curve 0.97 • Optimal cut off (0.3)		• Train predict final		• Y (converted) predict final		• TP • TN • FP • FN		• Accuracy • Sensitivity • Specificity

The ROC Curve ~1 (0.97) indicating a good predictive model.

Optimal cut off **0.3**

Train Data:
• Accuracy : 92.29%
• Sensitivity : 91.70%
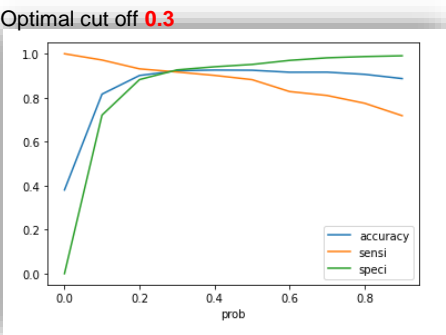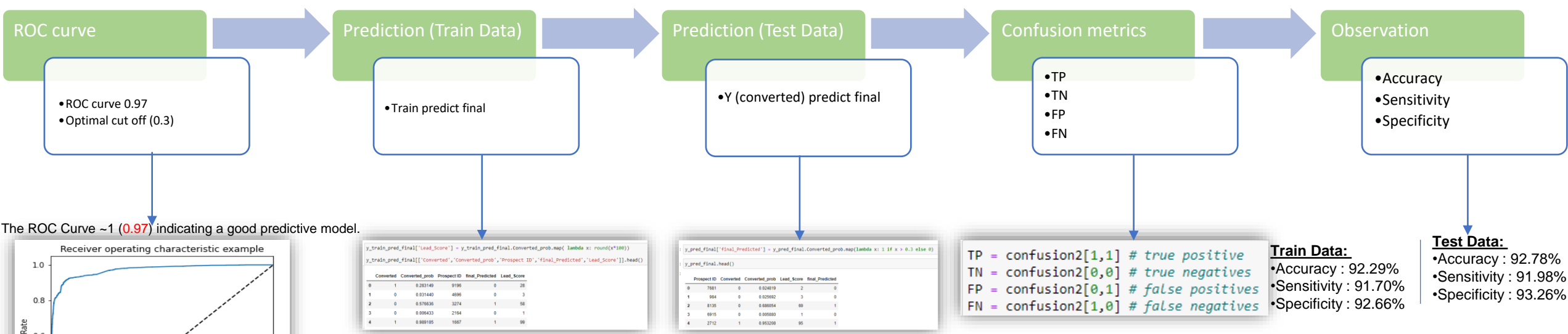• Specificity : 92.66%

Test Data:
• Accuracy : 92.78%
• Sensitivity : 91.98%
• Specificity : 93.26%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model.

Summary

1. Top three variables which could contribute high probability towards converting leads were;

- Tags_Closed by Horizon
- Tags_Lost to EINS
- Tags_Will revert after reading the email

2. Top three dummy variables which increase the probability of lead conversion were;

Tags
- ✓ Tags_Closed by Horizon
- ✓ Tags_Lost to EINS
- ✓ Tags_Will revert after reading the email

Lead Source
- ✓ Lead Source_Welingak Website

Last Activity
- ✓ Last Activity_SMS Sent

3. 0.3 is the optimum point to take it as a cut-off probability

4. The ROC Curve ~1 (0.97) indicating a good predictive model

5. Final Observation

<u>**Train Data:**</u>
Accuracy: 92.29%
Sensitivity: 91.70%
Specificity: 92.66%

<u>**Test Data:**</u>
Accuracy: 92.78%
Sensitivity: 91.98%
Specificity: 93.26%

Recommendation

✓ Working Professionals going for the course have high chances of joining it.

✓ Maximum number of leads are generated by Google and Direct traffic.

✓ Conversion Rate of reference leads and leads through welingak website is high.

✓ To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

✓ API and Landing Page Submission bring higher number of leads as well as conversion.

✓ In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

✓ 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.

End of the case Study, Thank you ☺

Name:- **Mohammad Shahid Rashid**
Mohammad.shahid.rashid@gmail.com
Batch :- DCS27

Name:- **Dharmendra Budha**
Dharmendra.budha@gmail.com
Batch :- DCS27