LEAD SCORING CASE STUDY

PRESENTED BY:

URVASHI

AGENDA

- The Purpose is to optimize the lead scoring mechanism based on their fit, demographics, behaviors, buying tendency etc. By implementing explicit & Implicit lead scoring modeling with the lead point system.

Goals of the Case Study

- There are quite a few goals for this 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. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Approach

- Source the data For analysis
- Reading & Understanding the data.
- Data Cleaning
- EDA
- Feature scaling
- Splitting the data into test & train dataset
- Prepare the data for modeling
- Model building
- Model evaluation-specificity & sensitivity or precision-recall
- Making predictions on the test set

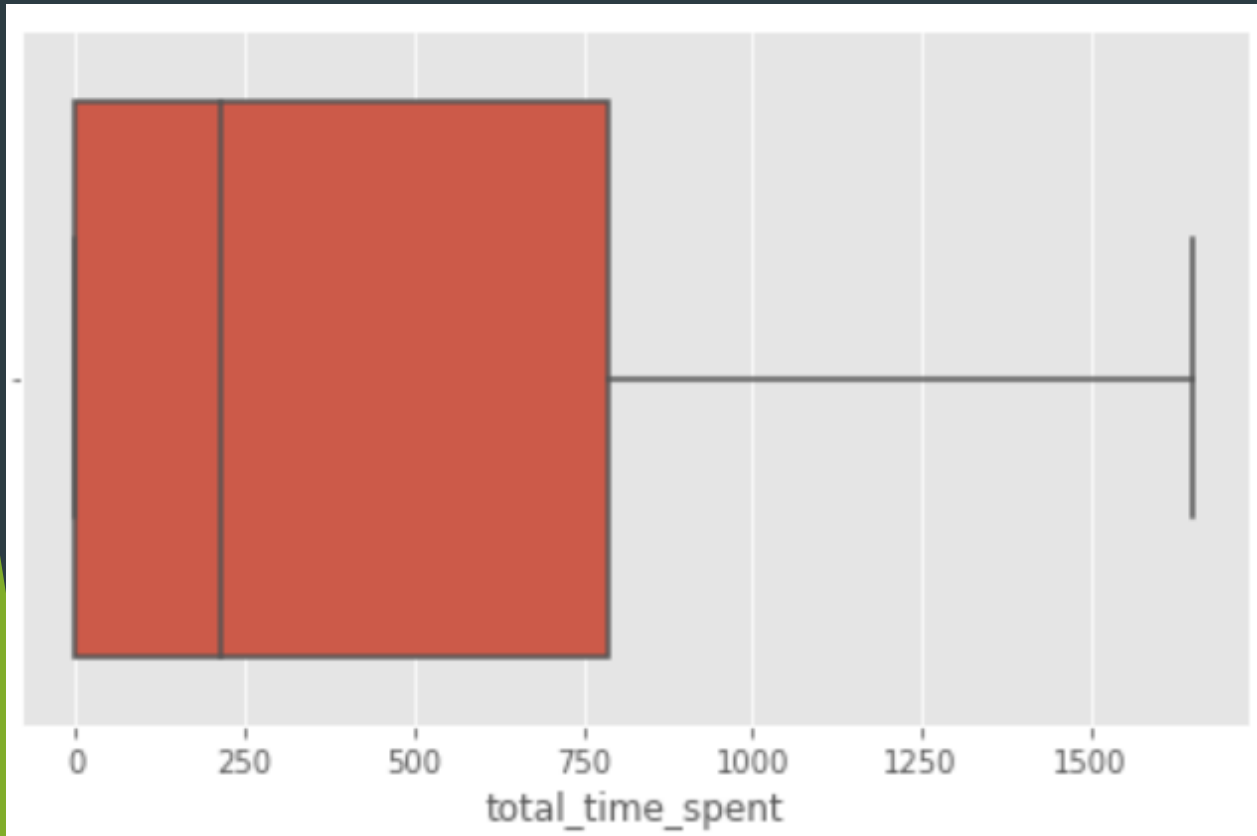
Data Sourcing, Cleaning and Preparation

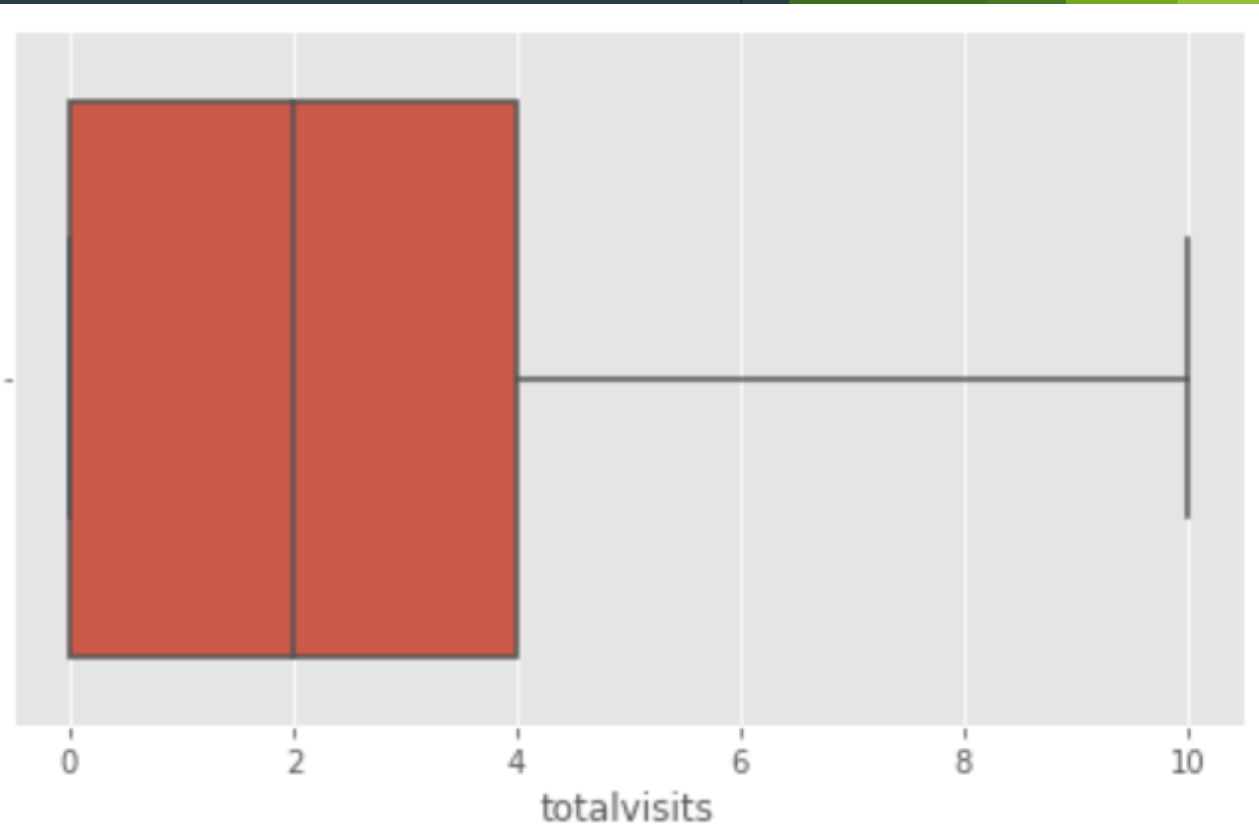
- Read the data from CSV File
- Outlier treatment
- Data cleaning -Handling Null Values & removing higher Null values data
- Removing Redundant columns in the data
- Imputing Null Values
- Exploratory data analysis-approx.
- Feature standardization

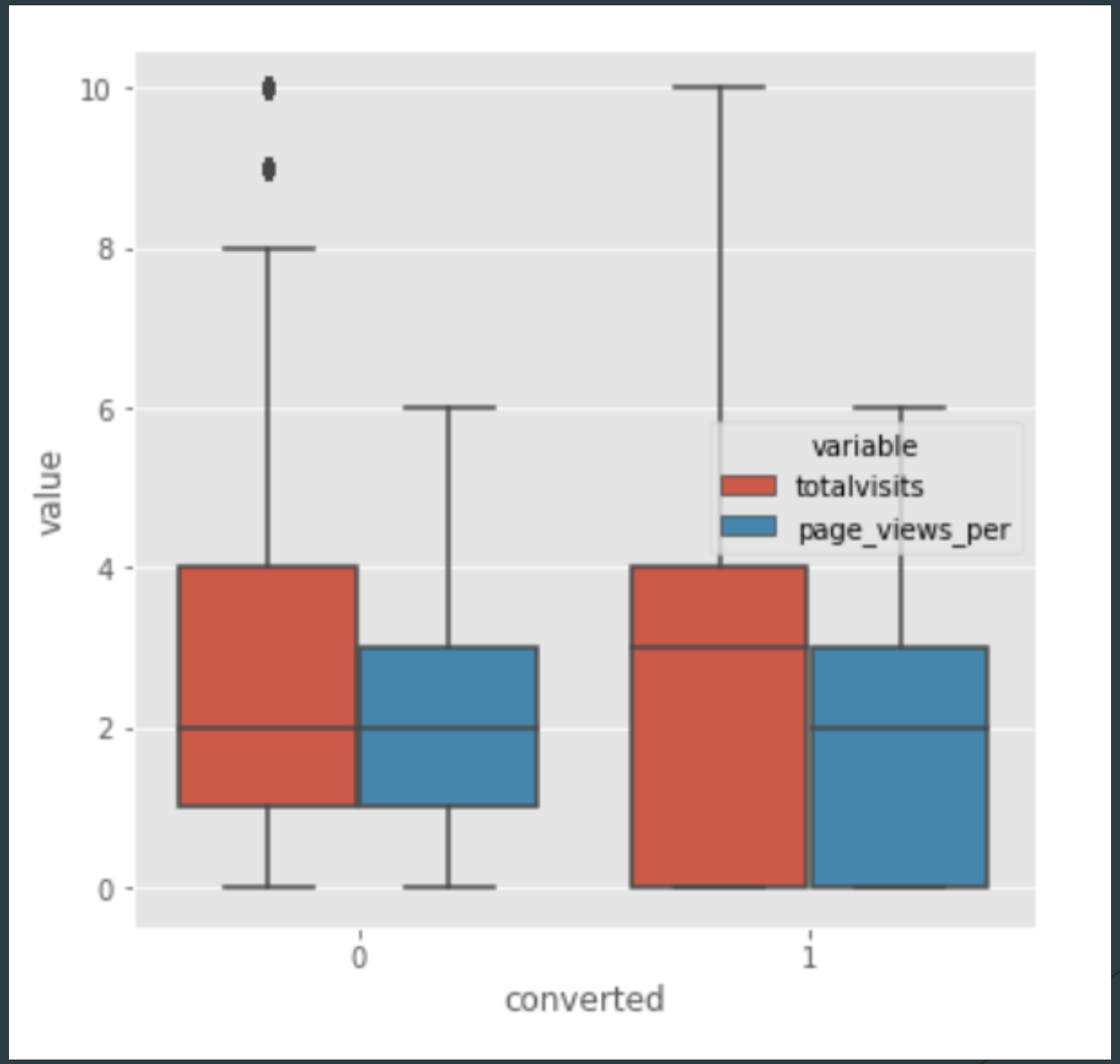
Outliers

- Total Visits, Total Time Spent on the Website, Page Views Per Visit have outliers

```
#Boxplots before outlier removal
num_df = clean_df[['converted', 'totalvisits',  'page_views_per', 'total_time_spent']]

def plot_boxes():
    plt.figure(figsize=(12, 4))
    plt.subplot(121)
    sns.boxplot(data=clean_df, x='total_time_spent')

    plt.subplot(122)
    sns.boxplot(data=clean_df, x='totalvisits')

    plt.tight_layout()
    plt.show()

    plt.figure(figsize=(6, 6))
    box_long = pd.melt(num_df.drop('total_time_spent', axis=1), id_vars='converted')
    sns.boxplot(x='converted', y='value', hue='variable', data=box_long)
    plt.show()

plot_boxes()
```

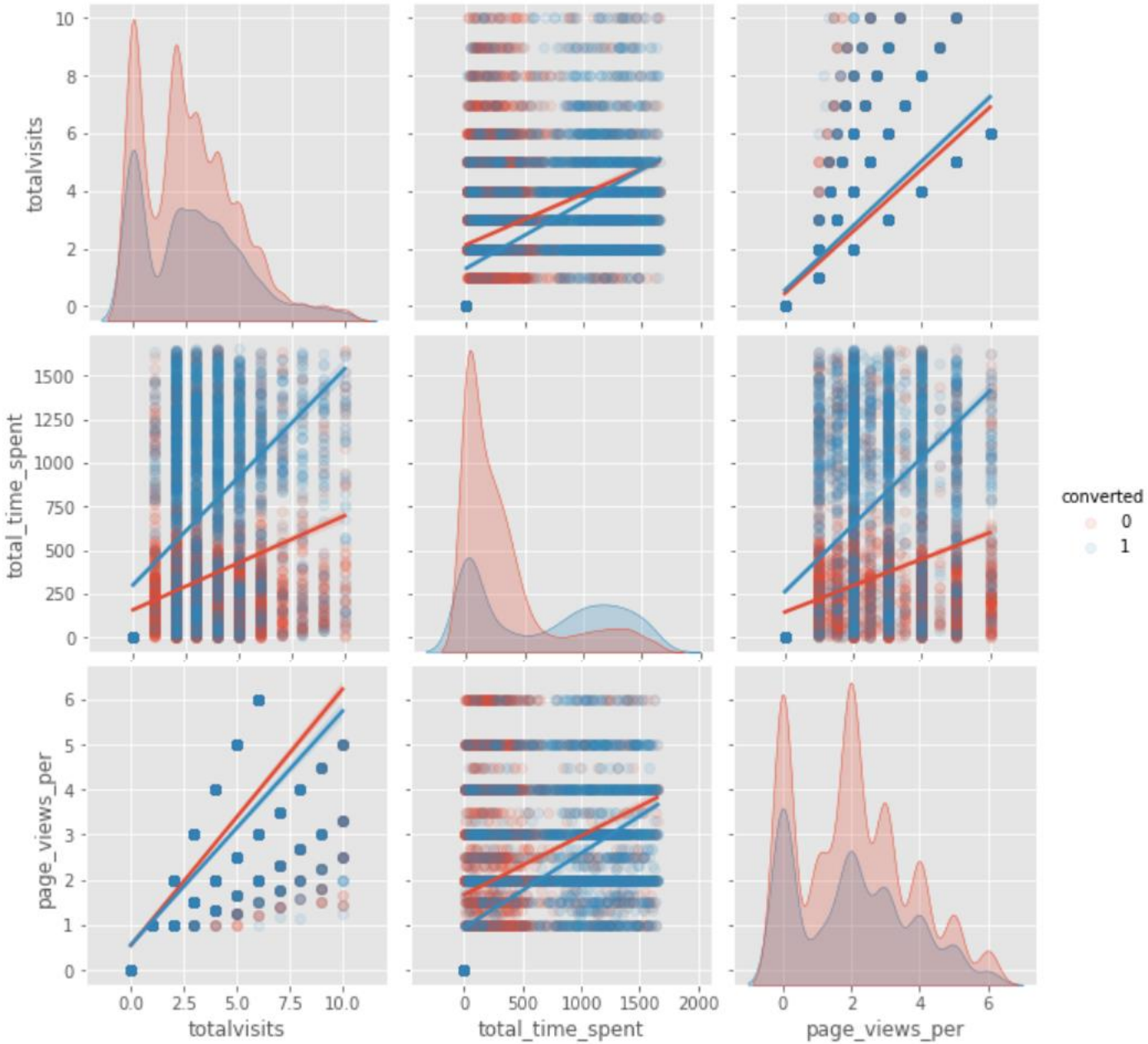

0
250
500
750
1000
1250
1500
total_time_spent


0
2
4
6
8
10
totalvisits


10
8
6
4
2
0
value
variable
totalvisits
page_views_per
0
1
converted

Data Analysis

```
sns.pairplot(data=clean_df, vars=clean_df.columns[1:4], hue='converted', kind='reg', height=3,
             plot_kws={'scatter_kws': {'alpha': 0.1}})
plt.show()
```



Data Preparation

- Converted Binary variables into 0 & 1
- Created dummy variables for categorical variables

Feature Scaling & Splitting Train & Test Sets

- Feature Scaling of Numeric Data
- Splitting data into Train & Test Set

Model building

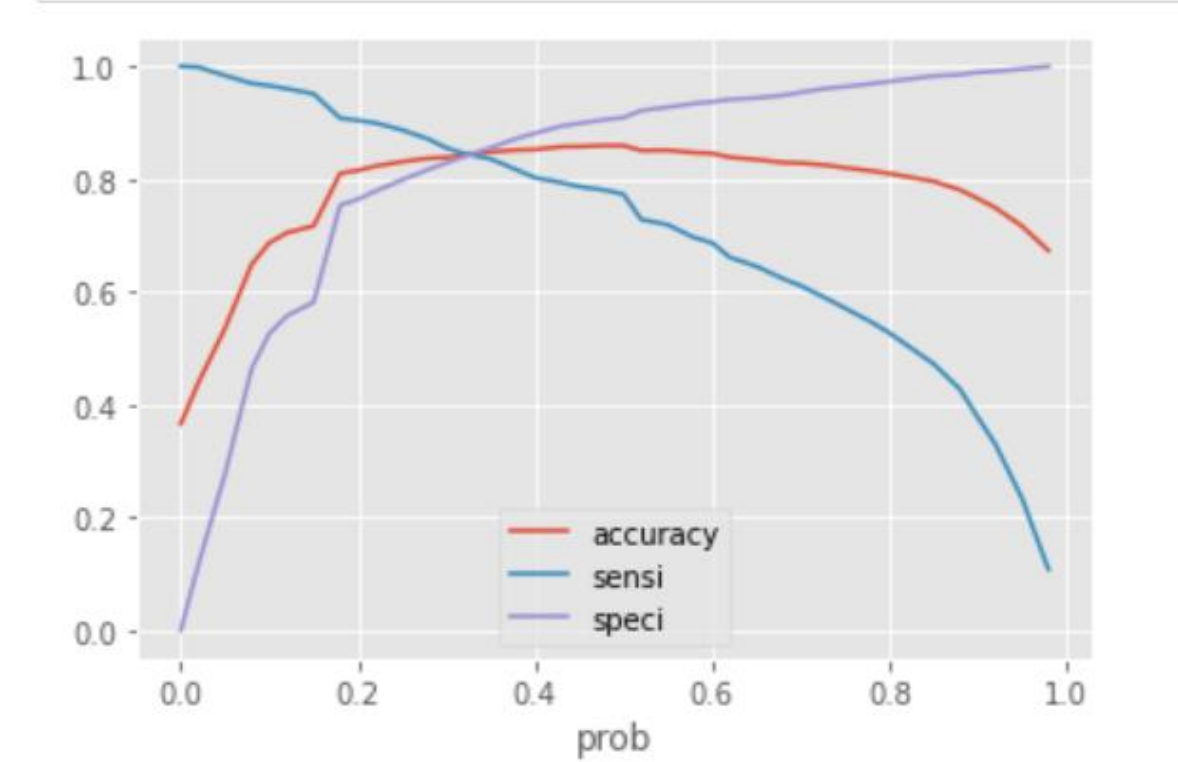
- Feature Selection using RFE
- Determined Optimal Model using Logistic Regression
- Calculated accuracy, sensitivity, specificity, precision & Recall & evaluate model

Variables Impacting the conversion rate

- Total Visits
- Total Time Spent on website
- Lead Source_Olark chat
- Lead Origin_Lead Add Form
- Lead Source Welingak Website
- Do Not Email
- Lead Source _Referral Sites

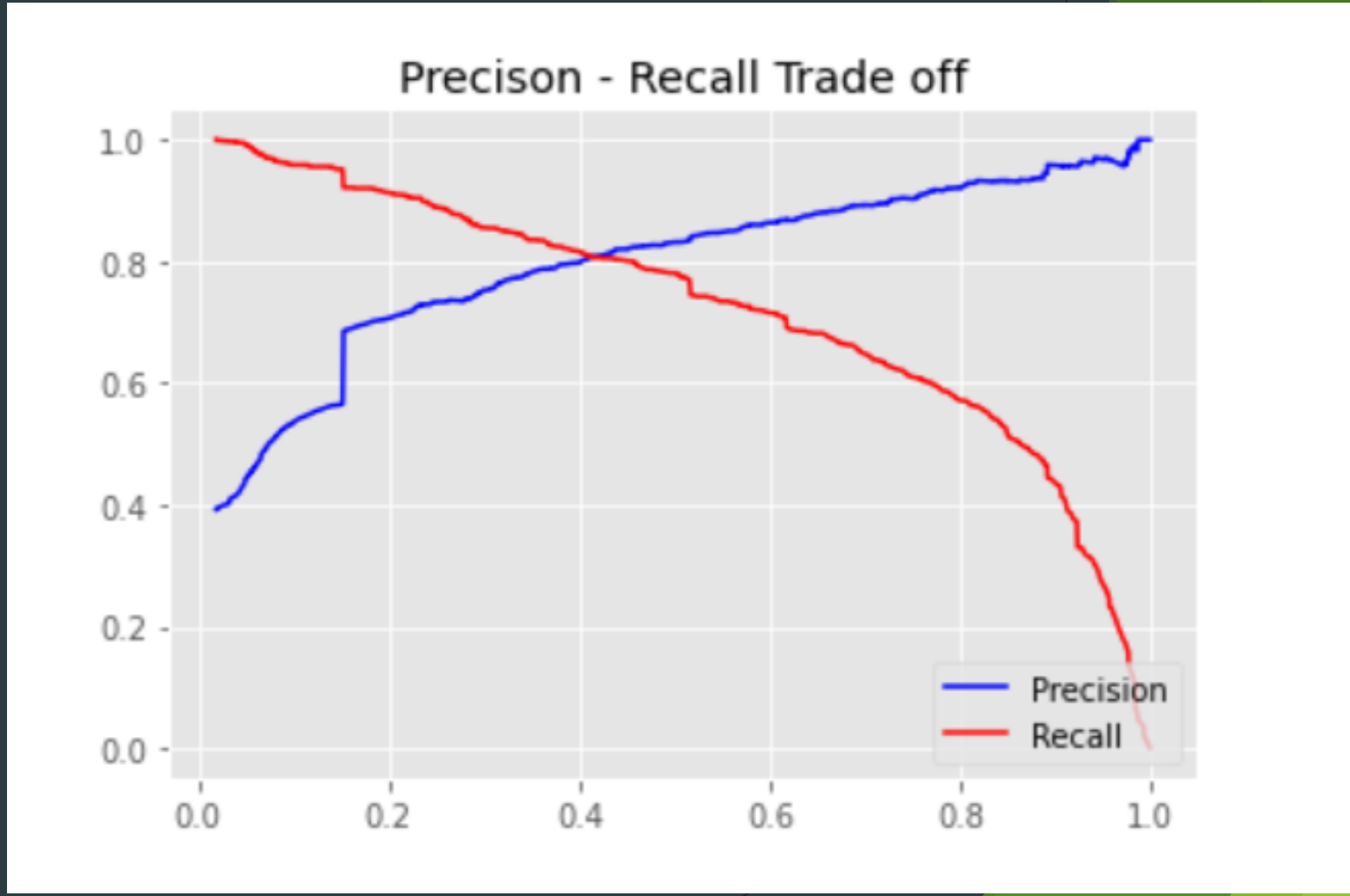
Model Evaluation-Sensitivity & Specificity on Train Data Set

- Graph depicts an optimal cutoff of 0.32 based on Accuracy, Sensitivity, Specificity



Model Evaluation
Precision & Recall on Train dataset

- Precision =76%
- Accuracy=85%



Model Evaluation
Sensitivity and Specificity on Test Data Set

- Accuracy= 85%
- Sensitivity= 85%
- Specificity= 85%

Result

- Accuracy, Sensitivity, and Specificity values of training and test set are close to the training set
- Accuracy, Sensitivity and Specificity values of training set are79%, 82%, 76% Respectively
- Accuracy, sensitivity & Specificity values of test are 78%, 81%,76% Respectively
- Conversion rate for Train & Test Dataset Is 82.7% & 80.8% Respectively
- We have done the prediction on the test set using cut off threshold from sensitivity & specificity metrics

Conclusion

- While we have checked both sensitivity-specificity as well as Precision & recall metrics, we have considered the optimal cut off based on sensitivity & specificity for calculating the final prediction
- Accuracy, Sensitivity & specificity values of test set are around 78%, 81%,76% which are approximately closer to Values calculated using Trained Data Set
- Lead Score Calculated for the conversion rate final model on Train & Test dataset is 82.7% &80.8% respectively.
- Hence, Overall Model seems to be Good

Summary

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion. First, sort out the best prospects from the leads you have generated. 'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted. Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects. Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.