

LEAD SCORING CASE STUDY


Tanvi Srivastava

Jitu Moni Das

PROBLEM STATEMENT

- ▶ X Education sells online courses to industry professionals
- ▶ The company markets its courses on several websites and search engines like Google. 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%.
- ▶ **Objective:**
 - ▶ X Education needs help in selecting 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%

METHODOLOGY

- ▶ Data Cleaning and Manipulation:
 - ▶ Duplicate Data Analysis
 - ▶ Null Value Analysis: Data imputation, Data Dropping
 - ▶ Outlier Analysis: Capping / Flooring of data
 - ▶ EDA:
 - ▶ Analysis of all categorical variables using bar charts
 - ▶ Analysis of all numeric variables using box plots
 - ▶ Correlation Analysis
 - ▶ Dummy Variables for Categorical Variables
 - ▶ Train-Test Split
 - ▶ Build Linear Regression Model using VIFs, RFE Analysis
 - ▶ Identify important coefficients and their impact
 - ▶ Metrics Analysis: Precision, Recall, Sensitivity, Specificity
- 

DATA CLEANING AND MANIPULATION

▶ Duplicate Data Analysis:

- ▶ No duplicates in Prospect ID and Lead Number
- ▶ Dropping these columns

▶ Null Value Analysis:

- ▶ Drop columns with null values $> 45\%$ and remove rows in columns where null values $< 10\%$
- ▶ Drop columns where one category is dominant ($> 75\%$) as this has data imbalance
- ▶ For other categorical columns with null values: Impute

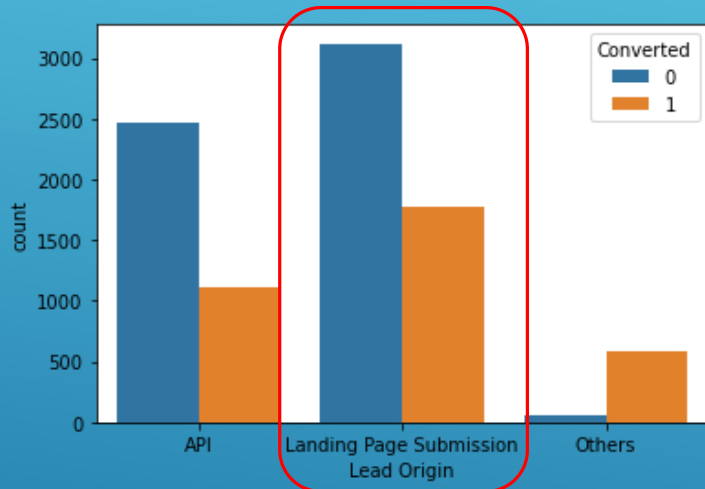
▶ Club small categories in columns with many categories

▶ Outlier Analysis

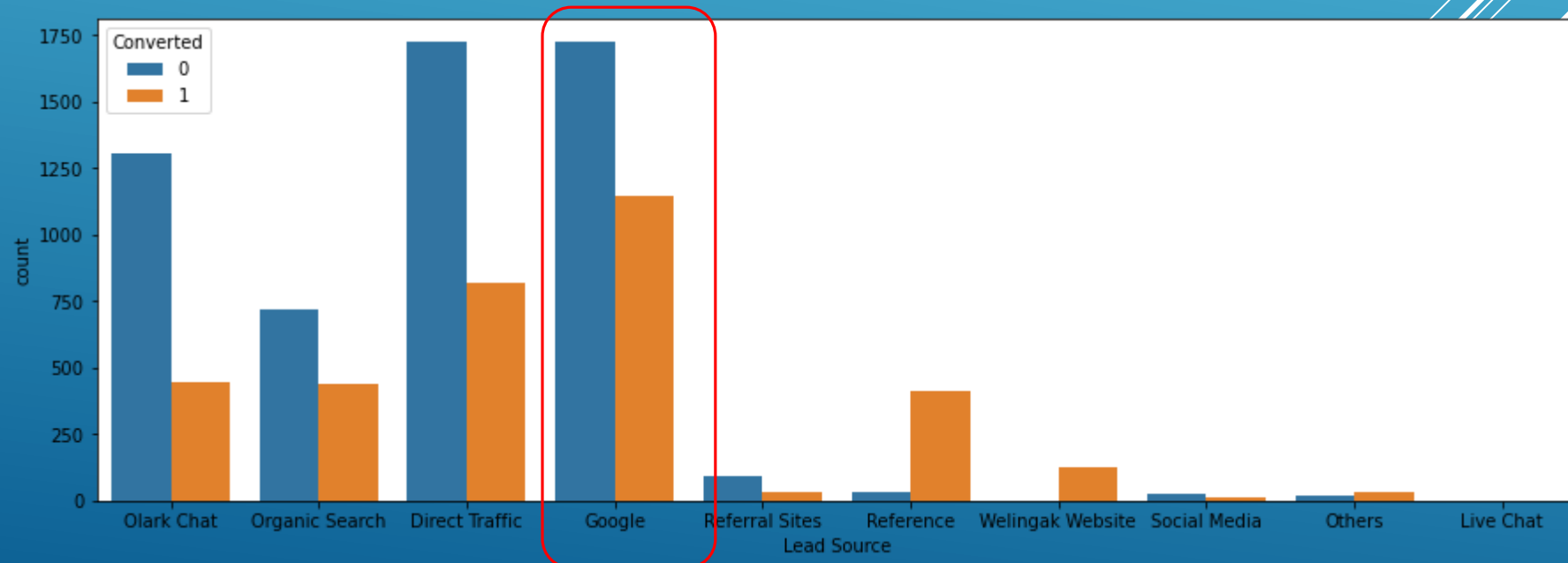
- ▶ Cap / Floor as necessary

EXPLORATORY DATA ANALYSIS

Categorical Columns: Analyzing each categorical column against Converted



Lead Origin

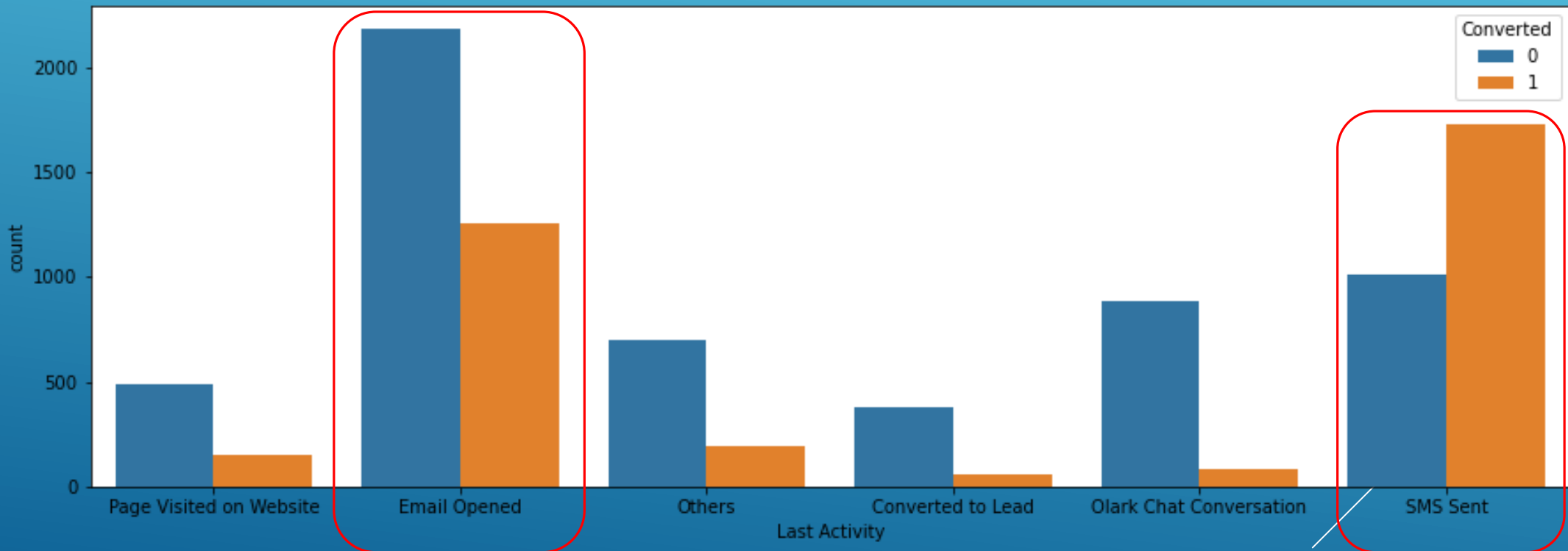


Lead Source

EXPLORATORY DATA ANALYSIS

Categorical Columns: Analyzing each categorical column against Converted

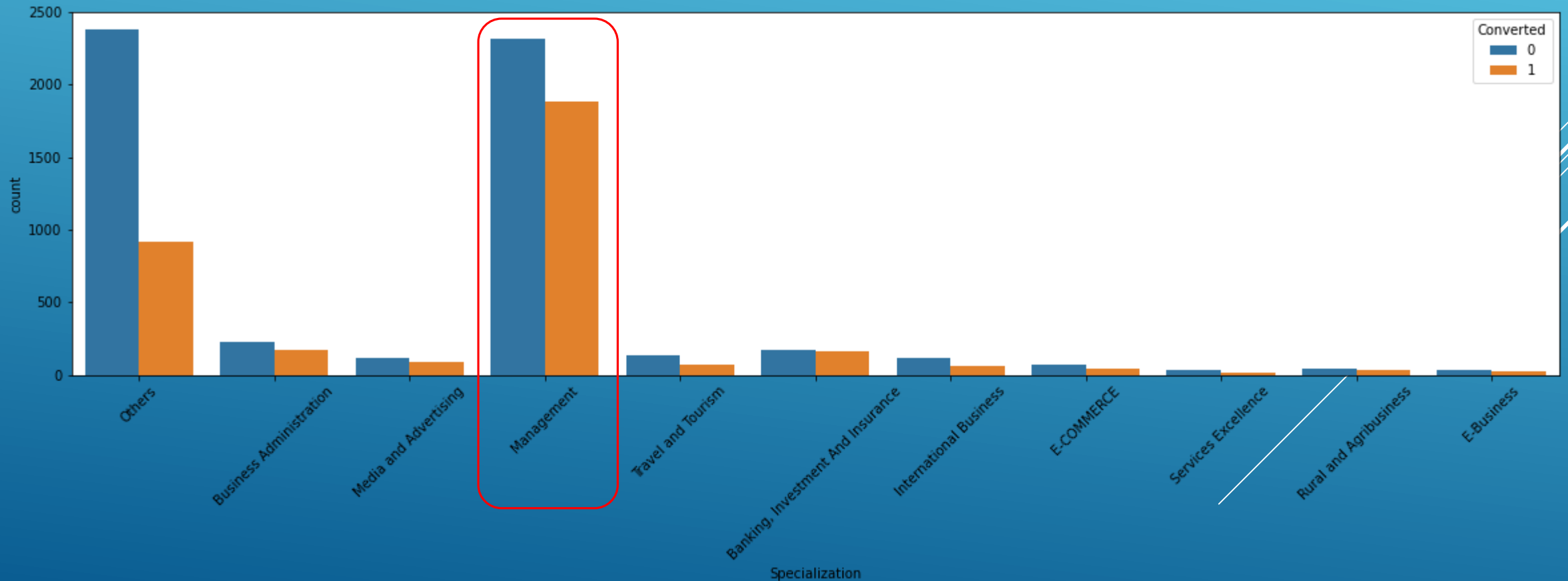
Last Activity



EXPLORATORY DATA ANALYSIS

Categorical Columns: Analyzing each categorical column against Converted

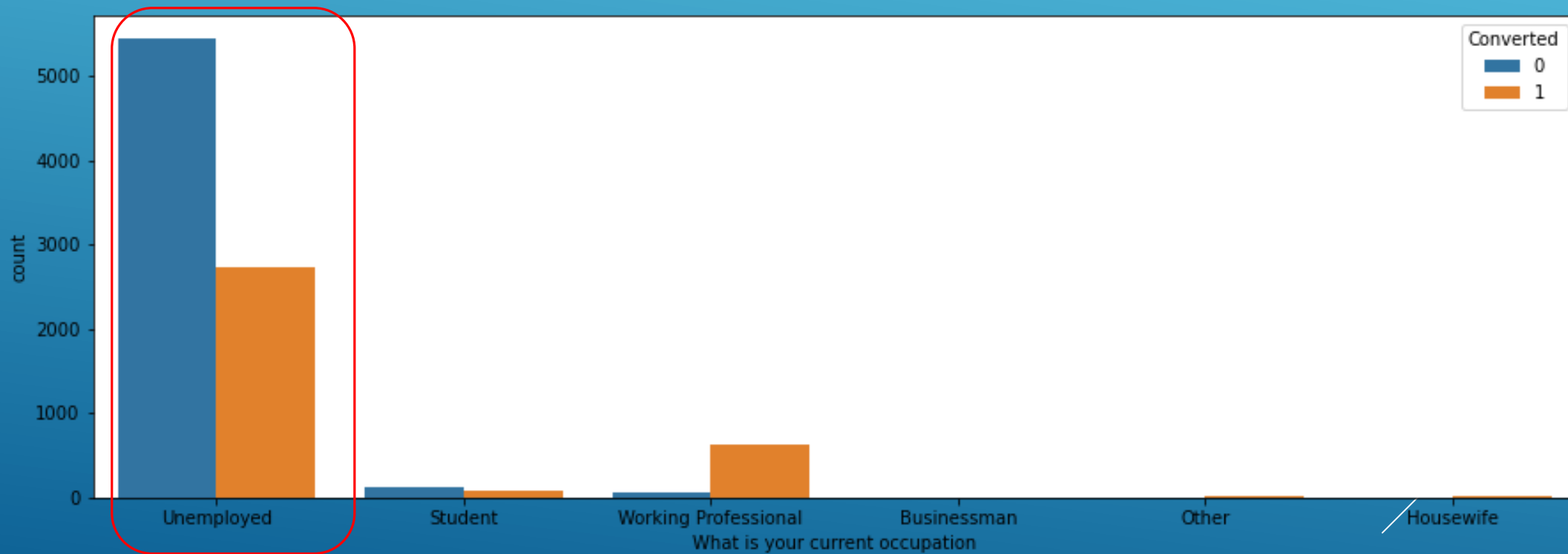
Specialization



EXPLORATORY DATA ANALYSIS

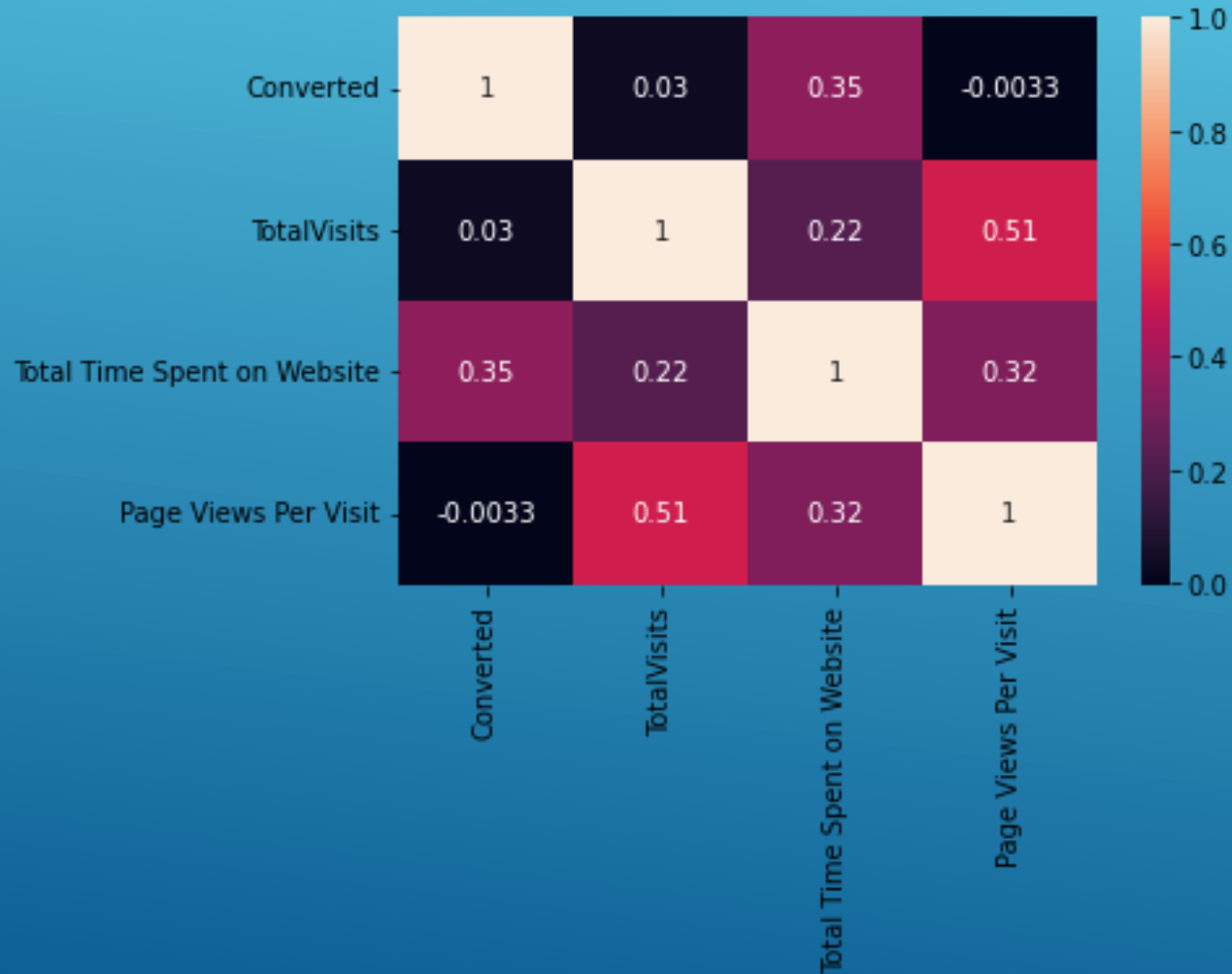
Categorical Columns: Analyzing each categorical column against Converted

What is your current occupation



EXPLORATORY DATA ANALYSIS

Numerical Columns: Heatmap



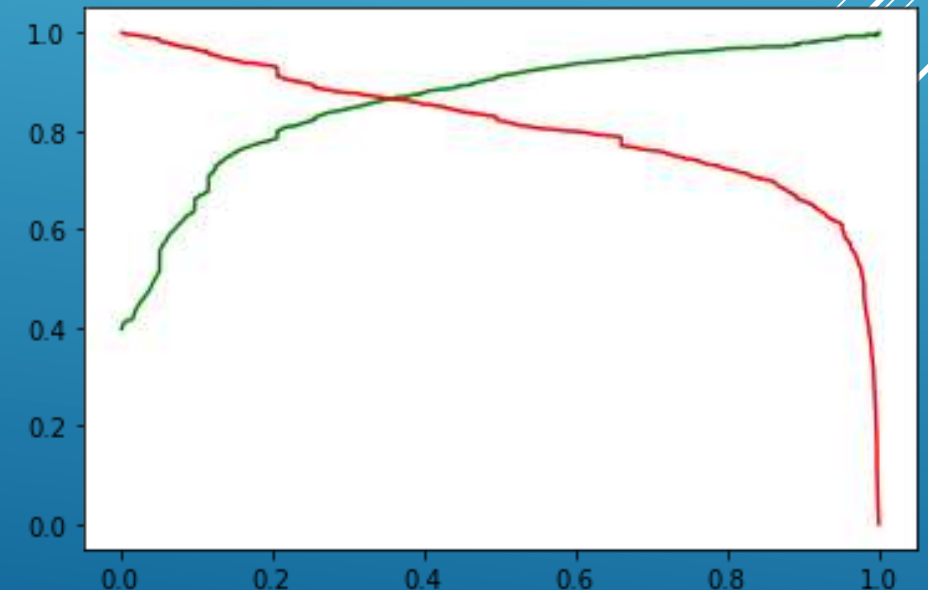
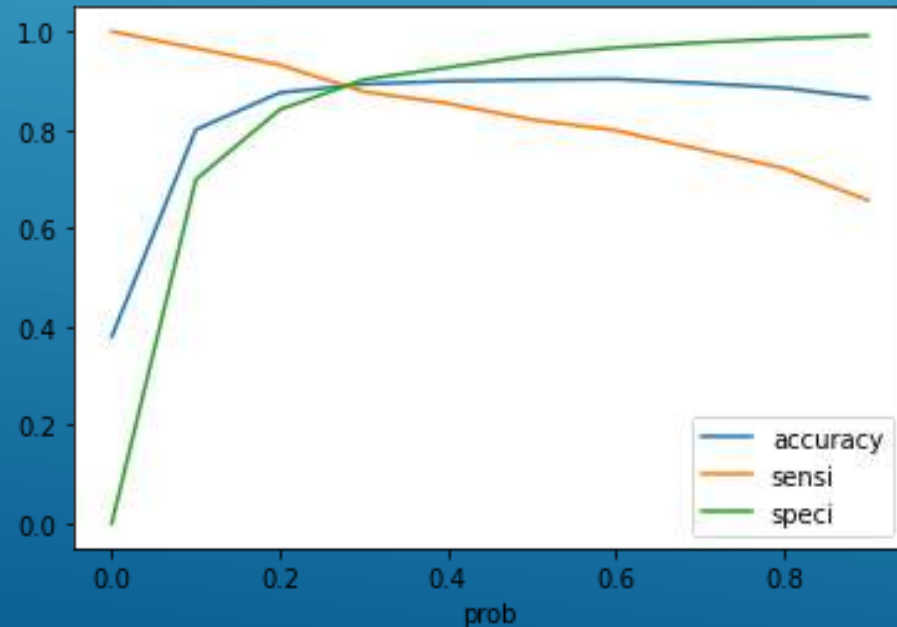
MODEL BUILDING INITIAL

- ▶ Post Data Manipulation and EDA:
 - ▶ Columns / Rows: 16 / 9103
 - ▶ Columns: Lead Origin, Lead Source, Do Not Email, Do Not Call, Converted, Total Visits, Total Time Spent on Website, Page Views Per Visit, Last Activity, Specialization, What is your current occupation, Tags, Lead Profile, City, A free copy of Mastering The Interview, Last Notable Activity
- ▶ Dummy Variables Creation
- ▶ Perform a Train-Test Split (70:30 Assumed)
- ▶ Perform Feature Scaling
- ▶ Running RFE with 15 variables as output
- ▶ Build Initial Model: Check VIFs and p-value more than 0.05
 - ▶ Drop columns with VIF more than 2.0
 - ▶ Columns Dropped: Last Activity_SMS Sent, Tags_Not Specified, Lead Profile_Lateral Student

MODEL BUILDING: FINAL

- ▶ Final Accuracy: 90%
- ▶ Metric Analysis: Sensitivity (82%), Specificity (95%), Precision (91%), Recall (82%)
- ▶ Find Optimal Cut-Off: 0.3 comes as optimal cut-off
- ▶ Make Predictions on Test Data
- ▶ Check Final Metrics for Test Data

ROC Curve



CONCLUSION

Factors which matter:

- ▶ Tags:
 - ▶ Will Revert After Reading the email, Closed by Horizon (+ve)
 - ▶ Switched Off, Ringing (-ve)
- ▶ Lead Source:
 - ▶ Welingak Website, Olark Chat (+ve)
- ▶ Lead Source:
 - ▶ Others
- ▶ Total Time Spent on Website