Lead Score Case Study

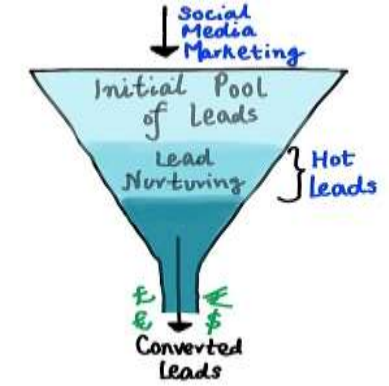
Team Members:

1. Pramod
2. Vineela.
3. Poornima.

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:



Lead Conversion Process - Demonstrated as a funnel

- 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.
- 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

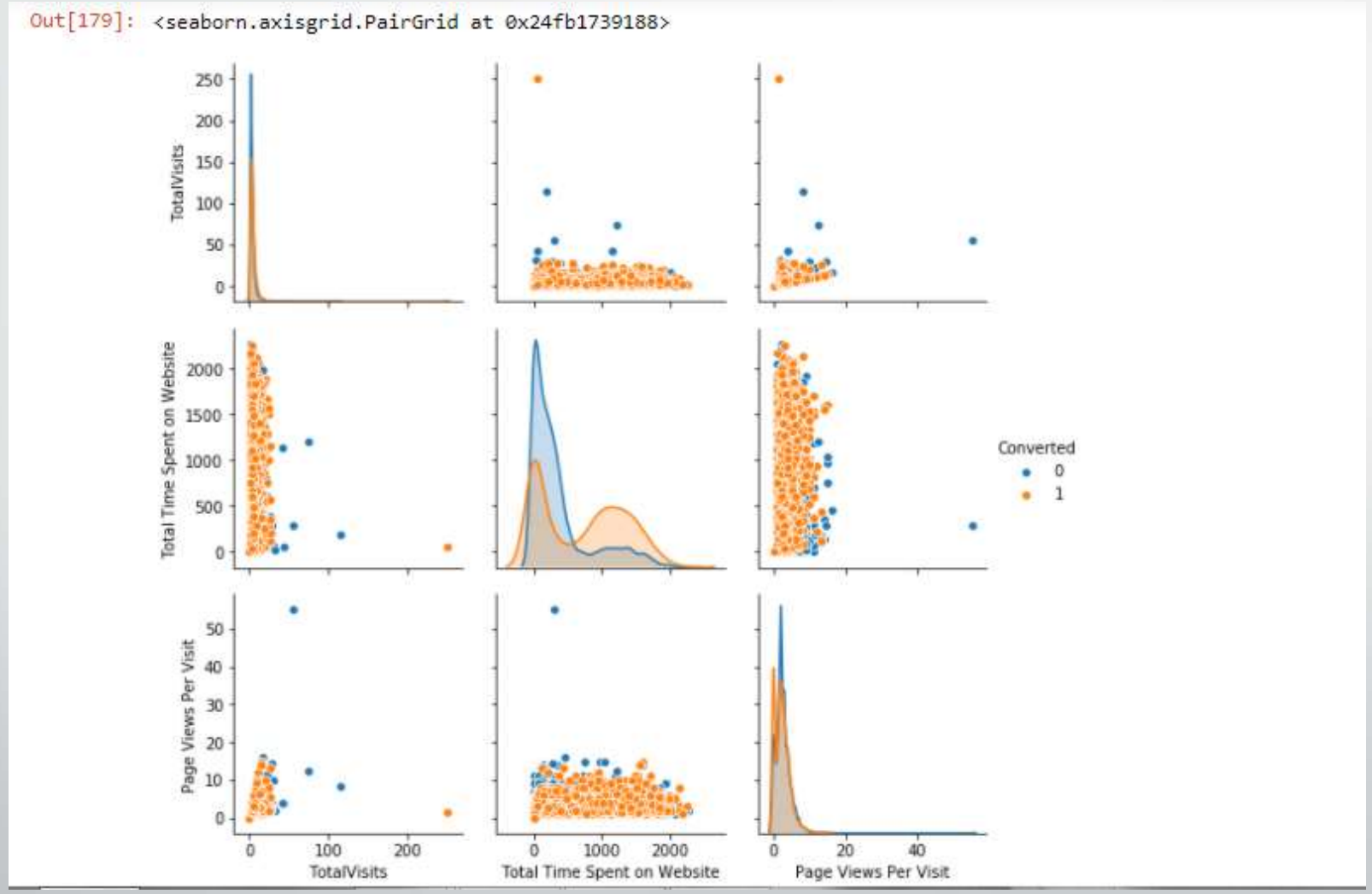
Solution Methodology

- Read and Understanding Data
- Data cleaning and data manipulation.
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Looking for Correlations
- Classification technique: logistic regression used for the model building.
- Model evaluation(checking P value and VIF) and predictions.
- Model presentation.
- Conclusions and recommendations.

Data Understanding, Cleaning and Preparation

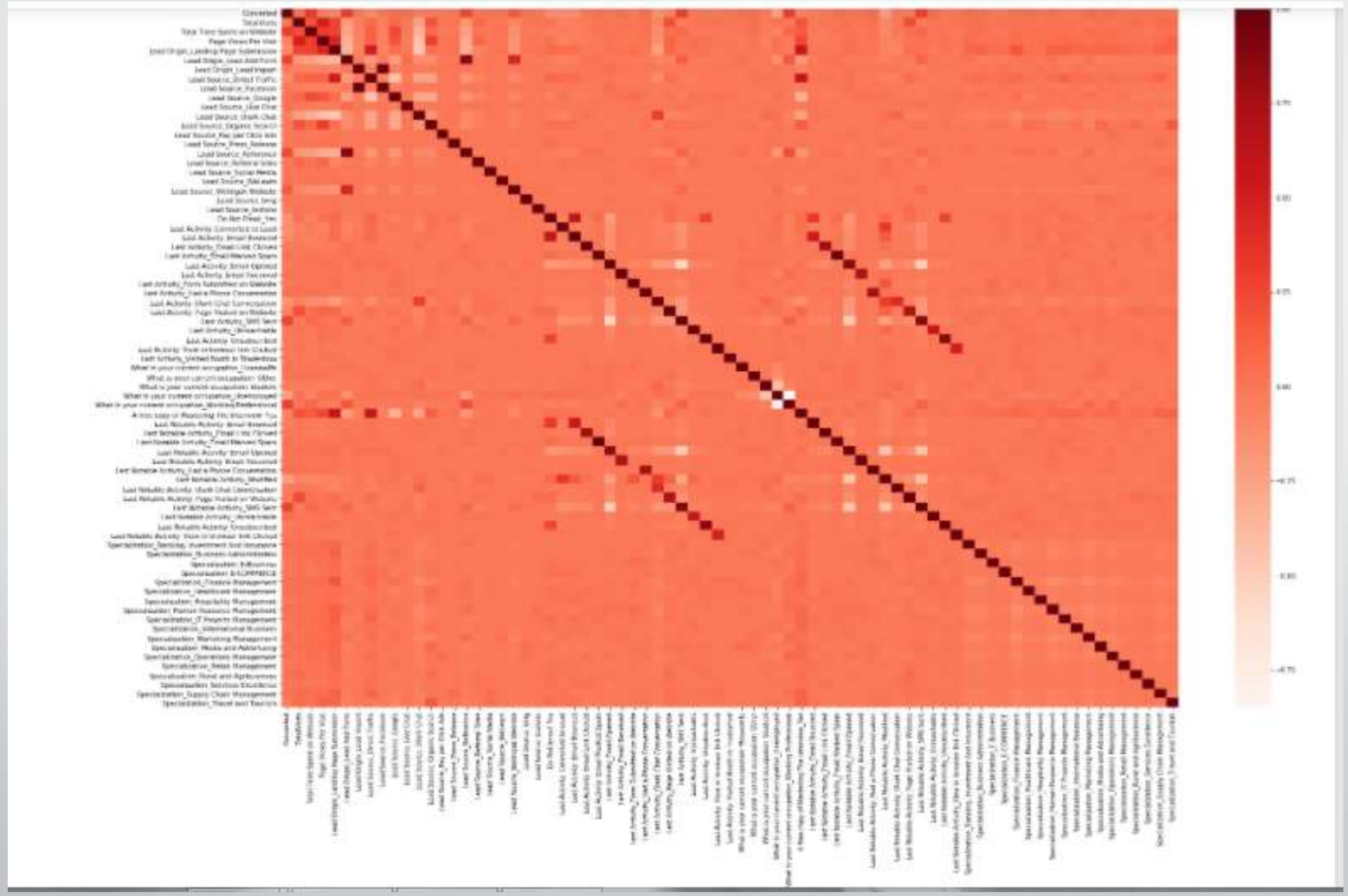
- Total Number of Rows =37, Total Number of Columns =9240.
- The column `City`, `Country` is not a potential columns as the problem statement is for online course.`Prospect ID`, `Lead Number` columns are ID columns which are not used for analysis
- The column `What matters most to you in choosing a course` has the input `Better Career Prospects` `6528` times while the other two inputs appear once twice and once respectively.
- The level called 'Select' means that the student had not selected the option for that particular column. These values can be considered as missing values and hence we need to identify the value counts of the level 'Select' in all the columns that it is present. Columns 'Lead Profile', 'How did you hear about X Education' and 'Specialization' has select level.The columns `Lead Profile`(4146) and `How did you hear about X Education`(5043) have a lot of rows which have the value 'Select' which are equalent to null so we can drop them.
- Few columns in which only one value was majorly present for all the data points. These includes `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`. As these columns won't help with analysis due to low variance we drop these columns.
- After dropping all the NULL values, We still have around 69% of the rows retained.

Data Visualization



Distribution of numerical of columns is currently right skewed. As we will use scaling not dealing with outliers currently.

Looking at the correlations



The number of variables are pretty high

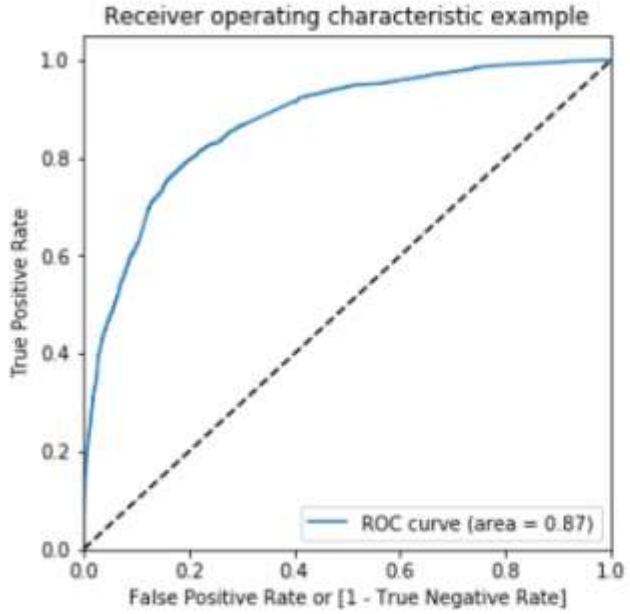
Model Building

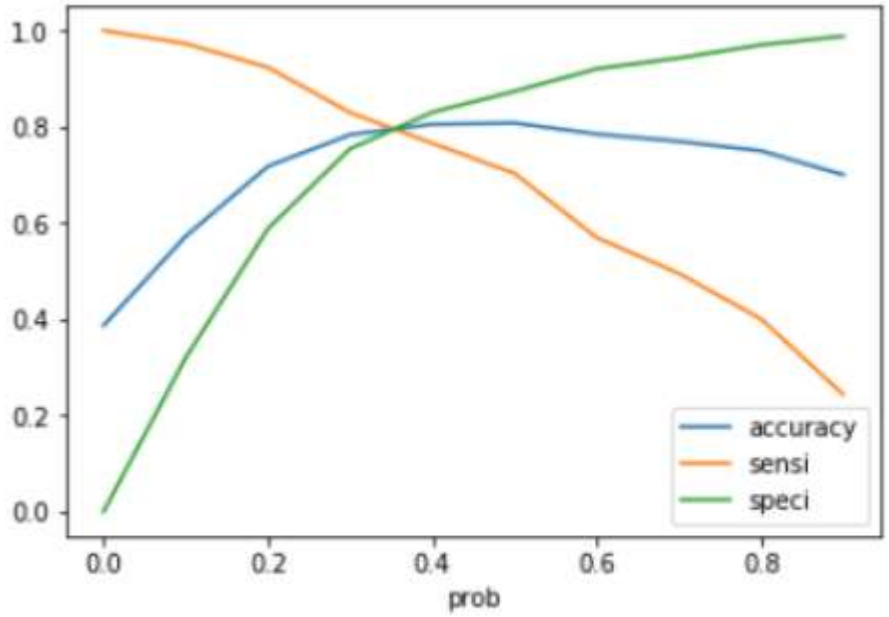
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Overall accuracy 79.1%

Model Evaluation

- Overall accuracy 78.1%

ROC Curve





- Finding Optimal Cut off Point
- Optimal cut off probability
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off around 0.4.

Conclusion

- It was found that the variables that mattered the most in the potential buyers are:

- The total time spend on the Website,Total number of visits.
- Observations on Train & Test dataset:
  - Train Data: Accuracy : 79.1%,Sensitivity : 81.0%,Specificity : 77.4%
  - Test Data: Accuracy : 78.1%Sensitivity : 76.0%,Specificity : 79.5%

- The Model is having 79.1% accuracy which is around 80% and we should be able to give the CEO confidence in making good calls based on this model