Lead Scoring Case-study

Study group participants:

- Roshan Chandru
- Dishi bhinde

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:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

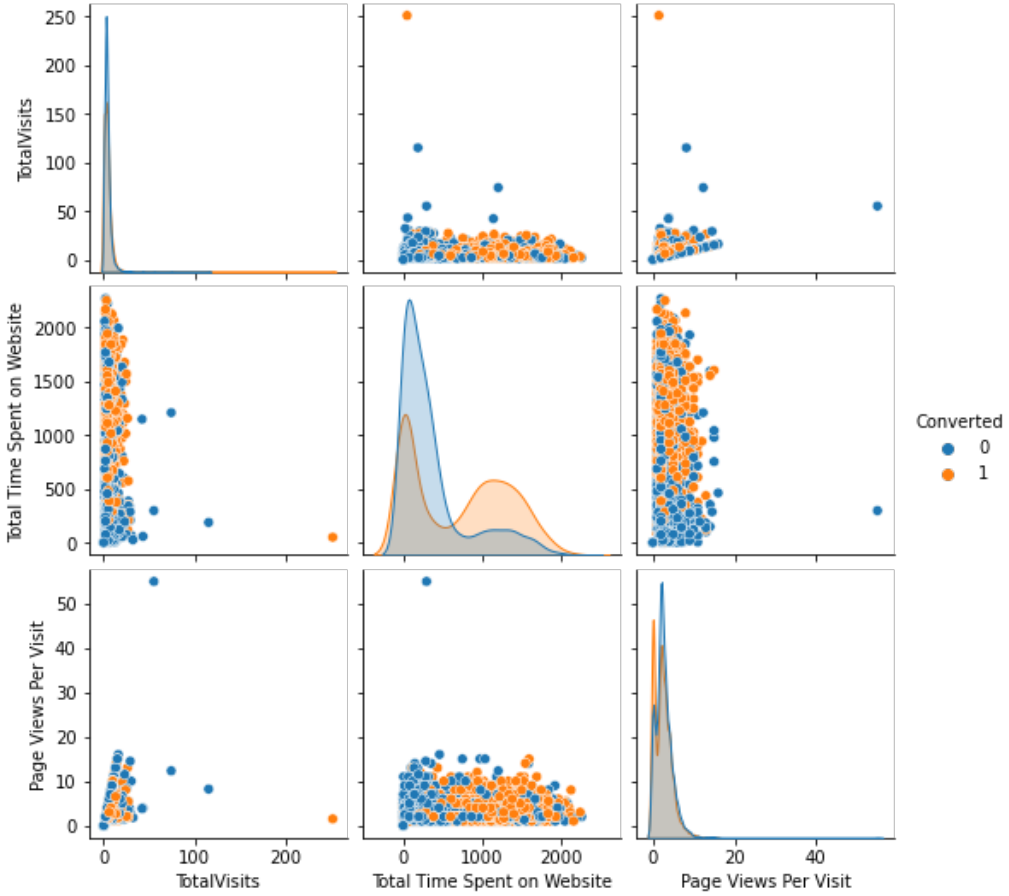
Solution Methodology

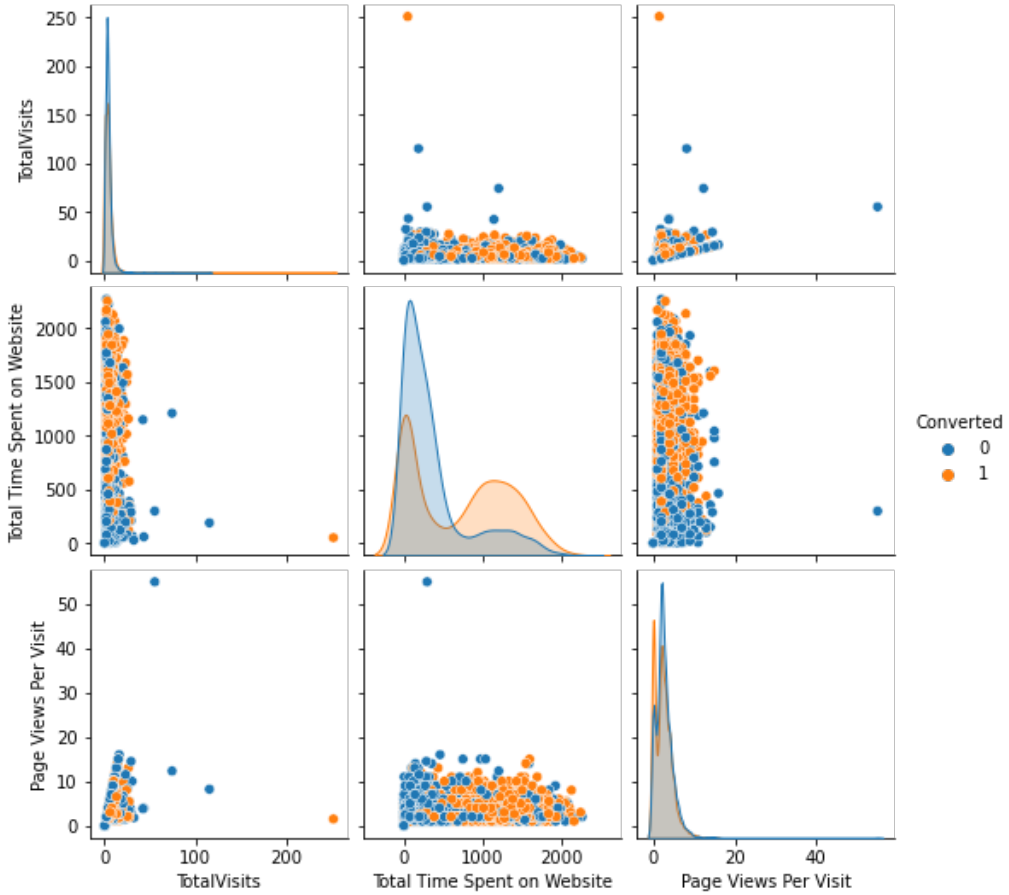
- Read and understand the data
- Clean the data
- Prepare the data for Model Building
- Model Building
- Model Evaluation
- Making Predictions on the Test Set

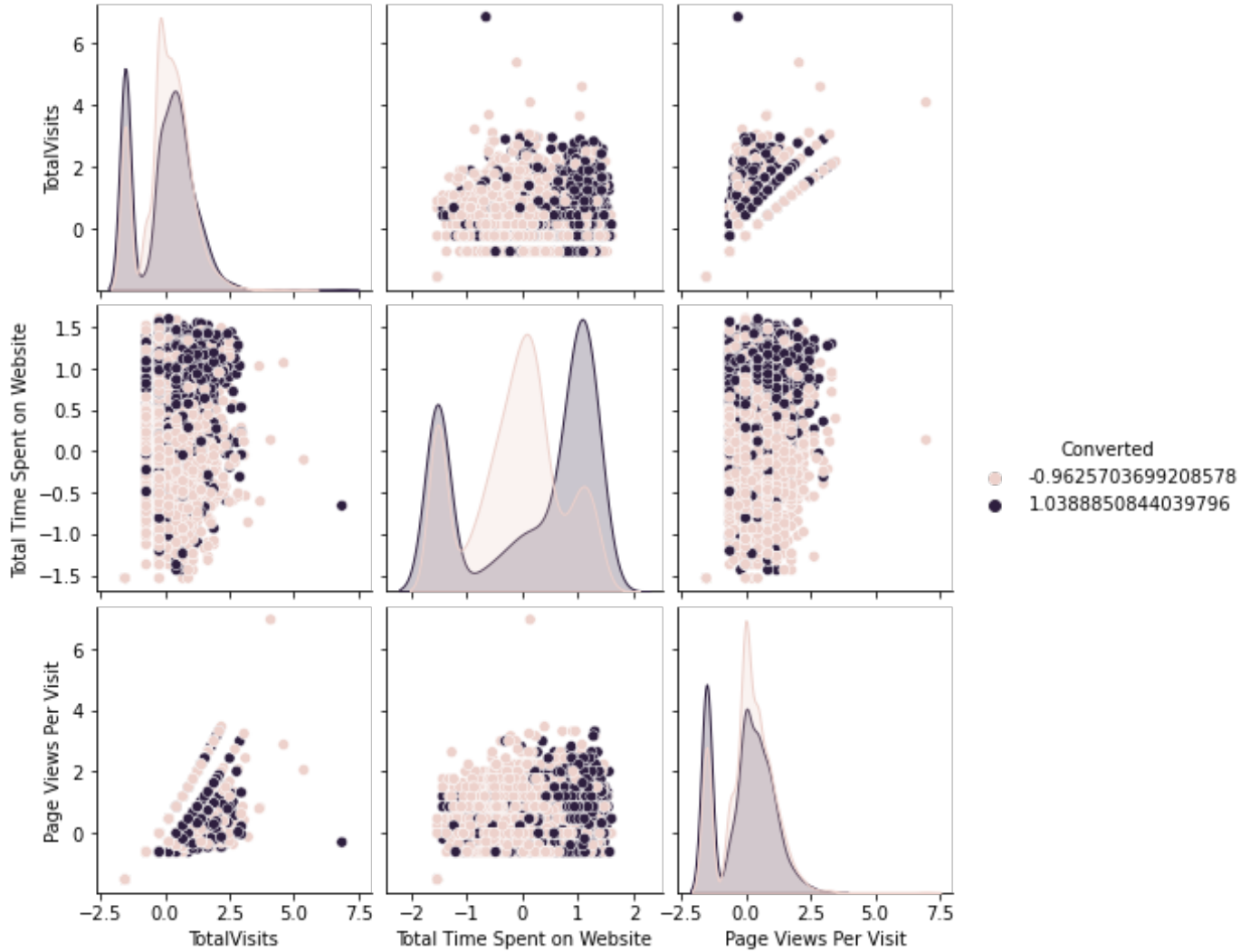
Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “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” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Preparing the data for modelling





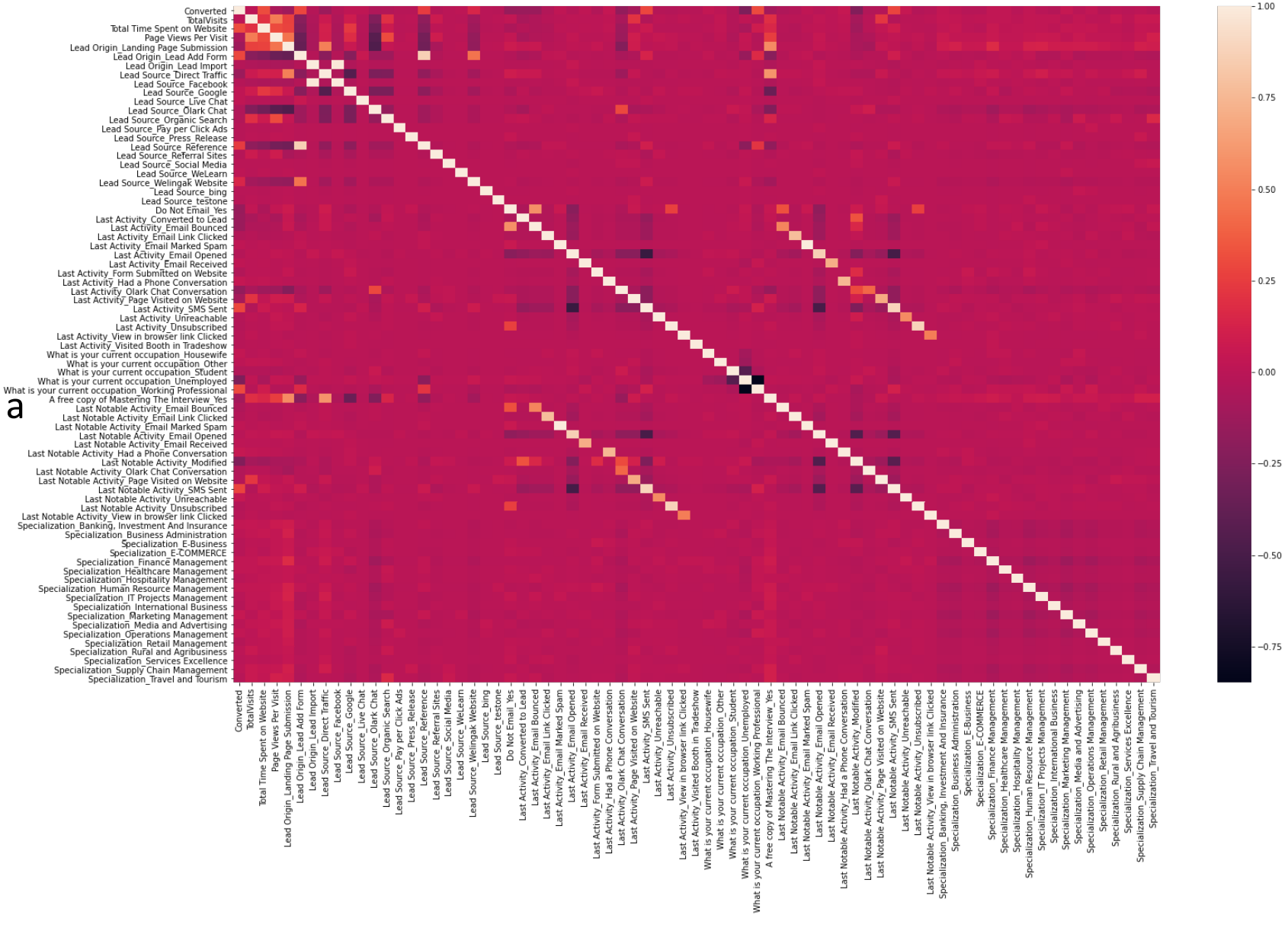

TotalVisits
Total Time Spent on Website
Page Views Per Visit
TotalVisits
Total Time Spent on Website
Page Views Per Visit
Converted
-0.9625703699208578
1.0388850844039796

Dummy variable Creation:

- The next step is to deal with the categorical variables present in the dataset. So first take a look at which variables are actually categorical variables.
- Check the columns which are of type 'object’
- Create dummy variables using the 'get_dummies' command
- Add the results to the master dataframe
- Creating dummy variable separately for the variable 'Specialization' since it has the level 'Select' which is useless so we
- Drop that level by specifying it explicitly
- Drop the variables for which the dummy variables have been created

Test-Train Split

- Put all the feature variables in X
- Put the target variable in y
- Split the dataset into 80% train and 20% test
- Import MinMax scaler
- Looking correlations by plotting it in a heatmap



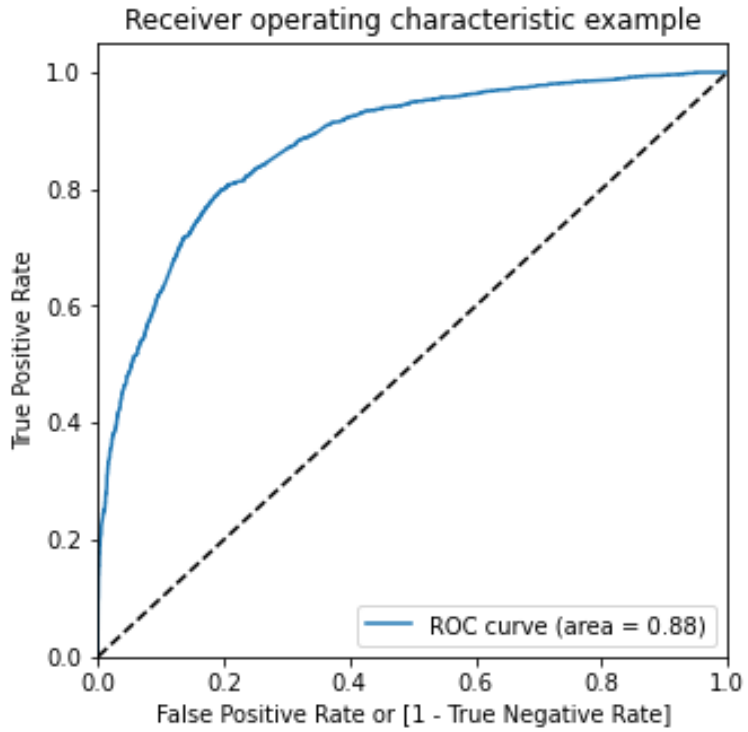
Model Building

- Import 'Logistic Regression' and create a Logistic Regression object
- Import RFE and select 15 variables
- Put all the columns selected by RFE in the variable 'col’ (Now you have all the variables selected by RFE and since we care about the statistics part, i.e. the p-values and the VIFs, let's use these variables to create a logistic regression model using statsmodels.)
- Select only the columns selected by RFE
- There are quite a few variable which have a p-value greater than 0.05. We will need to take care of them. But first, let's also look at the VIFs.
- Import 'variance_inflation_factor’
- Make a VIF dataframe for all the variables present (VIFs seem to be in a decent range except for three variables. Let's first drop the variable Lead Source_Reference since it has a high p-value as well as a high VIF.)
- Drop `What is your current occupation_Working Professional`.
- Refit the model with the new set of features
- All the p-values are now in the appropriate range. Let's also check the VIFs again in case we had missed something. Make a VIF dataframe for all the variables present

Model Evaluation

- Use 'predict' to predict the probabilities on the train set
- Reshaping it into an array
- Create a new dataframe containing the actual conversion flag and the probabilities predicted by the model
- Create confusion matrix
- Let's check the overall accuracy
- Let's evaluate the other metrics as well
- Calculate the sensitivity
- Calculate the specificity

Optimal Cut-off:



- Finding Optimal Cut off Point
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model.

Conclusion

- There are a lot of leads generated in the initial stage, 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 step is to 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.