LEAD SCORE

A CASE STUDY

BY TALLURI K SAKETH RAM & HIMABINDU EROLLA

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.
- The company gets leads on who may be interested for a course from various sources such as Online, or fill a form or through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Typical Lead Conversion rate at X Education is 30%. The lead conversion rate is poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- We need to help the company select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads

DETAILS OF DATA SETS

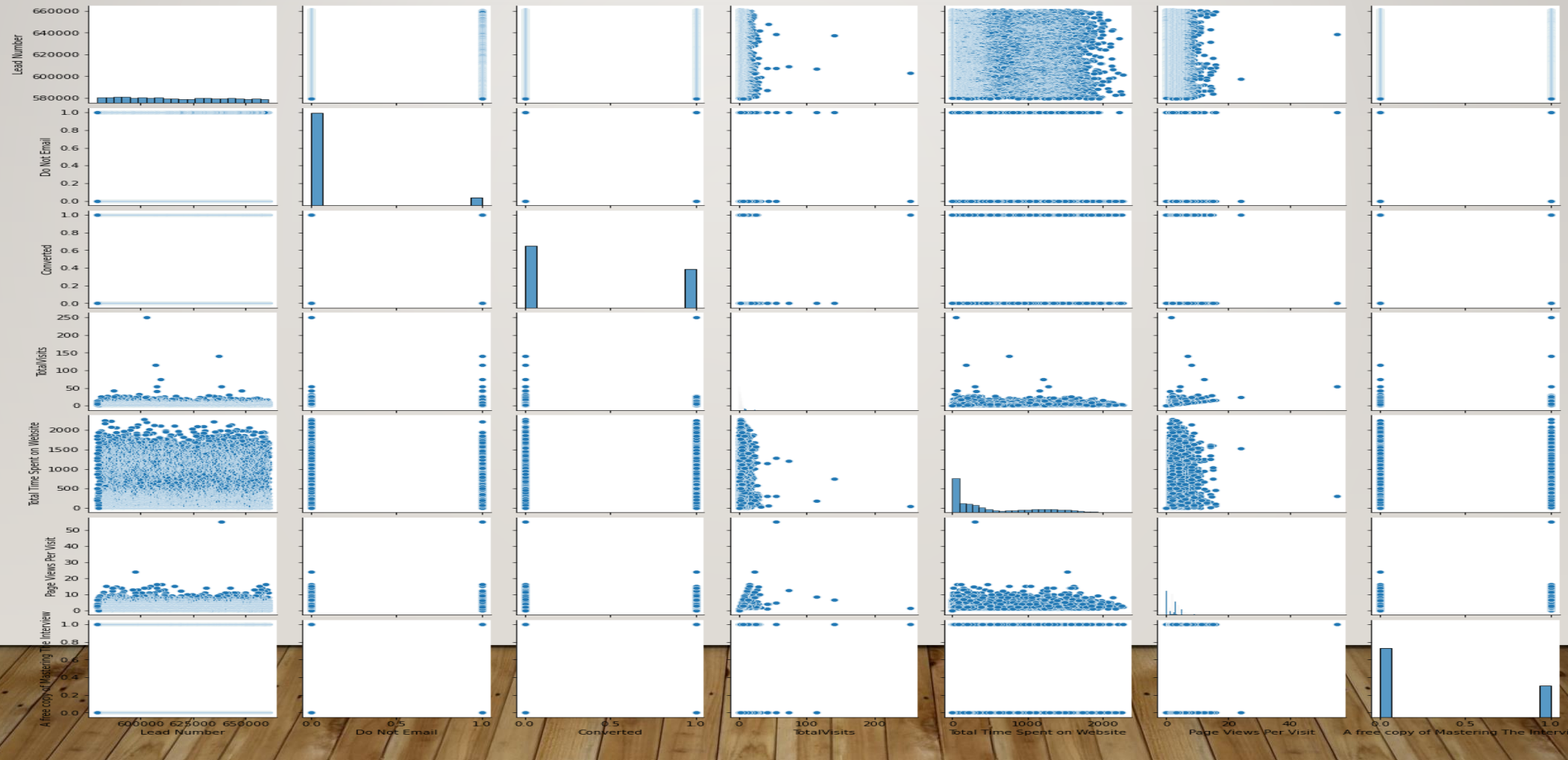
Dataset provided for this case had the following files:

1. *Leads.csv* - The data has the details about leads and information from online forms.
2. *Leads Data Dictionary.csv* – Data Dictionary for the given dataset.

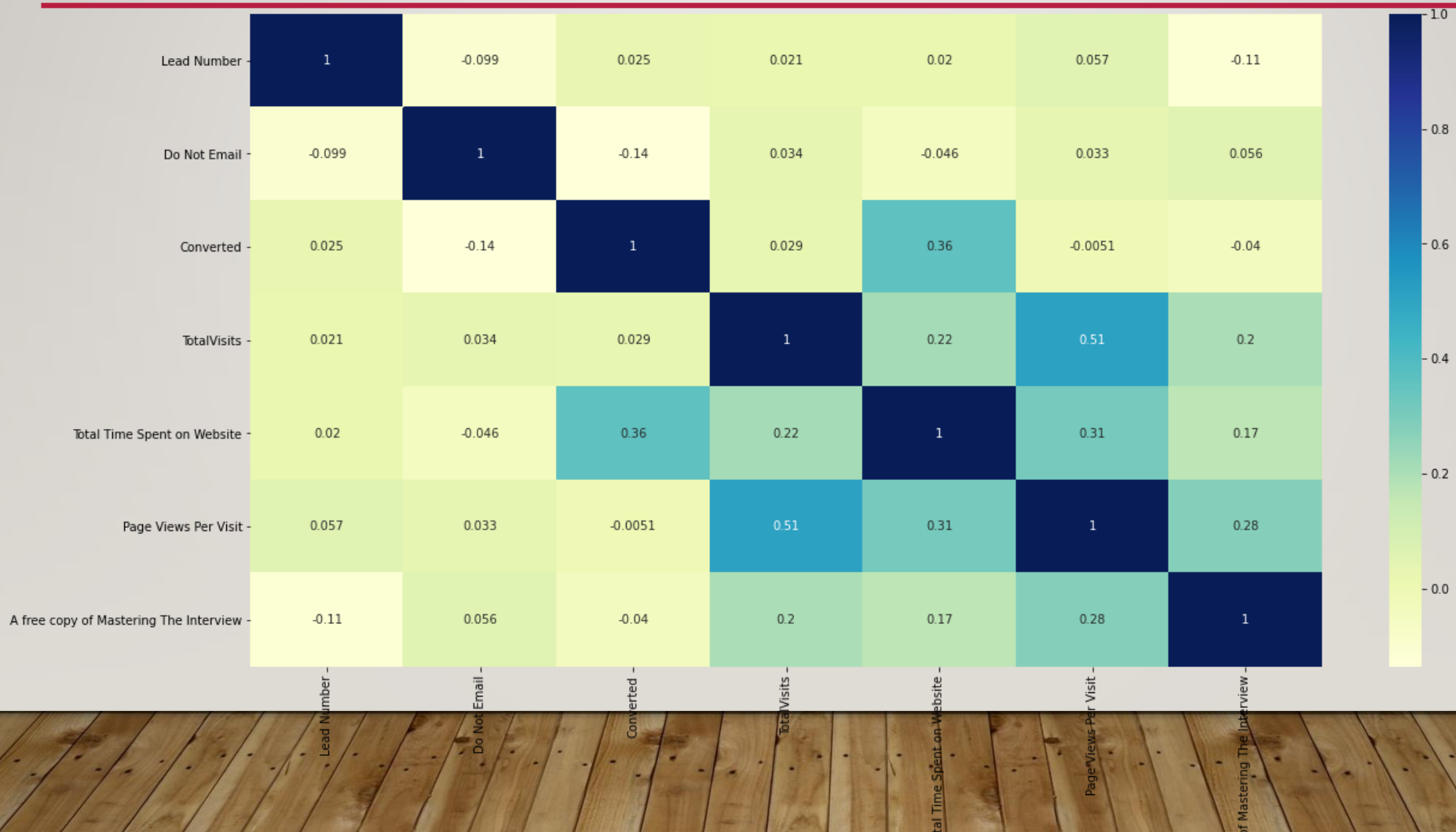
CLEANING DATA SET

- First we have checked the details like info on the data set and check for Null Values.
- Replace 'Select' cells to Null values in the Data Frame and remove the variables (columns) with more than 40% missing values.
- Next, we check the rows with most missing values with intent of removing rows with more than 70% Null Values.
- Next Stage will be the missing Value imputation. The columns with less number of missing values have to be imputed. The following process is followed for imputing
    - For Categorical Variables, missing values are imputed using Mode.
    - For Numerical Variables, missing values are imputed using Mean if the distribution of data is normal.
    - For Numerical Variables, missing values are imputed using Median if the distribution of data is not normal
- In many variables, the total values are very high like there are some 15 to 20 Variables also. We can reduce these based on the percentage of the value counts. These with minimal percentage are bunched into a single category of Others.
- Also, we checked the value counts on all the categorical variables and checked for Skewed Variables, which means the variable which has most of the variables in a single category. These variables are not much of a use for the model. So, we drop all these columns.

UNIVARIATE ANALYSIS



BIVARIATE ANALYSIS

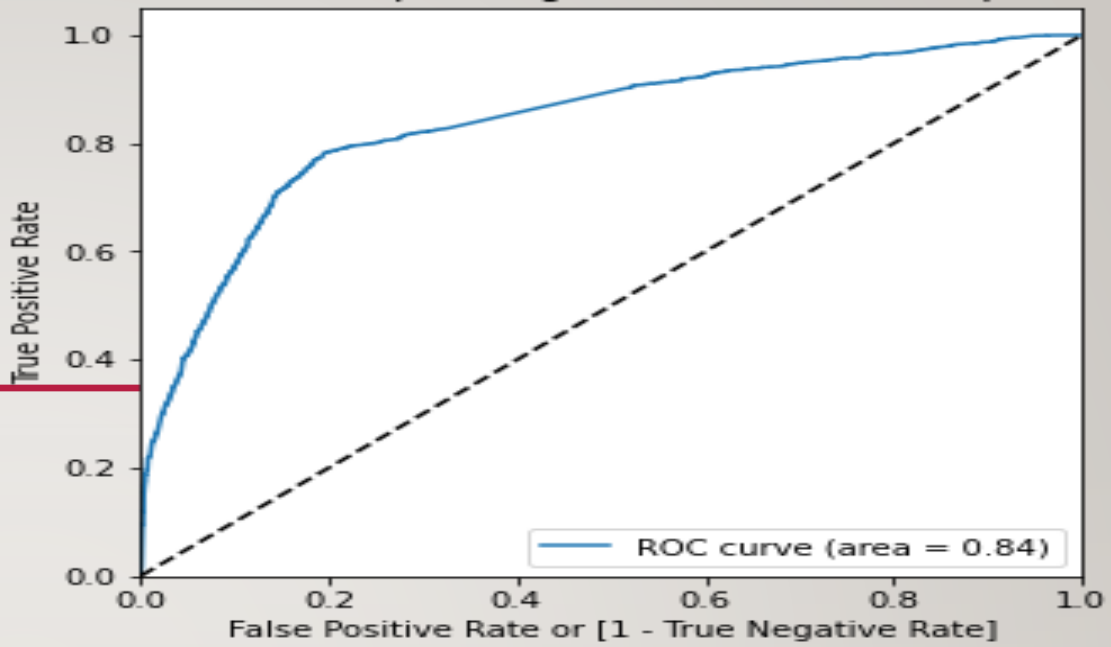


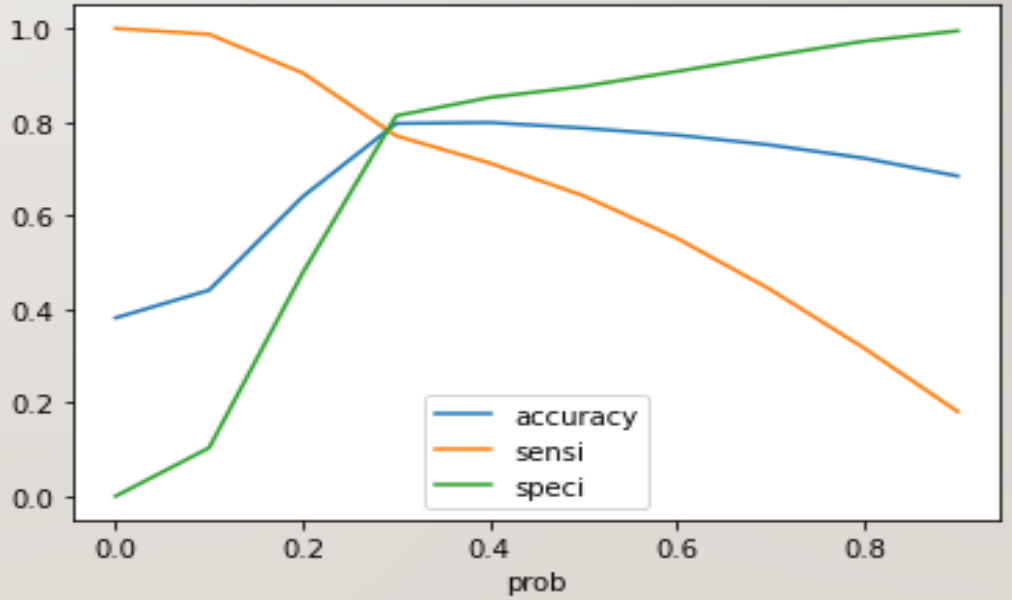
MODELING

- Next, we perform the binary mapping of the variables and also create dummy Variables.
- After the above step, we go for train and test split and Scaling of the variables using Standard Scaler.
- We start with the ranking of the variables using RFE.
- Next, we go for modelling of the data, create the predicted variables and make a Confusion Matrix and check the accuracy of the Model and check the VIF of the variables.
- Based on the P Value and VIF, we can start eliminating features which are not much helpful for the Model. We repeat this process till there is no much change in the accuracy of the model.

MODELING (CONTD.)

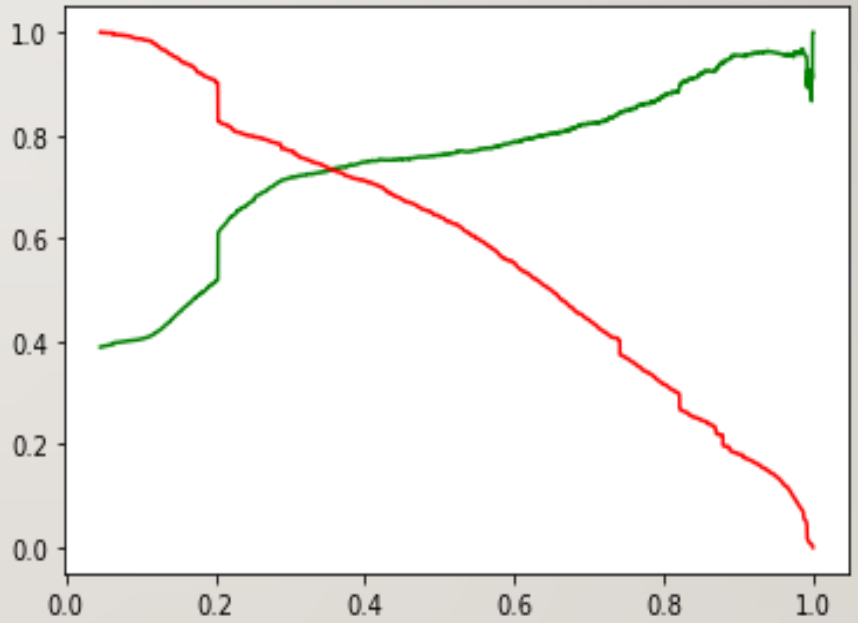
- Next we check Sensitivity and Specificity of the Model and plot the ROC Curve and find optimal cut-off point.
- The optimal cutoff point is 0.3
- Final Accuracy on the Train Data Set is 0.796
- Sensitivity – 0.77
- Specificity – 0.81





MODELING (CONTD.)

- Next, we check the Precision and recall values.
- Precision – 0.761
- Recall – 0.642
- Precision Recall Curve as seen in pic beside.
- Next, we move on to the test set and get the predictions of the test data set.
- Final Accuracy on the Test Data Set is 0.794
- Sensitivity – 0.715
- Specificity – 0.845



MODELING (CONTD.)

- After that, we concatenate the prediction values from train and test set and merge with the original data frame and create a score column by multiplying the predictions with 100.
- Hence, we also get the score on which lead is most likely to get converted.
- The main variables that are going into the model are as follows

'Do Not Email', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'A free copy of Mastering The Interview', 'Lead Origin_Landing Page Submission', 'Lead Origin_Others', 'Lead Source_Olark Chat', 'Lead Source_Others', 'Specialization_Finance Management', 'Specialization_Human Resource Management', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional','City_Other Cities', 'City_Thane & Outskirts'

BUSINESS ASPECTS

- Based on the Model we can see that Total Time Spent on Website, Lead Origin_Others,

What is your current occupation_Working Professional.

- The company should focus on the email campaign and social media campaigns to advertise and give information on the courses offered by the company and track the views and concentrate on the enquiries.
- The sales team of the company should focus more on the aggressive follow ups with the candidates and should highlight the advantages of taking the course professionally either that it helps for a promotion in current position or transition to a new role to convert more leads.

END