LEAD SCORE CASE STUDY

Group Name:

1. Sarvan Kumar Yadav
2. Rajasekhar Battula
3. Satish Kumar
4. Darshna

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.

- In order to increase the lead conversion rate, the company first should 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 METHODOLOGIES

- Following steps are performed as part of data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.

- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Scaling, dummification and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

DATA MANIPULATION PART -1

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.

- Removed the “Prospect ID” and “Lead Number” which are unique and are not necessary for the analysis.

- After checking for the value counts for some of the object type variables, we have found the variables “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc has not enough variance and hence are dropped.

- Conversion rate for the Recommended leads = 71.43 %, which is quite higher then the normal leads. But there is very few representations of this feature, therefore we have dropped this feature.

- Dropping the columns having more than 70% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

DATA MANIPULATION PART -2

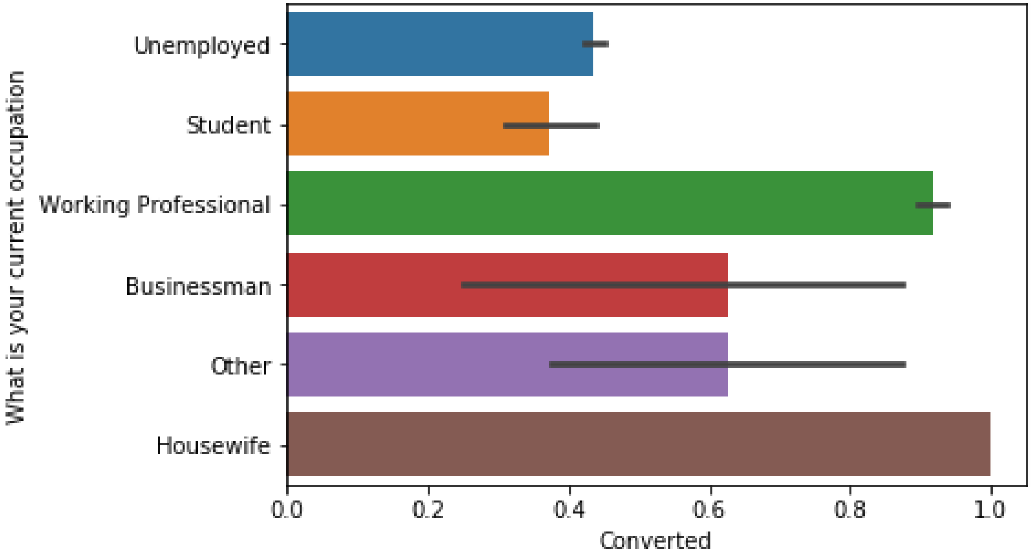
Now, we are left with total number of rows =22, total number of columns =9240.

- Some columns has **“Select”** as the value:

1. Specialization
2. City
3. Lead Profile
1. How did you hear about X Education

They are actually null value, so we impute the null value in place of “Select”

- There are 2690 missing values in the column “What is your current occupation” and these values are considered as a new feature and given the name as “Unknown”



DATA MANIPULATION PART -3

Total Number of Rows =20 , Total Number of Columns =9240

- Dropping the columns having more than 30% as missing value. (Specialization, Tags, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score and Asymmetrique Profile Score)
- Fraction of conversion is not matching with any of the items in the column. So, drop Lead Quality, Last Notable Activity.
- Outlier of upper boundary are removed for the features “TotalVisits” and “Page Views Per Visit”.
- Observed that major part of null values in "Page Views Per Visit", "TotalVisits" are Converted. So, imputing median values of them to null values.
- Missing values from the column “Lead Source“ are imputed using the field “Lead Origin’.
- Three columns are renamed for readability "Total Time Spent on Website“ → "Total Time Spent", "What is your current occupation“ → "current occupation", "A free copy of Mastering The Interview“ → "Mastering The Interview"

Number of retained rows=9233
Percentage of retained rows=99.92 %

DATA CONVERSION

- Converted binary categorical variable to numerical variable.
- Multilevel variable converting to dummy using One-hot encoding.

After these conversions the shape of the data is mentioned below

Total Number of Rows =52
Total Number of Columns =9233

MODEL BUILDING

- **Splitting the Data into Training and Testing Sets**

As you know, the first basic step for regression is performing a train-test split, we have selected 30% data for model validation.

- **Use RFE for Feature Selection**

Running RFE with 15 variables as output

- **Building Model**

Removing the variable whose p- value is greater than 0.05 and VIF value is greater than 3.

PREDICTIONS ON TRAIN DATA SET

Overall accuracy

Initial =81.76%
Overall the accuracy has dropped a bit but this is because of null values of "current occupation" column which is not explainable we can conveniently remove this column.
After removing columns with high p-values and unknown values =79.81%

MODEL RESULTS

- **METRICS BEYOND SIMPLY ACCURACY**

Confusion Matrix

```
[[3535  436]

[ 869 1623]]
```

Sensitivity =65.13%

Specificity=89.02%

False Positive Rate=10.98%

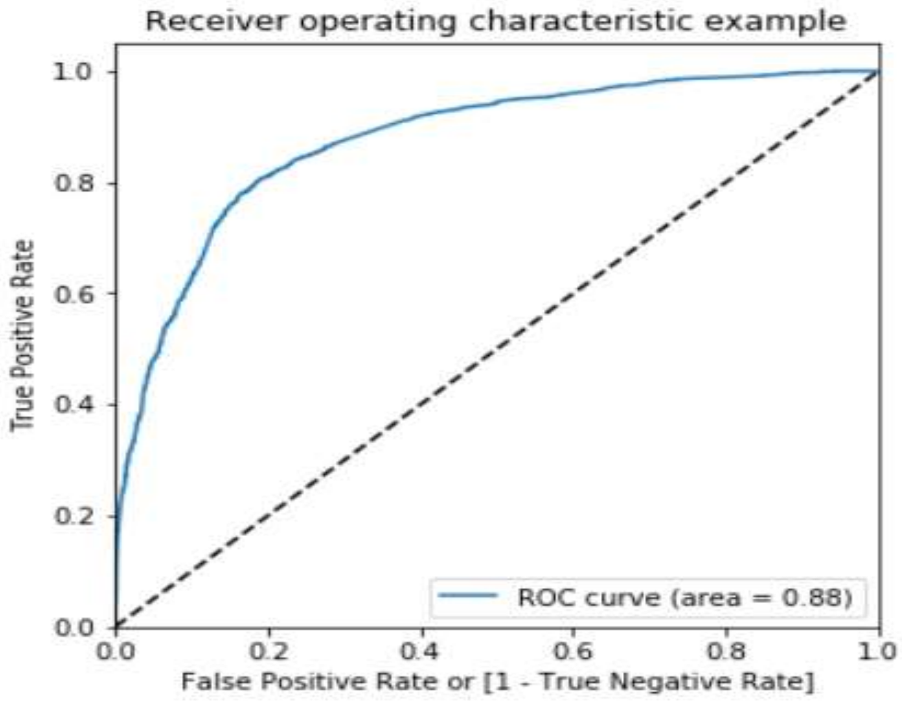
Positive predictive value=78.82%

Negative predictive value=80.27%

Here, Sensitivity value is less so to improve the value of Sensitivity we used ROC Curve.

- An ROC curve demonstrates several things: It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

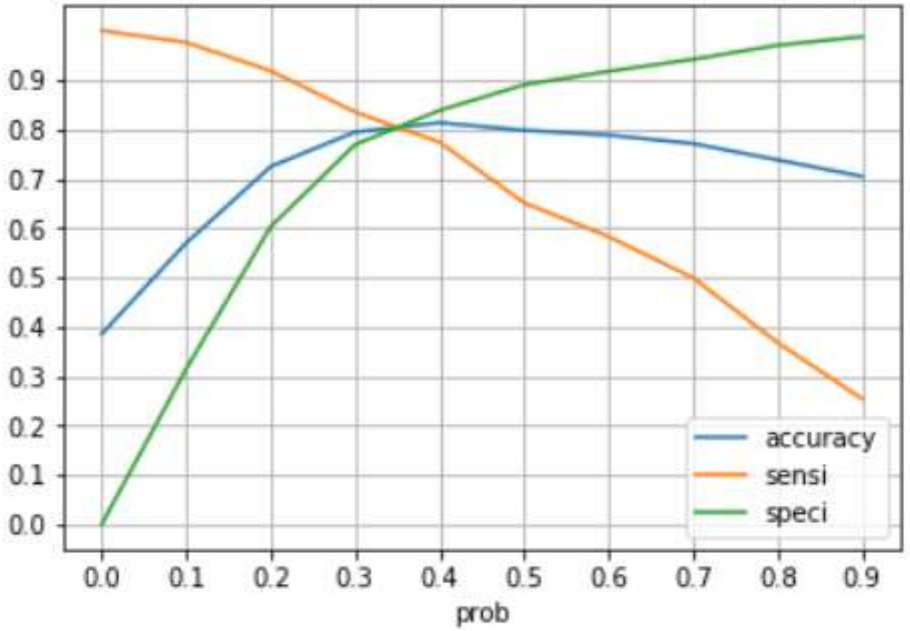
ROC CURVE



Finding Optimal Cutoff Point

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.

ACCURACY SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITIES



From the curve above, 0.35 is the optimum point to take it as a cutoff probability

Confusion matrix

[[3211, 760],

 [482, 2010]]

FINAL MODEL RESULTS

- **METRICS**

Accuracy=80.78%

Sensitivity=80.66%

Specificity=80.86%

False positive rate=19.14%

Positive predictive value =72.56%

Negative predictive value=86.95%

PREDICTIONS ON THE TEST SET

Accuracy : 81.51%

Sensitivity : 79.75%

Specificity : 82.61%

False Positive Rate :17.38%

Positive predictive value :74.19%

Negative predictive value :86.69%

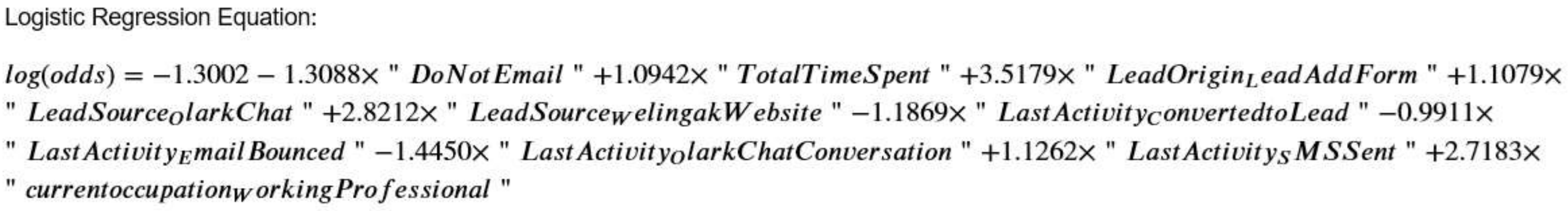
CONCLUSIONS AND SUGGESTIONS

➢Three most significant features are:

1. Lead Origin_Lead Add Form
2. Lead Source_Welingak Website
3. current occupation_Working Professional

➢ The X education should consider the above three features as driving features among all the other features as given by the equation below.

➢Recommended candidates have high conversion rate of 71.43% against about 60% when not recommended. But conclusive statement can not be made as it has very less data.



Logistic Regression Equation:

$$log(odds) = -1.3002 - 1.3088\times \text{"} DoNotEmail \text{"} + 1.0942\times \text{"} TotalTimeSpent \text{"} + 3.5179\times \text{"} LeadOrigin_LeadAddForm \text{"} + 1.1079\times \text{"} LeadSource_OlarkChat \text{"} + 2.8212\times \text{"} LeadSource_WelingakWebsite \text{"} - 1.1869\times \text{"} LastActivity_ConvertedtoLead \text{"} - 0.9911\times \text{"} LastActivity_EmailBounced \text{"} - 1.4450\times \text{"} LastActivity_OlarkChatConversation \text{"} + 1.1262\times \text{"} LastActivity_SMSSent \text{"} + 2.7183\times \text{"} currentoccupation_WorkingProfessional \text{"}$$