Lead Scoring Case Study

Results of analysis

Agenda

- ✓ what's the issue ?
- ✓ What did we do?
- ✓ What did we learn?
- ✓ Visualizations

X- Education is struggling to convert leads!!!

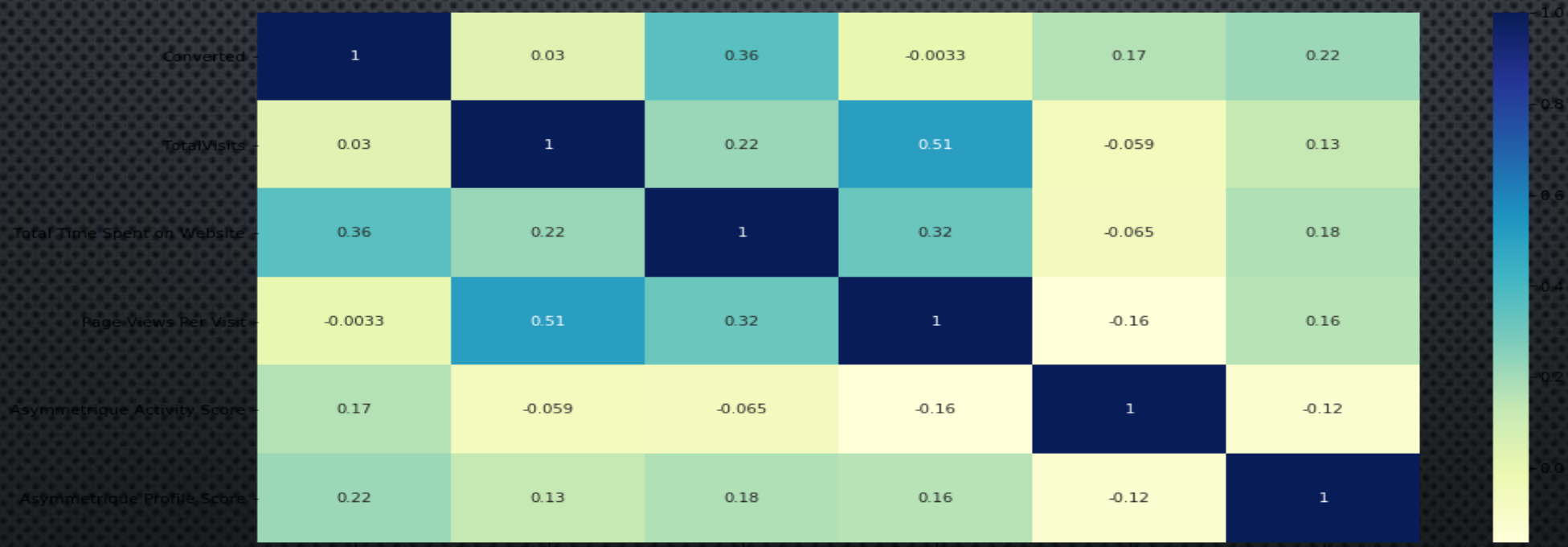
- Its an organization that specializes in providing professional courses, online.
- People come on their website and browse through for what they may like , these people are potential leads.
- Now, their lead conversion rate is at a dismal 30% and they want to raise it substantially.

THIS PROBLEM CAN BE SOLVED USING LOGISTIC REGRESSION

- THE DATA SET PROVIDED BY X EDUCATION HAS PREVIOUS DATA ON LEAD CONVERSION .
- THERE ARE OVER 9000 DATA POINTS HAVING 37 VARIABLES
- WE NEED TO FILTER OUT FEW VARIABLES IN ORDER TO RUN A GOOD LOGISTIC REGRESSION

Data exploration and manipulation

- There is a decent amount of correlation of converted i.e. our target variable with the other numerical features present in the data set. See the heatmap below



Data exploration and manipulation…

- On exploring- found features that doesn't make sense in predicting a good lead, so pruned the data set by dropping these-
  - Lead number, country, city, tag
- Variables with more then 30% nulls- wouldn't contribute in predicting, so dropped these as well-
  - Lead quality, asymmetric activity score, asymmetric profile score etc. full list is present in the attached python notebook.
- Lots of columns had values as "Select"- we decided to replace it with np.nan. Its as good as null really!

THE FINAL SET OF FEATURES

- Our dataset had 76 variables after removing non essential and null columns and converting yes/no columns to binary (0/1) as well as treating categorical variables with the help of dummies.
- We did use RFE with 13 variables to be used in our logistic regression model on an exploratory basis. We would update the feature count later depending on the outcome.

MODEL EVALUATION RESULTS

- ON TRAIN SET (70% OF THE ORIGINAL DATA WAS USED TO TRAIN THE MODEL)

Accuracy	**81.27%**
Sensitivity	**70.19%**
Specificity	**88%**
False Positive Rate	**11.89%**
Positive Predictive Value	**78.4%**
Negative Predictive Value	**82.75**

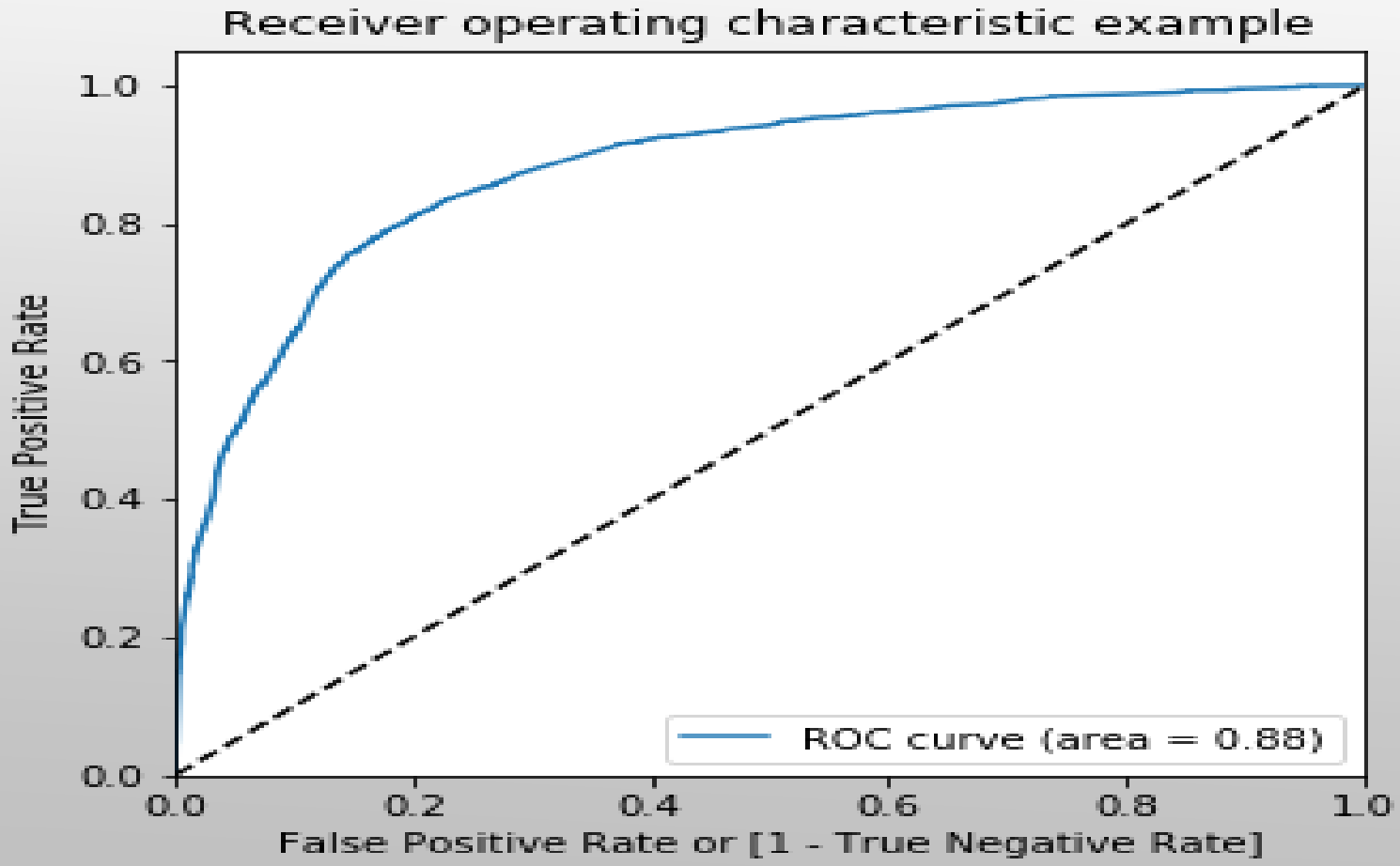
MODEL EVALUATION RESULTS

- ON TEST SET (30% OF THE ORIGINAL DATA WAS USED TO TRAIN THE MODEL)

Accuracy	**81.13%**
Sensitivity	**81.36%**
Specificity	**80.97%**
False Positive Rate	**19%**
Positive Predictive Value	**73.63%**
Negative Predictive Value	**86.93%**

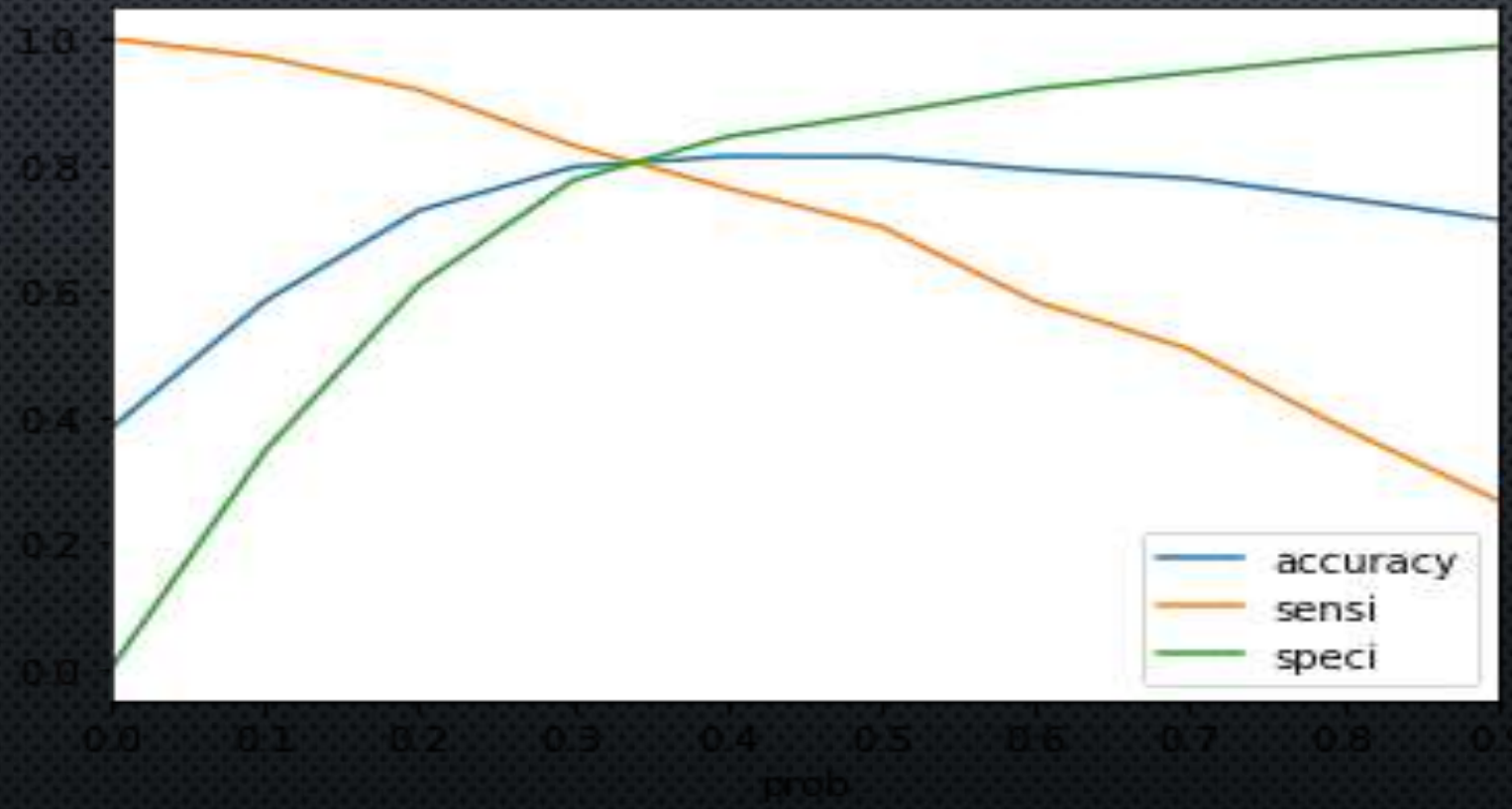
Important visuals

- ROC curve



Important visuals

- Selecting optimum cutoff to label conversion
- From the graph, 0.33 is the optimal Probability Cutoff.



CONCLUSIONS

TOTAL TIME SPENT ON THE WEBSITE IS A SIGNIFICANT INDICATOR TO IDENTIFY A GOOD SOLID LEAD

SOME LEAD SOURCES SUCH AS OLARK CHAT & WELINGAK WEBSITE ARE EFFECTIVE IN SENDING CONVERTIBLE LEADS. THEY SHOULD BE GIVEN DEDICATED COORDINATORS TO ENGAGE EFFECTIVELY WITH POTENTIAL CLIENTS

CUSTOMER IS MORE LIKELY TO CONVERT IF THEY HAVE FILLED THE ADD FORM ON THE PORTAL. IT'S A SIGNIFICANT LEAD ORIGIN AND SHOULD BE KEPT A VIGIL ON.

HOUSEWIVES AND WORKING PROFESSIONALS ARE MORE AMENABLE TO JOIN THE COURSES. THEREFOR LIAISE WITH THESE DEMOGRAPHICS, AGGRESSIVELY.

TALK TO THE PROSPECTS. TALKING HELPS IN CONVINCING THE POTENTIAL LEADS