X Education
Lead Scoring Case Study

Group:

Balakrishna Gadiyar

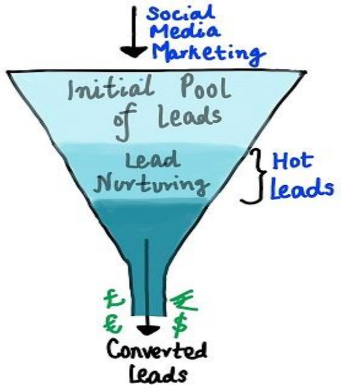
Deepak Padhan

Bishnu Agrawal

Abstract :

X Education sells online courses to industry professionals. Professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Professionals fill up a form for the course which contain their email address or phone number, are classified to be a lead. Moreover, the company also gets leads through past referrals.

Sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.



Problem Area -

The typical lead conversion rate at X education is around 30%, which is very low.

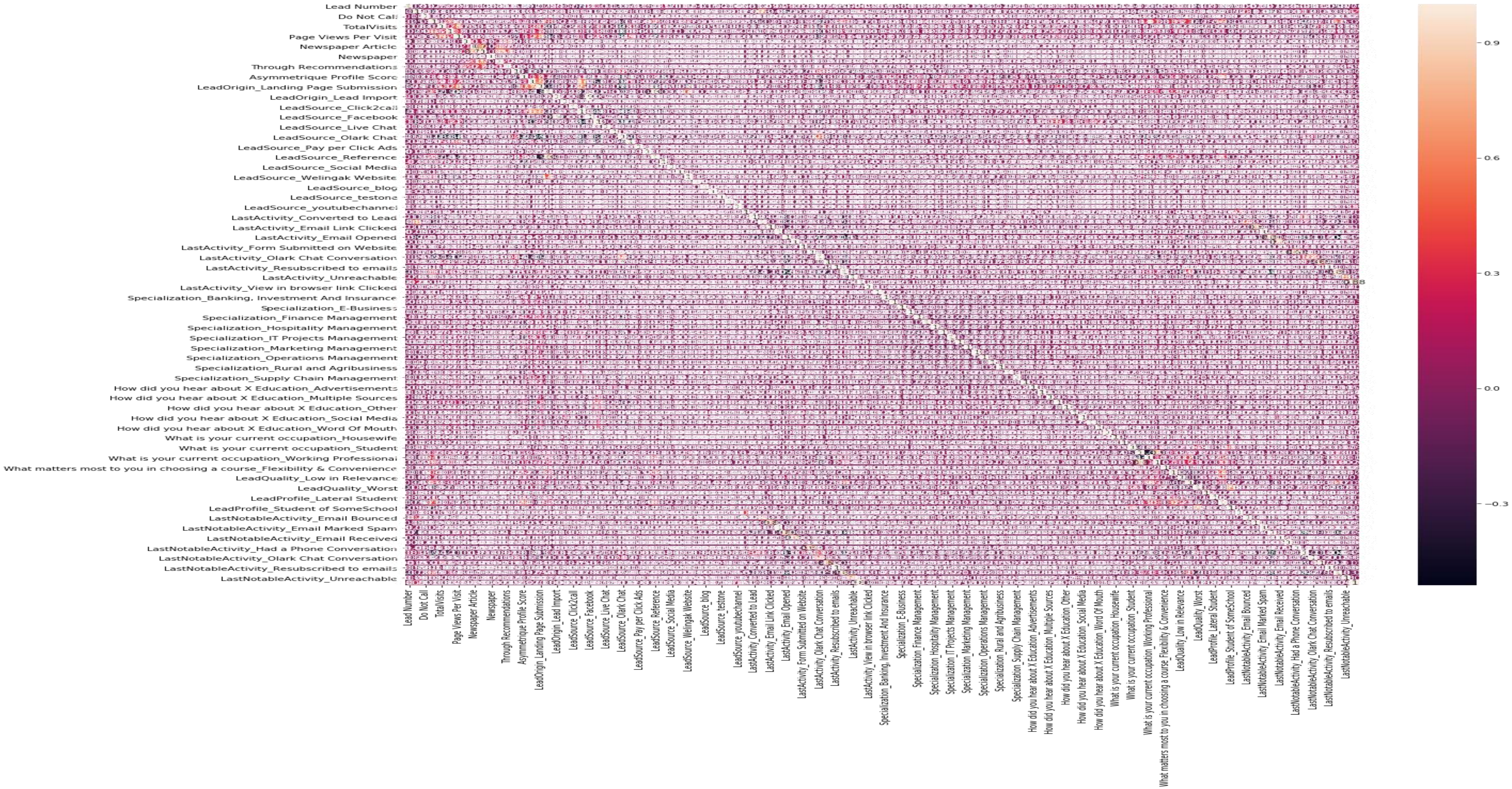
Objectives :

- Build a model to using available lead data, assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and lower score with lower convrsion chance.
- Business Recommendation - Target lead conversion rate to be around 80%.
- Additional asks - Provide answers to business questions on lead and its features

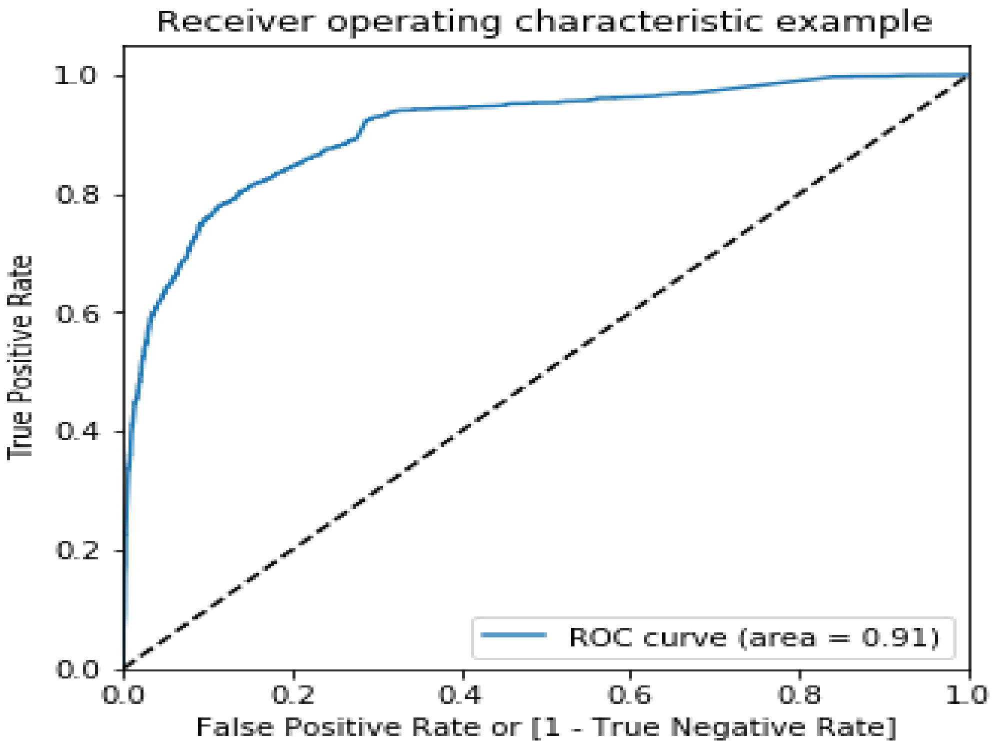
Technical Approach

- Understand the business process and problem
- Identify the key features in given lead data.
- Understand the data trends and record observations, decide the precision model to be logistic regression
- Perform EDA –
    - Import, Understand data frame
    - Treat null values
    - Treat missing/wrong values
    - Drop features with no variations or which doesn’t contribute value while building model
- Perform binary mapping for features with logical values, convert them to numbers
- Generate dummy variables for columns with more than 2 values, convert them to columns, drop original features
- Perform outlier analysis for the numeric columns and bring them in range.
- Prepare for training – Split the data into training and testing groups
- Scale the data before conducting RFE analysis
- Plot correlation map and remove highly correlated features from train set. Repeat the process until a stable data frame is obtained
- Build regression model and fit the train data and observe the parameters – p-value, coeffs,
- Select the right features through RFE recommended ranking
- Using statsmodel, asses the model and observe parameters,
- Generate predicted values and review the trends, compare actual vs predicted
- Generate confusion matrix, review model accuracy, turns out to be 80%
- Verify VIFs, check how the feature variables are correlated with each other
- Calculate sensitivity and specificity. Also calculate FPR, PPV, NPV
- Plot ROC curve and find optimal cut off point, calculate accuracy, sensitivity, probability and specificity for cut offs
- Calculate precision and recall values, identify thresholds
- Perform prediction on test sets
- Calculate accuracy score, sensitivity and specificity values.

Correlations among feature Variables :

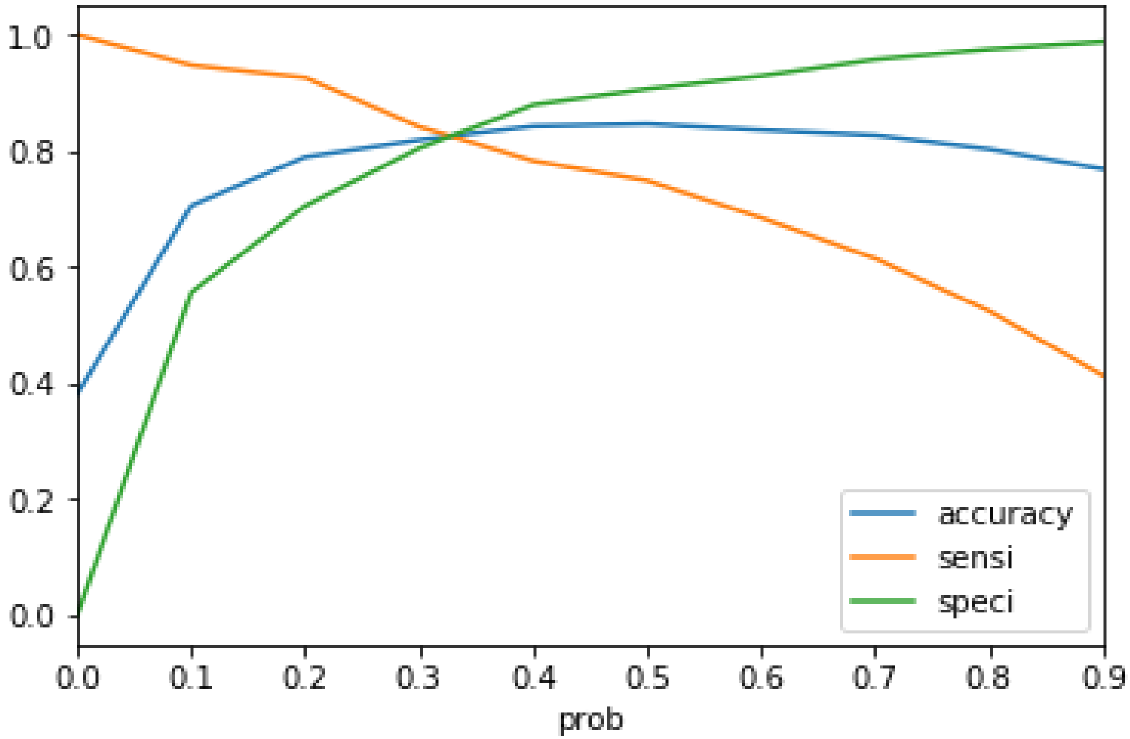


Logistic Regression - Visualizations

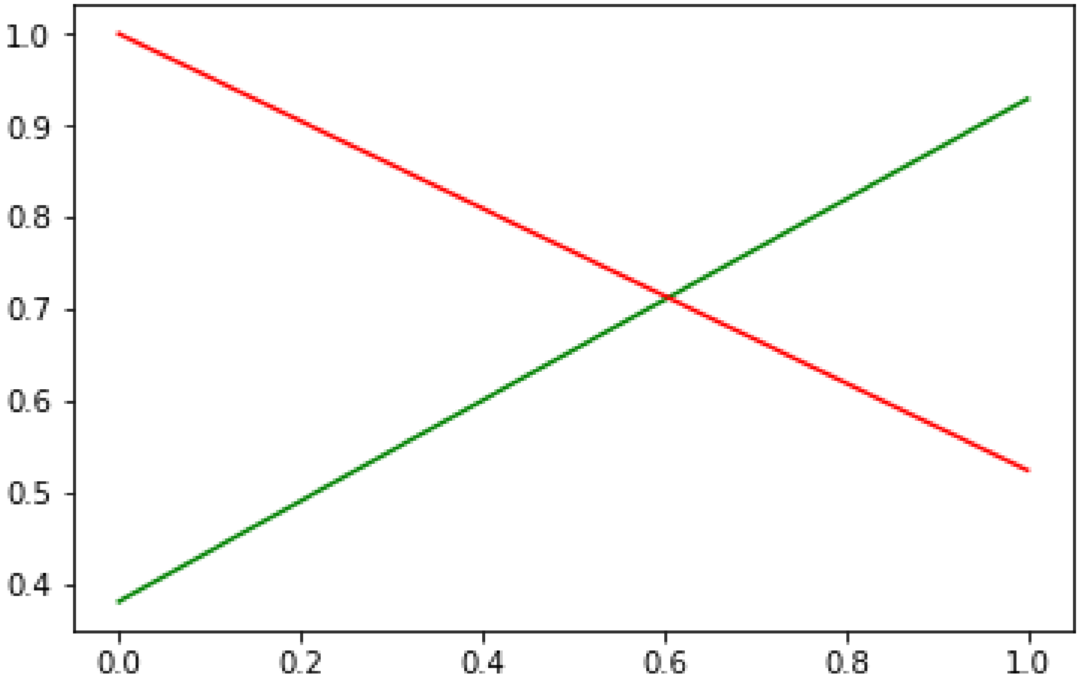


ROC curve demonstrates the tradeoff between sensitivity and specificity. Here the curve follows more closer the left-hand border and then the top border of the ROC space.

Logistic Regression - Visualizations



Accuracy Vs Sensitivity Vs Specificity



Threshold Curve

Conclusion

- Logistics Regression Model is suggested and developed for Lead scoring.
- Developed the Model which predicts the Lead with 80% accuracy. The model also has sensitivity of 52% and Septicity of 98% which is a symbol of a good model.
- X Education sales team can focus more on leads who has opted for emails
- Sales team can also focus on leads who are spending more time on website
- Sales team should also have a close look on leads origin features and target them accordingly