LEAD SCORING CASE STUDY

ROHAN GHOGARE

JYOTHISH J

LEAD SCORING CASE STUDY

Introduction:

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Statement:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted

Data Provided :

- Two Spread sheets containing all the information of the customer when they inquire or visited website. The other spread sheet gives information about what variable have gathered

SOLUTION APPROACH

Approach for case study:

- Load/ Read all data provided and identify the data structure, information about data like data types, missing values.
- Data Cleaning Techniques :
  - ❖ Investigate the missing values and impute that with suitable method viz, mean, mode, median etcc
  - ❖ Check outliers in data and take suitable action
  - ❖ Check data imbalance and treat that
- Data analysis
  - Cleaned data to investigate the relationship of variables with that of TARGET variable
  - Get the train and test set by splitting the cleaned data set
  - Use Recursive Feature Elimination (RFE) methodology from sklearn to get variable which are most effective
  - Get the matrices (TP,FP, TN and FN) and then calculate the necessary numbers such as sensitivity, specifivity, precision and recall at optimum probability value. This Optimum number can be fetch by building RoC curve
  - The above process need to done for both train and test data set
  - Finally Calculate the lead score at optimum probability value (lead score = predicted probability x100)

RESULTS

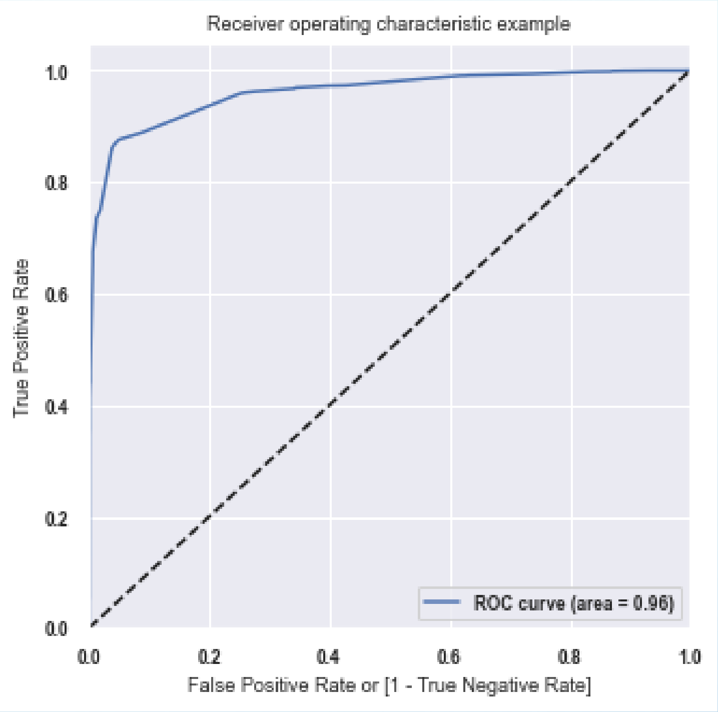
1. **MODEL 2 SHOWS THE OPTIMUM VIF FOR DIFFERENT VARIABLE AND THAT IS CHOSEN AS GOOD MODEL**

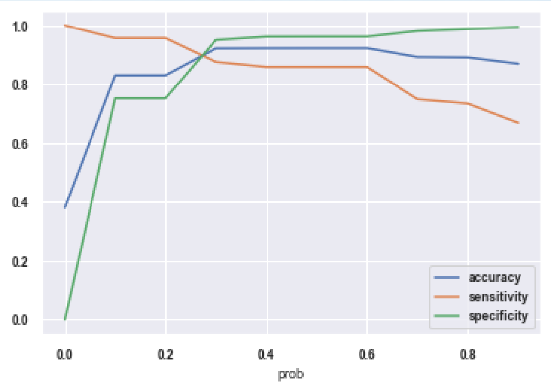
2. **ROC CURVE**

- The area under the curve is ~96% which is good model for prediction.
- Now based on this get the insight for specific parameters viz sensitivity, specificity, precision and recall

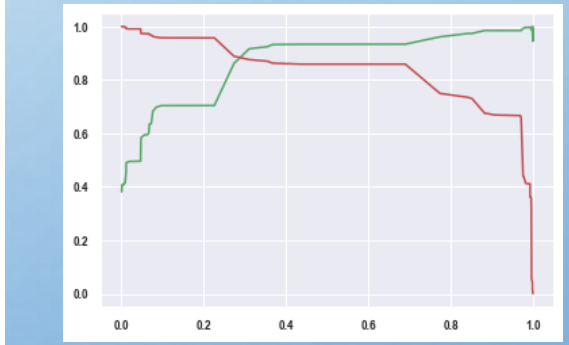
3. **OPTIMUM CURVE**

- The accuracy, sensitivity and specificity lines are intersecting at ~0.25 probability. So, we will proceed with this value.





Accuracy, sensitivity and specificity



Precision and recall

RESULTS

Sr no	Accuracy (%)	Sensitivity(%)	Specificity (%)	Precision (%)	Recall (%)
Train data	92	86	96	93	86
Test Data	91	90	91	87	90

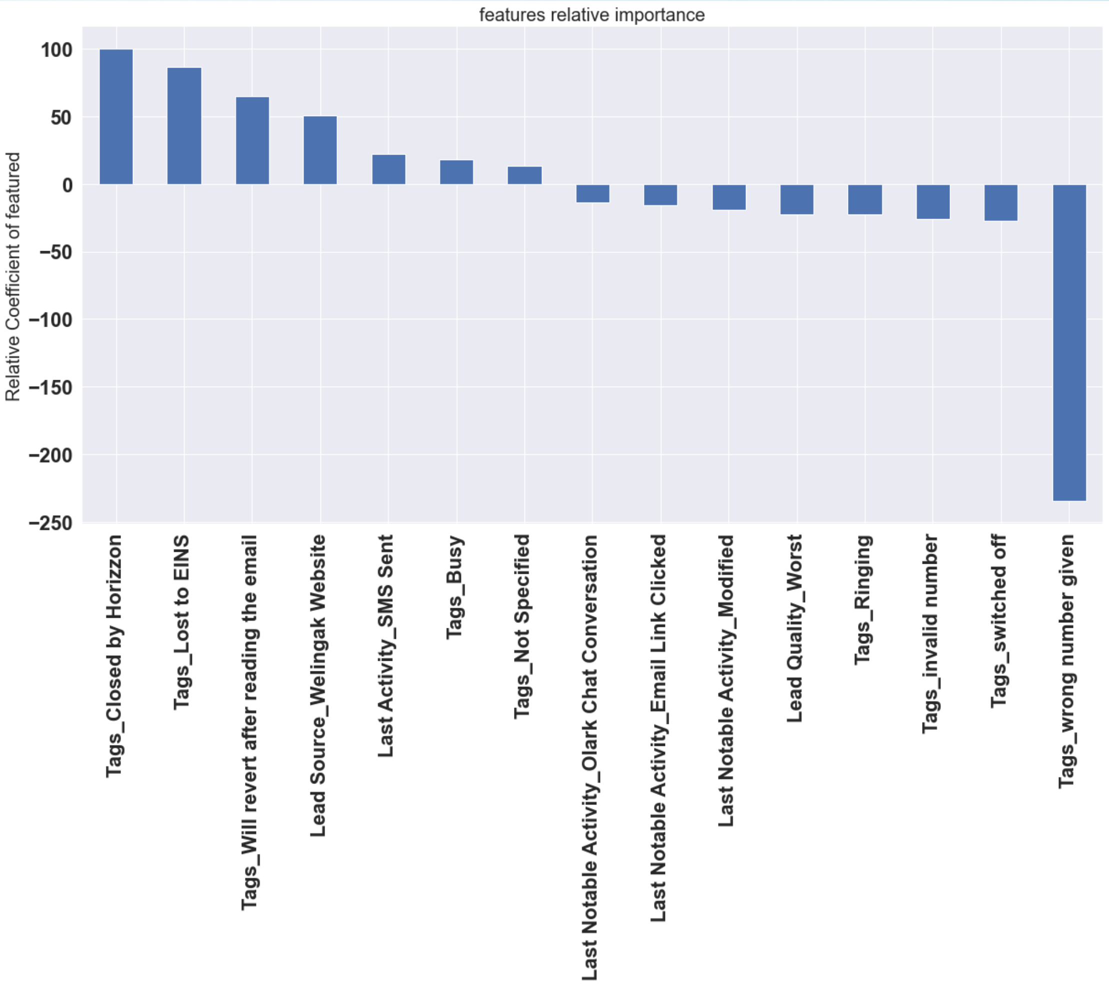
RESULTS : MAIN FEATURES

Top 5 variables which can fetch the lead:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tag_We will revert after reading the email
- Lead Source_Welingak Website

5 variables which need more attention for converting to lead:

- Tags_wrong number given
- Tag_switched off
- Tag_invalid number
- Tag ringing
- Lead Quality worse



CONCLUSION

- Some top leads for conversion need follow up
- Some features which relatively low importance need more focused from business
- Some of leads which talks about do not email, MIGHT BE or worse can be given low priority

THANK YOU