LEAD DATA CASE STUDY

ABHIJITH ANCHAN S

Modeling Demand for an Education company X content

About Education X Business

- Education company X sells online courses to industry profiles.
- The company Sales team effort is channeled to all the leads and finds it not to be very efficient.
- The company collects information about customers either by referrals or customers who have filled out the forms while consuming content online.
- Company wants to have new business plan assessing demand and the variables affecting it to focus on target customers who are highly likely to buy the course.

Goals of the Case Study:

- Build a model to assign a lead score between 0 and 100 to each of the leads.
- A higher score would mean that the lead is hot, i.e., Customers with a higher lead score have a higher conversion chance and customers with a lower lead score have a lower conversion chance

Approach followed

1.Load the necessary Python libraries such as pandas, numpy, matplotlib, seaborn, etc.

2.Load the loan application data provided in the case study into a pandas data frame.

3.Explore the structure of the data, such as the number of rows and columns, data types, missing values, etc.

4.Identify and handle missing data appropriately, either by removing columns with too many missing values or by imputing missing values with appropriate methods such as mean, median, or mode. Drop columns with more than 40% NA values. Read the Column description and eliminate some of the unwanted columns like previous id, flags, and name of accompanying person and their particulars. Regional default information.

5.Check for outliers in the data and decide whether to remove them or not based on the business context and the impact of outliers on the analysis.

6.Build a model with all columns as the baseline

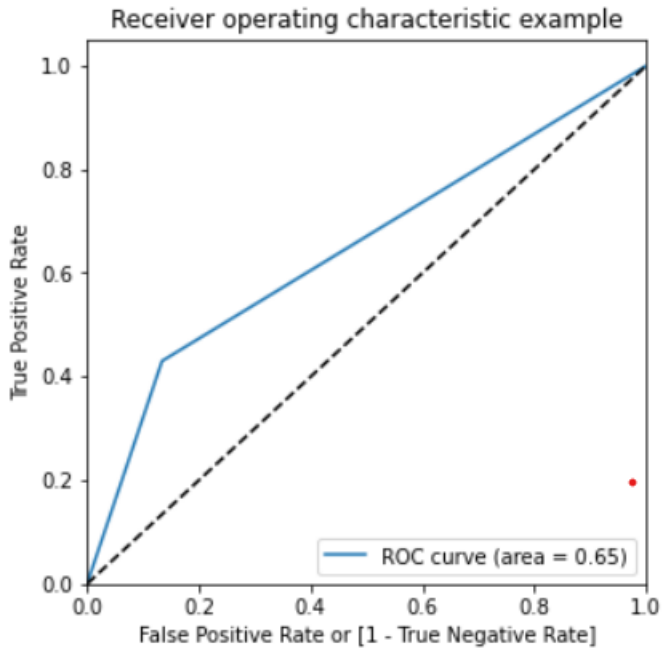
7.Do feature elimination using RFE

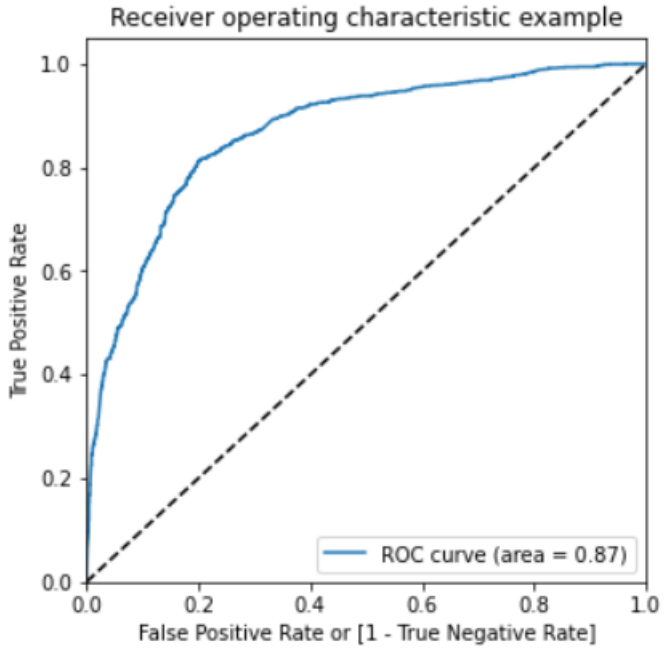
8.Rebuild models and assess to arrive at the best model based on AUC and recall

9.Assign the Score by calculating it based on probability and multiplied by 100.

10. Threshold to be selected based on higher recall as in this case, Recall parameters are critical,

Model selection using ROC curve





Curve 1 for Baseline model and Curve 2 for best fit model which has a higher area

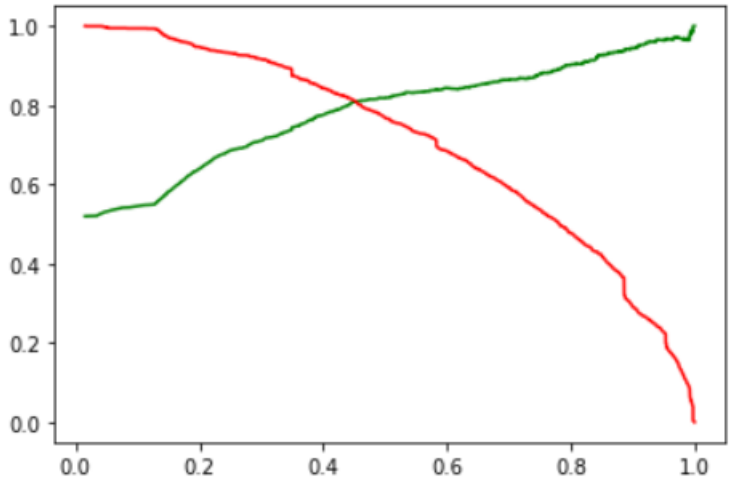
Threshold selection

Optimal Cutoff is observed at 0.43

We could also take a conscious call to have a different threshold to have a effective Recall increase.

Lead Scoring= probability*100

Customers with lead score >43 is high scoring customer most likely to buy



Overall model performs equally well both on train and test data set.

	Train dataset	Test dataset
•Sensitivity	0.825	0.807
•Specificity	0.77	0.75
•Precision	0.76	0.77
•Recall	0.81	0.80
•Accuracy	0.79	0.78

Most important variables to consider:

Do Not Email': If this variable is True or 1, then the log odds decrease by 1.39

'Total Time Spent on Website' log odd increase by 1.0514 for every unit of time spent on the website, which signifies interest in the content

Lead Origin Landing Page Submission- log odd decreases if a lead is collected from the landing page. That is a chance of entering information just to look at the content by the user

Lead Origin_Lead Add Form - if a customer lead is generated through the channel there is a high chance of being converted as log odds increase by 2.665

Last Activity_Converted to Lead and Last Activity_Email Bounced has a negative impact that is -1.4 and -2.4499.

Last Activity Sms sent. if this is the last activity then the log odds of converting increase by 0.9553

If its a working professional log odds of converted increase by 2.6884 compared to other occupation

'Last Notable Activity Email Bounced' and Last Notable Activity Unreachable' has a likelihood of conversion, so better try again

Recommendation

- It appears that customers with occupation are more likely to buy compared to others
- Customer who spend more time consuming contents are more likely to buy, keep them in touch .
- Leads generated through forms on landing page are least likely to purchase the course
- Do not contact Customers who have explicitly specified not be contacted over mail or call, as they are least likely to respon.
- Use key Salesperson to score on leads having scores 43 % to 90 % and others to 90% and above.
- Create more webinars as we see total time spent ineracting with comtent more is a chance of buying, could reduce need of coldcalling.