LEAD SCORING CASE STUDY

VIKHYAT NEGI

BUSINESS PROBLEM STATEMENT

Business Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many 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. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Goal:

To identify the features that contributes to predict Lead Conversion.

Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

OVERALL APPROACH

1. Importing Data.
2. Inspecting the Data Frame.
3. Data Preparation (Encoding Categorical Variables, Handling Null Values).
4. EDA.
5. Dummy Variable Creation
6. Test-Train Split.
7. Feature Scaling.
8. Looking at Correlations.
9. Model Building (Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-values).
10. Build final model.
11. Model evaluation with different metrics Sensitivity, Specificity.

UNDERSTANDING THE DATA & BASIC DATA CLEANUP

There are 37 columns (30 categorical and 7 Numeric) and 9240 observations in the dataset.

Select is present as a class in different columns like:

- ➢ Specialisation
- ➢ How did you hear about X Education
- ➢ Lead Profile
- ➢ City

As Select is not a valid class, we can conclude that the Select might be the default value set in the form dropdown and if the user has not selected any option from the dropdown, then the value remained as Select. We replaced Select with NaN.

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 - These columns have no missing data and have only one unique value. So, these columns have no variance and not helpful for our EDA or model building, hence we dropped these columns.

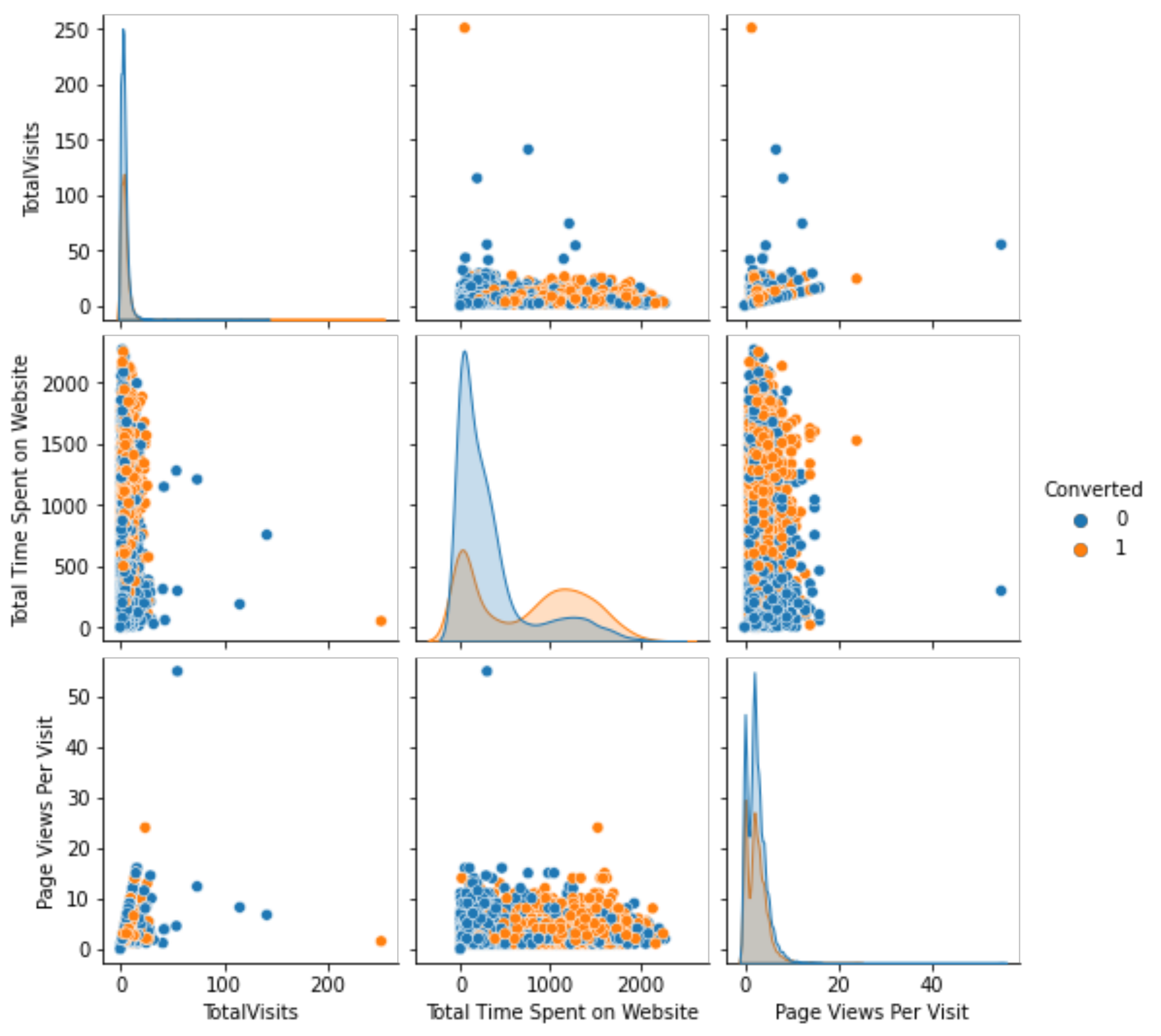
How did you hear about X Education, Lead Profile, Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Index , Asymmetrique Activity Score, Asymmetrique Profile Score – These columns have more than 40% missing value. So, we have dropped these columns from our EDA and model building.

There is no datapoint/ observation (rows) in our dataset having more than 70% missing values.

Performed missing value treatment using Business Understanding. For Specialisation and Occupation NaN values are replaced with a new category Not Disclosed.

We renamed What is your current occupation column to Occupation and What matters most to you in choosing a course to Reason_choosing for our convenience during EDA and Model building .

EDA



Median value of Total Time Spent on Website for converted Leads are considerably higher than the other group. Team should target those customers who are spending higher time on website. Those leads have higher odds of getting converted.

MODEL BUILDING

1. Recursive Feature Elimination (RFE) has been used to get top 15 features.

➢ Do Not Email : An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not. ➢ TotalVisits: The total number of visits made by the customer on the website.

➢ Total Time Spent on Website: The total time spent by the customer on the website.

➢ Page Views Per Visit: Average number of pages on the website viewed during the visits.

➢ Lead Origin_Landing Page Submission: Dummy variable for Landing Page category of the origin identifier with which the customer was identified to be a lead.

➢ Lead Origin_Other: Dummy variable for the Other category of the origin identifier with which the customer was identified to be a lead.

➢ Lead Source_Olark Chat: Dummy variable for the Olark Chat category of the source of the lead.

➢ Lead Source_Other Sources: Dummy variable for the Other category (other than Google, Direct Traffic, Olark Chat, Organic Search, Reference) of the source of the lead. ➢ Country_Other Countries: Dummy variable for the Other category (other than India and United States) of the country of the customer.

➢ Specialization_Domain Specialization: Dummy variable for Domain Specialization bin of Specialization variable.

➢ Specialization_Management Specialization: Dummy variable for Management Specialization bin of Specialization variable. ➢ Occupation_Other: Dummy variable for 'Other' category of customer's occupation.

➢ Occupation_Student: Dummy variable for 'Student' category of customer's occupation.

➢ Occupation_Unemployed: Dummy variable for 'Unemployed' category of customer's occupation.

➢ Occupation_Working Professional: Dummy variable for 'Working Professional' category of customer's occupation.

➢ City_Tier II Cities: Dummy variable for 'Tier II Cities' category of customer's city.

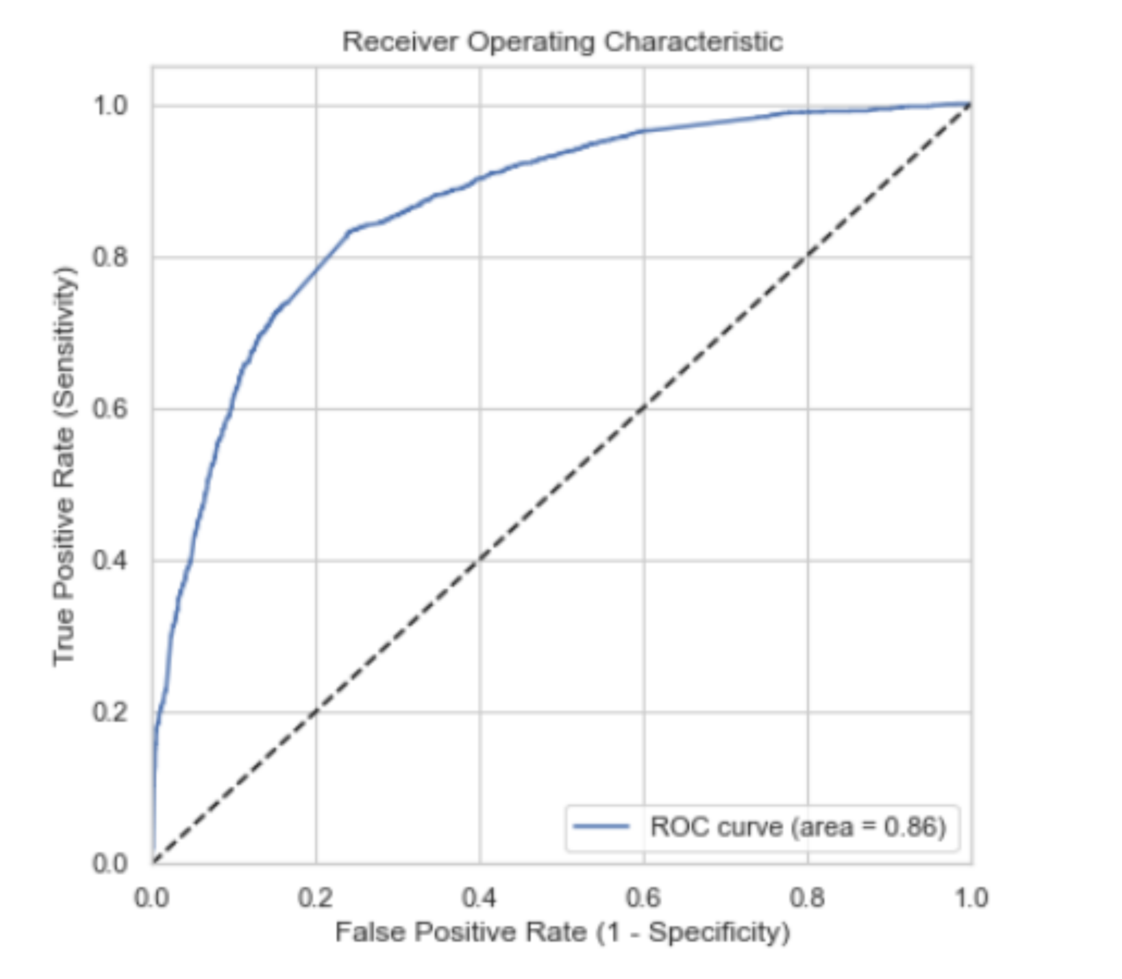
2. We built first Logistic Regression model using GLM (Generalised Linear Model) in stats models with these 15 features.

3. Then manually fine tuned the model to get statistically significant features (by checking the p-values) and removed multicollinearity (By checking Variance Inflation Factors) simultaneously. Accepted p-value is lower than .05 and accepted VIF is lower than 5.

4. p-values of all beta coefficients and VIFs have been checked and identified feature has been removed in next model building. We have also checked Overall model accuracy and Confusion Matrix after each new model, to understand how the new model is performing in compared to the previous one.

PREDICTION & MODEL EVALUATION : (ON TRAINING DATA - CUTOFF .5)

To predict the probability for all the observations in out training dataset and then used .5 as probability cut off to calculate our target variable 'Converted'. So, if probability is > .5 then 'Converted'= 1 (Yes) otherwise 0 (No).

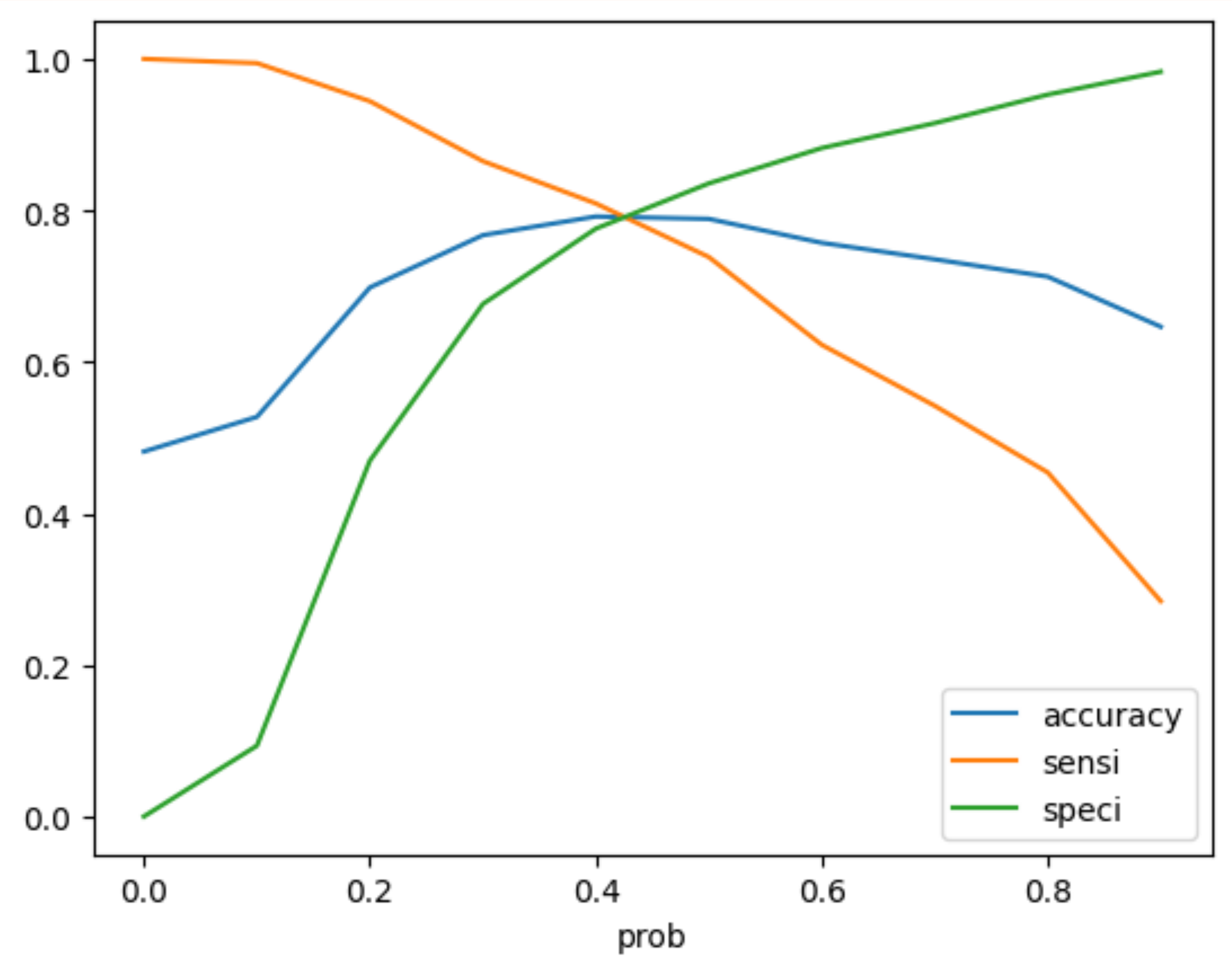


The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model.

By doing trade-off between Sensitivity-Specificity optimal probability cut-off value has been calculated.

FINDING OPTIMAL PROBABILITY CUTOFF & EVALUATING ON TRAIN DATA

In above plot, it’s visible that 0.42 is the optimal point to set as cutoff probability for our model.



Observations:

Sensitivity of our model has been increased without any significant reduction in overall accuracy. New Specificity is also in well accepted range.

PREDICTION & GENERATING LEAD SCORE (BUSINESS REQUIREMENT)

We calculated the probability on Test dataset and used cutoff =.42 to predict the Conversion_Prob (0,1). As per business requirement we have created a column Lead Score (between 0 to 100) of the leads. A higher score means hot lead (most likely to convert), lower score implies cold lead (mostly not get converted). We have multiplied the probability (Conversion_Prob) with 100 to generate the Lead Score.

CONCLUSION AND RECOMMENDATIONS

As per business requirement Lead Score (between 0 to 100) of the leads have been calculated by using this Logistic Regression model. A higher score means hot lead (most likely to convert), lower score implies cold lead (mostly not get converted).

This Lead Score would help to identify the hot leads faster and efficiently, that would result in decrease in lead conversion time and increase in lead conversion rate. Leads should be sorted in descending order according to their Lead Scores.

Phone calls or contact should be made to the leads having higher Lead Score first. Some special attentions should be provided to these hot leads (may be assigning a dedicated support SPOC for a small batch of hot leads that have higher Lead Scores), as there is a very high chance of Lead conversion.

Leads having medium Lead Score are also potentially good candidates for Lead conversion. They also should be contacted, and right questions should be asked, so that business can understand their requirements and problem areas and can take necessary actions. Few to mention: some changes in existing courses, introducing new courses, some change in class schedules, introducing easy financial options for fees etc. may help to successfully convert these leads.

Cold Leads should be contacted after business gets very good Conversion rate with the leads having High and Medium Lead scores. As chance of conversion is very less here, they could be part

THANK YOU