Lead score case study

By: Aastha

PROBLEM STATEMENT

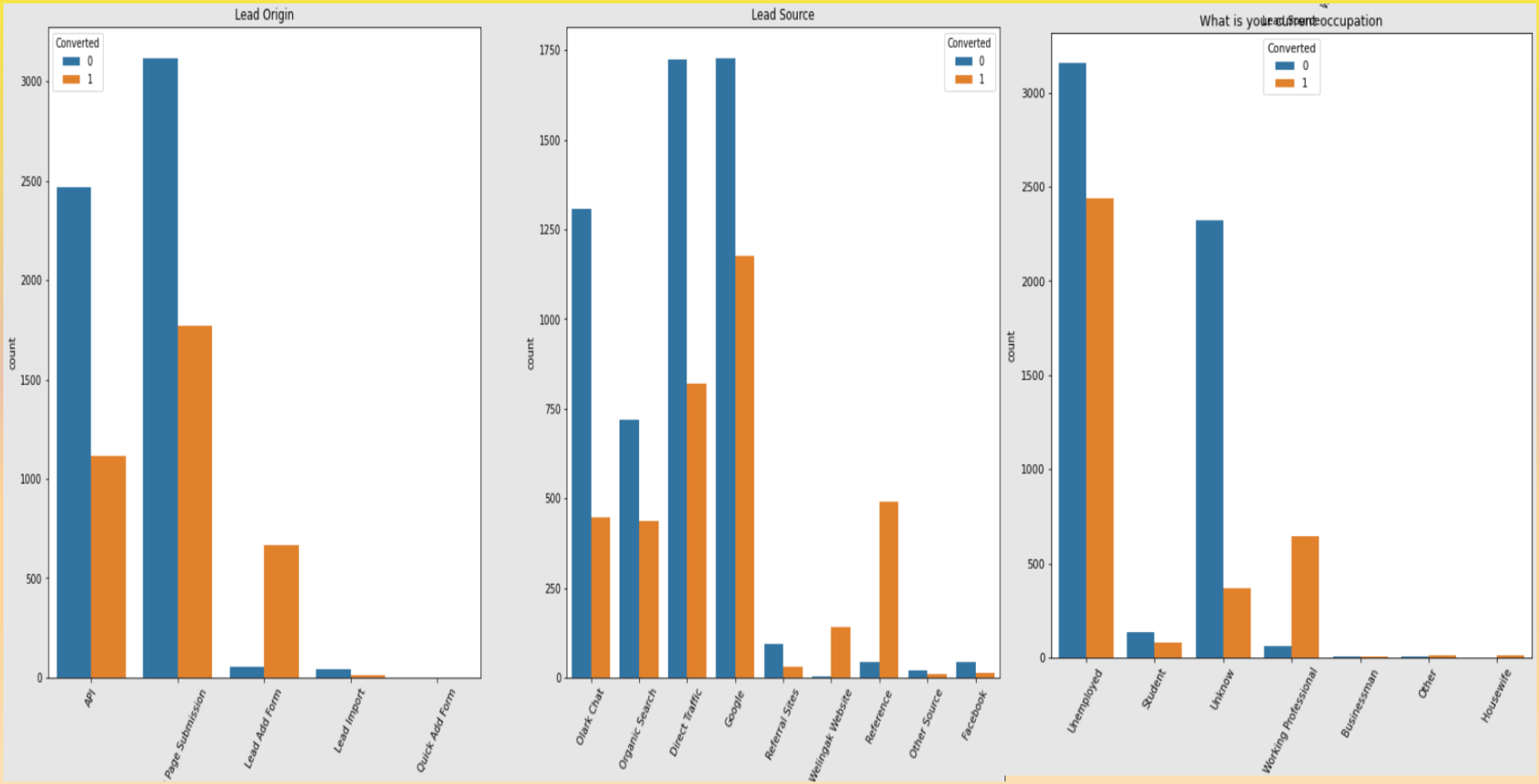
- An education company named X Education sells online courses to industry professionals.
- Company ask people fill up a form providing their email address or phone number, they are classified to be a lead. The typical lead conversion rate at X education is around 30%.
- The company wants to build a model where model 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%.

APPROACH TAKEN

- Data Cleaning
- Perform EDA
- Train test split and Scaling
- Model Building
- Model Evaluation
- Prediction on test set

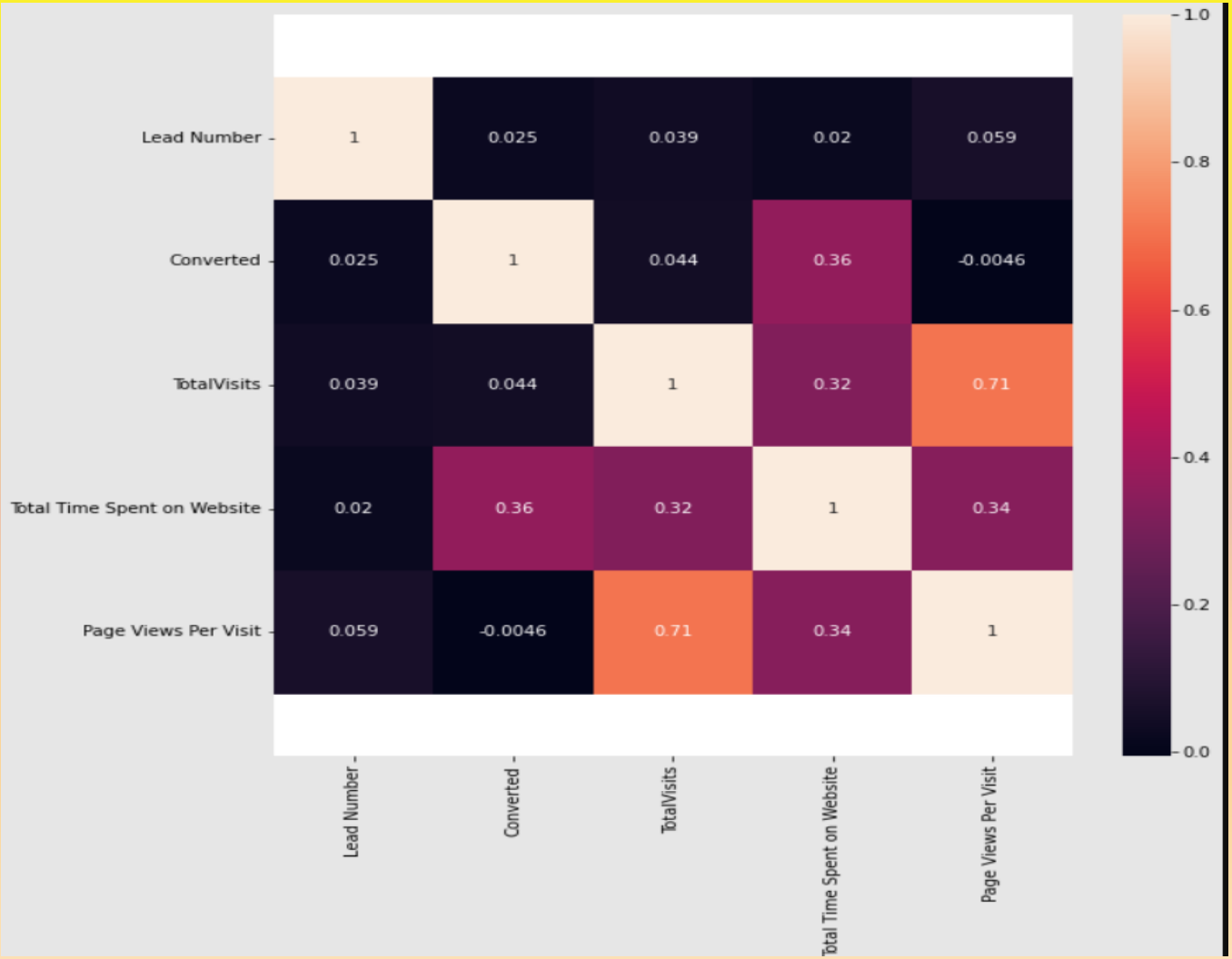
EDA

CATEGORICAL VARIABLE ANALYSIS WITH RESPECT TO TARGET VARIABLE:



NUMERICAL VARIABLE ANALYSIS

CORRELATION AMONG NUMERICAL VARIABLE:



ANALYSIS:

- We know the percentage of leads that are not converted is higher, each variable that is not converted has a high value. However, some categories in that variable, though, are succeeding in producing lead-positive results.
- Lead Add Form has a higher number of positive leads despite the low quantity of leads coming from this source.
- Reference and website like Welingak are performing exceptionally well as lead sources; their conversion rates are higher than those of competitors. Additionally, more people join with reference, which is quite natural. This might make enrolling in any course more assured.
- Compared to other occupations, working professionals appear to convert more frequently. Also, unemployed as well seems to show higher conversion rate.
- It appears that the lead conversion rate for specialization is close to 50% across all industries. As a whole, we can state that some of these categories would be helpful.

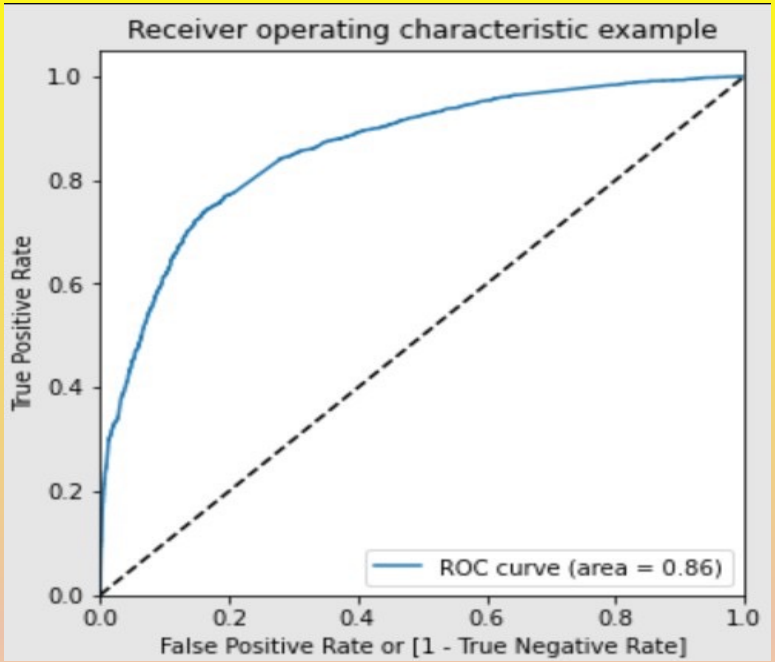
DATA PREPARATION:

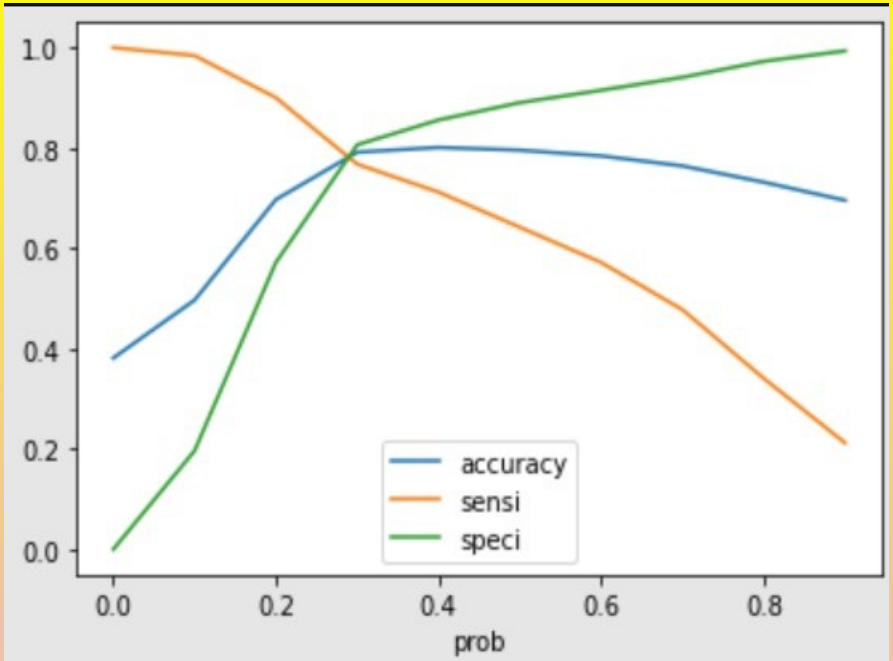
- Binary (Yes/No) variables are transformed to 1/0. ***Dummy variables*** were created for each categorical column. The original columns for which the dummies were created were dropped.
- ***Performed Test Train Split***. Two data frames, X and Y, were created. X is assigned with feature variables and Y with response/target variable i.e. Converted. The data has now been split into train values (70%) and test values (30%).
- ***Scaling***; All the three numerical columns from the train set are scaled using Standard scaler technique. TotalVisits, Total Time Spent on Website,Page Views Per Visit were on different scale value i.e. binary scale, so used rescaling to get the coefficients comparable with other variable coefficients.

MODEL BUILDING:

- We used logistic regression model train dataset. Feature selection is done by using RFE technique with 15 variables as output.
- Variables with high p values were dropped to remove unnecessary complexity in the model.
- Predict the value on the train set once we see that the p value is less than 0.05.
- We obtained an overall accuracy of 80% after importing metrics and checking the confusion matrix.

ROC CURVE:





- ROC curve shows the tradeoffs between sensitivity and specificity.
- Optimal point after plotting accuracy sensitivity and specificity is around 0.28, we took 0.3 as cutoff probability to build a model with good sensitivity.
- For Test Set, we obtained 79% overall accuracy and 77% sensitivity, which is good and in line with train set with overall accuracy of around 80%.

LEAD SCORE:

- Finally, we combined two train and test datasets that already contained a generated converted probability column.
- Added a column called Lead Score that denotes the likelihood that a customer will convert. Then multiplied probability by 100, the Lead Score variable, for the sales team's benefit.
- Final lead score:

	Lead Number	Lead Score
3115	2656	99.96
7014	3478	99.96
4891	8074	99.95
7213	6383	99.92
6187	7579	99.92

THANKYOU