LEAD SCORING CASE STUDY

VIJAY TEJA V

DRUVARAJ B

PROBLEM STATEMENT

An 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%.

SOLUTION METHODOLOGY

- Data cleaning and Data manipulation.
  1. Select has been converted to np.nan/null values
  2. Columns having more 35% null values have been removed.
  3. Impute values which have more than 10% by NA
  4. Dropped highly skewed categorical columns and combined the categories by others having count <10%
  5. After the above steps, removed the rows which had null values resulting with 98.52% of original rows
  6. Outliers in numerical values are handled by soft capping by 99^{th} percentile
- Data Preparation
- Splitting the dataset into Train and Test
- Creation of the model
- Model Evaluation and getting optimal threshold
- Inferences

DATA PREPARATION

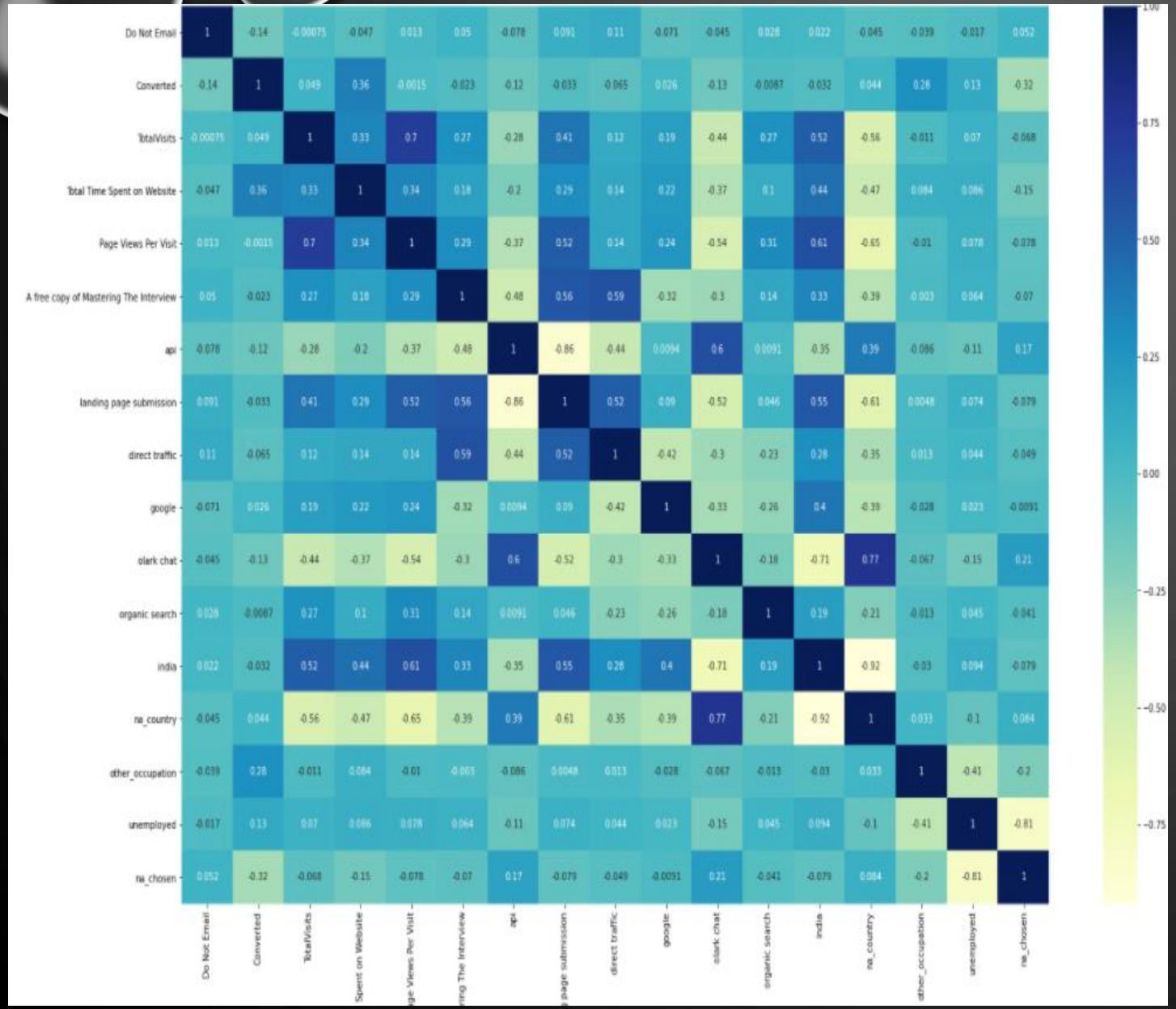
- Dummies for the categorical variables are created keeping the dropfirst as true
- The dummies created are concatenated to main dataset
- Lead Identifier and Categorical columns are removed as it is not required in the model creation
- After the above steps we are left with 9103 rows and 17 columns for the modelling.

TRAIN AND TEST DATASET

- Train and test data are split into 70 and 30 percent respectively from the main dataset.
- Normalized scaling is done on the train dataset.

```
In [43]: f_train,f_test = train_test_split(final, train_size=0.7, random_state=100)
         print(f_train.shape)
         print(f_test.shape)

         (6372, 17)
         (2731, 17)
```



CORRELATION ANALYSIS

- From the correlation plot, we could see that time spent on the website and other occupation are more postively correlated and na_chosen is more negatively corrleated with converted values.
- These variable contribute more to predicting the leads.

Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6363
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3007.6
Date:	Mon, 07 Dec 2020	Deviance:	6015.2
Time:	01:51:03	Pearson chi2:	6.57e+03
No. Iterations:	5		
Covariance Type:	nonrobust		

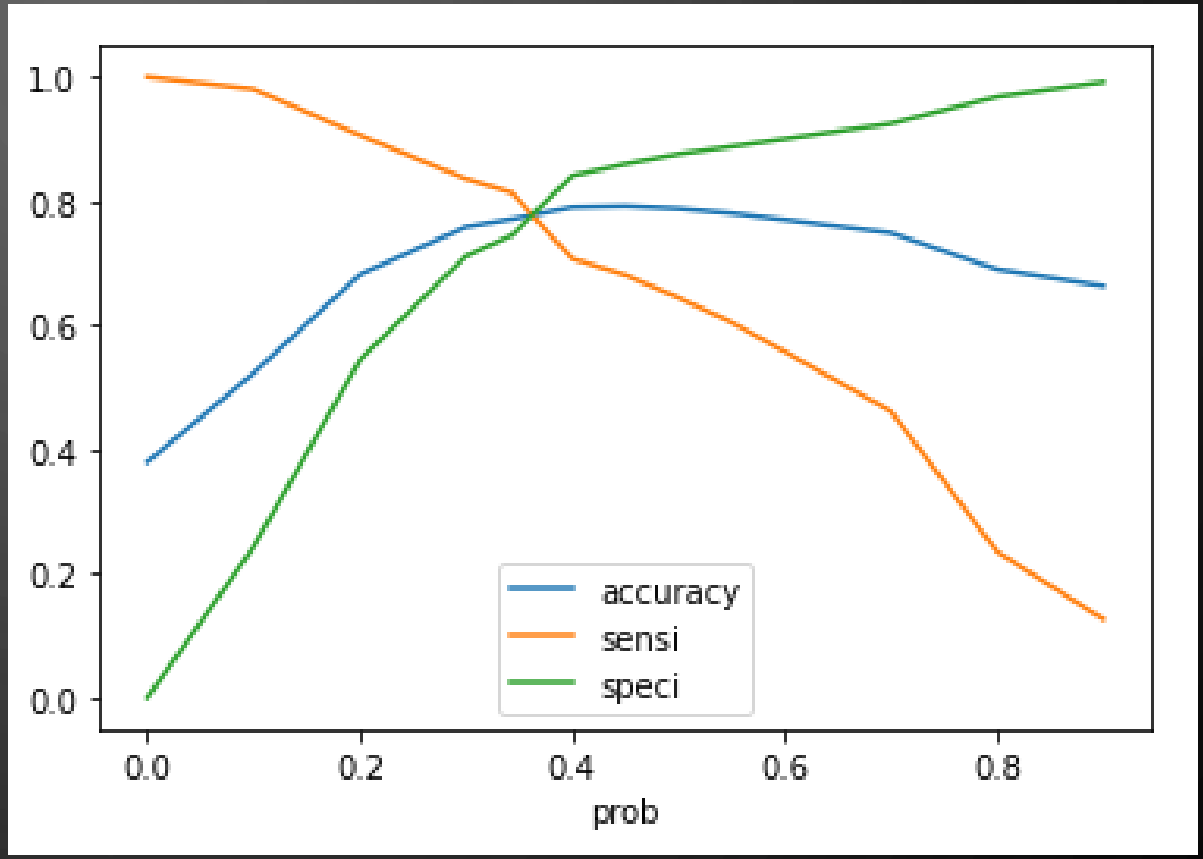
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3902	0.135	-2.896	0.004	-0.654	-0.126
Do Not Email	-1.2453	0.149	-8.380	0.000	-1.537	-0.954
TotalVisits	0.8863	0.198	4.468	0.000	0.497	1.275
Total Time Spent on Website	4.5590	0.160	28.524	0.000	4.246	4.872
google	0.1461	0.075	1.954	0.051	-0.000	0.293
olark chat	-1.8855	0.138	-13.657	0.000	-2.156	-1.615
na_country	3.1885	0.146	21.855	0.000	2.903	3.474
unemployed	-1.5633	0.118	-13.224	0.000	-1.795	-1.332
na_chosen	-2.9543	0.135	-21.922	0.000	-3.218	-2.690

FINAL MODEL

- These eight parameters help in predicting the probability of getting predicted.
- Low P-value and VIF has been achieved in this model

MODEL EVALUATION AND OPTIMAL THRESHOLD

- Optimal threshold is the point where accuracy, sensitivity and specificity intersect.
- From the graph, the cut off point is 0.342886
- The recall score obtained with this cut off value on the train and test dataset is 81.39% and 80.72% respectively



INFERENCES

- The model is finalized by the eight variables based on p(below 0.1) and VIF(below 5) values.
- The Linear equation used in the Logistic regression model(Sigmoid) is mentioned below.

y = -0.3902 -1.2453 * Do Not Email + 0.8863 * TotalVisits + 4.5590 * Total Time Spent On Website +0.1461 * google - 1.8855 * olark chat + 3.1885 *
na_country -1.5633 * unemployed - 2.9543 * na_chosen

- The Recall(Lead Conversion rate) Score from the finalized model achieved are 81.39% and 80.71% for the train and test datasets respectively


THANK YOU