LEAD SCORING CASE STUDY

- Akarsh Tyagi
-Sumit Chakraborty

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%.

TARGET

- Help the business select the most promising leads, i.e. the leads that are most likely to convert into paying customers

- 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 target lead conversion rate should be around 80%.

Steps Done to Build the Model

1. Collecting and reading the data
2. Data treatment
   - Filling the missing values with null or adequate value
   - Dropping columns having large number of null/missing values
   - Dropped those values with only one value.
3. Exploratory data analysis
   - Univariate data analysis
   - Multivariate data analysis
4. Data preparation
   - Forming adequate dummy variables for model building.
   - Splitting the data into train and test sets for training and testing the data.
   - Scaling the numerical variables using a standard scaler.
5. Model building
   - Summarizing the data
   - Feature selection using RFE
   - Assessing the model with StatsModels
   - Checked multicollinearity using the Value inflation factor method.
   - Creating Prediction
   - Model Evaluation
   - Checking precision and recall values
   - Plot ROC curve
   - Finding optimal cutoff point
   - Making prediction on test data set
6. Conclusion

READING & UNDERSTANDING THE DATA

```
# Let's see the head of our data
leads_df.head()
```

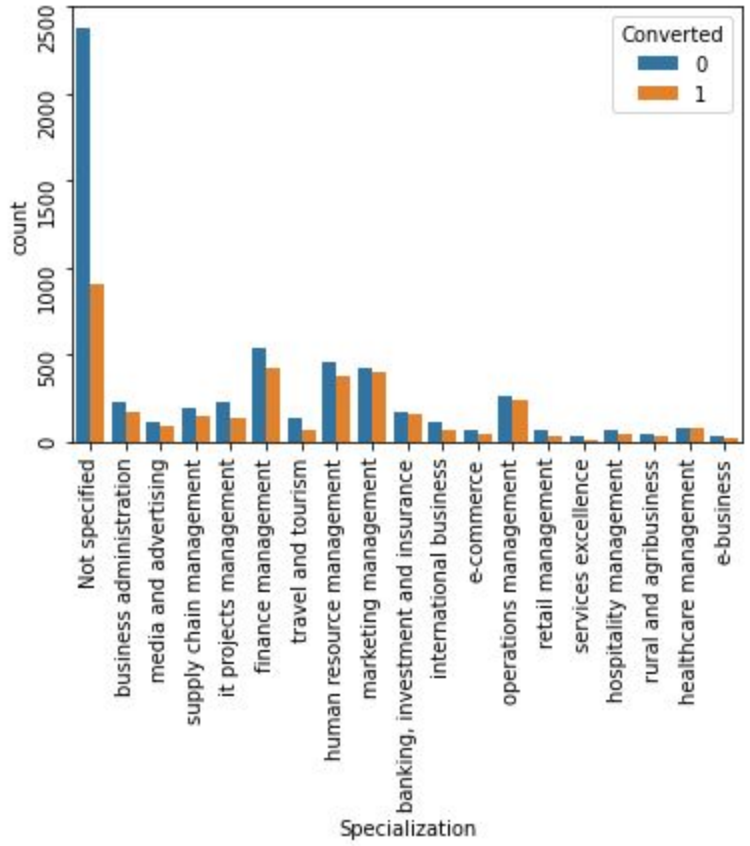
	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What yo curre occupati
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemploy
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemploy
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Stude
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth	Unemploy
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	India	Select	Other	Unemploy

Data given to us has 9240 rows and 37 columns.

CLEANING THE DATA

- We first converted all alphabetical strings to lower case for effective cleaning.
- We then replaced all select values to null as users has left it blank.
- We dropped all columns which had more than 35% null values apart from specialization as we think this column is required for business purpose.
- We dropped lead number and prospect id as it is present for identifying each rows. (Not required for analysis.)
- We filled the null values in 4 columns with not specified.
- We dropped the rows which had null values.
- Finally we left with 9074 rows and 21 columns and cleaned data.

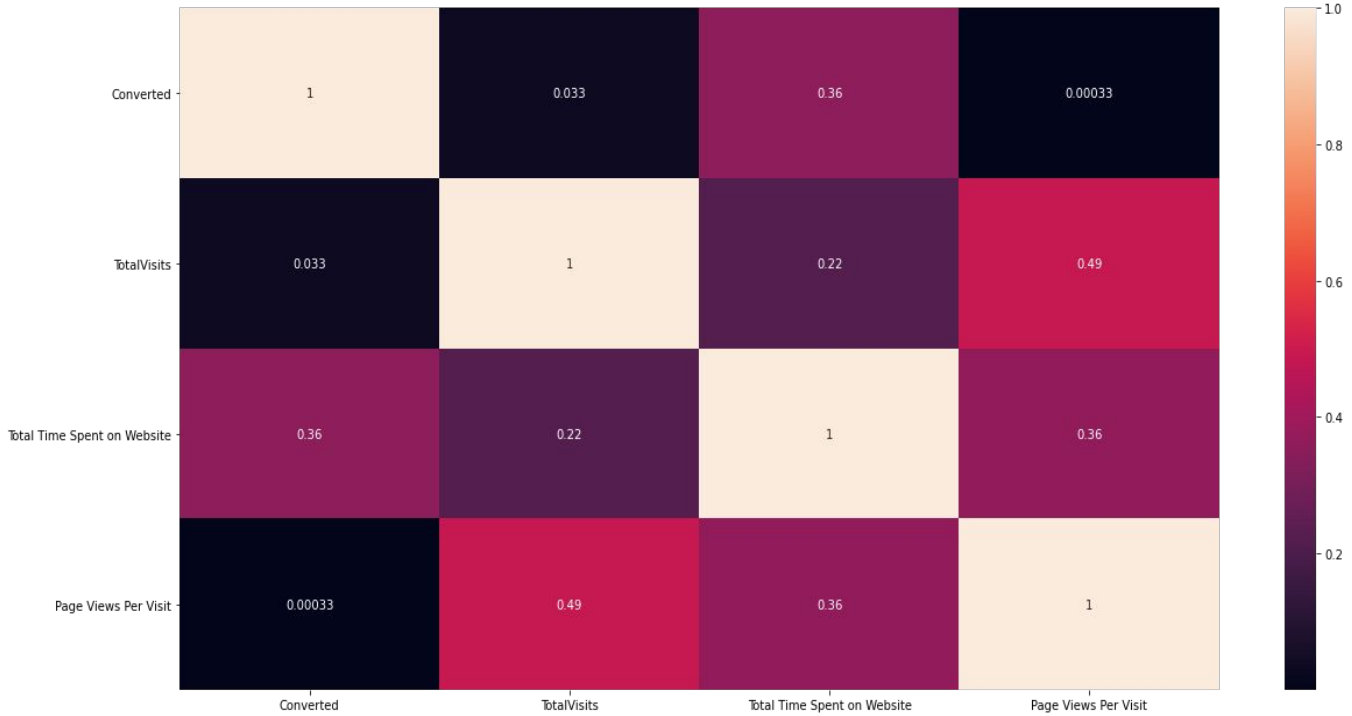
EDA - UNIVARIATE ANALYSIS



We tried doing some data analysis related to our target variable and we can see the people who has not selected their specialization, are not converted.

Among those who have selected specialization people with management specialization are more likely to get converted.

EDA-MULTIVARIATE ANALYSIS



As we can see here there are week correlation among numerical variables.

DATA PREPARATION

- Dummy variables were created for categorical variables.
- Variables for which dummy were created, were dropped.
- Data was split into Test & Train set in 70 : 30 ratio.
- Scaling was done using Standard Scaler for numeric variables.

MODEL BUILDING

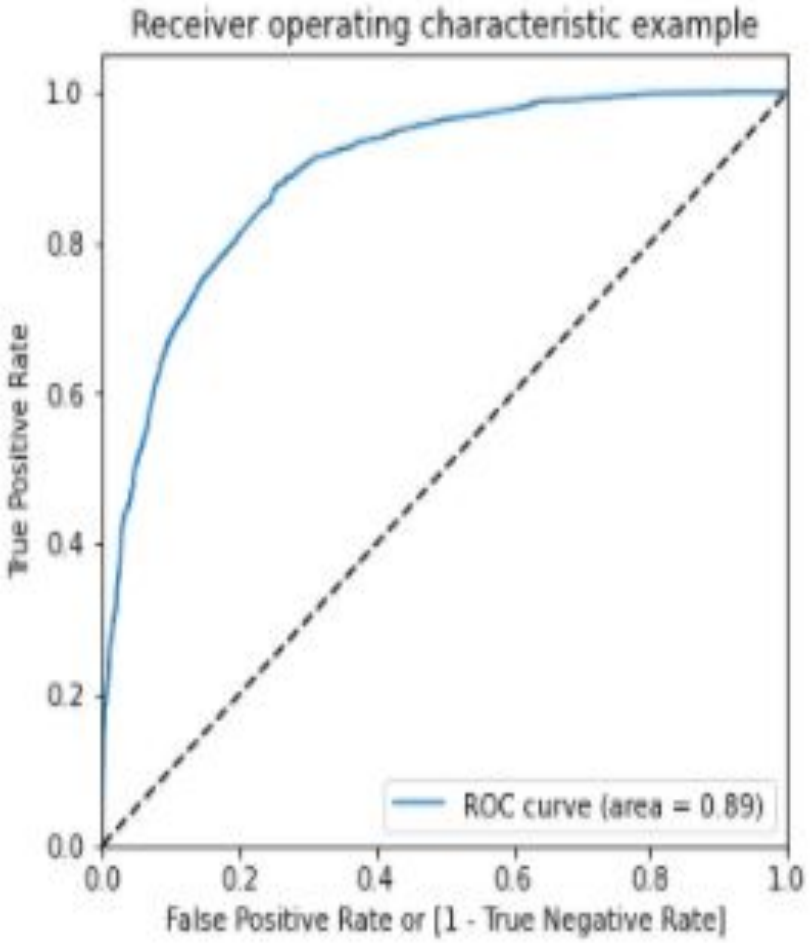
- We built the model using logistic regression model and selected top 15 features using RFE technique.
- Using the p-values and VIF values, variables were dropped where p > 0.05.
- Model was built using logistic regression, with binomial families applied on training data set.
- All the features have very low VIF values, which means there is no multicollinearity among the features.
- The overall accuracy of the model is 80.53%.

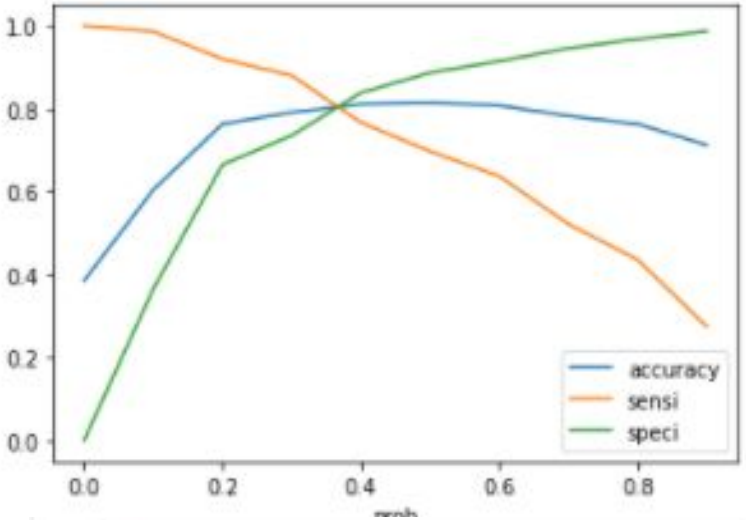
MODEL BUILDING (contd..)

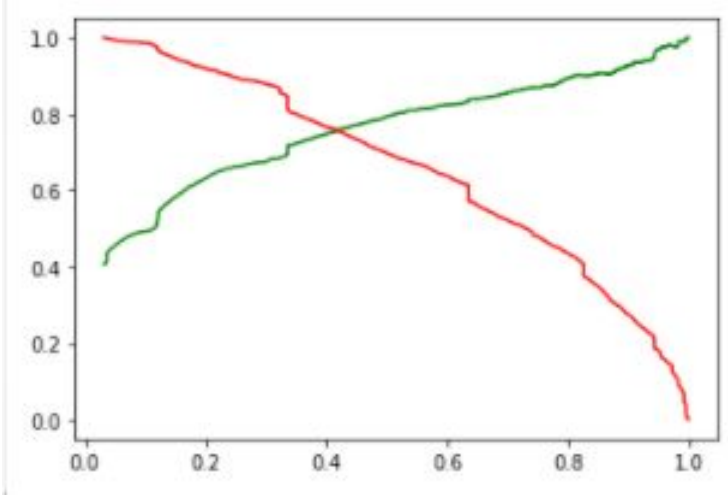
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3670	0.091	-25.878	0.000	-2.546	-2.188
Total Time Spent on Website	1.1450	0.041	27.693	0.000	1.064	1.226
Lead Origin_lead add form	2.2419	0.228	9.844	0.000	1.796	2.688
Lead Source_welingak website	2.0006	0.752	2.660	0.008	0.526	3.475
Do Not Email_yes	-1.8104	0.184	-9.838	0.000	-2.171	-1.450
Last Activity_converted to lead	-1.3229	0.224	-5.906	0.000	-1.762	-0.884
Last Activity_had a phone conversation	1.3030	1.171	1.113	0.266	-0.991	3.598
Last Activity_olark chat conversation	-1.3161	0.163	-8.078	0.000	-1.635	-0.997
Last Activity_sms sent	1.2392	0.076	16.374	0.000	1.091	1.388
Last Activity_unsubscribed	1.4738	0.471	3.127	0.002	0.550	2.398
Country_Not specified	1.4020	0.107	13.108	0.000	1.192	1.612
What is your current occupation_housewife	22.7474	1.59e+04	0.001	0.999	-3.12e+04	3.12e+04
What is your current occupation_working professional	2.4706	0.189	13.067	0.000	2.100	2.841
What matters most to you in choosing a course_better career prospects	1.2976	0.088	14.723	0.000	1.125	1.470
Last Notable Activity_had a phone conversation	2.2779	1.617	1.409	0.159	-0.891	5.446
Last Notable Activity_unreachable	2.0244	0.495	4.086	0.000	1.053	2.995

	Features	VIF
8	Country_Not specified	2.20
1	Lead Origin_lead add form	1.88
10	What matters most to you in choosing a course_...	1.81
6	Last Activity_sms sent	1.51
0	Total Time Spent on Website	1.35
5	Last Activity_olark chat conversation	1.34
2	Lead Source_welingak website	1.33
9	What is your current occupation_working profes...	1.20
3	Do Not Email_yes	1.12
7	Last Activity_unsubscribed	1.08
4	Last Activity_converted to lead	1.05
11	Last Notable Activity_unreachable	1.00

FINDING OPTIMAL CUTOFF POINT







FINDING OPTIMAL CUTOFF POINT

- We found the optimal cutoff from ROC Curve is 0.89.
- We first did accuracy, sensitivity and specificity analysis for different cutoffs and from analysis we can say that cutoff probability is 0.35.
- We also did precision recall analysis and after visualizing the trade off curve and we found that cutoff probability is 0.41.

METRICS BEYOND SIMPLE ACCURACY.

METRICES	DEFINITION	PERCENTAGE
1.Sensitivity	The proportion of observed converted leads that were predicted to be converted	80.08 %
2. Specificity	The proportion of observed non-converted leads that were predicted to be non-converted	80.87 %
3. False positive rate	The proportion of observed non-converted leads are predicted as converted leads	11.34 %
4. Positive predictive value	The converted lead predictive value tells you how often a converted lead test represents a true converted lead	79.37 %
5. Negative predictive value	The non- converted predictive value tells you how often a non-converted test represents a true non-converted lead.	82.37 %
6. Precision	Positive predictive value	75.33 %
7. Recall	Sensitivity	76.16 %

CONCLUSION

Top 3 Features which are most important

- Lead Origin
- Occupation
- What matters most to you in choosing the course

THANKYOU