Lead Scoring Case Study

By : Abdus Samad Abdullah

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 requires a model which we 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%.

Business Goal:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Step 1: Reading and understanding the data.

- Reading the leads-data file, and inspect the like identifying shape, info, describe, index, etc. to understand the data.
- Replace all Select fields with null values

Step 2: Exploratory Data Analysis(EDA).

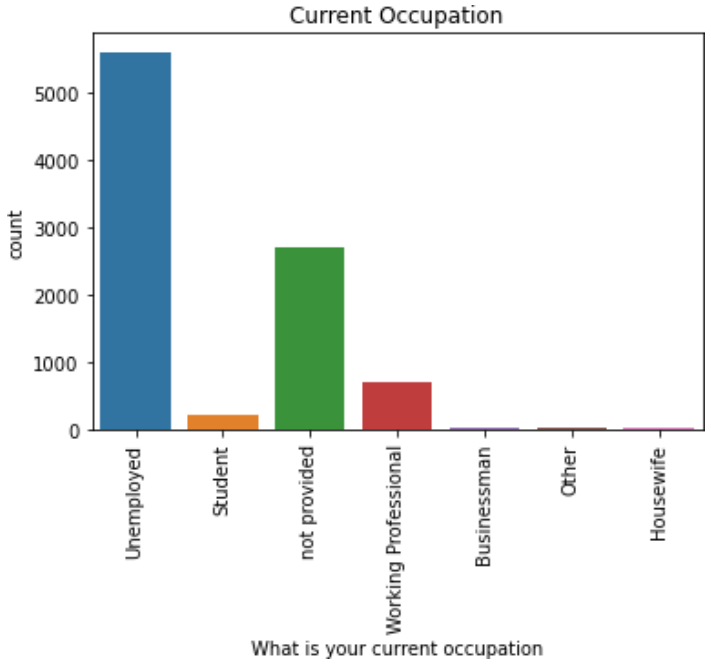
- Clean the data if any missing values.
  - Drop 8-10 columns which were having missing values greater than 40%
  - Impute the other categorical Variables with value as “not provided” or using mean , mode, median.

- Standardising the values.
  - We identified that a few variables are not scaled , we drop those columns which were highly skewed.

- Handling outliers.
  - We observed that there are outliers for few variables and we do have the flexibility of not removing the outliers, so we used capping technique to treat the outliers.

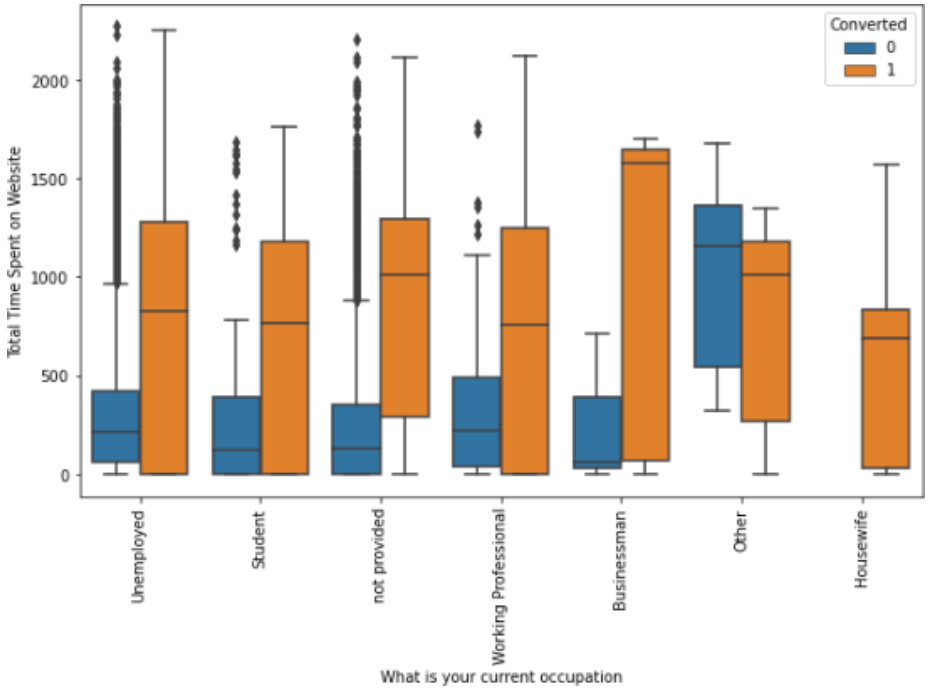
Step 3: Visualization

- **Univariate analysis.**
  - Visualizing single/individual variables.
  - To Understand the each variable frequencies we use countplot
  - Example for **What is your current occupation**:



  - We observed that the frequency of working professional is less in comparison to others so they have to be increased and transform as leads Likewise, visualized for all the variables and understood the variables individually.

- **Bivariate analysis.**
  - Visualizing multiple variables.
  - In bivariate analysis we Understand the each variable relation with other variable.
  - Example for 'What is your current occupation’ v.s ‘Total Time Spent on Website'



  - We observed that the variable working professional spend decent time on website but their conversion ratio is less. Likewise, we visualized for all the variables understood the variable relations.

Step 4: Dummy Variables

- The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. Most of the categorical variables were dumified.

Step 5: Test-Train Split

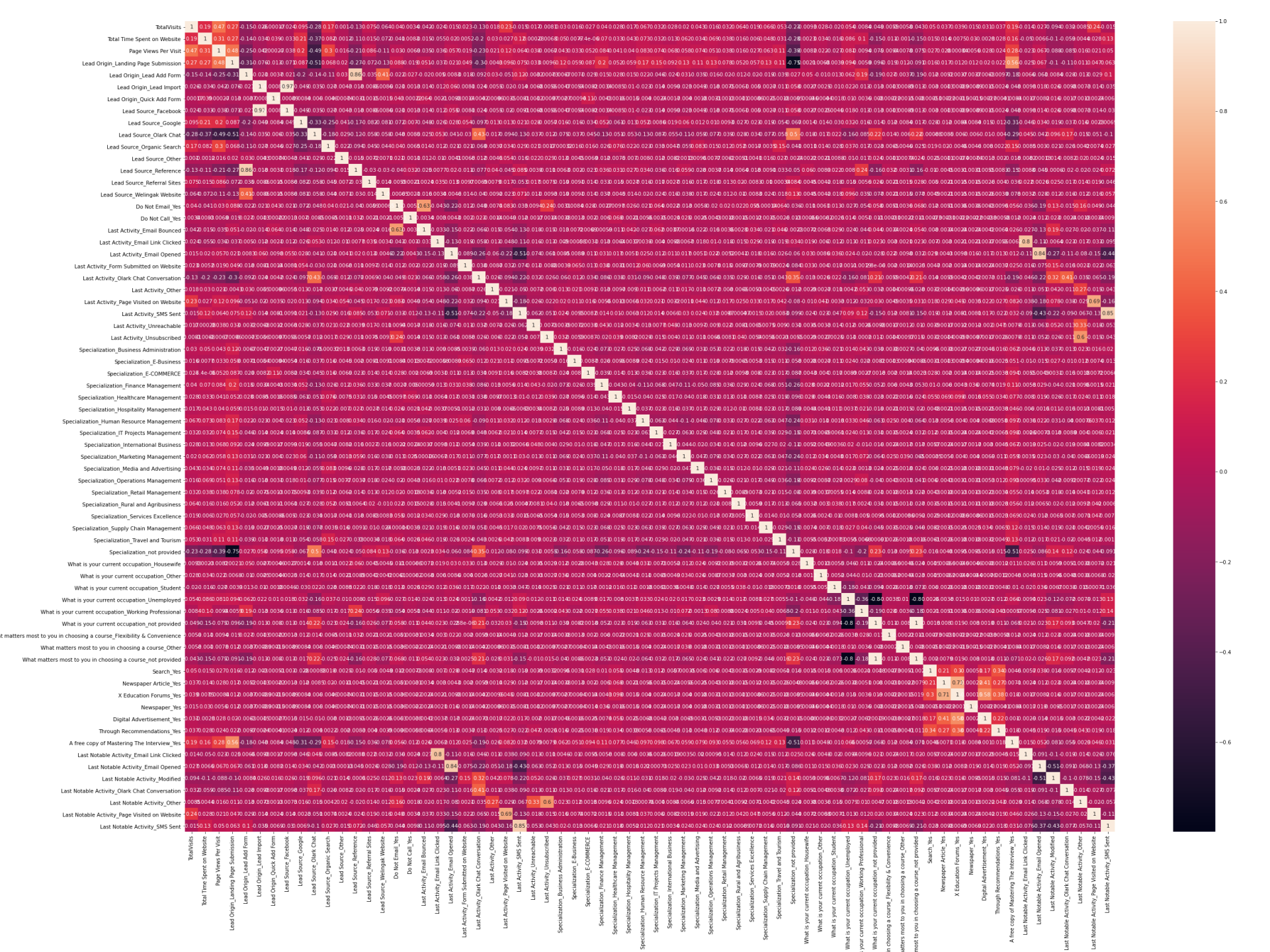
The split was done at 70% and 30% for train and test data respectively.

Step 6: Feature Scaling

To scale and standardize the variables we used Min-Max-Scaler.

Just A view of Step 7: Looking at Correlations

(Next page)


TotalVisits
Total Time Spent on Website
Page Views Per Visit
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Origin_Lead Import
Lead Origin_Quick Add Form
Lead Source_Facebook
Lead Source_Google
Lead Source_Olark Chat
Lead Source_Organic Search
Lead Source_Other
Lead Source_Reference
Lead Source_Referral Sites
Lead Source_Welingak Website
Do Not Email_Yes
Do Not Call_Yes
Last Activity_Email Bounced
Last Activity_Email Link Clicked
Last Activity_Email Opened
Last Activity_Form Submitted on Website
Last Activity_Olark Chat Conversation
Last Activity_Other
Last Activity_Page Visited on Website
Last Activity_SMS Sent
Last Activity_Unreachable
Last Activity_Unsubscribed
Specialization_Business Administration
Specialization_E-Business
Specialization_E-COMMERCE
Specialization_Finance Management
Specialization_Healthcare Management
Specialization_Hospitality Management
Specialization_Human Resource Management
Specialization_IT Projects Management
Specialization_International Business
Specialization_Marketing Management
Specialization_Media and Advertising
Specialization_Operations Management
Specialization_Retail Management
Specialization_Rural and Agribusiness
Specialization_Services Excellence
Specialization_Supply Chain Management
Specialization_Travel and Tourism
Specialization_not provided
What is your current occupation_Housewife
What is your current occupation_Other
What is your current occupation_Student
What is your current occupation_Unemployed
What is your current occupation_Working Professional
What is your current occupation_not provided
What matters most to you in choosing a course_Flexibility & Convenience
What matters most to you in choosing a course_Other
What matters most to you in choosing a course_not provided
Search_Yes
Newspaper Article_Yes
X Education Forums_Yes
Newspaper_Yes
Digital Advertisement_Yes
Through Recommendations_Yes
A free copy of Mastering The Interview_Yes
Last Notable Activity_Email Link Clicked
Last Notable Activity_Email Opened
Last Notable Activity_Modified
Last Notable Activity_Olark Chat Conversation
Last Notable Activity_Other
Last Notable Activity_Page Visited on Website
Last Notable Activity_SMS Sent
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

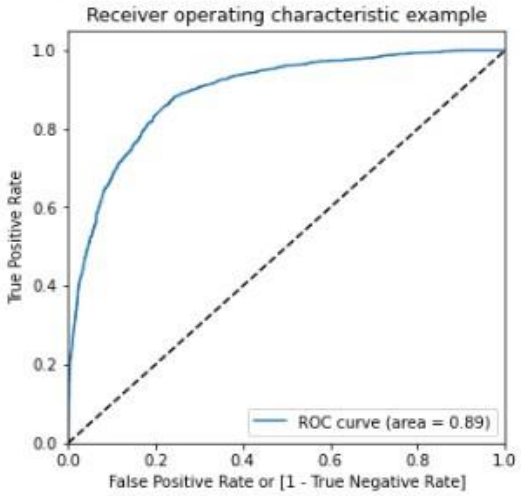
Step 8: Model Building

- Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
- So Finally, by dropping 5 variables final model was left with 15 Variables.

Step 9: Model Evaluation

- A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each. In fact sensitivity responded well with 81.4%

```
In [112]: # Call the ROC function
          draw_roc(y_train_pred_final.Converted, y_train_pred_final.Conversion_Prob)
```



The area under ROC curve is 0.89 which is a very good value.

Step 10 Prediction:

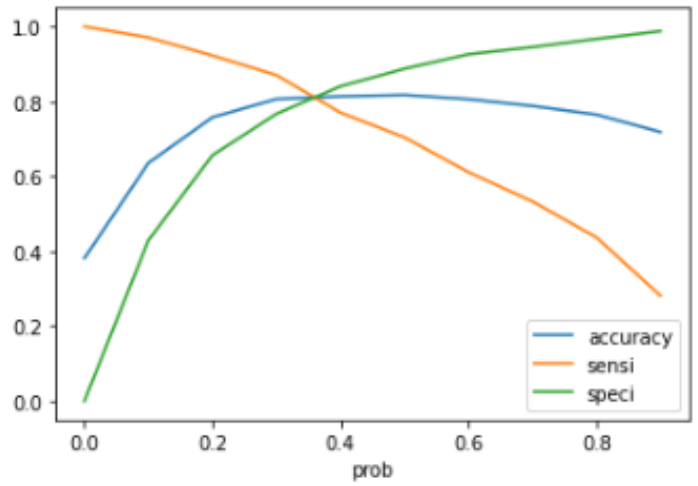
- Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

- **Precision – Recall:**
- This method was also used to recheck and a cut off of 0.41 was found with Precision around 74% and recall around 75% on the test data frame.

- Hence there is a decline so better the optimum cut off is 0.35 obtained from ROC we use cut off 0.35 for further analysis

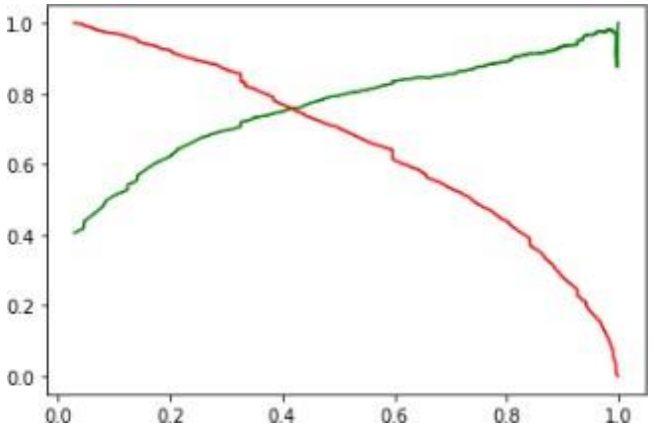
- Cutoff Plot from ROC

```
# Plotting it
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```



From the graph it is visible that the optimal cut off is at 0.35

Precision-Recall (Another metric optional)



Final Analysis:

- With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 80-81% for Test Data

- Hence The Sensitivity is 80% + which signifies that it is a decent model

- **Summary of the Final Logistic Regression Model Built :**

- **Which has final 15 variables used for the Model.**
- (In the Next Page)

Oep. 'U'a riable:	her' Ve	N-a . Ob se m a €sons:	*4•38
Monte l:	-mL11	Of Re sicJual s.	*W<
MocJel Ham ilyz	-> xc-m ol	DI MocJe l.	1*
Li n k F umction:	og-.	Sale.	1.CC-DC
Memeo.	S	Log-Likelizeru.	-2598.2
Oate:	I -' o . 3.- Eep <C-2C	Oevia noe.	fi1Z*.•1
Tirrsez	1 .- :C-g: D	F-'ea rsori chi2z	. . 11a+D*

n

	coef	std err	z	P>|z|	p.ozs	o.szsj
cons€	-2.5154	C-.128	-IP.G-4C	C-.CADC	-2.766	-2.264
TotaFv'isits	*.1C17	1.P11	:?.118	C-.C-D<	2.338	9.985
To€al Tirrze Spent on VUebsite	4.fi:Z1.-	C-.1•38	<.- .CA*	C-.CADC	4.207	4.864
LeacJ **Origin** Lead Add -arm	*. ?A *	C-.13*	to. ?	C-.C-DC	3.215	3.973
Leacl Socsrce_O€ark Chat	1. 4 D.-	C. IU	1s. 21	C-.C-DC	1.218	1.683
Least Source Welinga k VVebsite	<. Cir	C-..- 4*	..- 38	C-.C-D*	0.580	3.506
Oo l'do1 Errsail_Yes	-1. < w*	C-.1 - G	-C-.o'3*	C-.CADC	-1.579	-0.878
Last Activity Email Opemecl	C. fiP2*	C-.1D*	5.579	C-.CADC	0.384	0.800
Last *D-c v fity O t ier*	1. 7fi46	C-. 1*	:?.:?'3fi	C-.C-D1 Last	0.724	
Activity SMSSent	1.7Cw*	C-.1 GB	IN.c- 4	C-.C-3C	**1.GB**	**1.P2C**
Sjoec ia lizagion_H ospital ity Marsagerrsez €:	-C. G•1•3C	C-.:?28	- .c-3?	C-.C-D•1	-1.VG	-C-.:?D*
	-C. 4	C-.CAB	- .<-- .-	C-.C-D4	-C-.428	-C-.CU
	2.5095	C-.131	1:?.UC	C-.C-DC	2.135	.c- 4
	-1.CRZ*	C-.CAB	-1 .:?4C	C-.CADC	-1. 14	-C-.P1C
ccu last Notable Acgsvi€y_Modifiecl	-C. * fiwfi	C-.CAB	- .Z2C	C-.CADC	-C. 811	-C-.4'3*
Est Notable *X·câ•xfit' O t ier*	1.7o26	C-.:?1<	.C•3<	C-.CADC	1.C-g*	.4- *

Conclusion :

After Building the model we conclude that :

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- - 1. Total number of visits.
- - 2. The total time spend on the Website.
- - 3. Lead Origin through Add Form
- - 4. When the lead source was:
- a. Welingak Website
- b. Olark Chat
- - 5. When their current occupation is as a working professional.
- - 6. Last Activity of the user.

- **Recommendations**

- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses by following some different methods and approaches and increase the sales by :

- - Automated emails and SMS
- - Campaigns on Social Networking sites
- - Some Additional add-on prep courses

Thank you. ☺