LEADS SCORING CASE STUDY

SAI MEGHANA AND HARSHINI V BHAT

PROBLEM STATEMENT

- X education sells online courses. It gets a lot of leads but its conversion rate is very poor. For example if it gets 100 leads only 30 of them are converted.
- To make this process more efficient the company wishes to identify the most potential leads as 'HOT LEADS'.
- If they successfully identify the hot leads then their lead conversion rate would increase.

BUSINESS OBJECTIVE

- To know the which are the most promising leads and to build a model that identifies the most promising leads. To assign a lead score value between 0 and 100 to each lead so that the company can target potential leads.
- Deployment of the model for future use as well.

SOLUTION METHODOLOGY

- Data Cleaning and Data Handling
- EDA
- Creating Dummy Variables
- Feature scaling and data encoding
- Model building using classification technique i.e Logistic Regression
- Validation of the model
- Conclusions and recommendations

DATA HANDLING

- Inspecting the data and getting the statistical information

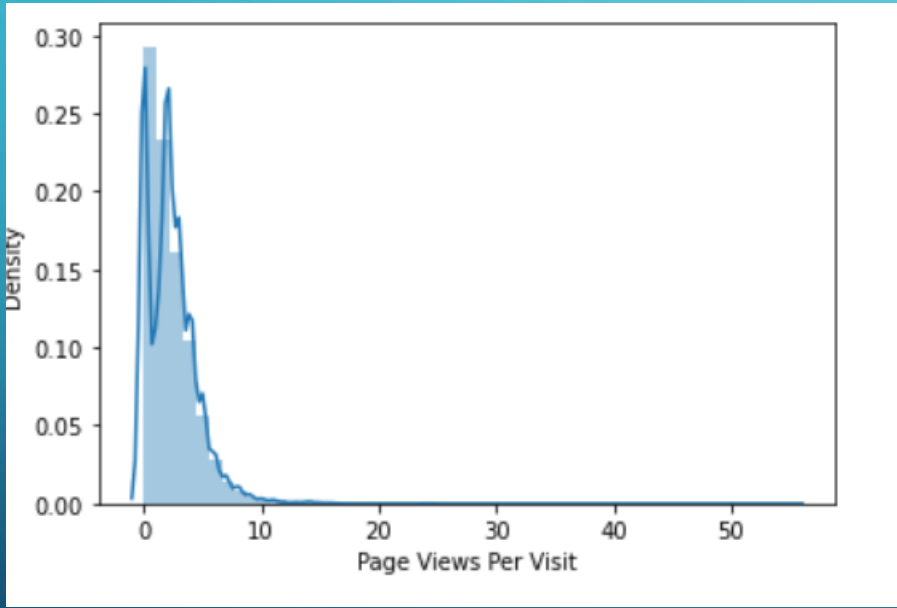
1. The data has 9240 datapoints and 37 variables

- Handling null values and missing values by imputing the values appropriately.

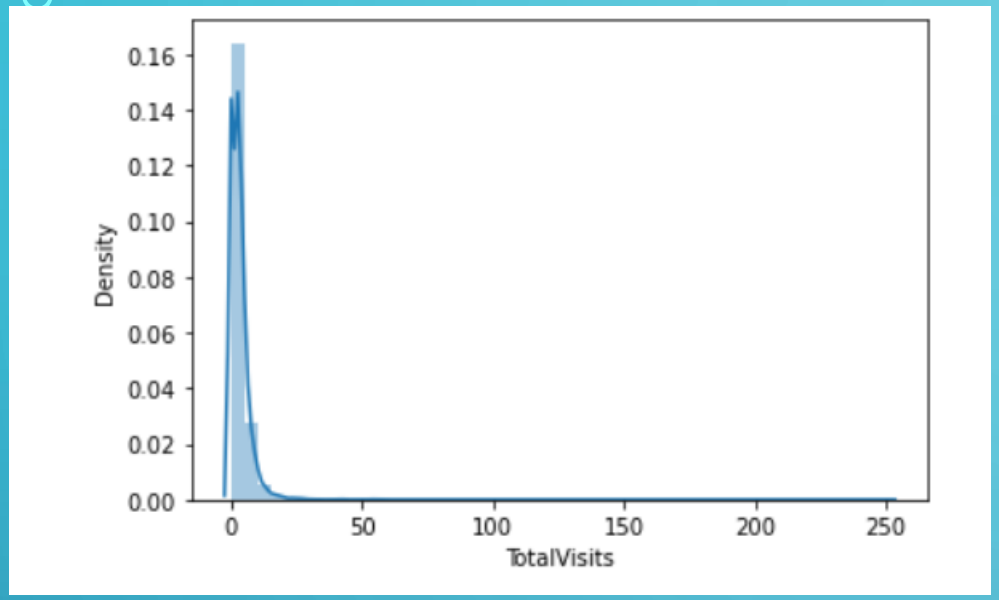
1. Converting the ‘select’ values in all the variables to NAN .
2. Dropping columns having more than 30% of missing values.
3. Checked for missing values in different variables one by one and imputed them with the mode of the variable accordingly.
4. Dropped columns such as ‘'What matters most to you in choosing a course’ as the values did not have any significant variance, hence it would not add any value for our analysis.

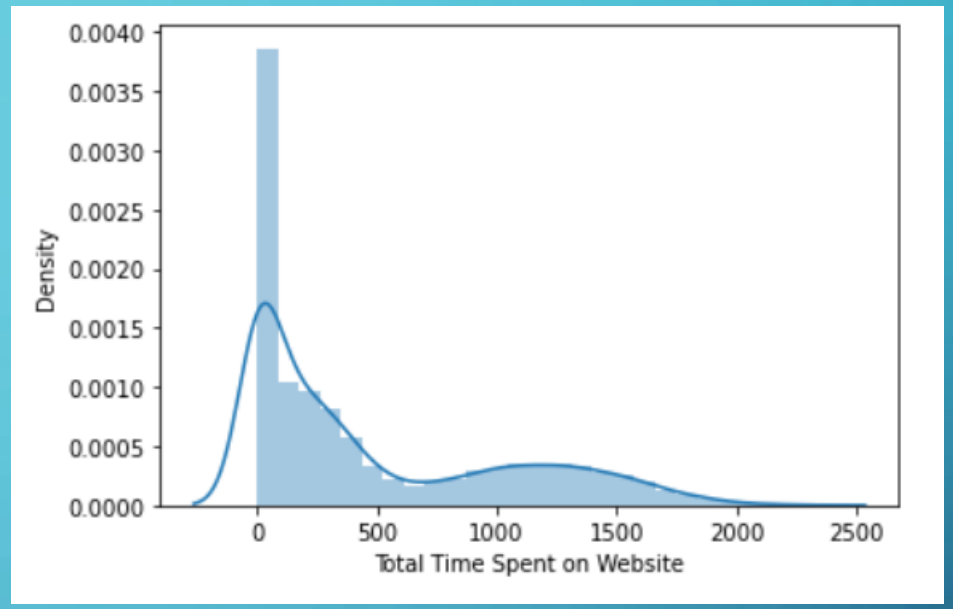
EDA

Univariate Analysis



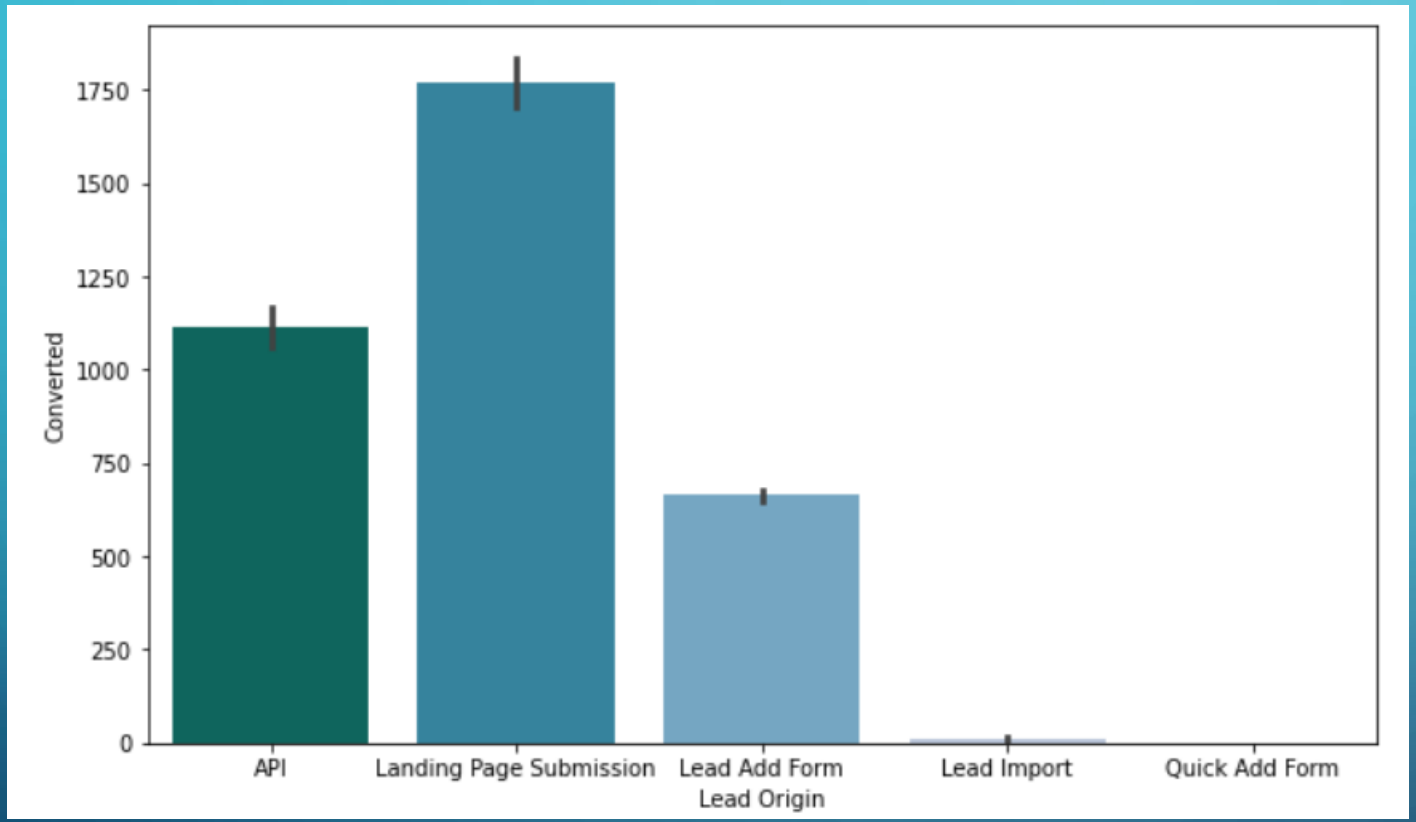
The plot gives the distribution of the variable 'Pages Views Per Visit



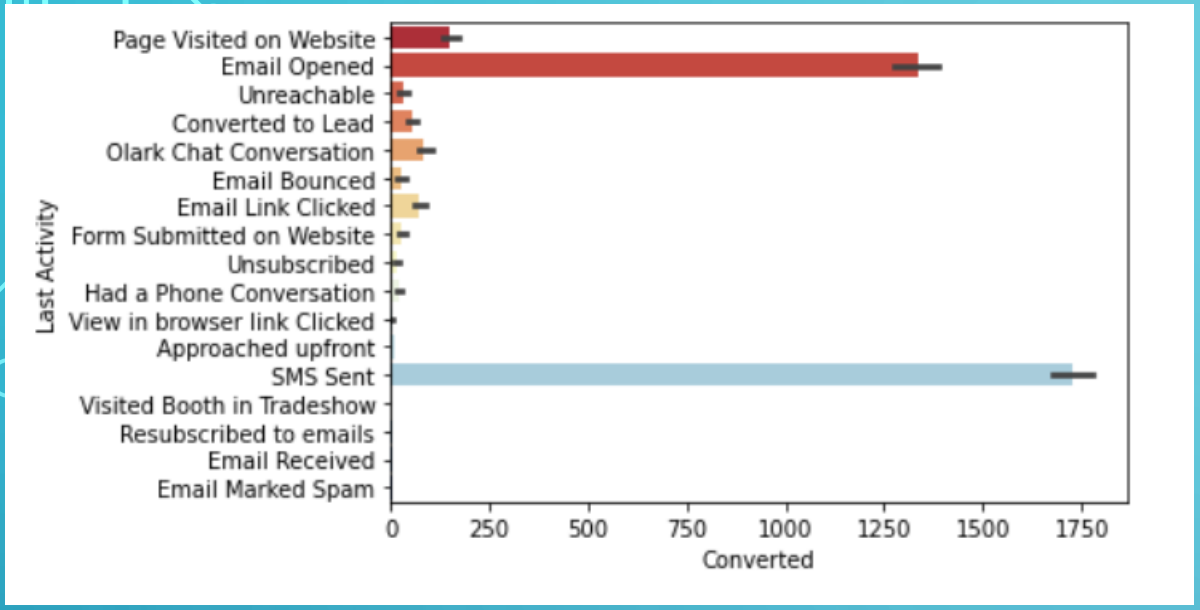


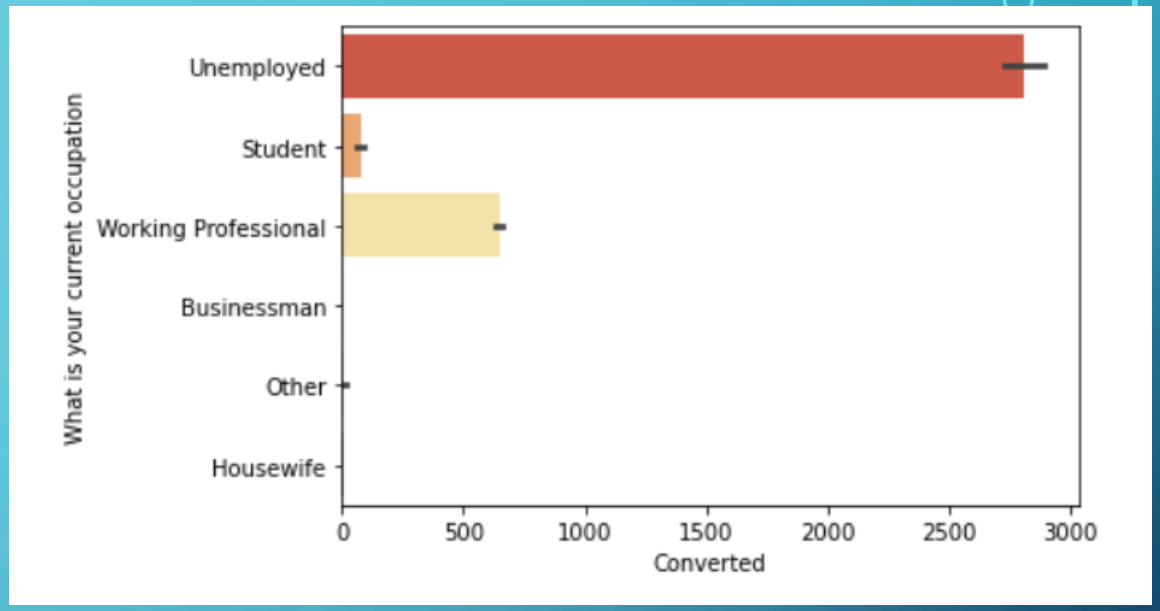
- The above two plots shows the distribution of the variables ‘TotalVisits’ and ‘Total Time Spent on Website

BIVARIATE ANALYSIS



The people who are landing on the submission page have a higher rate of conversion.





The above plots shows the categories such as ‘SMS sent’ and ‘unemployed’ that are more likely to be converted based on their Last activity and their current occupation.

DATA PREPARATION

- Creating dummy variables and concatenating them to the dataframe and then deleting the original variables to eliminate collinearity in the variables.(one-hot-encoding)
- converting numeric variables to normalized form.
- Total number of rows = 9240
- Total number of columns = 74

MODEL BUILDING

- Splitting the data into train set and test set
- The data is split in the ratio 70:30
- Scaling the numeric variables using standard scaler to reduce anomalies
- We use recursive feature elimination RFE technique for feature selection and select the top 15 features that are highly relevant in our model building.
- Building the first model with 15 features as output
- Successively building model by eliminating features with high p value and high VIF value greater than 5.

	Features	VIF
5	Last Activity_Olark Chat Conversation	1.89
3	Lead Source_Olark Chat	1.65
10	Last Notable Activity_Modified	1.51
2	Lead Origin_Lead Add Form	1.41
11	Last Notable Activity_Olark Chat Conversation	1.30
4	Lead Source_Welingak Website	1.23
1	Total Time Spent on Website	1.20
6	What is your current occupation_Working Profes...	1.14
0	Do Not Email	1.11
8	Last Notable Activity_Email Opened	1.10
7	Last Notable Activity_Email Link Clicked	1.02
12	Last Notable Activity_Page Visited on Website	1.02
9	Last Notable Activity_Had a Phone Conversation	1.00

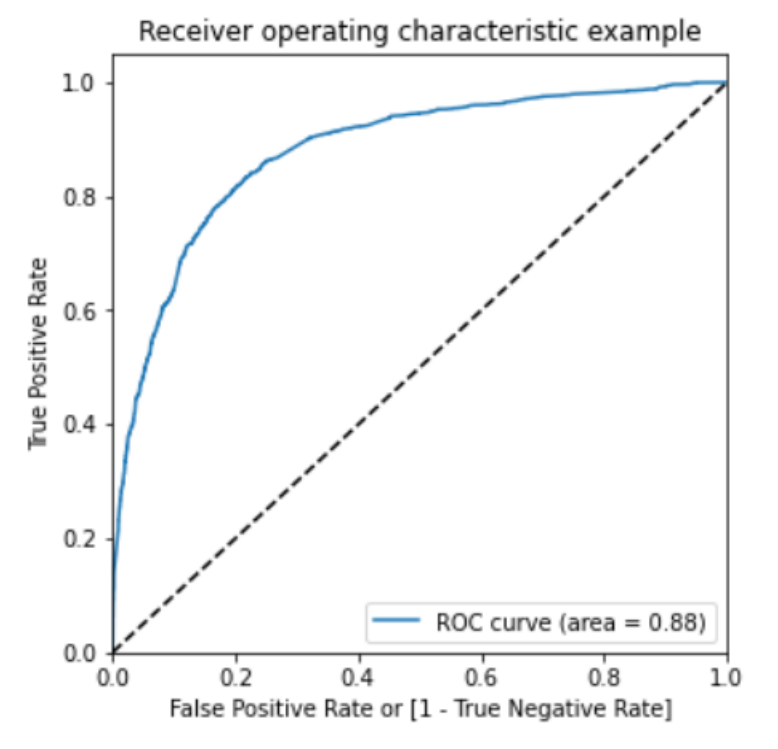
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4998	0.010	48.524	0.000	0.480	0.520
Do Not Email	-0.1765	0.018	-9.906	0.000	-0.211	-0.142
Total Time Spent on Website	0.1877	0.005	36.208	0.000	0.178	0.198
Lead Origin_Lead Add Form	0.5592	0.020	28.102	0.000	0.520	0.598
Lead Source_Olark Chat	0.1687	0.014	11.959	0.000	0.141	0.196
Lead Source_Welingak Website	0.1951	0.043	4.487	0.000	0.110	0.280
Last Activity_Olark Chat Conversation	-0.1225	0.020	-6.135	0.000	-0.162	-0.083
What is your current occupation_Working Professional	0.3447	0.018	19.040	0.000	0.309	0.380
Last Notable Activity_Email Link Clicked	-0.3067	0.036	-8.606	0.000	-0.377	-0.237
Last Notable Activity_Email Opened	-0.2225	0.013	-17.429	0.000	-0.247	-0.197
Last Notable Activity_Had a Phone Conversation	0.2398	0.113	2.124	0.034	0.019	0.461
Last Notable Activity_Modified	-0.2996	0.013	-23.342	0.000	-0.325	-0.274
Last Notable Activity_Olark Chat Conversation	-0.2831	0.040	-7.122	0.000	-0.361	-0.205
Last Notable Activity_Page Visited on Website	-0.2657	0.026	-10.117	0.000	-0.317	-0.214

From the above we can observe that all the P value are less than 0.05 and VIF values less than 5 which is the acceptable range

- Predicting the model

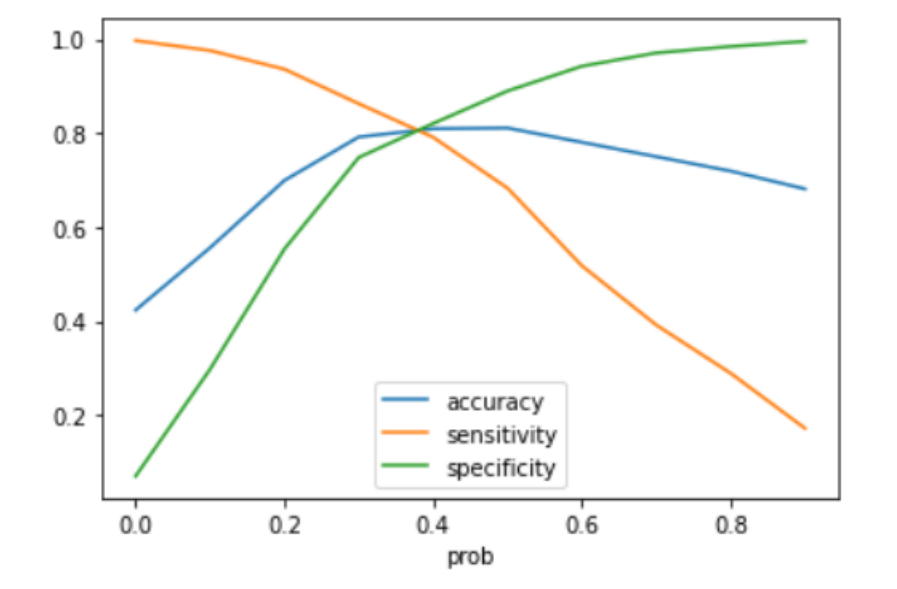
VALIDATING OUR MODEL

- ROC curve



As the curve is more towards the upper left corner covering 88% of the area, we can say that our model is good.

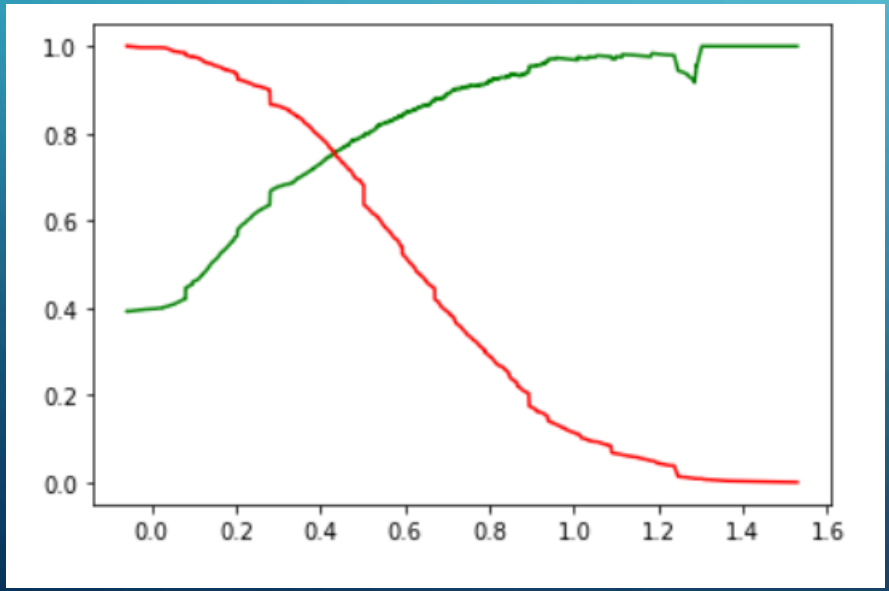
- Calculating specificity, sensitivity and accuracy with a random cut-off value as 0.5 and then obtaining the optimal cut-off value where we get balanced sensitivity and specificity.
- Clearly from the graph we can observe that 0.4 is the optimal cut-off point.



- Precision and Recall

Precision and recall plays an important role in building a more business oriented model. The recall percentage will be more valuable because it is okay if our precision is little low as it captures the less hot lead customers but we do not want to leave out any hot leads which are willing to get converted which is given by recall hence our focus should be more on recall than precision.

- The obtained precision and recall values are 73% and 79% respectively.



The graph shows that there seems to be trade off between Precision and Recall and the meeting point is nearly at 0.5

CONCLUSION

- The lead score is assigned to each lead based on their conversion probabilities.
- The variables that matter the most to realize the potential of the buyers

➢When the Lead origin is Lead Add Form

➢When value in What is your current occupation is Working Professional

➢When Last Notable Activity – Had a phone conversation

➢Total Time spent on website

RECOMMENDATIONS

X Education should focus more on those whose current occupation is working professionals as they are more likely to get converted and they are potential buyers.