LEAD SCORE ASSIGNMENT SUMMARY

The below mentioned steps are used for performing the analysis:

1. Data Cleaning

- The columns containing over 45% missing values have been removed as it would not be appropriate to use any statistical method to fill such a large amount of missing data. Please refer to the notebook for a complete list of the dropped columns.
- High cardinality columns, ID columns, and columns with constant values such as Prospect ID, Lead Number, Lead Profile, Magazine, etc. have been removed as they are not useful. Please refer to the notebook for a complete list of the dropped columns.
- There are some columns, such as City, Specialization, Tags, What matters most to you in choosing a course, What is your current occupation, and Country, which contain more than 20% missing values. Filling them with mean, median, or mode may not be appropriate, and dropping them would result in a significant loss of records. Since they are important by definition, we have denoted missing values in these columns with a new category 'Unknown'.
- The number of missing values is very low in TotalVisits, Page Views Per Visit, Last Activity, and Lead Source. Hence, we have used median and mode statistics to fill the missing values.
- We have also removed the columns Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations because they were heavily skewed towards the value of 'No' (0).

Following those steps, we were able to retain 97% of data. And this data is used for EDA.

2. EDA

We carried out exploratory data analysis (EDA) on the cleaned data by using a variety of plots and assessing both continuous and categorical variables. We performed univariate and bivariate analyses on the target variable to aid comprehension. The following are some of our conclusions:

- Total Visits has a high positive correlation with Page Views per Visit.
- High conversions are seen among people who opted for email subscription compared to those who did not.

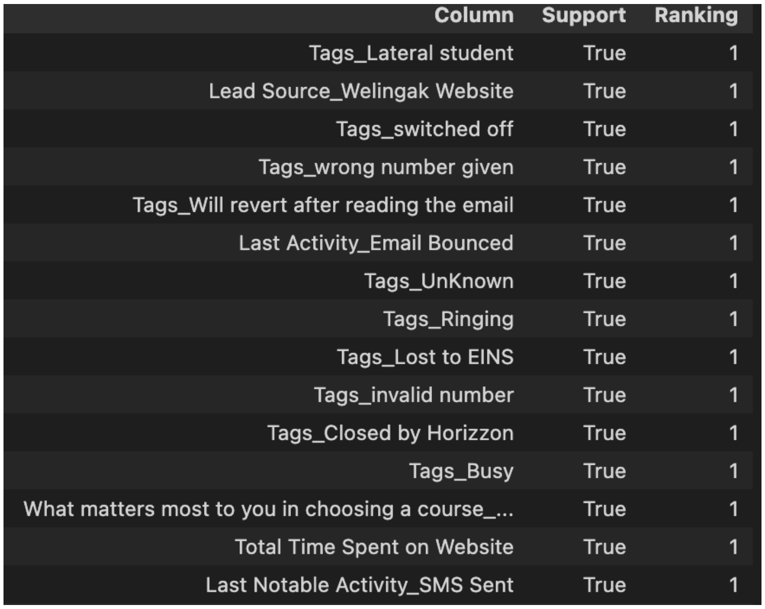
- Comparatively more conversions are seen among people who did not want a free copy of "Mastering the Interview".
- More time-consuming individuals are promising leads.
- There is no significant difference in the number of page views per visit between people who converted and those who did not.
- The most common lead sources are "Google", followed by "Direct Traffic" and "Olark Chat".
- "Google" has the highest conversion rate, followed by "Direct Traffic" and "Olark Chat".
- The most significant difference between converted and non-converted leads can be seen in the "Reference" lead source.
- "Email Opened" and "SMS Sent" are the most common last activities of customers.
- Those whose last activity was "SMS Sent" had the best conversion rate.
- Those whose specialization is unknown have the highest conversion rate.
- Other than "Unknown", high conversion rates are seen in "Finance Management", "Human Resource Management", and "Marketing Management".
- People who are unemployed have a higher conversion rate compared to working professionals.
- Those whose occupation is "Unknown" have a high number of non-conversions.
- Among "Working Professionals", a very high conversion rate is seen.
- Those whose last notable activity was "SMS Sent" have the best conversion rate, followed by "Modifier" and "Email Opened".

3. Dummy variable creation, train-Test split and scaling

- To prepare the data for machine learning algorithms, we performed OneHotEncoding by creating dummy variables for the final categorical columns in the dataset (after cleaning), including 'Lead Origin', 'Lead Source', 'Last Activity', 'Country', 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'Tags', 'City', and 'Last Notable Activity'.
- We used a 75-25 ratio to split the data into training and testing sets, respectively.
- To scale the numerical features, we used MinMaxScaler.

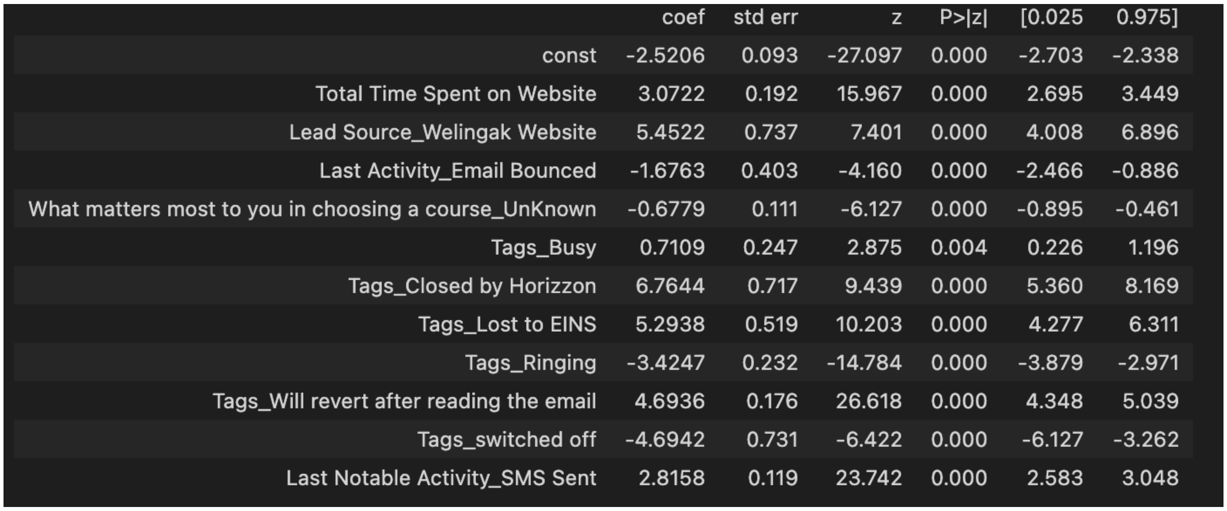
4. Model Building and Predictions

- Total of 15 features were selected using the RFE feature selection approach.



Column	Support	Ranking
Tags_Lateral student	True	1
Lead Source_Welingak Website	True	1
Tags_switched off	True	1
Tags_wrong number given	True	1
Tags_Will revert after reading the email	True	1
Last Activity_Email Bounced	True	1
Tags_UnKnown	True	1
Tags_Ringing	True	1
Tags_Lost to EINS	True	1
Tags_invalid number	True	1
Tags_Closed by Horizzon	True	1
Tags_Busy	True	1
What matters most to you in choosing a course_...	True	1
Total Time Spent on Website	True	1
Last Notable Activity_SMS Sent	True	1

- Five modeling attempts were done until the VIF and p-values were below allowable thresholds. Final Model:



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5206	0.093	-27.097	0.000	-2.703	-2.338
Total Time Spent on Website	3.0722	0.192	15.967	0.000	2.695	3.449
Lead Source_Welingak Website	5.4522	0.737	7.401	0.000	4.008	6.896
Last Activity_Email Bounced	-1.6763	0.403	-4.160	0.000	-2.466	-0.886
What matters most to you in choosing a course_UnKnown	-0.6779	0.111	-6.127	0.000	-0.895	-0.461
Tags_Busy	0.7109	0.247	2.875	0.004	0.226	1.196
Tags_Closed by Horizzon	6.7644	0.717	9.439	0.000	5.360	8.169
Tags_Lost to EINS	5.2938	0.519	10.203	0.000	4.277	6.311
Tags_Ringing	-3.4247	0.232	-14.784	0.000	-3.879	-2.971
Tags_Will revert after reading the email	4.6936	0.176	26.618	0.000	4.348	5.039
Tags_switched off	-4.6942	0.731	-6.422	0.000	-6.127	-3.262
Last Notable Activity_SMS Sent	2.8158	0.119	23.742	0.000	2.583	3.048

- Performed probability cutoff tuning and got 0.26 as threshold. And predictions were produced using that value.
- Final accuracy, sensitivity, and specificity for the test set were 91%, 92%, and 91% respectively. Recall was 92% and Precision was 85%.

5. Conclusion

The columns which played key role in modeling are listed above in the model summary. Below are the noteworthy ones:

- “Total Time Spent on Website”
- “Tags: Whether the Lead is tagged by”
    - "Ringing"
    - "Closed by Horizzon"
    - "Lost to EINS"
    - "Will revert after reading the email"
    - "Switched off"
- Lead Source_Welingak Website
- Last Notable Activity_SMS Sent
- Last Activity_Email Bounced