X Education Lead Scoring

By Vladimir Nikonov

Solution Strategy

1. Business problem understanding.
2. Mapping a business problem to a data science / machine learning problem.
3. Data sourcing, EDA, pre-processing.
4. Model building.
5. Model evaluation.

1. Business Problem Understanding

- An education company called *X Education* provides online courses to industry professionals.
- The main source of revenue for the company is the customers buying their courses.
- The typical lead conversion rate at *X Education* is **30%**.
- *X Education* requires a machine learning model, that will calculate the lead score. The leads with a higher conversion probability will get higher scores, while the leads with lower conversion probability will get lower score.

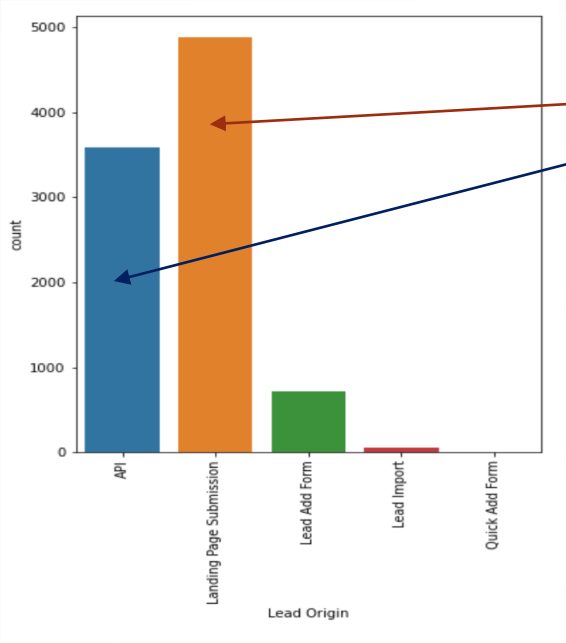
2. Problem Mapping

- Customer:
  - *X Education*'s sales team. They communicate with the leads, the model can help them identify the leads that they need to focus on communicating with.
- Expected ML Solution:
  - A Logistic Regression Model.
  - The model computes the lead score, which is basically the probability of conversion.
- Data Source:
  - Leads dataset from the past, provided by X Education.

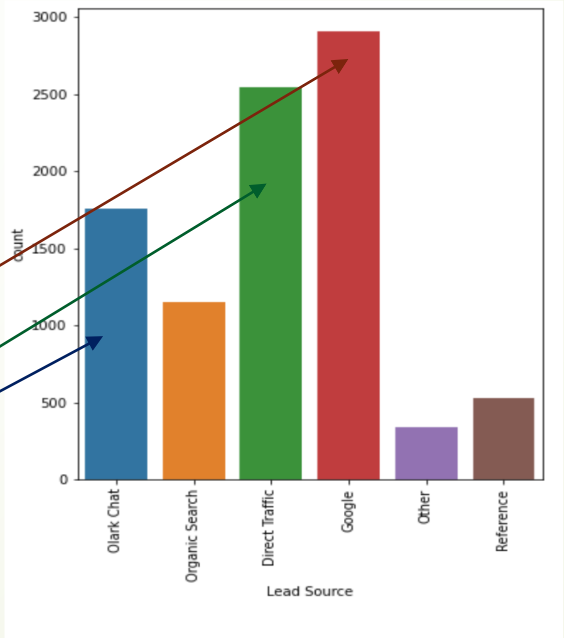
3. Data Sourcing, EDA, Pre-processing

- 3.1 Data Sourcing
  - The leads data set with 9240 rows and 37 columns came from X Education. No duplicate rows were found in the data set.
- 3.2 Data Cleaning
  - After removing the columns with high percentage of missing values (>= 40%) and imbalanced data, only 10 features were left (3 numerical and 7 categorical).
  - The rest of the missing values was imputed with most frequent values for categorical features and median values for numerical features.

- 3.3 EDA
- 3.3.1 Univariate Analysis

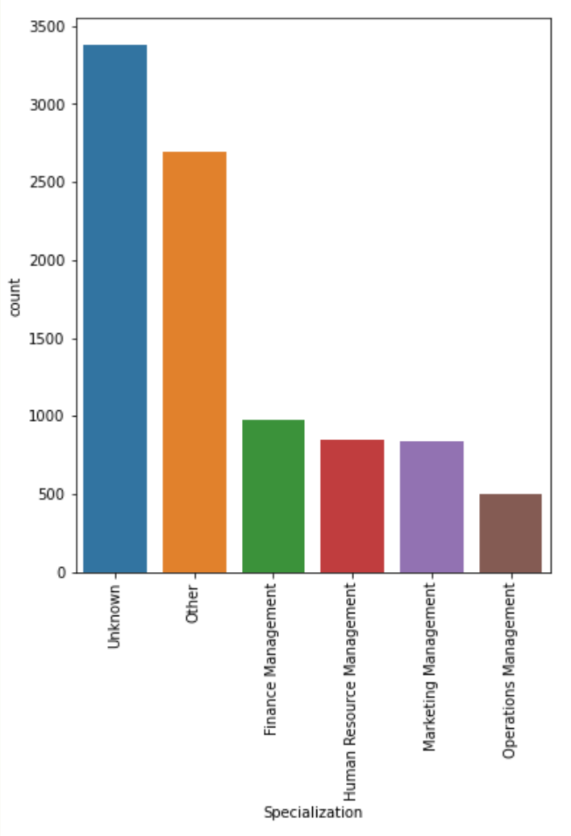


Leads mostly originate from landing a page submission or through the API.

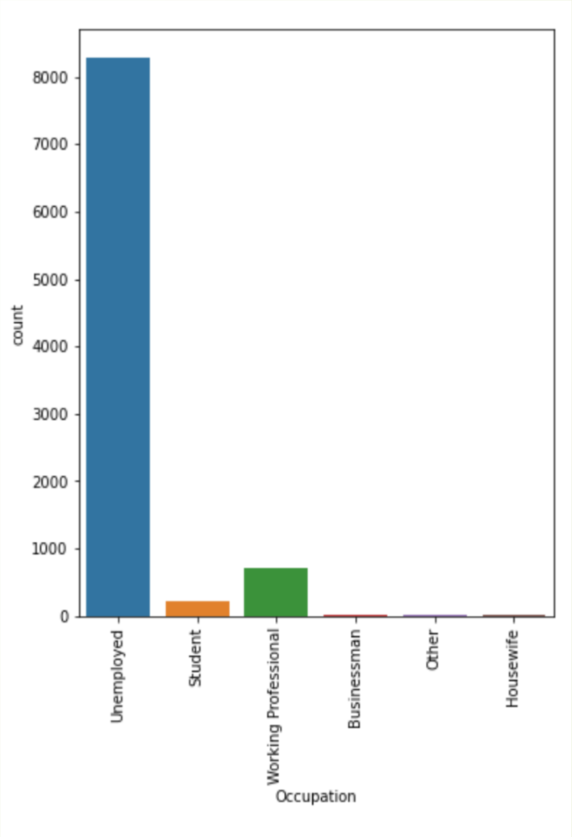


The three main sources of leads are:

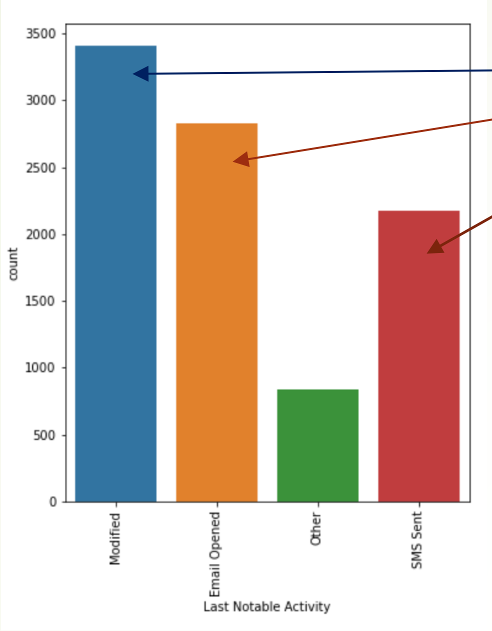
Google
Direct Traffic
Olark Chat



Out of known specializations, the ones connected to management are more popular with the leads.

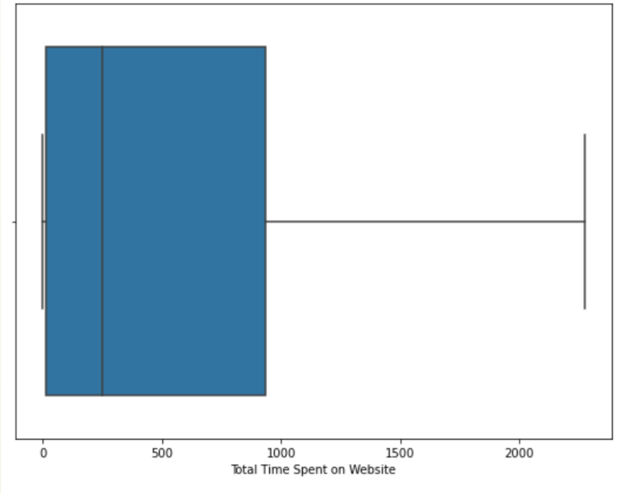


Most of the leads are unemployed.

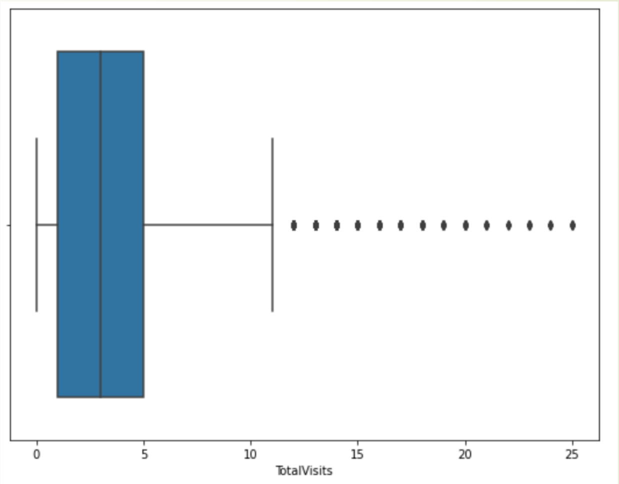


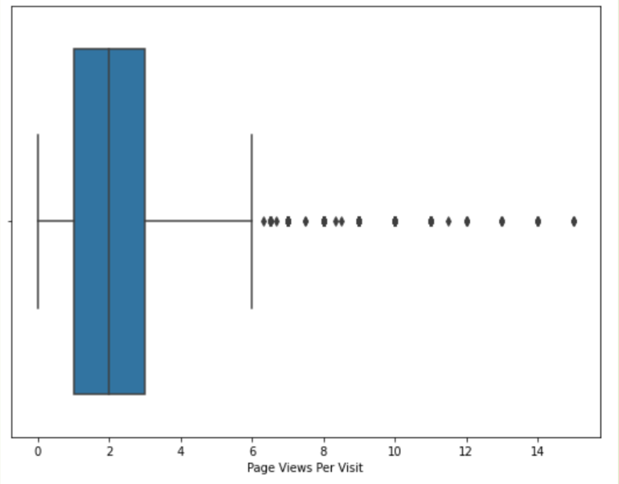
The most occurring last notable activity is Modified, with Email Opened and SMS Sent being second and third respectively.

75% of the leads spent less than 1000 seconds on X Education website

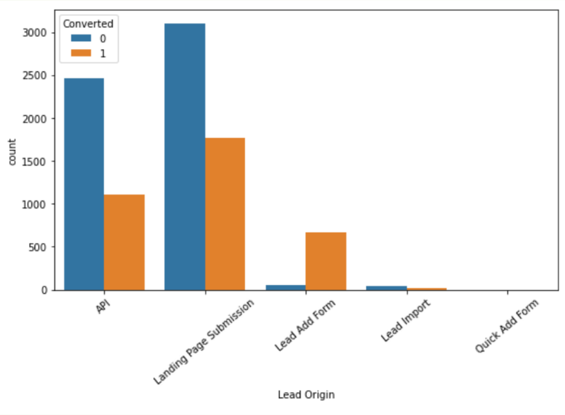


After getting rid of most outliers using the IQR method, we can see that 75% of the leads visit the website less than 5 times and view less than 6 pages per visit.

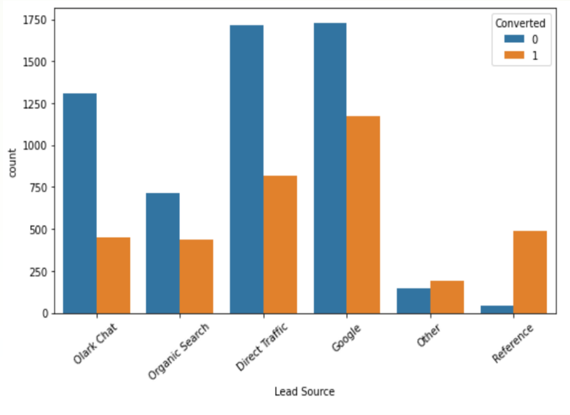




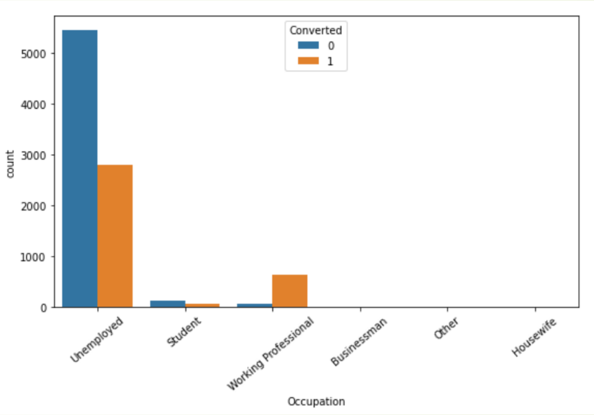
3.3.2 Bivariate Analysis



In the column 'Lead Origin', much more leads that originated from the lead add forms got converted than not. Lead add form can be a good indicator of a lead becoming converted.



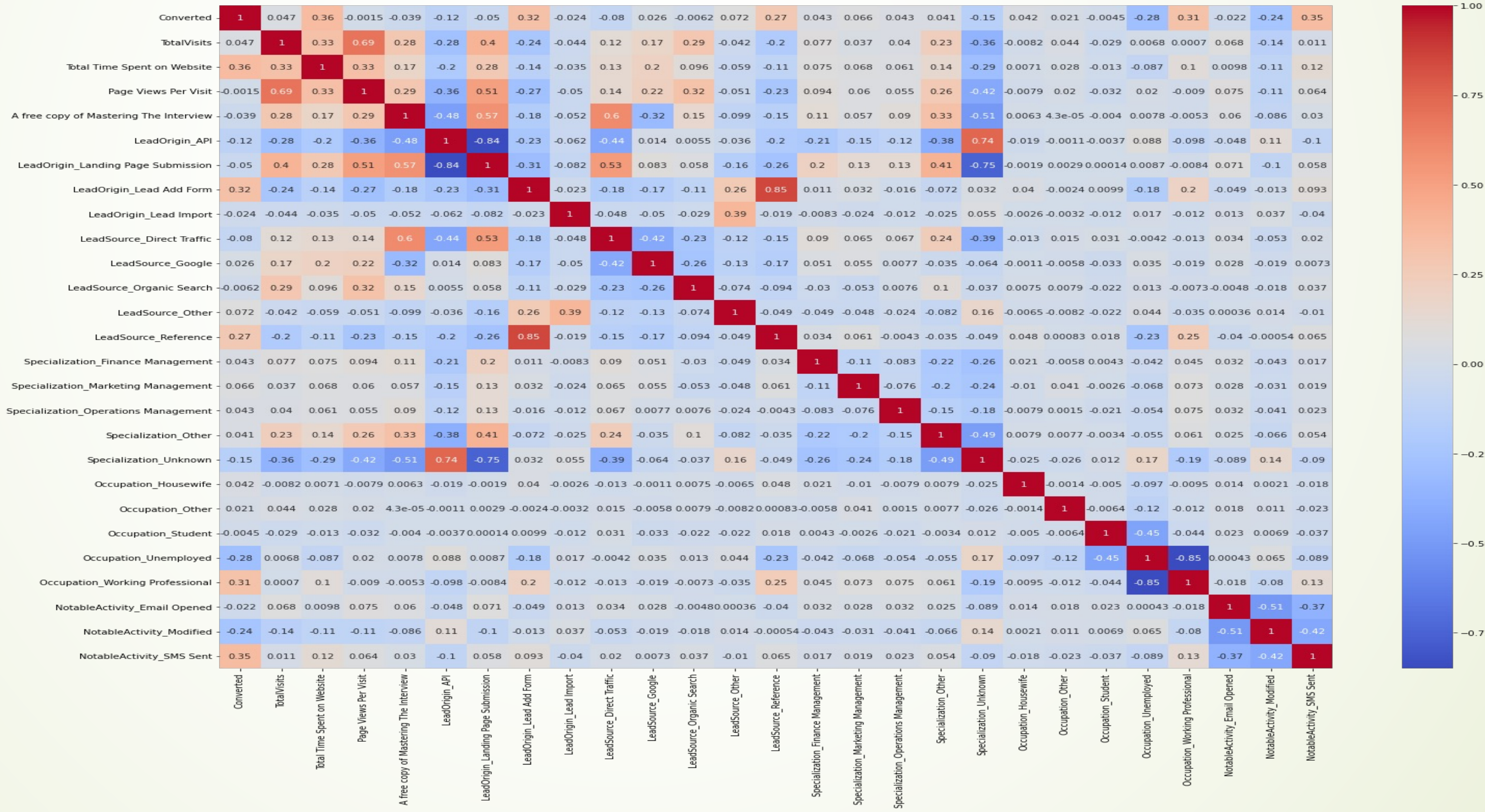
The percentage of converted leads is much higher than non-converted among the leads that were referred to X Education. Being referred can be a good indicator of a future conversion.



Most leads among working professionals got converted, which can be explained by their motivation to get better career prospects at the place of their current employment. Lead being a working professional can be a good indicator of a lead becoming converted.

3.4 Pre-processing

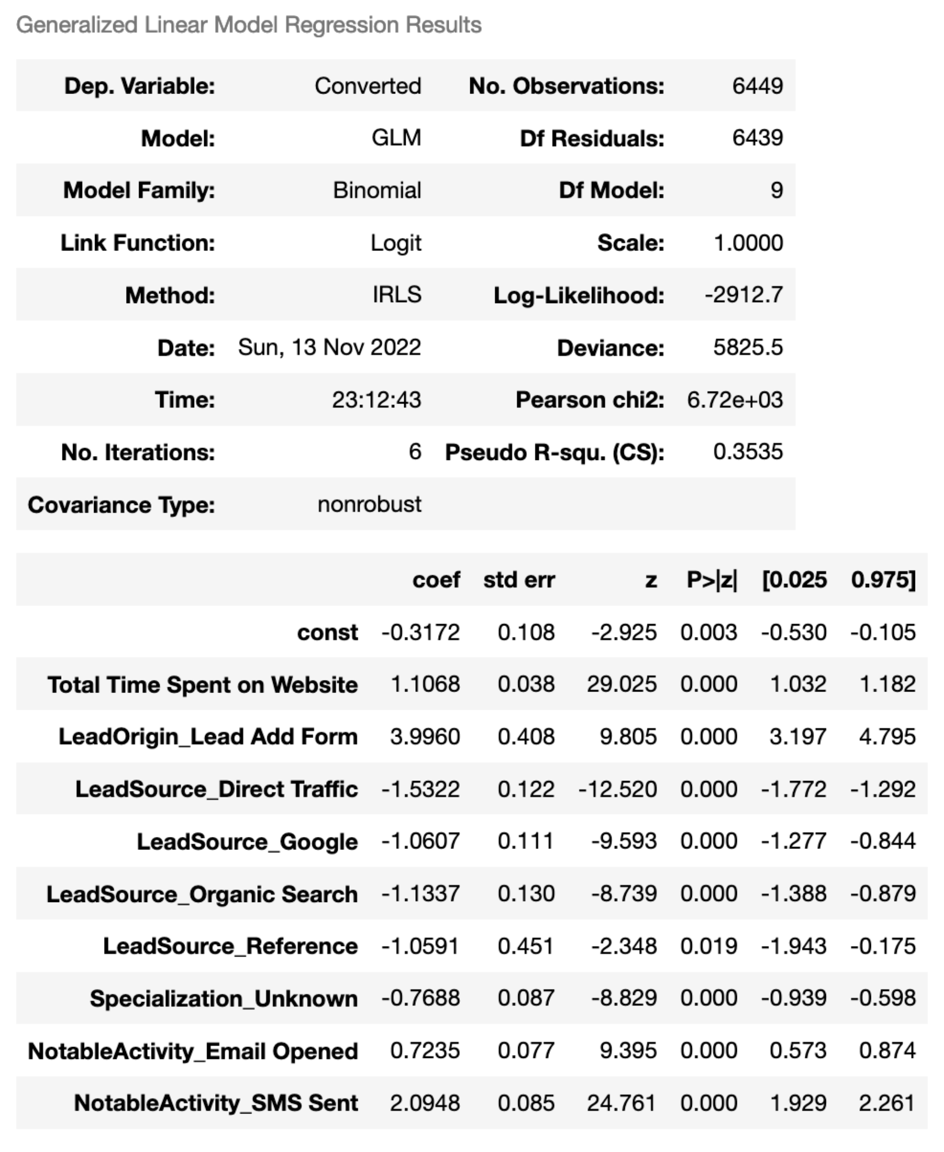
- Train-test split with 0.7 ratio was applied.
- The numeric variables were scaled using the Standard Scaler.
- The correlation matrix was drawn. Even though some variables had high correlation, it was decided not to remove them and let the EDA deal with them.



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	LeadOrigin_API	LeadOrigin_Landing Page Submission	LeadOrigin_Lead Add Form	LeadOrigin_Lead Import	LeadSource_Direct Traffic	LeadSource_Google	LeadSource_Organic Search	LeadSource_Other	LeadSource_Reference	Specialization_Finance Management	Specialization_Marketing Management	Specialization_Operations Management	Specialization_Other	Specialization_Unknown	Occupation_Housewife	Occupation_Other	Occupation_Student	Occupation_Unemployed	Occupation_Working Professional	NotableActivity_Email Opened	NotableActivity_Modified	NotableActivity_SMS Sent
Converted	1	0.047	0.36	-0.0015	-0.039	-0.12	-0.05	0.32	-0.024	-0.08	0.026	-0.0062	0.072	0.27	0.043	0.066	0.043	0.041	-0.15	0.042	0.021	-0.0045	-0.28	0.31	-0.022	-0.24	0.35
TotalVisits	0.047	1	0.33	0.69	0.28	-0.28	0.4	-0.24	-0.044	0.12	0.17	0.29	-0.042	-0.2	0.077	0.037	0.04	0.23	-0.36	-0.0082	0.044	-0.029	0.0068	0.0007	0.068	-0.14	0.011
Total Time Spent on Website	0.36	0.33	1	0.33	0.17	-0.2	0.28	-0.14	-0.035	0.13	0.2	0.096	-0.059	-0.11	0.075	0.068	0.061	0.14	-0.29	0.0071	0.028	-0.013	-0.087	0.1	0.0098	-0.11	0.12
Page Views Per Visit	-0.0015	0.69	0.33	1	0.29	-0.36	0.51	-0.27	-0.05	0.14	0.22	0.32	-0.051	-0.23	0.094	0.06	0.055	0.26	-0.42	-0.0079	0.02	-0.032	0.02	-0.009	0.075	-0.11	0.064
A free copy of Mastering The Interview	-0.039	0.28	0.17	0.29	1	-0.48	0.57	-0.18	-0.052	0.6	-0.32	0.15	-0.099	-0.15	0.11	0.057	0.09	0.33	-0.51	0.0063	4.3e-05	-0.004	0.0078	-0.0053	0.06	-0.086	0.03
LeadOrigin_API	-0.12	-0.28	-0.2	-0.36	-0.48	1	-0.84	-0.23	-0.062	-0.44	0.014	0.0055	-0.036	-0.2	-0.21	-0.15	-0.12	-0.38	0.74	-0.019	-0.0011	-0.0037	0.088	-0.098	-0.048	0.11	-0.1
LeadOrigin_Landing Page Submission	-0.05	0.4	0.28	0.51	0.57	-0.84	1	-0.31	-0.082	0.53	0.083	0.058	-0.16	-0.26	0.2	0.13	0.13	0.41	-0.75	-0.0019	0.0029	0.00014	0.0087	-0.0084	0.071	-0.1	0.058
LeadOrigin_Lead Add Form	0.32	-0.24	-0.14	-0.27	-0.18	-0.23	-0.31	1	-0.023	-0.18	-0.17	-0.11	0.26	0.85	0.011	0.032	-0.016	-0.072	0.032	0.04	-0.0024	0.0099	-0.18	0.2	-0.049	-0.013	0.093
LeadOrigin_Lead Import	-0.024	-0.044	-0.035	-0.05	-0.052	-0.062	-0.082	-0.023	1	-0.048	-0.05	-0.029	0.39	-0.019	-0.0083	-0.024	-0.012	-0.025	0.055	-0.0026	-0.0032	-0.012	0.017	-0.012	0.013	0.037	-0.04
LeadSource_Direct Traffic	-0.08	0.12	0.13	0.14	0.6	-0.44	0.53	-0.18	-0.048	1	-0.42	-0.23	-0.12	-0.15	0.09	0.065	0.067	0.24	-0.39	-0.013	0.015	0.031	-0.0042	-0.013	0.034	-0.053	0.02
LeadSource_Google	0.026	0.17	0.2	0.22	-0.32	0.014	0.083	-0.17	-0.05	-0.42	1	-0.26	-0.13	-0.17	0.051	0.055	0.0077	-0.035	-0.064	-0.0011	-0.0058	-0.033	0.035	-0.019	0.028	-0.019	0.0073
LeadSource_Organic Search	-0.0062	0.29	0.096	0.32	0.15	0.0055	0.058	-0.11	-0.029	-0.23	-0.26	1	-0.074	-0.094	-0.03	-0.053	0.0076	0.1	-0.037	0.0075	0.0079	-0.022	0.013	-0.0073	-0.0048	-0.018	0.037
LeadSource_Other	0.072	-0.042	-0.059	-0.051	-0.099	-0.036	-0.16	0.26	0.39	-0.12	-0.13	-0.074	1	-0.049	-0.049	-0.048	-0.024	-0.082	0.16	-0.0065	-0.0082	-0.022	0.044	-0.035	0.00036	0.014	-0.01
LeadSource_Reference	0.27	-0.2	-0.11	-0.23	-0.15	-0.2	-0.26	0.85	-0.019	-0.15	-0.17	-0.094	-0.049	1	0.034	0.061	-0.0043	-0.035	-0.049	0.048	0.00083	0.018	-0.23	0.25	-0.04	-0.00054	0.065
Specialization_Finance Management	0.043	0.077	0.075	0.094	0.11	-0.21	0.2	0.011	-0.0083	0.09	0.051	-0.03	-0.049	0.034	1	-0.11	-0.083	-0.22	-0.26	0.021	-0.0058	0.0043	-0.042	0.045	0.032	-0.043	0.017
Specialization_Marketing Management	0.066	0.037	0.068	0.06	0.057	-0.15	0.13	0.032	-0.024	0.065	0.055	-0.053	-0.048	0.061	-0.11	1	-0.076	-0.2	-0.24	-0.01	0.041	-0.0026	-0.068	0.073	0.028	-0.031	0.019
Specialization_Operations Management	0.043	0.04	0.061	0.055	0.09	-0.12	0.13	-0.016	-0.012	0.067	0.0077	0.0076	-0.024	-0.0043	-0.083	-0.076	1	-0.15	-0.18	-0.0079	0.0015	-0.021	-0.054	0.075	0.032	-0.041	0.023
Specialization_Other	0.041	0.23	0.14	0.26	0.33	-0.38	0.41	-0.072	-0.025	0.24	-0.035	0.1	-0.082	-0.035	-0.22	-0.2	-0.15	1	-0.49	0.0079	0.0077	-0.0034	-0.055	0.061	0.025	-0.066	0.054
Specialization_Unknown	-0.15	-0.36	-0.29	-0.42	-0.51	0.74	-0.75	0.032	0.055	-0.39	-0.064	-0.037	0.16	-0.049	-0.26	-0.24	-0.18	-0.49	1	-0.025	-0.026	0.012	0.17	-0.19	-0.089	0.14	-0.09
Occupation_Housewife	0.042	-0.0082	0.0071	-0.0079	0.0063	-0.019	-0.0019	0.04	-0.0026	-0.013	-0.0011	0.0075	-0.0065	0.048	0.021	-0.01	-0.0079	0.0079	-0.025	1	-0.0014	-0.005	-0.097	-0.0095	0.014	0.0021	-0.018
Occupation_Other	0.021	0.044	0.028	0.02	4.3e-05	-0.0011	0.0029	-0.0024	-0.0032	0.015	-0.0058	0.0079	-0.0082	0.00083	-0.0058	0.041	0.0015	0.0077	-0.026	-0.0014	1	-0.0064	-0.12	-0.012	0.018	0.011	-0.023
Occupation_Student	-0.0045	-0.029	-0.013	-0.032	-0.004	-0.0037	0.00014	0.0099	-0.012	0.031	-0.033	-0.022	-0.022	0.018	0.0043	-0.0026	-0.021	-0.0034	0.012	-0.005	-0.0064	1	-0.45	-0.044	0.023	0.0069	-0.037
Occupation_Unemployed	-0.28	0.0068	-0.087	0.02	0.0078	0.088	0.0087	-0.18	0.017	-0.0042	0.035	0.013	0.044	-0.23	-0.042	-0.068	-0.054	-0.055	0.17	-0.097	-0.12	-0.45	1	-0.85	0.00043	0.065	-0.089
Occupation_Working Professional	0.31	0.0007	0.1	-0.009	-0.0053	-0.098	-0.0084	0.2	-0.012	-0.013	-0.019	-0.0073	-0.035	0.25	0.045	0.073	0.075	0.061	-0.19	-0.0095	-0.012	-0.044	-0.85	1	-0.018	-0.08	0.13
NotableActivity_Email Opened	-0.022	0.068	0.0098	0.075	0.06	-0.048	0.071	-0.049	0.013	0.034	0.028	-0.0048	0.00036	-0.04	0.032	0.028	0.032	0.025	-0.089	0.014	0.018	0.023	0.00043	-0.018	1	-0.51	-0.37
NotableActivity_Modified	-0.24	-0.14	-0.11	-0.11	-0.086	0.11	-0.1	-0.013	0.037	-0.053	-0.019	-0.018	0.014	-0.00054	-0.043	-0.031	-0.041	-0.066	0.14	0.0021	0.011	0.0069	0.065	-0.08	-0.51	1	-0.42
NotableActivity_SMS Sent	0.35	0.011	0.12	0.064	0.03	-0.1	0.058	0.093	-0.04	0.02	0.0073	0.037	-0.01	0.065	0.017	0.019	0.023	0.054	-0.09	-0.018	-0.023	-0.037	-0.089	0.13	-0.37	-0.42	1

4. Model Building

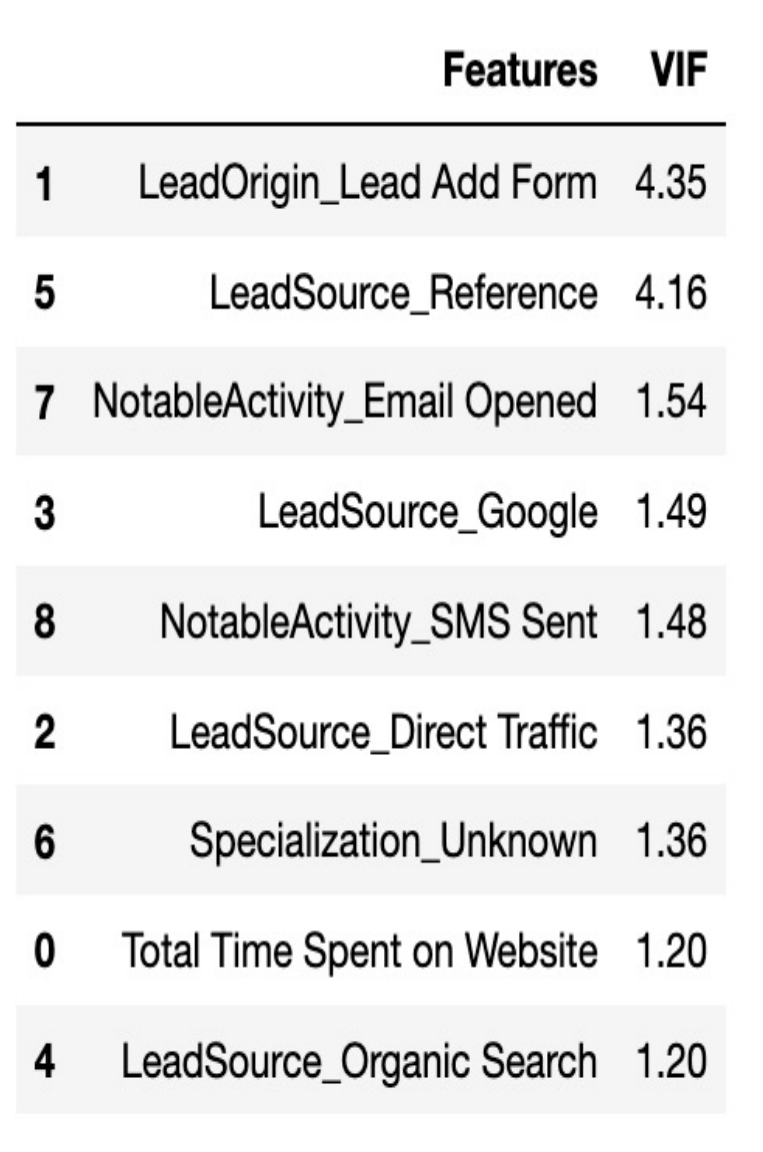
- RFE was performed to select 15 most important features.
- Logistic regression model was built using statsmodels.api.
- Multicollinearity and significance of the variables were checked using the VIF scores and p-values. If a variable was insignificant or had high multicollinearity, it was removed and the process was repeated.
- In the end, 9 features were left in the final model.



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6449
Model:	GLM	**Df Residuals:**	6439
Model Family:	Binomial	**Df Model:**	9
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2912.7
Date:	Sun, 13 Nov 2022	**Deviance:**	5825.5
Time:	23:12:43	**Pearson chi2:**	6.72e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3535
Covariance Type:	nonrobust		

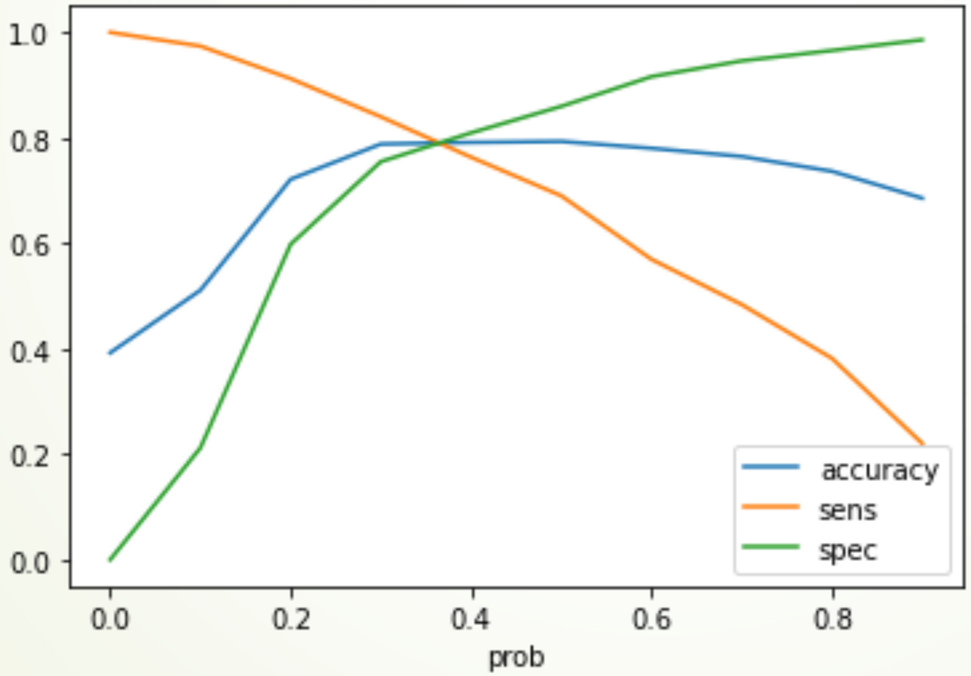
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.3172	0.108	-2.925	0.003	-0.530	-0.105
Total Time Spent on Website	1.1068	0.038	29.025	0.000	1.032	1.182
LeadOrigin_Lead Add Form	3.9960	0.408	9.805	0.000	3.197	4.795
LeadSource_Direct Traffic	-1.5322	0.122	-12.520	0.000	-1.772	-1.292
LeadSource_Google	-1.0607	0.111	-9.593	0.000	-1.277	-0.844
LeadSource_Organic Search	-1.1337	0.130	-8.739	0.000	-1.388	-0.879
LeadSource_Reference	-1.0591	0.451	-2.348	0.019	-1.943	-0.175
Specialization_Unknown	-0.7688	0.087	-8.829	0.000	-0.939	-0.598
NotableActivity_Email Opened	0.7235	0.077	9.395	0.000	0.573	0.874
NotableActivity_SMS Sent	2.0948	0.085	24.761	0.000	1.929	2.261



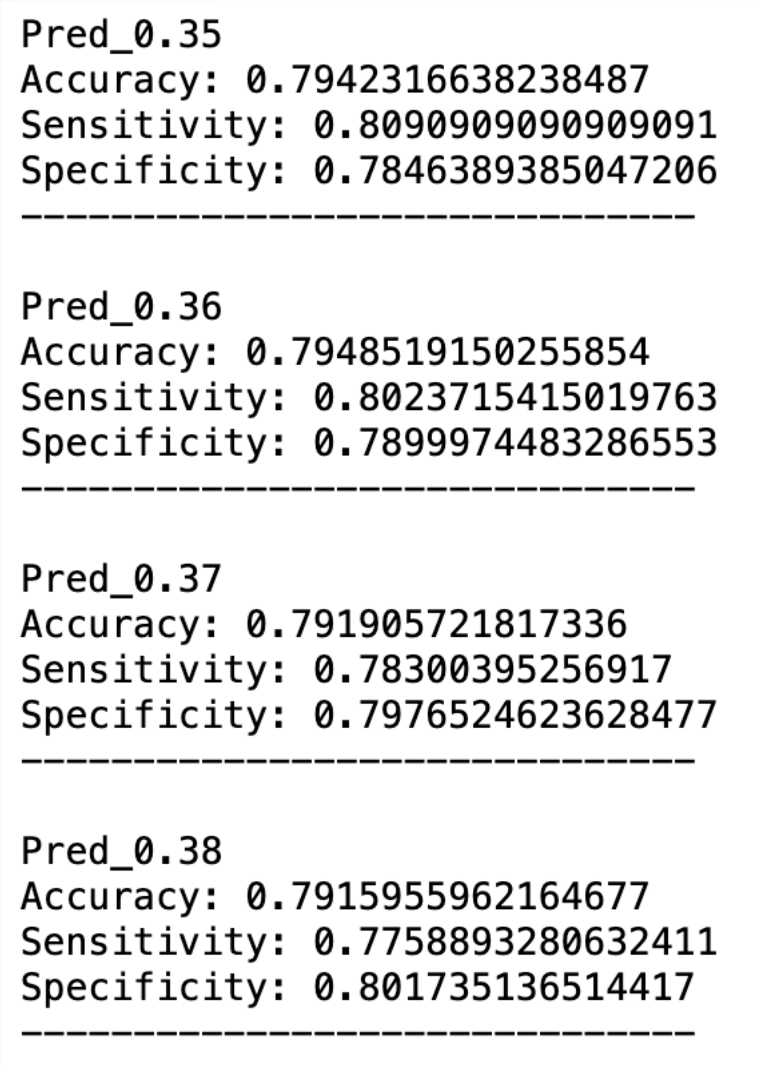
	Features	**VIF**
1	LeadOrigin_Lead Add Form	4.35
5	LeadSource_Reference	4.16
7	NotableActivity_Email Opened	1.54
3	LeadSource_Google	1.49
8	NotableActivity_SMS Sent	1.48
2	LeadSource_Direct Traffic	1.36
6	Specialization_Unknown	1.36
0	Total Time Spent on Website	1.20
4	LeadSource_Organic Search	1.20

5. Model Evaluation

- In order to evaluate the model’s performance, an optimal cutoff for probability had to be chosen.
- In order to do that, a plot of accuracy, sensitivity and specificity for all cutoffs was plotted.



- Judging by the plot, the optimal cutoff was somewhere between 0.35 and 0.38.
- In order to decide the exact cutoff, accuracy, sensitivity and specificity were calculated for the probabilities of 0.35, 0.36, 0.37, 0.38.



```
Pred_0.35
Accuracy: 0.7942316638238487
Sensitivity: 0.8090909090909091
Specificity: 0.7846389385047206
------------------------------

Pred_0.36
Accuracy: 0.7948519150255854
Sensitivity: 0.8023715415019763
Specificity: 0.7899974483286553
------------------------------

Pred_0.37
Accuracy: 0.791905721817336
Sensitivity: 0.78300395256917
Specificity: 0.7976524623628477
------------------------------

Pred_0.38
Accuracy: 0.7915955962164677
Sensitivity: 0.7758893280632411
Specificity: 0.801735136514417
------------------------------
```

0.35 was chosen as the optimal cutoff with the accuracy of ~0.79, sensitivity ~0.81 and specificity ~0.78

- Finally, the predictions were made on the test set.
- The model performed well on the test set with the following scores:
  - Accuracy – 0.78
  - Sensitivity – 0.81
  - Specificity – 0.76
- The most important variables for identifying the hot leads are:

- LeadOrigin_Lead Add Form
- NotableActivity_SMS Sent
- LeadSource_Direct Traffic
- LeadSource_Organic Search
- Total Time Spent on Website
- LeadSource_Google
- LeadSource_Reference
- Specialization_Unknown
- NotableActivity_Email Opened