Logistic Regression Report:

To create this model we need to keep in mind the end goals.

We are primarily going to classify the lead details from based on existing dataset to identify hot leads and cold leads

We have used univariate classification using logistic regression to do this classification and identify hot leads.

To implement this solution we follow the below steps:

Read the Data and do EDA:

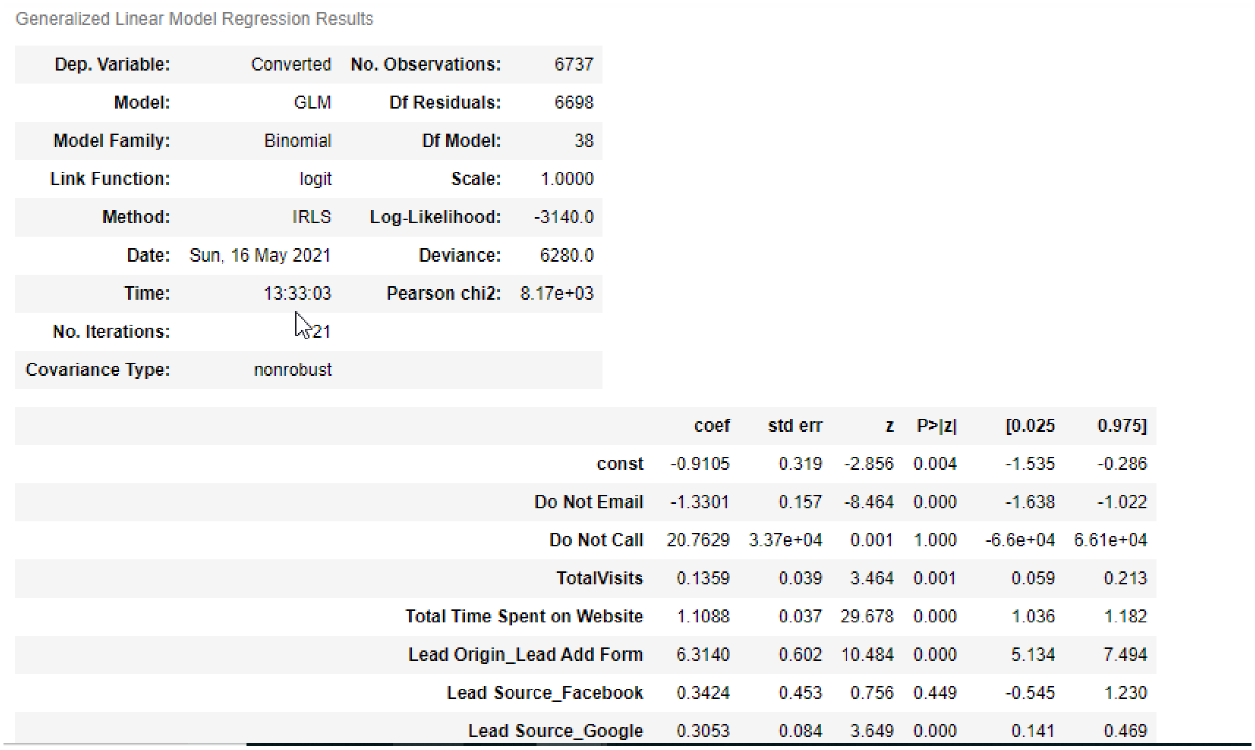
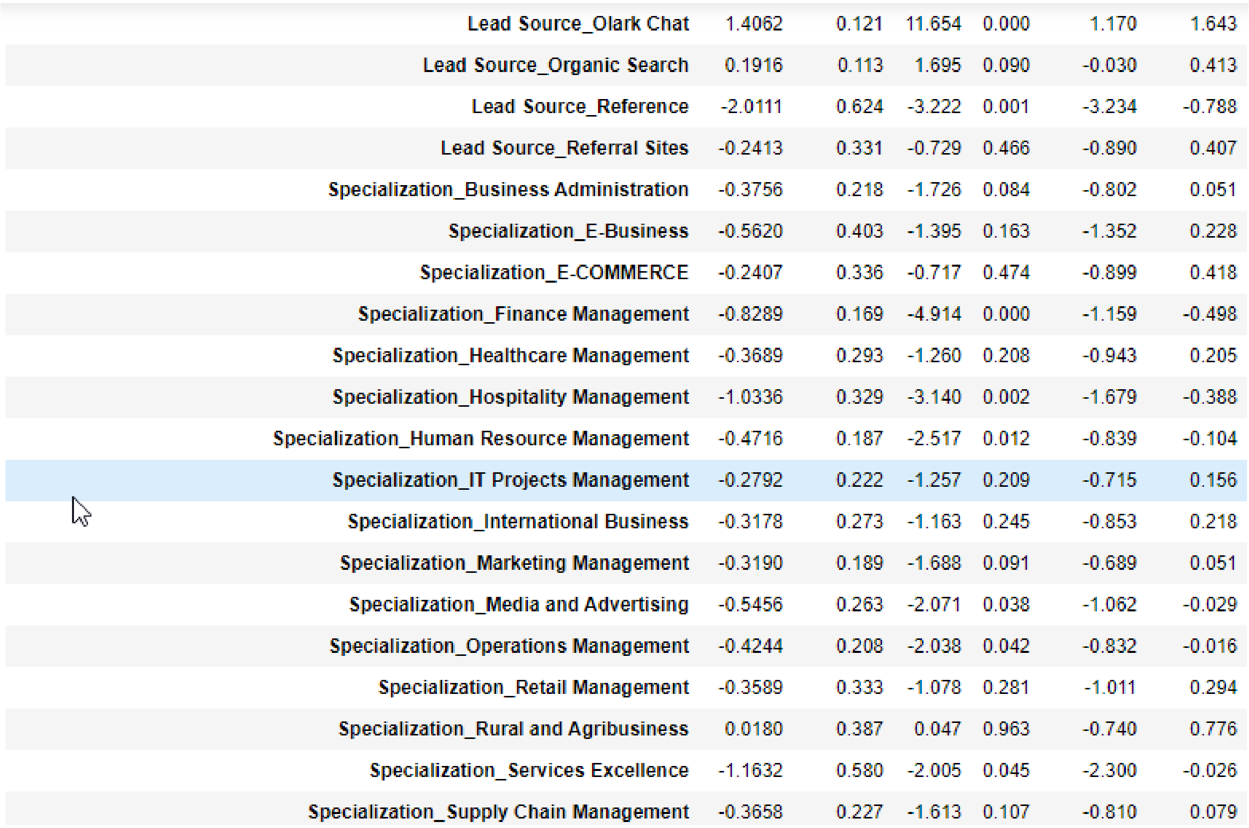
1. Identify and Handle all columns which have “Select” as column value and replace them with null values
2. Null value handling – here we have dropped all columns > 40% null values and imputed other lower % null values with “mode” method – we have used mode method instead of mean as due to the presence of low frequency high values the average gets skewed to higher value whereas that might not be the right data to impute this with.
3. Dropped all rows for columns where distribution of data is highly skewed e.g. where a particular value is available only 1 time as opposed to >100 for other values

Once this is done we are replacing Yes/No with 1/0 values and create dummy variables using one hot encoding for all the categorical values. Once we create dummy variables all original categorical columns are removed.

Once the categorical variables are removed we do outlier treatment for the numeric variables

Once all the above steps are done, we split the dataset and do GLM modelling for logistic regression using stats modelling to get an idea of how these features are affecting the model

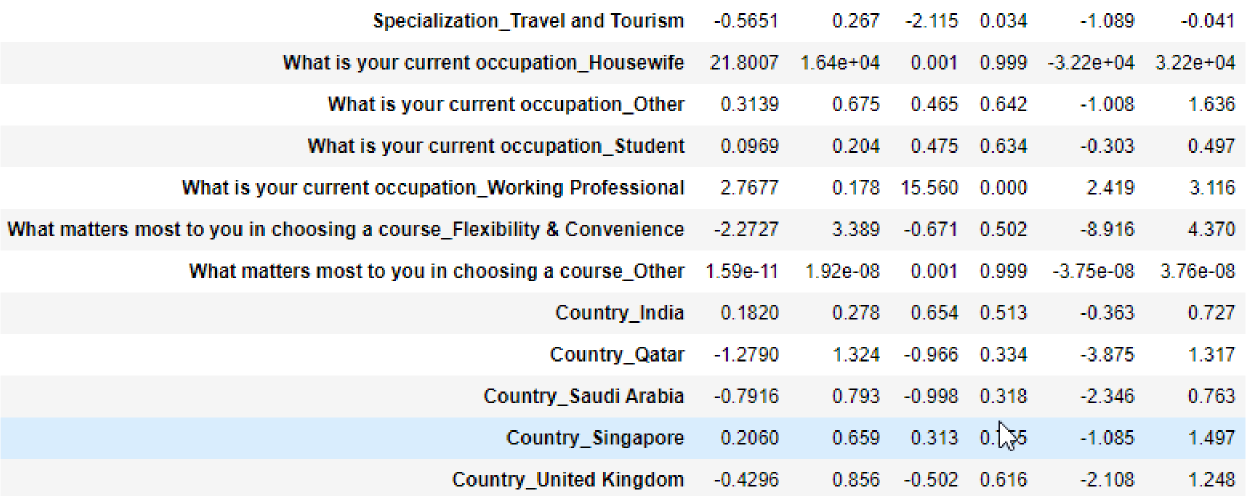
Following are the observations from first model:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6737
Model:	GLM	**Df Residuals:**	6698
Model Family:	Binomial	**Df Model:**	38
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3140.0
Date:	Sun, 16 May 2021	**Deviance:**	6280.0
Time:	13:33:03	**Pearson chi2:**	8.17e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9105	0.319	-2.856	0.004	-1.535	-0.286
Do Not Email	-1.3301	0.157	-8.464	0.000	-1.638	-1.022
Do Not Call	20.7629	3.37e+04	0.001	1.000	-6.6e+04	6.61e+04
TotalVisits	0.1359	0.039	3.464	0.001	0.059	0.213
Total Time Spent on Website	1.1088	0.037	29.678	0.000	1.036	1.182
Lead Origin_Lead Add Form	6.3140	0.602	10.484	0.000	5.134	7.494
Lead Source_Facebook	0.3424	0.453	0.756	0.449	-0.545	1.230
Lead Source_Google	0.3053	0.084	3.649	0.000	0.141	0.469
Lead Source_Olark Chat	1.4062	0.121	11.654	0.000	1.170	1.643
Lead Source_Organic Search	0.1916	0.113	1.695	0.090	-0.030	0.413
Lead Source_Reference	-2.0111	0.624	-3.222	0.001	-3.234	-0.788
Lead Source_Referral Sites	-0.2413	0.331	-0.729	0.466	-0.890	0.407
Specialization_Business Administration	-0.3756	0.218	-1.726	0.084	-0.802	0.051
Specialization_E-Business	-0.5620	0.403	-1.395	0.163	-1.352	0.228
Specialization_E-COMMERCE	-0.2407	0.336	-0.717	0.474	-0.899	0.418
Specialization_Finance Management	-0.8289	0.169	-4.914	0.000	-1.159	-0.498
Specialization_Healthcare Management	-0.3689	0.293	-1.260	0.208	-0.943	0.205
Specialization_Hospitality Management	-1.0336	0.329	-3.140	0.002	-1.679	-0.388
Specialization_Human Resource Management	-0.4716	0.187	-2.517	0.012	-0.839	-0.104
Specialization_IT Projects Management	-0.2792	0.222	-1.257	0.209	-0.715	0.156
Specialization_International Business	-0.3178	0.273	-1.163	0.245	-0.853	0.218
Specialization_Marketing Management	-0.3190	0.189	-1.688	0.091	-0.689	0.051
Specialization_Media and Advertising	-0.5456	0.263	-2.071	0.038	-1.062	-0.029
Specialization_Operations Management	-0.4244	0.208	-2.038	0.042	-0.832	-0.016
Specialization_Retail Management	-0.3589	0.333	-1.078	0.281	-1.011	0.294
Specialization_Rural and Agribusiness	0.0180	0.387	0.047	0.963	-0.740	0.776
Specialization_Services Excellence	-1.1632	0.580	-2.005	0.045	-2.300	-0.026
Specialization_Supply Chain Management	-0.3658	0.227	-1.613	0.107	-0.810	0.079



Specialization_Travel and Tourism	-0.5651	0.267	-2.115	0.034	-1.089	-0.041
What is your current occupation_Housewife	21.8007	1.64e+04	0.001	0.999	-3.22e+04	3.22e+04
What is your current occupation_Other	0.3139	0.675	0.465	0.642	-1.008	1.636
What is your current occupation_Student	0.0969	0.204	0.475	0.634	-0.303	0.497
What is your current occupation_Working Professional	2.7677	0.178	15.560	0.000	2.419	3.116
What matters most to you in choosing a course_Flexibility & Convenience	-2.2727	3.389	-0.671	0.502	-8.916	4.370
What matters most to you in choosing a course_Other	1.59e-11	1.92e-08	0.001	0.999	-3.75e-08	3.76e-08
Country_India	0.1820	0.278	0.654	0.513	-0.363	0.727
Country_Qatar	-1.2790	1.324	-0.966	0.334	-3.875	1.317
Country_Saudi Arabia	-0.7916	0.793	-0.998	0.318	-2.346	0.763
Country_Singapore	0.2060	0.659	0.313	0.755	-1.085	1.497
Country_United Kingdom	-0.4296	0.856	-0.502	0.616	-2.108	1.248

In the cursory view we notice the positive impact based on coefficients – working professional, total time spent and lead origin ad form.

Now based on RFE we do feature selection and see that the above mentioned features are selected.

Now based on ROC model we calculate the threshold probability for identifying positive predictions being 30% giving us an optimum ROC level with area under curve being around 0.84

The model developed gives below observations:

Final Observation:

Train Data:

Accuracy: 79.70%

Sensitivity: 76.51%

Specificity: 81.70%

Test Data:

Accuracy: 79.61%

Sensitivity: 70.68%

Specificity: 85.17%

The model seems to predict the conversion rate very well