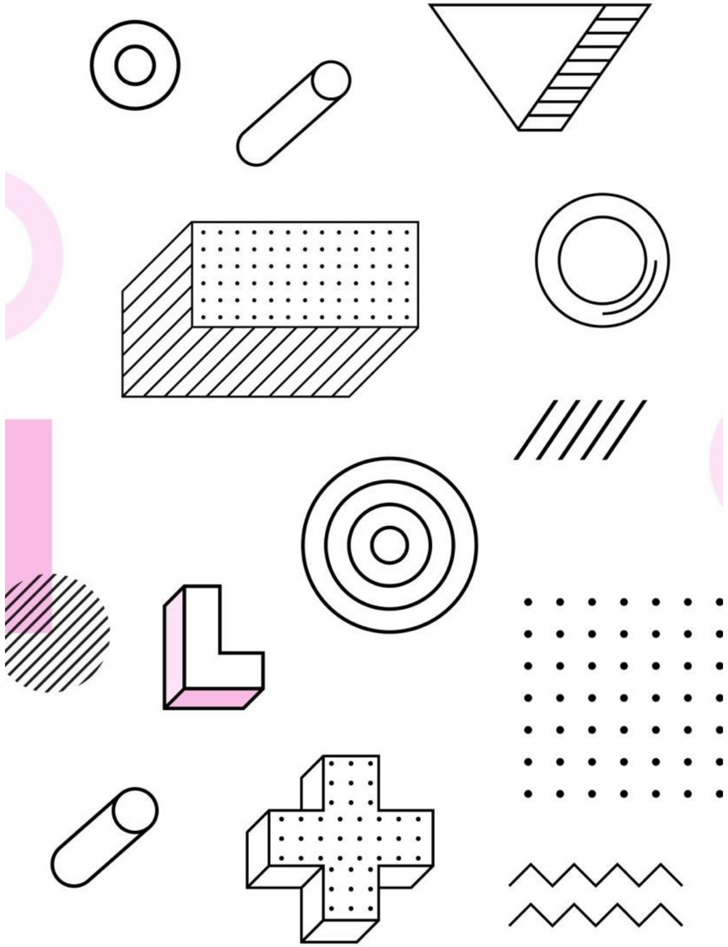


Lead Scoring Case Study

BY- ANURAG AGARWAL & SAYI TTEJA VEPA

Index

- Problem Statement
- Data Preparation
- Model Building and Validation
- Final Model
- Recommendations

Problem Statement

- X Education, an education company that sells online courses to industry professionals, has appointed us to help them select the most promising leads.
- The company requires us to build a model wherein we need 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.
- The CEO has given a ballpark of the target lead conversion rate to be around 80% (currently at 30%).
- There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

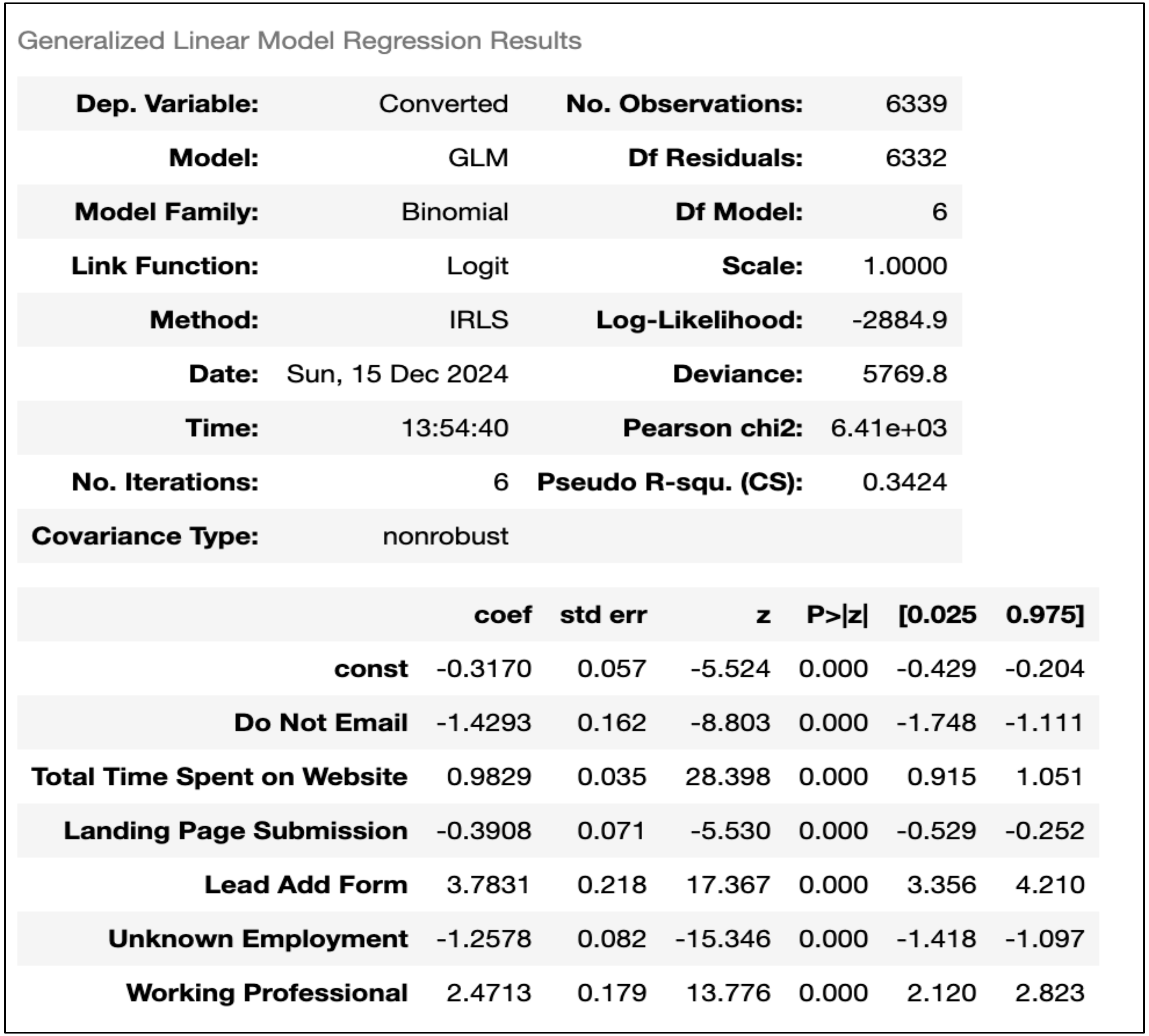
Data Preparation

- Dropped columns with large percent of null values
- Dropped categorical columns with only all or most values of single category
- Dropped unnecessary columns such as Prospect ID, Last Activity, Country, Specialization etc.
- Converted binary variables Yes/ No to 0/1 for conversion to numerical variables
- Converted categorical variables to dummy variables
- Handled outliers in columns by capping till 99.5th percentile
- Performed test train split and scaled features

Model Building and Validation

- Used statsmodel to create the first logistic regression model on the train data set and dropped irrelevant columns
- Used Recursive feature elimination (RFE) to select the 7 most important features
- Checked VIF of features to eliminate multicollinearity
- Assessed the model with remaining columns and removed any columns with high P values
- Utilized the precision-recall curve to find out the optimal cut-off for lead score for required precision value of 80 percent.
- Checked the precision value of the model against test data set to check if it meets the required criteria.

Final Model

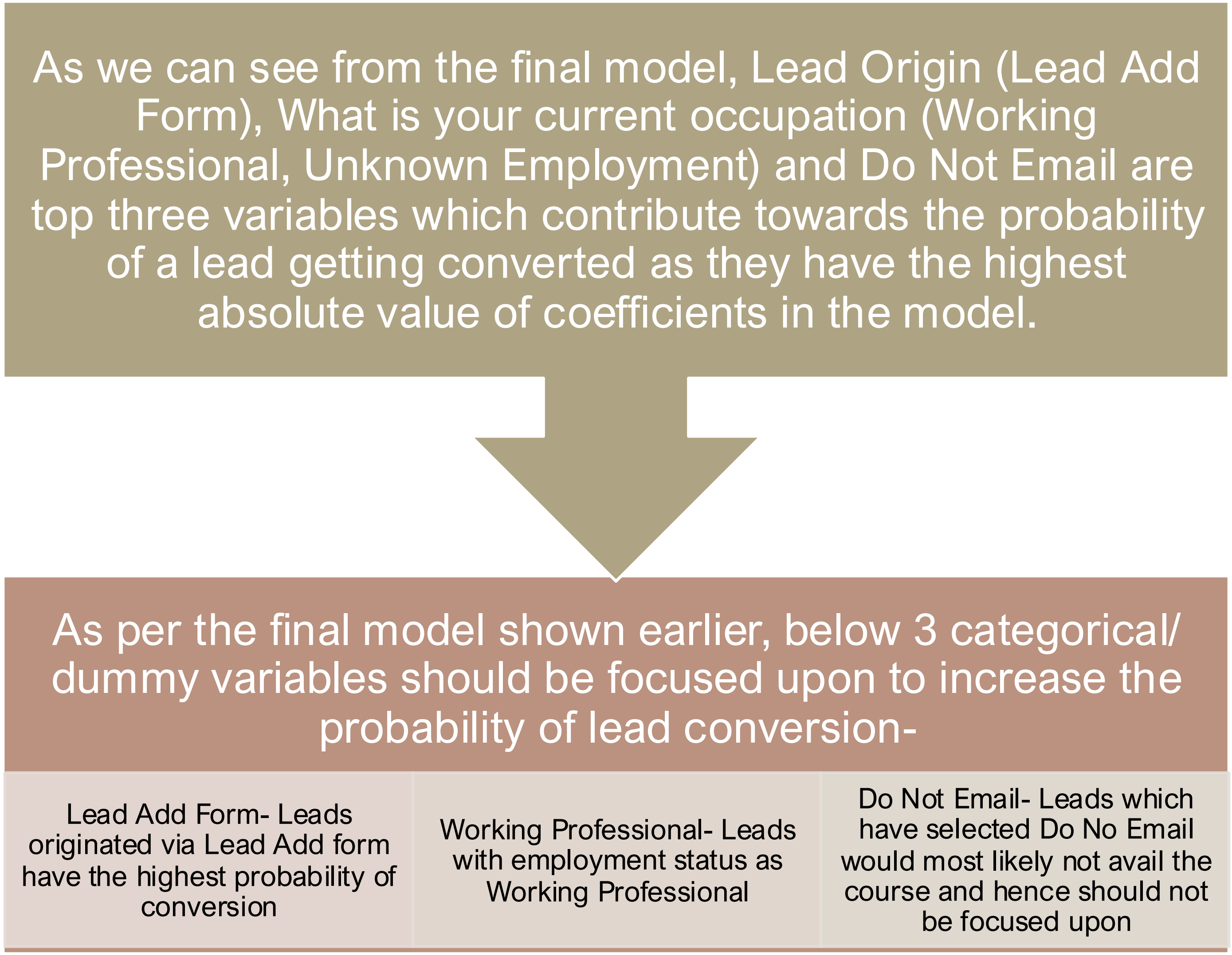


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6339
Model:	GLM	**Df Residuals:**	6332
Model Family:	Binomial	**Df Model:**	6
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2884.9
Date:	Sun, 15 Dec 2024	**Deviance:**	5769.8
Time:	13:54:40	**Pearson chi2:**	6.41e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3424
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.3170	0.057	-5.524	0.000	-0.429	-0.204
Do Not Email	-1.4293	0.162	-8.803	0.000	-1.748	-1.111
Total Time Spent on Website	0.9829	0.035	28.398	0.000	0.915	1.051
Landing Page Submission	-0.3908	0.071	-5.530	0.000	-0.529	-0.252
Lead Add Form	3.7831	0.218	17.367	0.000	3.356	4.210
Unknown Employment	-1.2578	0.082	-15.346	0.000	-1.418	-1.097
Working Professional	2.4713	0.179	13.776	0.000	2.120	2.823

Recommendations



As we can see from the final model, Lead Origin (Lead Add Form), What is your current occupation (Working Professional, Unknown Employment) and Do Not Email are top three variables which contribute towards the probability of a lead getting converted as they have the highest absolute value of coefficients in the model.

As per the final model shown earlier, below 3 categorical/ dummy variables should be focused upon to increase the probability of lead conversion-

Lead Add Form- Leads originated via Lead Add form have the highest probability of conversion

Working Professional- Leads with employment status as Working Professional

Do Not Email- Leads which have selected Do No Email would most likely not avail the course and hence should not be focused upon

Recommendations

Strategy to increase conversion by additional interns: Since we are having additional resources in our team in form of interns, we suggest lowering the threshold of lead score from the existing value of 60 (for the required conversion rate of 80%). Since, as per the precision-recall curve, optimum cut off point is 40, we suggest taking the cut-off point as 40 to increase the number of probable leads we can call.

Strategy to minimize useless phone calls during quarter end: Given we need to reduce the number of useless phone calls, we need to maximize the conversion rate i.e. precision value for this case. We suggest taking the cut-off point for lead score as 80 which would result in precision value of 88% and hence result in only c. 12% incorrect calls.