Lead Scoring Case Study

Presented by: Le Thinh Phat

Problem Statement

Company Background:

- X Education, an online course provider for industry professionals, generates many leads daily through various marketing channels like websites, Google, and referrals.
- When potential customers (leads) show interest, such as by filling out a form, they are passed on to the sales team for further communication (calls, emails).

Problem Statement

Current Challenge:

- Despite generating a large number of leads, the company struggles with a low conversion rate of around 30% (i.e., out of every 100 leads, only about 30 convert into paying customers).
- This inefficient process results in wasted effort, as the sales team spends time on leads that are unlikely to convert.

Objective

- X Education wants to optimize its sales process by identifying "Hot Leads", which are the most promising leads with a higher likelihood of conversion.
- By focusing on these potential leads, the company aims to increase the conversion rate to around 80%.

Approach & Methodology

- Analysis Problem Statement.
- Loading data.
- Exploring Data with Column Structure/Column Types/Value Types.
- Data Cleaning and Preparation:
    - Handle missing value:
        - Drop columns with more than 3000 missing values.
        - Drop “City” and “Country” columns because they would not be used in analysis.
        - Drop columns with uninformative categorical data
        - Drop rows with null values in “What is your current occupation”, “TotalVisits”, “Lead Source”, “Specialization”

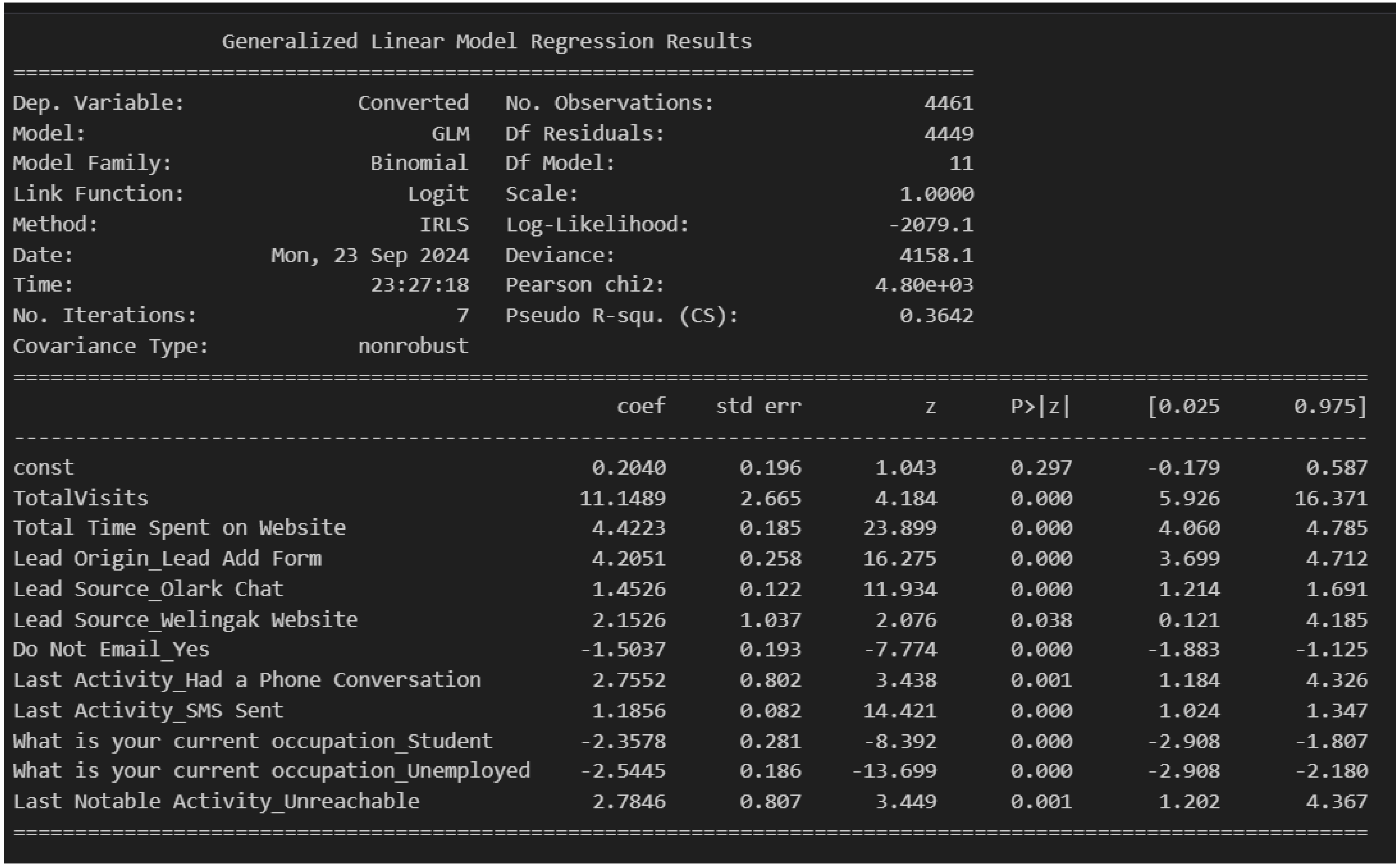
Approach & Methodology

- Categorical Variable Handling:
  - Created dummy variables for categorical columns.
- Training and Test Split: 70% for training, 30% for testing.
    - Model Selection: Logistic Regression was chosen due to the binary nature of the target variable “Converted”.
- Scaling:
    - Scale the three numeric features: “TotalVisits”, “Total Time Spent on Website”, “Page Views Per Visit”.
    - Fit and transform the training data

Approach & Methodology

- Use the Logistic Regression for Model Building
- Use VIF and P-values for Model Validation:
    - Variance Inflation Factor (VIF):
        - Ensured no multicollinearity.
        - All VIF values were below 5.
    - P-values:
        - All significant features had p-values < 0.05.
- Model Evaluation.
- Making Predictions on the Test Set

Explain the results



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4461
Model:                            GLM   Df Residuals:                     4449
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2079.1
Date:                Mon, 23 Sep 2024   Deviance:                       4158.1
Time:                        23:27:18   Pearson chi2:                 4.80e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3642
Covariance Type:            nonrobust
============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       0.2040      0.196      1.043      0.297      -0.179       0.587
TotalVisits                                11.1489      2.665      4.184      0.000       5.926      16.371
Total Time Spent on Website                 4.4223      0.185     23.899      0.000       4.060       4.785
Lead Origin_Lead Add Form                   4.2051      0.258     16.275      0.000       3.699       4.712
Lead Source_Olark Chat                      1.4526      0.122     11.934      0.000       1.214       1.691
Lead Source_Welingak Website                2.1526      1.037      2.076      0.038       0.121       4.185
Do Not Email_Yes                           -1.5037      0.193     -7.774      0.000      -1.883      -1.125
Last Activity_Had a Phone Conversation      2.7552      0.802      3.438      0.001       1.184       4.326
Last Activity_SMS Sent                      1.1856      0.082     14.421      0.000       1.024       1.347
What is your current occupation_Student    -2.3578      0.281     -8.392      0.000      -2.908      -1.807
What is your current occupation_Unemployed -2.5445      0.186    -13.699      0.000      -2.908      -2.180
Last Notable Activity_Unreachable           2.7846      0.807      3.449      0.001       1.202       4.367
============================================================================================================
```

This is the Generalized Linear Model Regression Results

P-values:

- P-values indicate the significance of each feature (independent variable) in predicting the dependent variable “Converted”.

- Lower p-values (< 0.05) indicate statistically significant predictors, meaning that changes in these variables are associated with changes in the likelihood of conversion.

Explain the results

- Significant Predictors (P < 0.05):
  - Total Visits (0.000): Highly significant; more visits increase the likelihood of conversion.
  - Total Time Spent on Website (0.000): Highly significant; more time on the website significantly correlates with conversion.
  - Lead Origin_Lead Add Form (0.000): This lead origin type is significant in predicting conversion.
- And with the Positive Coefficients: Increase the probability of conversion.
  - Total Visits (11.1489) and Total Time Spent on Website (4.4223) show that more visits and time increase the likelihood of conversion.
  - Lead Origin_Lead Add Form (4.2057): this one demonstrate a notably higher conversion rate, indicating a clear intent to enroll or seek more information.

Conclusion

In order to increase the probability of lead conversion, the top variables to prioritize for enhancing are:

1. Lead Origin_Lead Add Form: Both the RFE and logistic regression results indicate that this variable has a significant impact on conversion. Focusing on leads generated through this source is likely to produce better outcomes.

2. Last Activity_SMS Sent: The model shows that leads receiving an SMS have a higher likelihood of conversion, making SMS outreach an effective strategy.

3. Lead Source_Olark Chat: This source is linked to higher conversion rates, suggesting that engaging with leads through live chat positively influences conversion.