In this assignment, I addressed a critical business problem for X Education, an education company offering online courses to industry professionals. The company faced a low lead conversion rate of about 30% and sought to improve this by identifying “hot leads”—those with the highest potential for conversion into paying customers. The objective was to develop a predictive model to score leads based on their likelihood of conversion, helping the sales team focus on high-potential leads and ultimately increase the conversion rate to around 80%.

My approach:

Step 1: Data Cleaning and Preparation

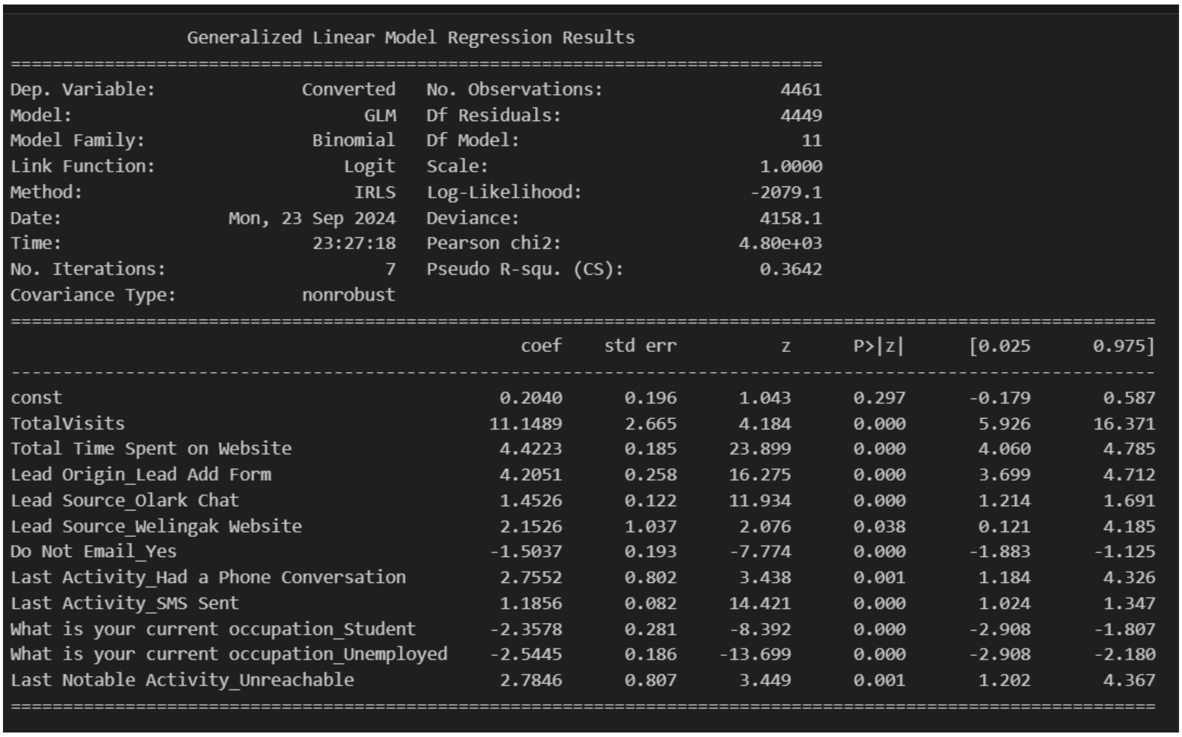
The initial dataset contained various features such as lead source, lead activity, and demographic information like occupation. I started by handling missing values and irrelevant features. Columns with a significant percentage of missing values, like "City" and "Country," were removed. Additionally, columns with dominant values, such as "Do Not Call" and "Magazine," were also dropped, as they provided little predictive value.

After data cleaning, about 69% of the original rows were retained, ensuring that the dataset was still representative. I then created dummy variables for categorical columns, followed by splitting the data into 70% training and 30% test sets. I scaled the numeric features, including “TotalVisits,” “Total Time Spent on Website,” and “Page Views Per Visit,” to standardize the values. The chosen model for prediction was logistic regression, suitable for the binary nature of the target variable, “Converted.”

Step 2: Model Building

Using logistic regression, I aimed to identify key features influencing conversion and assign a lead score to each prospect. Initially, all features were included in the model. Through feature refinement using p-values and variance inflation factor (VIF) analysis, I eliminated insignificant features and addressed multicollinearity.

Key features retained in the final model included:



```
                 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
============================================================================================================
```

Step 3: Model Evaluation

The final logistic regression model performed well, successfully identifying leads most likely to convert based on their online behavior and interaction with the company’s marketing efforts. Model performance was evaluated using statistical metrics like pseudo R-squared and log-likelihood, which indicated a reasonable fit to the data.

Step 4: Making Predictions on the Test Set

To assess the model’s generalizability, I applied the trained logistic regression model to the test set. The results showed the model's ability to accurately predict conversion likelihood in unseen data, further validating its effectiveness.

Key Learnings

This assignment reinforced the importance of data cleaning and preprocessing, especially when handling missing values and removing redundant features. I also deepened my understanding of logistic regression and feature selection techniques such as p-values and VIF. Additionally, I gained valuable insight into how predictive modeling can address real-world business challenges, driving operational efficiency and improving performance metrics, like lead conversion rates.