Problem & Hypothesis

Utilizing the research dataset of Online Purchasing Intention from the UC Irvine Machine Learning Repository, the research study proposed the question, "Which attributes within the captured Google Analytics metrics could contribute to the possibility of a customer purchasing an item?".

Null hypothesis: The logistic regression model shows that the predictor variable of "Revenue" and the explanatory variables do not have a statistically significant relationship.
Alternate Hypothesis: The logistic regression model shows a statistically significant relationship between the predictor variable "Revenue" and explanatory variables.

Evaluation Metrics: To reject the null hypothesis, the optimized logistic regression model accuracy score of greater than 80% and statistically significant attributes were noted. A p-value of less than 0.05 marked statistical significance. To fail to reject the hypothesis, the model would have a p-value greater than 0.05, which meant there was no statistically significant relationship between the target and explanatory variables.

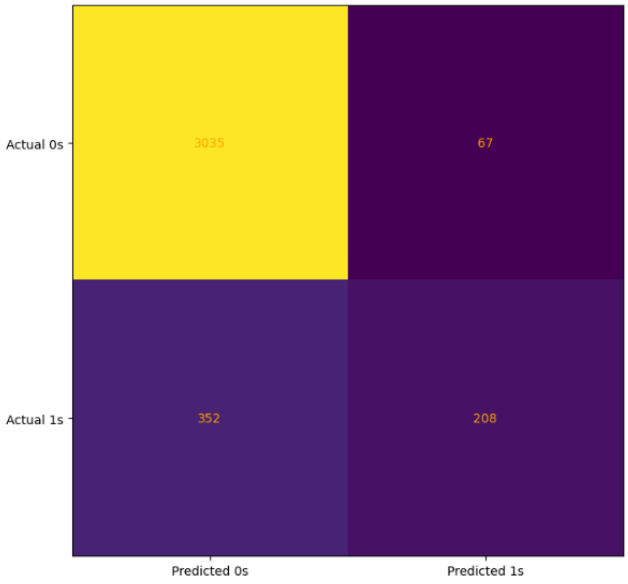
With the increasing rise of online purchasing in current society, using Google Analytics within e-commerce offered insights to increase profitability. The metrics captured the user behavior within the website. These included the number of times each page was visited, the time duration in seconds a user interacted with the page, or an item was purchased i.e., revenue was obtained. Logistic Regression could identify if any of these metrics provided a significant relationship between whether an item was purchased or not. This would be an ideal method for narrowing the scope of the metrics to be used for strategic marketing.

Data Analysis Process

The initial regression model utilized the target variable of "Revenue" along with all the explanatory variables within the prepared dataset. The initial regression model showed the various evaluation metrics identifying the goodness of fit measurement. The model revealed a LLR p-value of 0.0 deeming the overall model statistically significant and a pseudo r-squared measure of 30.34%.

```
Optimization terminated successfully.
        Current function value: 0.302019
        Iterations 8
                          Logit Regression Results
==============================================================================
Dep. Variable:                Revenue   No. Observations:                12205
Model:                          Logit   Df Residuals:                    12191
Method:                           MLE   Df Model:                           13
Date:                Sat, 15 Jul 2023   Pseudo R-squ.:                 0.3034
Time:                        11:45:32   Log-Likelihood:                -3686.1
converged:                       True   LL-Null:                       -5291.3
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Administrative                   0.0052      0.017      0.310      0.756      -0.027       0.038
Administrative_Duration         -0.0003      0.001     -0.463      0.644      -0.001       0.001
Informational                    0.0204      0.026      0.787      0.431      -0.030       0.071
Informational_Duration           0.0002      0.000      0.747      0.455      -0.000       0.001
ProductRelated                   0.0047      0.002      2.095      0.036       0.000       0.009
ProductRelated_Duration          0.0003   5.13e-05      5.188      0.000       0.000       0.000
BounceRates                    -11.4239      4.365     -2.617      0.009     -19.979      -2.869
ExitRates                      -12.4520      2.410     -5.167      0.000     -17.175      -7.729
PageValues                       0.0794      0.002     33.910      0.000       0.075       0.084
SpecialDay                      -0.8830      0.216     -4.086      0.000      -1.307      -0.459
Weekend                          0.1431      0.070      2.043      0.041       0.006       0.280
VisitorType Other               -0.7309      0.521     -1.404      0.160      -1.751       0.289
VisitorType Returning_Visitor   -0.4543      0.088     -5.155      0.000      -0.627      -0.282
const                           -2.0319      0.098    -20.722      0.000      -2.224      -1.840
================================================================================================
```

The train_test_split was generated with a 70:30 ratio. Seventy percent of the dataset went to training, with thirty percent allocated to test data. The logistic regression score had an ideal success of 88.13%. The regression model was summarized with the following prediction results: True positive predictions: 3035, True negative predictions: 208, False positive predictions: 352, and False negative predictions: 67. This was plotted within a confusion matrix for graphical visualizations. The prediction accuracy score of the model was 89% as shown in the classification report.

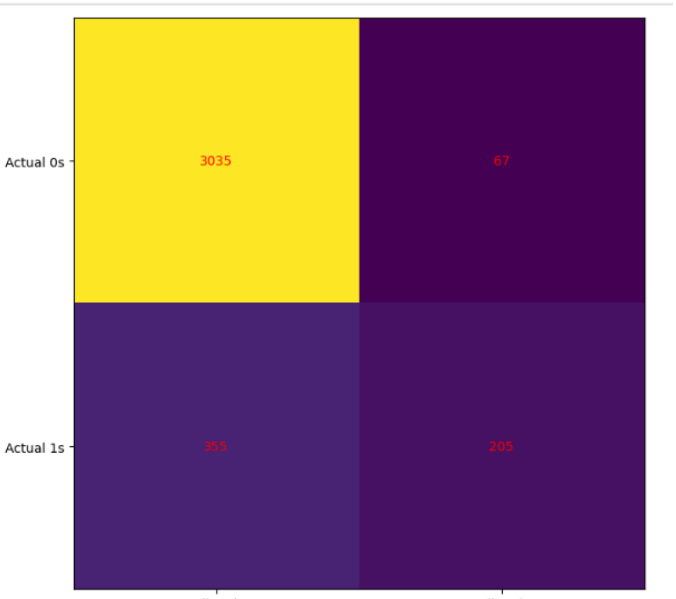


	precision	recall	f1-score	support
0	0.90	0.98	0.94	3102
1	0.76	0.37	0.50	560
accuracy			0.89	3662
macro avg	0.83	0.67	0.72	3662
weighted avg	0.87	0.89	0.87	3662

The optimized model was reduced via the backward stepwise elimination method. This reduction kept only the variables with p-values of less than 0.05. It reduced the variables from the initial model's thirteen to eight key attributes.

```
Optimization terminated successfully.
         Current function value: 0.302230
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                Revenue   No. Observations:                12205
Model:                          Logit   Df Residuals:                    12196
Method:                           MLE   Df Model:                            8
Date:                Sat, 15 Jul 2023   Pseudo R-squ.:                 0.3029
Time:                        11:46:51   Log-Likelihood:                -3688.7
converged:                       True   LL-Null:                       -5291.3
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
ProductRelated                   0.0051      0.002      2.329      0.020       0.001       0.009
ProductRelated_Duration          0.0003   5.05e-05      5.392      0.000       0.000       0.000
BounceRates                    -11.1620      4.352     -2.565      0.010     -19.692      -2.632
ExitRates                      -12.5706      2.379     -5.285      0.000     -17.232      -7.909
PageValues                       0.0795      0.002     34.035      0.000       0.075       0.084
SpecialDay                      -0.8830      0.215     -4.098      0.000      -1.305      -0.461
Weekend                          0.1486      0.070      2.124      0.034       0.012       0.286
VisitorType Returning_Visitor   -0.4317      0.087     -4.949      0.000      -0.603      -0.261
const                           -2.0585      0.094    -21.923      0.000      -2.243      -1.874
================================================================================================
```

The same train_test_split ratio was utilized to provide an adequate comparison between the modeling. The optimized logistic regression score was slightly higher at 88.47%. The optimized model's prediction results were summarized as follows: True positive predictions: 3035, True negative predictions: 205, False positive predictions: 355, and False negative predictions: 67. The prediction accuracy score was slightly lower at 88%.



	precision	recall	f1-score	support
0	0.90	0.98	0.93	3102
1	0.75	0.37	0.49	560
accuracy			0.88	3662
macro avg	0.82	0.67	0.71	3662
weighted avg	0.87	0.88	0.87	3662

Overall, the optimized modeling was deemed successful based on the z-statistics found within the logit summary along with the higher model score. The z-statistics of the optimized model provided a better statistical fit due to the strength of the predictors. A z-statistic value further from zero meant the stronger its role was as a predictor variable. These statistics were larger in the optimized model. For example: "PageValues" with a z-statistic value of "34.035" within the Optimized model, but "33.91" within the initial.

Outline of Findings

As noted previously, the optimized model's prediction accuracy score was 88%. This exceeds the evaluation metric threshold of 80%. It deemed the optimized model an effective logistic regression model. The second metric of success was the identification of attributes with statistical significance. This was noted by the eight explanatory variables with a p-value of less than 0.05. A further reduction to narrow the scope of contributing attributes was identified with the usage of z-statistics. All of these metrics were needed to successfully reject the null hypothesis.

Limitations

The main limitations of this study were related to the time period of the dataset. The data was limited to a one-year period. There was not a clear identification of what year was captured or even if this was a typical year in general. E-commerce has grown vastly in the past decade, especially during the COVID-19 pandemic. The lack of specifics to the actual year captured within the data could restrict the resulting outcomes and attributes.

Another limitation was some minor issues with the data. The data was noted as a one-year period to capture all the annual holidays. Unfortunately, the dataset was missing some key observations such as the sessions within January and April. This would lose data that could be provided by holiday sales in those months including after Christmas sales, New Year's, and Easter.

Proposed Actions

The key attributes could lead the e-commerce company to a few proposed actions. These actions could include the following:

- The investigation of contributions based on Page Values
- Identification of key features and page duration
- Various Sales/Retention discounts and advertisements.

The website pages along with associated revenue could be ranked accordingly. This could prompt analysis of those specific duration metrics and their strengths to help in marketing brainstorming. The marketing team could initiate further analyses on attributes and how a user's page duration could contribute to sales. Sales and retention teams could increase retention and sales efforts via promos and targeted advertisements.

Expected Benefits of Study

The expected benefits of this logistic regression model could relate to the identification of key attributes for improved business strategies in various departments within the e-commerce company. As noted previously, this would provide various insights into which attributes would contribute to revenue occurring. For example, the marketing team would focus on strategies to increase revenue for low-performing pages based on their specific page values. Similarly, the sales team can focus attention on the high revenue months and revenue obtained during holidays to suggest seasonal and holiday pricing. Lastly, the customer retention team can provide targeted campaigns to increase customer retention of returning customers via promo codes and other discounting techniques.

Sources

Kumar, B. (2023, March 6). Google Analytics for Ecommerce in 2023 (Complete Guide). Retrieved from Shopify: https://www.shopify.com/blog/14681601-google-analytics-for-ecommerce-a-beginners-guide

Miyasato, Kenny. (2020, April 5). Classification Report: Precision, Recall, F1-Score, Accuracy.Received from Medium: https://medium.com/@kennymiyasato/classification-report-precision-recall-f1-score-accuracy-16a245a437a5

Sakar, C., & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset. Retrieved from UCI Machine Learning Repository: https://doi.org/10.24432/C5F88Q