Logistic Regression

Predictive analytics and classification frequently employ this kind of statistical model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. The natural logarithm of odds or the log odds are other names for this. (Swaminthan,2018)

When a dependent variable is dichotomous, the proper regression analysis to use is logistic regression (binary). The logistic regression is a predictive analysis, just like all regression analyses. To describe data and explain the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables, we employ logistic regression. Logistic regression is used when the dependant variable is categorical. (Thanda,2022)

For Example,

- To predict whether a SMS is spam (1) or true (0)
- Whether a tumour is malignant (1) or not (0)

The different types of Logistic regression include:

- Binary logistic Regression. Is the statistical method used to forecast the relationship between the independent variable (X) and the binary dependent variable (Y). (IBM,2022) ,(Thanda,2022)
- Multinominal logistic regression. Used when a categorical dependent variable has two or more unordered levels. (Swaminthan,2018)
- Ordinal logistic regression. Used when Y, the dependent variable, is ordered, is used.

Regression analysis comes in a variety of forms, as does logistic regression. Regression models should be chosen carefully depending on the dependent and independent variables in your data. (IBM,2022)

Advantages of logistic regression:

- Particularly in the context of machine learning, logistic regression is substantially simpler to implement than other approaches. (Thanda,2022)
- When the dataset can be linearly separated, logistic regression is effective.
- Logistic regression offers insightful information. (Swaminthan,2018)

Disadvantages of logistic regression:

- A continuous result cannot be predicted by logistic regression. (Swaminthan,2018)
- Between the predictor (independent) variables and the dependent (dependent) variables, logistic regression presupposes linearity. (Thanda,2022)
- If the sample size is too small, the accuracy of logistic regression may be compromised. (Thanda,2022)

Reason why the chosen Dataset is Appropriate for Logistic Regression

Logistic regression is the correct type of analysis to use when you're working with binary data. You know you're dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as "yes" or "no", "pass" or "fail", and so on). (Thanda,2022)

For a dataset to be appropriate for logistic regression, the following must be true:

- The dependent variable is binary or dichotomous.
- The independent variables should be linearly related to the log odds.
- Quite high sample sizes are necessary for logistic regression, the larger the sample size the more reliable the results will be.
- There shouldn't be much multicollinearity amongst the predictor variables if any at all.

The dataset chosen fulfils all the above categories and is why it is most appropriate in this regard to be used on logistic regression.

(Swaminthan,2018)

Analysis that will be conducted on the dataset

Imports and Libraries used:

- Pandas
- NumPy
- Seaborn
- Matplotlib
- Datetime
- Os
- Sklearn

Exploratory Data Analysis

- Looking at the distribution of counts between sarcastic and non-sarcastic comments.
- Looking if there is a relationship between length of comments and the comment being sarcastic.
- Looking if some users are more sarcastic that others.
- Looking to see if the day of the week makes a difference on comments being sarcastic and not sarcastic.

Data Pre-Processing

- Most likely there will be comments missing and some comments will be unusable. We will remove all null comments and in doing so dropping the corresponding rows.
- We will convert the time into a datetime object so that it is useable.

Overfitting and Underfitting

Ways to deal with underfitting:

- Remove noise from data.
- Increasing model complexity.
- Increase duration of training.

Ways to deal with overfitting:

- Increase training data.
- Reduce model complexity.

Model training

Then the model will be trained according to logistic regression, a vectorizer will be used to build bigrams, put a limit on maximal number of features and minimal word frequency. We will be using a multinominal logistic regression; the reason why is using this is stated in the above section. A pipeline will be used, to allowing for the transformation and correlation of data into a model, which can subsequently be examined to provide outputs.

Interpret the results

- Using accuracy score
- Using a confusion matrix
- Using a classification report

Evaluation of model and improvements to model

We evaluate the model we will be using:

- An accuracy score: One parameter for assessing classification models is accuracy. The percentage of predictions that our model correctly predicted is known as accuracy. The following is the official definition of accuracy: Number of accurate guesses equals accuracy number of guesses overall. In our case we got an accuracy score of 0.72167, which is equal to 72.1% before improvements.
- A confusion matrix: The performance of the classification models for a certain set of test data is evaluated using a matrix called the confusion matrix. Only after the true values of the test data are known can it be determined. The first and last quadrant must be a high number and the second and third quadrant should be a much lower number for the model to be relatively accurate. In our case the accuracy is correct.
- A classification report: In machine learning, a classification report is a performance evaluation metric. The precision, recall, F1 Score, and support of your trained classification model are displayed using this method. In our model the average output is around 72% which is very accurate when you look at the accuracy score that we used above.

The improvements that we put in place to increase accuracy:

- We removed null values, such as null comments.
- Used a Function transformer: Some common techniques used by other sklearn estimators are provided through the FunctionTransformer (e.g., fit and transform). This has the advantage that arbitrary, stateless transforms can be added to a sklearn Pipeline, which combines many processing steps.
- This estimator concatenates the results after applying a series of transformer objects in parallel to the input data. Combining multiple feature extraction techniques into a single transformer is advantageous.

Results of the model

The results of the accuracy and report are not too bad, however as we can see the model is around 70% accurate. Meaning that around 1 in 5 can be predicted inaccurately, which in some other case would not be a good enough, for accuracy in something like cancer prediction or health prediction, which needs to be very high such are 98% - 99% so that people aren't diagnosed wrong. In terms of what we are trying to predict which is sarcasm the accuracy is good enough.

Ways I would have increased the model's performance:

- Looking for class imbalance.
- Optimizing the log loss and f1-scores.
- Performing a grid search to tune hyperparameters.

Alternative to Multinominal Logistic regression: Discriminant function analysis, which is more effective than multinomial logistic regression and necessitates the fulfilment of these conditions, is an alternative. Given that the analysis does not make these assumptions, multinomial logistic regression is in fact more frequently utilised than discriminant function analysis.

References

IBM (2022) *What is logistic regression?*, *IBM*. Available at: https://www.ibm.com/za-en/topics/logistic-regression (Accessed: October 24, 2022).

Swaminathan, S. (2018) *Logistic regression - detailed overview*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc (Accessed: October 25, 2022).

Thanda, A. (2022) *What is logistic regression? A beginner's guide [2022]*, *CareerFoundry*. Available at: https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#:~:text=Logistic%20regression%20works%20well%20for,of%20data%20from%20each%20other (Accessed: October 24, 2022).