*Journal of Statistics and Computer Science*
*Vol. 1, No. 1, 2022, pp. 11-21*

*URL : www.arfjournals.com*
*https://DOI: 10.47509 /JSCS.2022.v01i01.02*

# Determining Brand Attributes for a Consumer Packaged Goods (CPG) Brand from Imbalanced Binary Data

**Chiranjit Dutta[*], Balaji Raman[†], Venu Gorti[‡] and Kamal Sen[§]**
[*]*University of Connecticut; chiranjit.dutta@uconn.edu*
[†]*Cogitaas AVA; balaji.raman@cogitaas.com*
[‡]*Cogitaas AVA; venu.gorti@cogitaas.com*
[§]*Cogitaas AVA; kamal.sen@cogitaas.com*

*Received March 13, 2022 • Revised April 22, May 22&25 2022 • Accepted May 30, 2022 • Published July 15, 2022*

***Abstract:*** Classification of imbalanced binary data set is challenging due to a large presence of one class, typically more than 90% of data. Few examples of imbalanced data sets are: fraudulent activity of credit cards, conversion rates of online advertisements, clinical diagnostic tests of rare diseases etc. Standard classification techniques like logistic regression assumes that the underlying data set is evenly distributed. Applying these to imbalanced data set results in a classifier with poor prediction accuracy for minority class. Machine learning methods like SMOTE (Synthetic Minority Over-Sampling Technique) address this issue by oversampling the minority class, i.e it creates synthetic samples of the minority class instead of sampling with replacement. However, there are no standard rules to determine sample size of the oversampled minority class. Moreover, it is difficult to draw causal inference from a machine learning approach. In this paper, we discuss applications of novel statistical methodologies like generalized linear model using GEV-links, power links and latent factor models to a respondent level survey data. Objective of the study is to identify demand enhancing attributes for a CPG (Consumer Packaged Goods) brand in an emerging market.

***Keywords:*** Imbalanced binary response, GEV, power and reverse power link, SMOTE, brand imagery.

**To cite this article**

Chiranjit Dutta, Balaji Raman, Venu Gorti & Kamal Sen. (2022). Determining Brand Attributes for a Consumer Packaged Goods (CPG) Brand from Imbalanced Binary Data. *Journal of Statistics and Computer Science*. Vol. 1, No. 1, pp. 33-43. *https://DOI: 10.47509 /JSCS.2022.v01i01.02*

## 1. Introduction

In CPG firms, the key objective of most brand managers is to drive brand sales and growth. In this respect, firms conduct consumer surveys to infer consumer's perspective of brands, wherein respondents are asked series of questions, answers to which are usually binary, to associate attributes to a brand. These surveys are necessary to identify attributes that improves a brand's equity, enhance its pricing power, drive consumer demand, and build product differentiation. Further, these surveys are important to extract attributes that are necessary to participate in a category, attributes that are necessary to be in a consumer's consideration set, attributes to maintain market share and attributes that differentiates a brand from its competitors which can be used to build market share. For premium, niche brands and for newly launched brands, the number of "users" are significantly fewer as compared to number of "non-users" of the brand. Such datasets are known as imbalanced dataset (Chawla *et al.*, 2002). In imbalanced binary data the classes are not approximately equally represented. In this article, we discuss one such imbalanced data from a survey and evaluate different classification techniques to extract key imagery statements that drive the sales of a premium shampoo brand in an emerging market. We refer to these key imagery statements as Demand Enhancing Attributes (DEA). Among the different techniques discussed are the more recent statistical techniques like Generalized Extreme Value regression, (Wang and Dey, 2010) and power and reverse power link regressions, (Bazán *et al.*, 2017)

**Table 1: Sample data from the survey**

| *Y* | *Makes my hair beautiful* | *Makes my hair strong* | *Is good value for money* | *Is affordable* |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 1 | 0 |

These techniques are able to handle skewness better and allow inference on explanatory variables. We have compared predictions from these models with more commonly used machine learning techniques for handling imbalance binary data like random forests (Breiman, 2001) and Synthetic Minority Over-sampling Technique (SMOTE), Chawla *et al.* (2002).

A class of generalized logistic models with a two-parameter family to handle skewness in binary data was proposed by Stukel (1988). Later, many model improvements were proposed, for details see Wang and Dey (2010). Generalized extreme value distribution (GEV) as a link function was first proposed by Wang and Dey (2010), where the authors have applied their model to a B2B electronic payment system adoption.

Bazán *et al.* (2017) proposed a more general approach for asymmetric link functions and have applied their model for a data set in insurance domain. Different techniques for modelling imbalanced binary data and for comparing predictive performance has been discussed in Van der Paal (2014). More recently Yin *et al.* (2020) have proposed an ensemble learning approach with adaptive weight adjusting component, Lin *et al.* (2020) have used deep reinforcement learning method for imbalanced binary classification and Munkhdalai *et al.* (2020) have proposed a deep neural network architecture along with Gumbel distribution as an activation function for predicting imbalanced binary data.

The rest of this article is as follows, we begin with a brief discussion of data in Section 2, followed by a discussion on different methods used in Section 3. Modeling results are presented in Section 4 and end with insights in Section 5.

## 2. Data

Data is from a consumer survey conducted across major urban centres in an emerging market. This survey was conducted by the brand in collaboration with a market research agency. Along with information on brand awareness and usage, data on brand image perceived by a consumer is collected through questions on attribute(s) associated with a brand. For example, below are few questions on attribute(s) to which the answer is either a "Yes" or a "No":

- Are you aware of brand X?
- Have you used it in the last 3 months?
- Does it make your hair beautiful?
- Is it a good value for money?
- Is it a modern and trendy brand?

There are 18 such questions on physical and personality aspects of the brand. Detailed list is given in Section 6. A total of 635 consumers were surveyed. For our focal brand, a majority (93%) of respondents were non-users.

## 3. Methodology

The objective is to identify a model which will identify significant imagery statements or attributes along with a good predictive power. We consider a group of models for this purpose. Our benchmark model is logistic regression, ideally the performance of any complex model should be better than our benchmark model. Other models considered can be divided into two groups, a) Generalized Linear Models (GLM), McCullagh and Nelder (2019), with appropriate link functions to handle skewness like Generalized Extreme Value (GEV) link (Wang and Dey, 2010), power, and reversal power link (Bazán *et al.*, 2017) and b) machine learning models like random forests, SMOTE, XGBoost (Hastie *et al.*, 2009). In this section, we describe briefly all the models considered.

### *3.1. Logistic Regression*

Following Bazán *et al.* (2017), let $Y = (Y_1, Y_2, \ldots, Y_n)^T$, be a $n \times 1$ vector of $n$ independent binary random variables with probability $p_i = P(Y_i = 1)$, $i = 1, \ldots, n$. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$,

$i = 1, \ldots, n$ be a $k \times 1$ vector of covariates and take $\mathbf{X}$ to denote a $n \times k$ design matrix with rows $\mathbf{x_i^T}$, let $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^T$ be a $k \times 1$ vector of regression coefficients. Assume $p_i$ = Prob($Y_i = 1$), therefore $p(Y_i = 0)$ is $1 - p_i$. In the GLM framework

$$p_i = E(Y_i|x_i) = G(\mathbf{x_i^T}\beta);\ i = 1, \ldots, n, \qquad (1)$$

Where $G$ is a cumulative distribution function (cdf) and $G^{-1}$ determines the link function.

One of the most common methods for classification is logistic regression. The model is defined by taking the link function in (1) as

$$G(x) = \ln\left(\frac{x}{1-x}\right)$$

Thus, the probability of success is given by

$$p_i = E(Y_i|x_i) = \frac{\exp(x_i^T\beta)}{1+\exp\left(x_i^T\beta\right)}$$

### *3.2. Generalized Linear Regression with GEV-link*

Following Coles *et al.* (2001), we define the family of Generalized Extreme Value (GEV) distributions as

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

defined on the set $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, with $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$, where $\mu$ is the location parameter, $\sigma$ is the scale parameter and $\xi$ the shape parameter.

For handling skewed binary data, Wang and Dey (2010) proposed the use of a flexible link function based on GEV distribution for the binary regression model. GEV-link has a flexible parameter which determines the skewness of the data, $\xi$ in (2).

With GEV link the probability of success is given by:

$$E(Y_i|x_i) = P(Y_i = 1) = 1 - F_{GEV}(-x_i^T\beta) \qquad (2)$$

where $F_{GEV}(\cdot)$ represents the standard cdf of GEV distribution at $\mu = 0$, $\sigma = 1$, with unknown shape parameter $\xi$. Wang and Dey (2010) have used Bayesian approach for inference. The model likelihood is given by

$$P(y \mid X, \beta, \xi) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

with the priors being $\beta_j \sim \mathcal{N}(0, 100)$ and $\xi \sim \mathcal{N}(0, 1)$. Assuming $\beta$ and $\xi$ are mutually independent, the posterior is given by

$$\pi(\beta, \xi|y, X) \propto p(y|X, \beta, \xi)\pi(\beta)\pi(\xi). \qquad (3)$$

The posterior samples are obtained by applying the Metropolis Hastings algorithm (Gelman *et al.*, 2013), the proposal density used is a symmetric distribution, $N(0, 1)$, as discussed in Wang and Dey (2010). The mean of the posterior samples have been used for the inference and highest posterior density intervals are obtained at 5% level of significance to check the significance of the variables.

### *3.2.1. Generalized linear regression with power and reversal power link*

Bazán *et al.* (2017) proposed a new class of links for binary regression called positive exponent links. These links have an additional parameter for handling skewness. For this class of link functions, the probability of success is given by

$$p_i = E(Y_i|x_i) = F_\lambda(\eta_i) = F_\lambda(x_i\beta) \tag{4}$$

Where, $F_\lambda$ is the cumulative density functions (cdf) of a standard power or a standard reversal power distribution.

#### *Power distribution*

Following Bazán *et al.* (2017), let $S$ be univariate random variable. It follows a power distribution $S \sim \mathcal{P}(\mu, \sigma^2, \lambda)$, with location, scale and shape parameters given by, $\mu \in \Re$, $\sigma^2 > 0$, $\lambda > 0$ respectively, if the density is of the form

$$f_p(s \mid \mu, \sigma^2, \lambda) = \frac{\lambda}{g} g\left(\frac{s-\mu}{\sigma}\right)\left[G\left(\frac{s-\mu}{\sigma}\right)\right]^{\lambda-1} \tag{5}$$

where $G(\cdot)$ is any absolutely continuous cdf (cumulative distributive function) and $g(\cdot)$ an unimodal and log concave pdf (probability density function) with support in real line, $\Re$, known as the baseline distribution. A standard power distribution is obtained by putting $\mu = 0$ and $\sigma = 1$ in (5).

#### *Reversal distribution*

Assume $S \sim F(\cdot)$. The distribution of $S$ satisfies the reversal property if the cdf of $-S$ can be written as $-S \sim G(\cdot) = 1 - F(-\cdot)$. Here, the distribution of $G(\cdot)$ is known as the reversal distribution of $F(\cdot)$. For more details on power and reversal power distributions see Bazán *et al.* (2017). For our analysis we consider power Cauchit and reverse power Cauchit.

The power Cauchit link is given by

$$F_\lambda(z) = \left(\frac{1}{\pi}\tan^{-1}(z) + \frac{1}{2}\right)^\lambda$$

and the reversal power Cauchit link is given by:

$$F_\lambda(z) = 1 - \left(-\frac{1}{\pi}\tan^{-1}(z) + \frac{1}{2}\right)^\lambda$$

Here we use Bayesian methods to estimate the model parameters as discussed on Bazán *et al.* (2017). The likelihood function is given by

$$L(\beta, \lambda|y, X) = \Pi F_\lambda(x_i^T\beta)^{y_i}(1 - F_\lambda(x_i^T\beta))^{1-y_i}$$

The prior are as follows, $\beta_j \sim \mathcal{N}(0, 100)$ and $\lambda \sim \mathcal{N}(0, 1)$. Assuming $\beta$ and $\lambda$ are mutually independent, the posterior is given by

$$\pi(\beta, \lambda|y, X) \propto p(y|X, \beta, \lambda)\pi(\beta)\pi(\lambda).$$

### 3.3. Logistic Regression with Latent Factor Scores Using SMOTE

So far we have considered models under GLM framework. Here, we begin discussion of machine learning models starting with a method combining Synthetic Minority Over-sampling Technique, (SMOTE), Chawla *et al.* (2002) and Multidimensional Item Response Theory, (MIRT), Andreis and Ferrari (2014). MIRT identifies latent variables and thus helps in dimension reduction, SMOTE handles skewness. There are many explanatory dichotomous variables in the data and most of them are highly associated with each other. Therefore, dimension reduction is necessary and hence we apply multidimensional item response theory, which is analogous to the traditional factor analysis, to the explanatory variables to obtain latent factor scores in fewer dimensions.

#### *Multidimensional Item Response Theory (MIRT)*

MIRT models are popular in the fields of psychometrics and ability assessment (Andreis and Ferrari, 2014). Following Reckase (2009), given a set of response variables $X_1, X_2, \ldots, X_p$, the basic idea of latent variable analysis is to find a smaller set of latent variables $\theta_1, \theta_2, \ldots, \theta_q (q < p)$ that contains essentially the same information (Reckase, 2009). Let $X_{ij}$ be a random variable which assumes the following

$$x_{ij} = \begin{cases} 0, & i^{\text{th}} \text{ respondent said "No" to } j^{\text{th}} \text{ item} \\ 1, & i^{\text{th}} \text{ respondent said "Yes" to } j^{\text{th}} \text{ item} \end{cases}$$

The two-parameter logistic model can be expressed as:

$$P(X_{ij} = 1 \left| \theta_i, a_j, d_j \right| = \frac{\exp\left(a_j^T \theta_2 + d_j\right)}{1 + \exp\left(a_j^T \theta_i + d_j\right)}$$

where $d_j$ is an item-specific intercept term, $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iq})$, $a_j = (\boldsymbol{a}_{j1}, \boldsymbol{a}_{j2}, \ldots, \boldsymbol{a}_{jq})$, $i = 1, 2, \ldots., p$, $j = 1, 2, \ldots, N$. Here, $N$ denotes the number of respondents and $p$ denotes the no. of items and $a_j^T\theta_i = \sum_{l=1}^{q} a_{jl}\theta_{il} + d_j$. The latent factor score is a numerical value that indicates a person's relative standing on the latent dimensions. It is calculated as the mode of the posterior distribution (Embretson and Reise, 2013).

The latent factor scores are computed as

$$\theta_s \sim \mathcal{N}(0, 1)$$

$$\hat{\theta}_s = \arg\max P(X_i|\theta_i)P(\theta_i)$$

These latent factor scores are used as covariates in logistic regression. As the data is highly skewed, we apply SMOTE to artificially create synthetic samples of minority class.

### *SMOTE*

SMOTE, proposed by Chawla *et al.* (2002) is an over-sampling approach in which the minority class is over-sampled by creating synthetic examples. The algorithm is as follows

For each minority class instance '*c*'

- Choose *k* nearest neighbours by *K*-Nearest Neighbor algorithm (Hastie *et al.*, 2009).
- Randomly pick 1 neighbor (for 100% replication of minority class) from the neighbors obtained in Step 1 and call it '*n*'.
- Create a new minority class instance '*r*' using *c*'s feature vector and *n*'s feature vector.

The balanced data is used to identify the significant latent factors via logistic regression. Using the interpretation of traditional factor analysis, the loading weights more than 0.5 are used to identify the significant explanatory variables from the significant latent factors.

### *3.4. Random Forests and XGBoost*

We have compared the above mentioned methods with two popular machine learning algorithms which we describe briefly. Random forests proposed by Breiman (2001) is a popular ensemble learning methods. It is a flexible and easy to use algorithm for both classification and regression. It combines many decision trees to obtain a more accurate prediction and overcomes issues of overfitting. For more details on the methodology refer Breiman (2001), Hastie *et al.* (2009). Chen and Guestrin (2016) proposed XGBoost, a scalable tree boosting system. It is popularly known as extreme gradient boosting and it implements the gradient boosting decision tree algorithm. Boosting is a combination of many weak classifiers. More details can be found in Hastie *et al.* (2009).

## 4. Modeling Results

For analysis, we chose *R* (R Core Team (2018) and Ihaka and Gentleman (1996)) as our computing environment and R Studio (R Studio Team, 2015) as our Integrated Development Environment (IDE) for R. For model comparison, we use measures like sensitivity, specificity and AUC (Hastie *et al.*, 2009) and see Section 6. Table 2 displays these metrics for all the models. Here, we observe that although XGBoost has the highest AUC, the sensitivity is significantly lesser compared to others. Based on the values of sensitivity, specificity, AUC and giving importance to probabilistic interpretation of covariaties, we recommend model with reversal power cauchit link when compared to random forests. The estimated reversal cauchit link model can be written as:

$$\hat{p}_i = 1 - \left(0.5 + \frac{1}{\pi}\arctan(-(-1.888 + 0.869X_1 + 1.042X_2 + 0.986X_3))\right)^{0.147}$$

where $X_1$ : "Brand that cares for my hair"; $X_2$ : "Is a brand I trust" and $X_3$ : "Gives me salon like hair".

## 5. Conclusion

In this article we discuss modeling of imbalanced binary data for a problem in marketing domain. The basic aim of this study is to identify the demand enhancing attributes or DEAs of a premium shampoo brand in an emerging market. We recommend the use of a GLM with power and reversal power link as it has comparable predictive power to popular ML methods used in modeling of skewed data and in addition it provides a probabilistic interpretation of covariates. For the shampoo brand the key DEAs are *Is a brand that cares for my hair*, *Is a brand I trust* and *Gives me salon like hair*. The demand enhancing attributes are of great value to a brand manager, since it allows him/her to understand consumer's perception about the brand, provide levers to fight competition by sending out appropriate marketing messages. Thus, the brand managers should focus on these attributes in brand advertisements to drive the competitive advantage. Performing similar analysis for competitor brands will help the brand manager in identifying common attributes across the category in which the brand is present and if the scores of the brand is low compared to competition, then it is a sign that although the brand is conveying appropriate message through its advertisements, it must work on improving the scores. Further, this study provides attributes that distinguishes a brand from its competitors. If the brand manager follows the traditional proportion method, he/she might miss out on the significant demand enhancing attributes.

**Table 2: Predictive measures for model comparison**

| *Method* | *Sensitivity* | *Specificity* | *AUC* |
|---|---|---|---|
| Logistic regression | 0.2 | 0.94 | 0.68 |
| Logistic regression with latent factor scores using SMOTE | 0.4 | 0.80 | 0.64 |
| GLM with GEV-link | 0.3 | 0.96 | 0.67 |
| GLM with Power Cauchit link | 0.4 | 0.89 | 0.65 |
| GLM with Reversal Power Cauchit link | 0.4 | 0.89 | 0.70 |
| Random Forest | 0.6 | 0.74 | 0.66 |
| XG Boost | 0.1 | 0.95 | 0.72 |

**Table 3: Confusion matrix**

| | Predicted | |
|---|---|---|
| | Yes | No |
| Actual | | |
| Yes | True Positive (TP) | False Negative (FN) |
| No | False Positive (FP) | True Negative (TN) |

## 6. Appendix

### *Evaluation Metrics*

In this study we use the three commonly used measures such as specificity, sensitivity and AUC to measure the predictive performances of the above methods. We use the confusion matrix shown in Table 3 to explain these terms. Consider a binary response with classes "Yes" and "No". Here, True Positive (TP) is the number of instances where the predicted value is "Yes" and the actual value is also "Yes". False Negative (FN) is the number of instances where the predicted value is "No" whereas the actual value is "Yes". Similarly, one can define False Positive (FP) and True Negative (TN).

- **Sensitivity:** The proportion of positive class that has been correctly classified as positive.

**Table 4: List of questionnaire items**

| Attributes |
|---|
| Gives me salon like hair |
| Repairs damaged hair |
| Reduces hair fall |
| Is a premium brand |
| Contains natural ingredients |
| Is affordable |
| Is easy to lather lathers well |
| Makes my hair strong |
| Is a modern trendy brand |
| Makes my hair beautiful |
| Nourishes my Hair |
| Is good value for money |
| Is effective against dandruff |
| Is a glamorous brand |
| Is an expert in hair care |
| Is a brand I trust |
| Is a brand that cares for my hair |
| Makes me feel attractive |

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- **Specificity:** The proportion of negative class that has been correctly classified as negative.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **AUC:** This is the area under the ROC (Receiver Operating Characteristics) curve. ROC is a twodimensional curve representing a compromise between True Positive and False Positive rate. Whereas, the area under ROC curve or AUC is used in assessing the accuracy of the classifiers. Classifiers giving high values of AUC are considered to be best.

### *Questionnaire items*

This consumer survey consisted of a series of questionnaire on attribute(s) associated with the brand. Each consumer was asked about whether they associate each of the following 18 attributes with the brand or not. Below is the exhaustive list of attributes which were asked to the consumer by the surveyor.

### *Acknowledgement*

We would like to thank two anonymous reviewers for providing useful suggestions towards improving our article.

## *References*

Andreis, F. and Ferrari, P. A. (2014). Multidimensional item response theory models for dichotomous data in customer satisfaction evaluation. *Journal of Applied Statistics*, 41(9): 2044–2055.

Bazán, J., Torres-Avilés, F., Suzuki, A., and Louzada, F. (2017). Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry*, 33(1): 22–34.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1): 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22$^{nd}$ acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning:data mining, inference, and prediction*, volume 2. Springer.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3): 299–314.

Lin, E., Chen, Q., and Qi, X. (2020). Deep reinforcement learning for imbalanced classification. *Applied Intelligence*, 50(8): 2488–2502.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Munkhdalai, L., Munkhdalai, T., and Ryu, K. H. (2020). Gev-nn: A deep neural network architecture for class imbalance problem in binary classification. *Knowledge-Based Systems*, 194: 105534.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, 83(402): 426–431.

Van der Paal, B. (2014). A comparison of different methods for modelling rare events data. *Master of Statistical Data Analysis thesis, Universiteit Gent*.

Wang, X. and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *The Annals of Applied Statistics*, 4(4): 2000–2023.

Yin, J., Gan, C., Zhao, K., Lin, X., Quan, Z., and Wang, Z.-J. (2020). A novel model for imbalanced data classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6680–6687.