MODELS FOR FITTING CORRELATED NON-IDENTICAL BERNOULLI RANDOM VARIABLES WITH APPLICATIONS TO AN AIRLINE DATA PROBLEM

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Andres Perez
May 2021

Examining Committee Members:

Dr. Edoardo Airoldi, Dissertation Advisory Chair, Department of Statistical Science
Dr. Marcus Sobel, Department of Statistical Science
Dr. Kenichiro McAlinn, Department of Statistical Science
Dr. Richard Souvenir, Department of Computer and Information Sciences

ABSTRACT

Our research deals with the problem of devising models for fitting non-identical dependent Bernoulli variables and using these models to predict future Bernoulli trials.

We focus on modelling and predicting random Bernoulli response variables which meet all of the following conditions:

1. Each observed as well as future response corresponds to a Bernoulli trial.
2. The trials are non-identical, having possibly different probabilities of occurrence.
3. The trials are mutually correlated, with an underlying complex trial cluster correlation structure. Also allowing for the possible partitioning of trials within clusters into groups. Grouped trials would share a higher correlation with their grouped counterparts.
4. The probability of occurrence and correlation structure for both observed and future trials can depend on a set of observed covariates.

A number of proposed approaches meeting some of the above conditions are present in the current literature. Our research expands on existing statistical and machine learning methods to allow for all of the conditions to be met.

We propose three extensions to existing models that make use of the above conditions. Each proposed method brings specific advantages for dealing with correlated binary data. The proposed models allow for within cluster trial grouping to be reflected in the correlation structure. We partition sets of trials into groups either explicitly estimated or implicitly inferred. Explicit groups arise from the determination of common covariates; inferred groups arise via imposing mixture models. The main motivation of our research is in modelling and further understanding the potential of introducing binary trial group level correlations. In a number of applications, it can be beneficial to use models that allow for these types of trial groupings, both for improved predictions and better understanding of behavior of trials.

The first model extension builds on the multivariate probit model. This model makes use of covariates and other information from former trials to determine explicit trial groupings and predict the occurrence of future trials. We call this the Explicit Groups model.

The second model extension uses mixtures of univariate probit models. This model predicts the occurrence of current trials using estimators of parameters supporting mixture models for the observed trials. We call this the Inferred Groups model.

Our third method extends on a gradient descent boosting algorithm which allows for correlation of binary outcomes called WL2Boost. We refer to our extension of this algorithm as GWL2Boost.

Bernoulli trials are divided into observed and future trials; with all trials having associated known covariate information. We apply our methodology to the problem of predicting the set and total number of passengers who will not show up on commercial flights using covariate information and past passenger data.

The models and algorithms are evaluated with regards to their capacity to predict future Bernoulli responses. We compare the models proposed against a set of competing existing models and algorithms using available airline passenger no-show data. We show that our proposed algorithm extension GWL2Boost outperforms top existing algorithms and models that assume independence of binary outcomes in various prediction metrics.

ACKNOWLEDGEMENTS

I would like to thank Dr. Marc Sobel for his support from the early stages of defining this research problem. His dedication, extended number of hours, feedback and interest were key for this work to be fulfilled. Dr. Sobel showed patience, care and genuine interest in working with me. His timeless advice has helped far way beyond the academic work presented.

I want to thank the rest of the members of my committee: Dr. Edoardo Airoldi, Dr. Kenichiro McAlinn and Dr. Richard Souvenir, mainly for believing in me and giving me the opportunity to accomplish the final stage of a life-long dream of mine to get a PhD in a field that is very special to me. Their feedback, continued support and interest in this work, served in many ways, mostly as motivation to keep pushing and doing the best work I can.

I want to thank the airline, which shall remain anonymous for confidentiality reasons. Their collaboration and support were key to the success of this research.

I also want to thank the staff and faculty of the Department of Statistical Science at Temple, for also giving me the opportunity to be a part of the program and for helping me so many invaluable skills.

I would also like to thank Dr. Jesse Frey and Dr. Michael Posner from Villanova University Department of Mathematics and Statistics (and many other beloved teachers from my undergradute and Masters degrees) for stimulating me to pursue a PhD and believing that I could do it.

Finally, I want to thank my family and friends who supported and believed in me through all the tough moments of this long process.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND MOTIVATION

When fitting statistical models to observed Bernoulli trial data, the binomial probability distribution is the standard choice. The binomial probability distribution assumes a collection of independent and identically distributed Bernoulli trials. The standard assumption of independent and identically distributed Bernoulli trials is unrealistic for many data applications. Alternative methods or approaches to fitting the standard binomial probability distribution are sometimes needed. Yu and Zelterman (2002)[59] proposed a distribution for the sums of dependent exchangeable Bernoulli random variables. For some applications, as we illustrate further below, the identical trials assumption is not applicable. While some existing models assume non-identical correlated Bernoulli trials, they often assume equicorrelation amongst all pairs of trials, an example is the model proposed by Gupta and Tao (2010)[23]. When fitting models to non-identical correlated Bernoulli trial data and predicting future trial data it is advantageous to analyze each trial separately; we adopt this viewpoint.

The main motivation for our research is to allow for complex correlation structures in binary responses in prediction/classification models. Specifically, correlation structures dependent on groups of trials. We illustrate our motivation with the practical example of correlated binary data from an international airline. Airlines seek to predict which booked passengers will not show up. Each booked passenger could be a Bernoulli trial, either showing up to a flight or not. If we assume that every passenger has the same probability of showing up or not, and that passengers show up or not independent of each other, then the total sum of passengers could be treated as a binomal random variable.

Data as well as previous research (Lawrence et al. (2003)[29], Sveinsdottir (2019)[52]) have shown that booked passengers have different probabilities of showing up or not based on their covariates. Which is also intuitive. Classification algorithms and models such as logistic regression could be used to allow for different probabilities based on covariates, if we still assume that each passenger is independent of each other.

Previous research in airline no-show prediction models have not used passenger correlation. Airline industry experts largely believe that certain groups of passengers are dependent, and there are very intuitive examples that would reflect this. Such examples include: (i) couples that fly together are expected to have a higher correlation of showing up together or not than unrelated passengers; (ii) passengers on connecting flights are another example of possibly correlated passengers; and (iii) the sensitivity of older passengers to issues involving a digital communication channel. Non-explicit groups, such as passengers with close but non-identical covariate information, can also be modelled. We do this by assuming an underlying mixture model.

Within cluster correlation of trials would be expected in this application. Meaning, passengers belonging to a same flight (cluster) might be correlated. However, allowing for further groupings of passengers might help identify a more realistic and stronger correlation structure between them. Two unrelated passengers belonging to a same flight might be correlated with each other (factors such as weather, traffic, etc. could lead to within flight, or cluster, correlation), but two related (or grouped) passengers would be expected to have a higher correlation with each other.

In the application we analyze, we aim to predict the total number of booked passengers that will not show up for a number of flights. We use historical data of past flights for a major international airline to fit the model extensions proposed. The historical data corresponds to what we refer to as observed Bernoulli trials. We use the historical data to make inferences about future Bernoulli trials (i.e., those where success/failure have not occurred yet).

We refer to particular sets of trials (i.e., flights) as clusters. In this case, the probability of not showing up to a flight depends on a number of passenger and flight-based factors. For example, passengers that are frequent flyers might behave differently than their non-frequent flying counterparts. Many other trial (i.e., passenger) and cluster (i.e., flight) level covariates might help to accurately estimate trial probabilities of success. Personal information provides some examples of trial level covariates while flight time and weather are examples of flight level covariates.

To better illustrate our motivation, let's assume we have a flight with passengers per Table 1.1. Where S1 represents the passengers flying solo (passengers A1 and B2), S2 the passengers flying in pairs (C1/C2 and D1/D2), and S3 the trios (E1/E2/E3 and F1/F2/F3).

S1	**S2**	**S3**
A1	C1/C2	E1/E2/E3
B1	D1/D2	F1/F2/F3

Table 1.1: Example of Passenger Groups

The correlation structure possibilities that arise can be of interest in modelling. We could assume that all pairs of passengers (groups in S2) share a specific correlation, meaning the correlation between C1 and C2 would be the same as the correlation between D1 and D2. We call this correlation the intragroup correlation of pairs. Using historical data from pairs of passengers, this intragroup correlation of pairs could be estimated and used in predicting a future trial's (belonging to a pair) outcome. Trios could also share a specific intragroup correlation, which could differ from or be the same as the intragroup correlation of pairs. The correlation between passengers not belonging to the same group, meaning the intergroup correlation, may not exist. It could also be present, and intuitively, we would assume it would be smaller than any intragroup correlation. This intergroup correlation would be shared between any pair of trials (passengers) not belonging to a same group. For example, the pairs of passengers A1/B1, A1/C2, C2/D1, are some of the passengers that would be share this intergroup correlation with each other.

This particular example illustrates the usefulness of models that allow for complex inter and intragroup correlation structures within clusters of binary trials, while also allowing for varying probabilities of success by trial. Such models could have various other applications, where our assumptions would seem to intuitively hold as well.

One such example is Finance, specifically credit risk modelling, where there has been much research showing improvements in capturing loan portfolio risk when allowing for correlation between individual loan defaults (Das et al. (2007)[14], Schonbucher (2001)[48], Duffie et al. (2009)[17], Zazzara (2001)[60]). In this application we could allow for further complex correlation structures than just allowing for within cluster trial correlation. Loans could be grouped by explicit groups (ex. Industry, Geography) or non-explicit groups, and allowing for intragroup correlation.

Let's imagine we have a portfolio of loans with twelve loans, per Table 1.2. Four loans (A, B, C and D) are for clients that live in State 1, four other loans (E, F, G and H) are for clients that live in State 2 and the remaining loans are for clients that live in State 3. States 1 and 2 belong to a same region, for example West Coast. We also assume that all these loans were originated around the same time period.

It is common to assume that there is a general cluster (loan) level correlation shared by all loans of this portfolio, especially if they were all originated around a same time period. National level factors (ex. a pandemic) could affect the correlation of defaults between any two loans. We could also allow for

State 1	**State 2**	**State 3**
A	E	I
B	F	J
C	G	K
D	H	L

Table 1.2: Example of Groups of Loans in a Portfolio

an intra group correlation that is shared by pairs of loans belonging to a same state. This intragroup correlation could be the same for all States, but they could also be different. For example, loans from State 1 might show stronger intragroup correlation than loans from State 2. State level factors (ex. regulation changes at a state level) could impact all loans belonging to the same state. Loans belonging to the same region, but not the same State (ex. loans A and E) could share a regional correlation. Regional factors (ex. weather phenomena such as a hurricane) could affect all states in a same region.

These examples demonstrate how complex correlation structures, based on binary trial groupings, could be of interest.

In some applications, groups of trials have a (relatively) straightforward inter and intragroup correlation structure. In such settings, we propose models which explicitly assign trials to groups, and make use of this structure for fitting and predicting responses.

In some applications, groups of trials can have a more complicated inter and intragroup correlation structure. The grouping of trials that share higher correlations, might not be explicitly determined a priori. In such settings, we propose mixture models where the correlations and trial groups are estimated implicitly.

We also propose a method for introducing this group correlation structure in a gradient descent boosting algorithm which allows for correlation of binary outcomes. We show that this proposed algorithm extension outperforms top existing out-of-the-box algorithms and models that assume independence of binary outcomes in various prediction metrics.

We fit airline passenger data and predict future passenger no-show behavior with our proposed models which allow for complex trial group correlation structures. Each proposed model provides different benefits.

There are a variety of approaches and proposed models and algorithms satisfying some of the aforementioned data conditions. We build on these existing models and adapt them to fit the motivation outlined in this chapter. We outline these in chapter 2 below.

CHAPTER 2

LITERATURE REVIEW

We describe some models and algorithms currently used for analyzing Bernoulli trials. We present existing literature for modelling both the total sums of correlated binary random variables, or individual binary trials. Both of which are relevant to our airline application.

If we assume that the Y_i's are independent and identically distributed (iid) random variables with common success probability $p = P(Y_i = 1)$, where $Y_i \sim Bernoulli(p)$, then $U = \sum_{i=1}^{n} Y_i$ would have a standard binomial distribution B(n, p). The iid assumption has been proven inaccurate in many practical applications as we mentioned in the previous chapter.

The beta-binomial distribution (BBD) introduces a measure of dependence between Bernoulli trials and is a common method (see Moran(1968)[36]) found in the literature for modelling binary data with overdispersion/underdispersion due to positive/negative association between trials. Skellam (1948)[50] showed how modelling the 'fluctuation' of the probability parameter p using BBD introduces correlation between trials. Skellam (1948)[50] also shows that, using the notation, $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, for the BBD, the probability mass function of the sums of binary data, u, would take the form:

$$P(U = u) = \binom{n}{u} \frac{B(\alpha + u, \beta + n - u)}{B(\alpha, \beta)}, \; u = 0, 1, ..., n \qquad (2.1)$$

The BBD is very useful for modelling correlated binary data. Griffiths

(1973)[22] derived maximum likelihood as well as moment estimators for the parameters of BBD, and applies his results to fitting household incidence of disease data. Kemp and Kemp (1956)[25] used the BBD in the analysis of quadrant point data. Chatfield and Goodhart (1975)[10] applied the BBD to model consumer purchasing behavior. Williams (1975)[54] used BBD to analyze clustered binary data, letting the success probability of each cluster, or litter, vary in a teratogenic application to determine the effect of a chemical treatment on birth rates in rats. Williams et al. (1987)[56] showed another approach for analyzing correlated binary data using the BBD. Williams et al. (1988)[55] showed the implications of the bias found in the estimators that maximized the beta-binomial likelihood in teratological applications.

The expected variance of the standard binomial random variable is $np(1-p)$. Overdispersion occurs in binary data whenever the observed variance is greater than this expected variance. Underdispersion is less common in binary data, occurring when the observed variance is less than the expected variance. Prentice (1986)[43] extended previous work on the BBD to show that it can fit either overdispersion or underdispersion in binary data allowing for positive or negative correlations. It is important to note that the BBD assumes equicorrelation between trials within clusters.

Modeling individual probabilities of success for binary data can also be of interest in some applications. Common statistical models such as logistic regression are still widely used, for example in modelling loan default risk (Costa et al. (2020)[12] and Kwofie et al. (2015)[28] are examples). See Menard (2002)[34] for a well referenced book on applied logistic regression. Variations on the logistic regression when standard conditions are not met, such as the fuzzy logistic regression are also still widely used and researched (see Sohn et al. (2016)[51] and Pourahmad et al. (2011)[42]). Logistic regression and other statistical classification models for binary data typically apply to uncorrelated trials. Other models allowing for individual trial modelling, and for correlation between trials have been proposed. Most assume equicorrelation between trials.

Prentice (1986)[43] regresses transformed cluster level occurrence probabilities $\mathbf{p}= (p_1, ..., p_f)$ on covariates W_f using cluster level parameters β. In this case the probabilities of occurrence p for the trials in each cluster are assumed to be equal. Crowder(1978)[13] regresses the transformed probabilities $\mathbf{p}$ on cluster level covariates, while assuming a common covariance parameter ψ for each given cluster. Williams (1982)[57] proposed an alternative method for modelling the common correlation parameter. Ochi and Prentice (1984)[40] propose a probit regression model for equicorrelated binary data. All of these approaches, including that proposed by Kupper and Haseman (1978)[27] and Haseman and Kupper (1979)[24] assume either a shared individual trial probability p_f or equicorrelation between trials for each given cluster f .

Pires et al. (2012)[41] propose a Bayesian correlated binomial regression model in which individual and cluster level covariates are used. This correlated binomial regression model is based on a model introduced by Luceño (1995)[31] and Luceño and Ceballos (1995)[32]. Diniz (2010)[16], Luceño (1995) [31] and Luceño and Ceballos (1995)[32] propose a variety of Bayesian approaches which introduce prior distributions for the probability and correlation parameters. These models assume a common probability for trial success and a common covariance parameter ψ between trials. The probability mass function for these models can be seen as a mixture of two binomials. These Bayesian correlated binomial models are strong candidates for fitting binary data, but again, equicorrelation is assumed.

Witt (2014)[58] proposes another model for correlated binary variables. He also uses a generalized correlation parameter and a common success probability for the trials and assumes conditions insuring stationarity. Altham (1978)[4] proposed two additional generalizations of the binomial distribution both of which assume mutual equicorrelation between trials. George and Bowman (1995)[20] proposed a procedure for estimating the moments of the sum of Bernoulli trials for exchangeable, correlated, binary data. Bowman and George(1995)[8] and George and Kodell (1996)[21] also presented models for exchangeable binary data. Kuk (2004)[26] used a log-log-link function to reformulate the model proposed by George and Bowman as a generalized linear model. Kupper and Haseman(1978)[27], apply this correlated binomial model to toxicological experiments. Stefanescu et al.(2003) showed how approximate MLE estimates of parameters should be adjusted when different sized clusters of binary data are assumed. Prentice (1988)[44] presented approaches for modeling correlated binary data, and allowing each observation to have its own set of covariates. Zhao and Prentice (1990)[62] used a quadratic exponential model in analyzing correlated binary data.

Ensemble methods such as random forests and boosting provide some of the best out of the box algorithms for prediction/classification of binary data. Ensemble methods such as boosting combine many weak learners to form a strong final learner (see Schapire (1990) [47] and Freund and Schapire (1997)[18] for examples of influential papers on the subject). Common ensemble methods, including boosting, however typically assume independent outcomes.

Adewale et al. (2010)[1] proposed an algorithm called WL2Boost which allows for correlated binary outcomes. This algorithm introduces a covariance matrix in a loss function within a functional gradient descent boosting algorithm (Friedman (2001)[19] and Bühlmann and Yu (2003) [9]). They applied WL2Boost to match-pair binary responses in a clinical study application, allowing for the correlation between matched pairs. Recent research on matched case-control studies in clinical applications use ensemble classification methods (Adler et al. (2011)[2], Adler et al. (2011)[3], Balasubramanian et al. (2014)[7]) and the WL2Boost algorithm is still referenced as one of the most valuable, accurate and efficient methods for such applications (Liang et al. (2018)[30], Balasubramanian et al. (2014)[7]). The WL2Boost algorithm can extend to more general combinations of correlation between binary responses not just matched-pair binary responses. Our third proposed method extends by applying small variations to the WL2Boost algorithm. We apply this method to the airline prediction problem as well. This also serves as a practical example of applying algorithms that allow for correlation between binary outcomes beyond clinical matched case-control applications.

CHAPTER 3

MODEL SETUP

We define a trial, as a Bernoulli event, and a cluster as a set of individual trials. We use the terminology ‘observed’ for trials whose outcome is known; and ‘future’ trials correspond to those with (as yet) unknown outcomes. We assume both observed and future trials have associated known covariates. Each cluster can have a different number of trials, but for the purpose of simplification we assume n trials per cluster.

In general settings we use the notation $Y_{f,i}$ for the i’th Bernoulli trial (i=1,...,n) in cluster $f \in \{1, ..., F\}$. Let $X_{f,i}$= $(x_{f,i,1}, x_{f,i,2}, ..., x_{f,i,M})$ denote the M covariates associated with trial $Y_{f,i}$, and V_f= $(v_{f,1}, v_{f,2}, ..., v_{f,L})$ denote the L covariates associated with cluster f. In the sequel we frequently combine these into a single set X_i of covariates. To simplify notation, we write, Y_i for the i’th Bernoulli trial and X_i for the covariates associated with trial i and drop the cluster index wherever possible. We use the notation g ($g \in \{1, ..., G\}$) for groups based on common features when they are available.

3.1 Multivariate Probit Model

In Dey et al. (2000)[15] a variety of Bayesian models for non-identical Bernoulli trials are presented. Once observed, trials are partitioned into groups. We then can build conjoined models, one for each group, which accurately reflect the probability of occurrence and dependence between trials.

We make particular use of the multivariate probit models below for this purpose. The multivariate probit model is a distribution for nonidentical, dependent Bernoulli observations, $Y_1, ..., Y_n$. The multivariate probit model is discussed by Ashford and Sowden (1970)[6]. Amemiya (1974)[5] present parameter estimation methods for the bivariate case. Chib and Greenberg (1998)[11] proposed more general (do not require simplification assumptions), simulation based, inference methodologies for the multivariate probit model.

In what follows $\phi(Z|\mu, \Sigma)$ refers to the multivariate normal density with mean vector μ and correlation matrix Σ. Σ must be in correlation form for identifiability reasons as noted by Chib and Greenberg (1998)[11]; we sometimes refer to it as a correlation matrix in the context of the multivariate probit model extension proposed. Then the joint probability of the dependent Bernoulli observations, $Y_1, ..., Y_n$ is given by:

$$P(Y_1, ..., Y_n) = \int_{Z_1 \in A(Y_1)} ... \int_{Z_n \in A(Y_n)} \phi(Z|\mu, \Sigma) \quad dZ_1 \; ... \; dZ_n \tag{3.1}$$

$$\mu = (\mu_1, ..., \mu_n); \qquad \Sigma = (\sigma_{i,j})_{1 \le i,j \le n} \tag{3.2}$$

$$Z \sim \mathcal{N}(\mu, \Sigma) \qquad \text{with density} \qquad \phi(Z|\mu, \Sigma) \tag{3.3}$$

$$A(Y_i) = \left\{ \begin{matrix} Z_i > 0 & \text{if } Y_i = 1 \\ Z_i < 0 & otherwise \end{matrix} \right\} \qquad i = 1, ..., n \tag{3.4}$$

Generalizations to mixtures of normals will be considered. We apply the generalization of this result to models in which covariate information is used to characterize the structure of Bernoulli data.

Suppose we have covariate information X_i for each observation. We use the notation, $X = (X_1, ..., X_n)$ for the matrix of covariates, $W = (W_1, ..., W_n)$ for a vector of parameters, XW for the corresponding matrix product, and Σ for the Gaussian covariance matrix. Then the joint distribution of the dependent Bernoulli variables, given the parameters W and Σ, and covariates X, is given by:

$$P(Y_1, ..., Y_n) = \int_{Z_1 \in A(Y_1)} ... \int_{Z_n \in A(Y_n)} \phi(Z|XW, \Sigma) \quad dZ_1 \; ... \; dZ_n \quad (3.5)$$

$$Z \sim \mathcal{N}(XW, \Sigma) \qquad \text{with density} \qquad \phi(Z|XW, \Sigma) \quad (3.6)$$

$$A(Y_i) = \begin{Bmatrix} Z_i > 0 & \text{if } Y_i = 1 \\ Z_i < 0 & otherwise \end{Bmatrix} \qquad i = 1, ..., n \quad (3.7)$$

Roughly, groups of trials are identified in this setting, by paths through the grid containing only non-zero elements of the inverse covariance matrix Σ^{-1}. We generalize this further below. The (marginal) correlation between Y_i and Y_j depends only on the local parameters $\mu_i, \mu_j, \sigma_{i,j}$; it is given by:

$$\rho_{i,j} = \frac{P(Y_i = 1, Y_j = 1) - P(Y_i = 1)P(Y_j = 1)}{\sqrt{P(Y_i = 1)P(Y_j = 1)(1 - P(Y_i = 1))(1 - P(Y_j = 1))}} \quad (3.8)$$

$$= \frac{\int_{Z_i>0} \int_{Z_j>0} \phi \left((\mu_i, \mu_j); \begin{pmatrix} 1 & \sigma_{i,j} \\ \sigma_{j,i} & 1 \end{pmatrix} \right) - \Phi(\mu_i)\Phi(\mu_j)}{\sqrt{\Phi(\mu_i)\Phi(\mu_j)(1 - \Phi(\mu_i))(1 - \Phi(\mu_j))}}$$

We will refer to $\{\sigma_{i,j}\}$ as the normal correlation parameters, and $\sigma_{i,j} = \sigma_{j,i}$. Bernoulli variables can be designed to have a correlation and probability structure which accurately reflects that of the data. By assuming a hierarchical model for the parameters, we can flexibly impose dependence between trials and assign probabilities. As was shown by Chib in Dey et al. (2000)[15] and Chib and Greenberg (1998)[11], it is straightforward to estimate parameters using Markov Chain Monte Carlo (MCMC) techniques. We discuss generalizations of this result below. We explore a particular case in which the latent variables can be partitioned into groups having the property that their associated covariance matrix Σ has a common variance and a common correlation.

3.2 Mixture Models using Univariate Probit Models

Another approach which we take involves modelling each trial using a univariate probit model, outlined below, and assuming that some of the parameters underlying that model are distributed according to a mixture model. The variables distributed according to a mixture model include the covariates X_i. The univariate probit model assumes that Bernoulli variables are conditionally independent; each modelled using a probit regression. Bernoulli variables are assumed to be conditionally independent given the parameters; the dependence arises from the assumed distributions of W and X_i. We assume in this case that:

$$\begin{aligned} P(Y_i = 1|W, X_i) &= \int_{Z>0} \phi(Z|X_iW, 1)dZ \qquad (3.9) \\ &= \Phi\left(\frac{X_iW}{1}\right) \qquad (3.10) \end{aligned}$$

We assume additionally that the X_i's are distributed as a mixture of multivariate normal distributions.

3.3 WL2Boost Gradient Descent Boosting Algorithm

The third approach we take builds on the WL2Boost gradient descent algorithm proposed by Adewale et al. (2010)[1]. Given that we make small adaptations on WL2Boost and to avoid redundancy we present the full details of the algorithm along with our adaptations at once in chapter 7. For now, we summarize the algorithm's approach. Bühlmann and Yu (2003) [9] discussed using the least squares loss as the loss function within gradient descent boosting (known as L2Boost). Adewale et al. (2010)[1] introduced a variance-covariance matrix **V** into this least squares loss function. This matrix **V** introduces the correlation structure between binary observations and serves as a weighting matrix in the least squares loss function. For this reason, the authors called their algorithm WL2Boost, as in weighted least squares boosting.

CHAPTER 4

AIRLINE DATA APPLICATION

Airlines lose potential profit for every empty seat on a flight that takes off (Lawrence et al. (2003)[29]). If a passenger that has booked a flight does not show up, then this is considered a ‘no-show’. Since airlines expect a number of passengers to not show up for the flight, airlines overbook flights to maximize revenues. This means airlines sell more tickets than they have seats available. This poses a risk, since sufficiently high levels of overbooking could lead to issues such as customer dissatisfaction, brand damage (especially during social media times where passenger complains increase in scope and reach) and revenue impact, the latter because overbooked passengers can also represent other costs to airlines (ex. rebooking or hotel fees). Thus, having accurate predictions of the total number of passengers that will not show up to a given flight can be very profitable and there have been several approaches for modelling passenger no-show behavior, both in industry and academically (Lawrence et al. (2003)[29], Neuling et al. (2004)[39], Zenkert (2017)[61], Morales and Wang (2017)[35]).

The operative aspects of optimizing overbookings and managing such models and related topics, are out of scope for our research. Most research and models in this area do not consider correlations between passenger no-shows. Our research focuses on providing insights into the correlation (and complex correlation structures) that can be exhibited in this type of data. Such insights and models, could be incorporated into existing overbooking practices by airlines and other related research. Airline data is one example of how models allowing for complex trial group correlation structures can provide benefits such as better fit and prediction of data, as well as key insights into trial behavior. We hope that with the airline example, we can demonstrate the potential for other applications. As previously mentioned, loan default risk modelling is an example of another application where such complex group correlation would intuitively seem to exist.

The assumption of independent identically distributed trials has limited application to data. We present the airline data application, where these assumptions have been shown not to hold. When counting the number of passengers booked to ride on commercial airline flights who do not show up, it is not realistic to assume that the events of passengers showing up are mutually independent and identical Bernoulli trials. Implicit and explicit factors specific to each passenger and each flight can have an important influence on the probabilities that the booked passengers show up or not. A passenger that frequently flies on the same route has a different probability of showing up than his/her counterpart who is an infrequent flier. Additional factors like age, method of booking, flight class, etc. could also have an impact on passenger probabilities of showing up. As such, we would expect a model that allows for different probabilities of success for trials based on observation covariates to be more suitable to fit this data.

In the airline application, trials distinguishing whether passengers show up or not are clearly correlated and their inter-correlations can be complex. In chapter 1 we illustrated how the correlation structures of clusters (flight) of passengers can also have underlying group level correlations. There are many explicit passenger groups (e.g., families, business groups, etc.) whose intragroup correlations differ from their intergroup correlations. For example, we would expect family members (or any passengers that have booked together) to show higher correlation for showing up together or not than individual passengers. As such, we would also expect that models that allow for passenger correlations, and that can differentiate levels of correlation based on trial groupings, to fit this data better. Such models could give valuable insights into the correlation relationships between al passengers in a cluster (flight) vs. correlations of passengers in a specific group.

To take account of this, we propose to partition passengers into groups; passengers belonging to a same group share the same correlation as passengers grouped in a same group size. For example we would have a correlation for passengers that fly in pairs which is shared by all passengers flying in pairs with their booked counterpart. Our model extension Explicit Groups model and our extension to the WL2Boost algorithm make use of this method for explicitly grouping trials.

Alternatively, by way of comparison, we consider Mixture Models in which inter and intra group correlations are analyzed by assuming a mixture model which implicitly partitions passengers into groups. We assume, in this case, that flight covariates and associated additional parameters are distributed according to a mixture model. The parameters underlying the mixture model are estimated and used, together with the current trial covariates, to estimate the success probabilities of future Bernoulli trials whose covariates are known.

In the context of the airline, knowing which passengers will not show is of interest. Hence, we are interested in models/algorithms that can classify and provide predicted probabilities for individual trials/passengers. In terms of overbooking optimization, airlines are more interested in having predictions for the total sum of passengers that will not show up to a given flight. In the context of loan default modelling, individualized prediction probabilities are very important, since financial institutions modelling these loans can better manage specific risks during the lifetime of a loan. The methods proposed allow for individual observation modelling and prediction. Via these individualized predictions, the models also predict expected sums of total observations. We show that our extension on WL2Boost provides the best results in both individual predicted probabilities, and even more so in predictions of sums of total passengers. We also simulate the total number of no-show passengers for a set of future flights to further demonstrate that allowing for correlation in binary responses can have a very valuable impact in better capturing overdispersion of highly correlated data.

CHAPTER 5

EXPLICIT GROUPS MODEL

We seek to estimate, for a variety of passengers who have booked flights, which passengers will not show up. The methodology given below provides estimates of probabilities that each passenger will not show up. The sums of these probabilities for each given flight provide an estimate of the expected total numbers of passengers who will not show up on the flight. We partition the passengers into groups based on their covariates. We use generalized linear models to regress the Bernoulli variables against covariate information. We assume that each group g is linked to a set of covariates. We also assume that each latent variable belonging to group g, shares a known variance σ_g^2 and intragroup correlation a function of ψ_g. The variance σ_g^2 can be set to 1 for a model that does not assume general intergroup correlation between trials. We estimate these parameters for the trials with known result (observed trials) and use them to estimate probabilities for the trials with unknown result. We briefly outline this below.

5.1 General Formulation with explicit groups: no covariates

All passengers $\{1, ..., n\}$ for a given flight are partitioned into k nonoverlapping groups: $G = \{g_1, ..., g_k\}$. Passengers within each group are assumed to have a common intragroup correlation. Some groups may share this correlation with one another. In specific data applications, the model can be easily adapted to allow for this. For this notation we assume, for the moment, that the groups are known and reflect the number of passengers flying together using a covariate that identifies passengers that booked together. Groups with the same intragroup correlation are taken to belong to a set S. We assume that the size of set S_u is n_u (u=1,...,k). Dispensing, for the moment, with the notation u, we index the observations in set S_u by $Y_{g,i}$; $i \in g$; $g \in S_u$; $i \in 1, ..., n_u$.

This can be generalized to other types of grouping. Passengers could also be grouped by other shared covariates, that don't necessarily assign passenger groups to sets. If we partitioned passengers into the groups based on class, for example Business and Economy, then the notation of sets S can be dropped. As we see in chapter 8 there is strong evidence to assume intragroup correlation based on this grouping. As such, we focus our examples with this grouping option.

Using the standard indexing, we define:

$$A(Y_i) = \begin{Bmatrix} Z_i > 0 & \text{if } Y_i = 1 \\ Z_i < 0 & \text{if } Y_i = 0 \end{Bmatrix}$$

and the multivariate normal density is given by:

$$\phi(Z|\mu, \Sigma) \propto \exp\left\{-(1/2)(Z-\mu)'\Sigma^{-1}(Z-\mu)\right\}$$

When grouping by groups of passengers booked together and their respective sizes, we take M_u to be the $n_u \times n_u$ matrix that is described by having 0's along the diagonal and 1 for elements corresponding to observations that belong to a same group.

Below, we take I_u to be the identity matrix of size n_u. We assume that for all passengers $Y_i; i \in 1, ..., n_u$ belonging to groups of size u in S_u, the joint probability is given by:

$$P(Y_1 = i_1, ..., Y_{n_u} = i_{n_u}) = \int_{A_{i_1}} ... \int_{A_{i_{n_u}}} \phi\left(Z|\mu_u, (I_u + \psi_u M_u)\right) dZ \tag{5.1}$$

We assume that $\psi_u < 1$, that the correlation matrix $\Sigma(\psi) = I_u + \psi_u M_u$ is invertible, and that groups are (conditionally) mutually independent. We can also allow for a general correlation between ungrouped trials, and this is easily incorporated by adapting the covariance matrix through the prespecified structure of M matrix.

More complex correlation structures of groups and hierarchies of groups can also be incorporated into this framework. The corresponding ψ parameters and M matrix can be added, layered, removed or adjusted to allow for different combinations and levels of trial grouping correlation. Following the example of the loan portfolio in chapter 1, we could have a ψ parameter for general ungrouped trials (cluster level correlation), another ψ for trials belonging to a first level of groups (ex. Region) and a final ψ for trials belonging to a second level of groups (ex. State).

One of the main benefits that we see of using this model proposed, is that experts in specific industries can incorporate their know-how as to potential correlation structures based on specifically defined groups a priori and incorporate this into the model structure and correlation parameter estimations. Chib in Dey et al (2000)[15] provides a general model in terms of correlation structure, by allowing for every pair of trials to have a specific correlation. While this provides flexibility, it is unrealistic to put into practice for the type of applications we have mentioned. Placing trials into groups that are expected to share a given correlation allows to estimate correlation parameters (both inter and intragroup) from historical data, and apply these parameters for probability estimation of future grouped trials. For example, we could estimate the correlation of all historical passengers that have flown in pairs, and use this pair intragroup correlation when predicting the probability of no-show for a future pair.

All trials within a given group $g \in S_u$ would share a common normal correlation. Because of the marginalization properties of the multivariate normal distribution, the marginal probability that, for each group g, $Y_{g,\bullet}$ takes on a binary vector $V \in \{0,1\}^{n_g}$ is:

$$P(Y_{g,\bullet} = V) = \int_{A(Z_{g,\bullet})} \phi\left(Z_{g,\bullet} | \mu_u, \Sigma(\psi_u)\right) dZ$$

where ϕ is the standard multivariate normal density, $Z_{g,\bullet}$ is the vector of latent Z variables associated with trials in group g. We then have the following theorem:

Theorem 1. *The correlation between Bernoulli variables $Y_{g,i}$ and $Y_{g,j}$, $g \in S_u$, is:*

$$\begin{aligned} \rho_u[i,j] &= \frac{P(Y_{g,i} = 1, Y_{g,j} = 1) - (P(Y_{g,i})P(Y_{g,j}))}{\sqrt{(}P(Y_{g,i})(1 - P(Y_{g,i})))\sqrt{(}P(Y_{g,j})(1 - P(Y_{g,j})))} \qquad (5.2) \\ &= \frac{\int_{Z_i>0}\int_{Z_j>0} \phi\left((Z_i, Z_j)|\ (\mu_{u,i}, \mu_{u,j}),\ \begin{pmatrix} 1 & \psi_u \\ \psi_u & 1 \end{pmatrix}\right) - (P(Y_{g,i})P(Y_{g,j}))}{\sqrt{(}P(Y_{g,i})(1 - P(Y_{g,i})))\sqrt{(}P(Y_{g,j})(1 - P(Y_{g,j})))} \end{aligned}$$

We use the notation $\Sigma[g]$ for the $g \times g$ part of the covariance associated with group g. In the above setting, in which there are groups $g_1, ..., g_m$ which partition the data, the auxiliary variables $\mathbf{Z} = (Z_1, ..., Z_N)$ are distributed according to the multivariate normal density

$$\mathbf{Z} \sim \phi\left(\mathbf{Z}|\mu, \Sigma\right) \qquad (5.3)$$

$$\Sigma[g] = \begin{pmatrix} 1 & \rho_g & .... & \rho_g \\ .. & ... & ..... & .... \\ .. & \rho_g & .... & 1 \end{pmatrix}; \qquad g = g_1, ..., g_m \qquad (5.4)$$

$$\mu = (\mu_{g_1}, ..., \mu_{g_m}) \qquad (5.5)$$

Theorem 2. *With standard results involving pattern matrices (Rowe (2002)[46]) we can show that the conditional distribution of each auxiliary variable Z_i in group g with group size n_g given the remaining auxiliary variables is:*

$$(Z_i|Z_{-i}) \sim$$

$$\phi\left\{\mu_g + \sum_{j\in g;\ j\neq i}(Z_j-\mu_j)\frac{\rho_g}{1+(n_g-2)\rho_g}, \sqrt{\frac{(1-\rho_g)(1+(n_g-1)\rho_g)}{1+(n_g-2)\rho_g}}\right\}$$

Theorem 3. *Using our formulation, supposing that the model parameters Ω have been estimated by $\widehat{\Omega}$ from the training data, we can predict the probabilities of trials $j_1, ..., j_l$ for the test data corresponding to a group g by maximizing the probabilities, for $j_1, ..., j_l \in g$,*

$$P(Y_{j_1} = i_1, ..., Y_{j_l} = i_l) = \int_{A[i_1]} ... \int_{A[i_l]} \phi(Z|\widehat{\Omega})dZ \tag{5.6}$$

over $i_1, ..., i_l \in \{0, 1\}$.

Theorem 2 can be used to simulate all auxiliary variables Z_i. This can be useful when group sizes are large as calculating the exact probabilities using Theorem 3 can become unmanageable to evaluate in an algorithm with a large number of individual trials for large groups.

We can then use Theorems 2 and 3 to make predictions about test data from training data noting that the estimated probability $\widehat{P}(Y_i = 1)$ that $Y_i = 1$ is:

$$\widehat{P}(Y_i = 1) = \Phi\left\{\frac{\mu_g + \sum_{j\in g;\ j\neq i}(Z_j-\mu_j)\frac{\rho_g}{1+(n_g-2)\rho_g}}{\sqrt{\frac{(1-\rho_g)(1+(n_g-1)\rho_g)}{1+(n_g-2)\rho_g}}}\right\} \tag{5.7}$$

5.2 Bayesian Formulations: No Covariates

We juxtapose a variety of Bayesian formulations. In order to ensure parameter identifiability, we assume below that all covariance matrices Σ are correlation matrices with ones on the diagonal. One simple non-hierarchical formulation assumes standard priors for the parameters and known hyperparameters:

$$\mu_u \sim \mathcal{N}(\mu_{\text{mean}}, \tau^2_{\text{mean}}) \tag{5.8}$$

$$\psi_u \sim \left(\frac{\exp(\eta)}{1+\exp(\eta)} | \eta < \eta_0 \right); \eta \sim \mathcal{N}(\mu_\psi, \sigma_\psi) \tag{5.9}$$

$$\eta_0 = \max_{\eta<\eta_0} \Sigma(\frac{\exp(\eta)}{1+\exp(\eta)}) \text{ is invertible} \tag{5.10}$$

$\Sigma(\frac{\exp(\eta)}{1+\exp(\eta)})$ represents the covariance matrix with the value $\frac{\exp(\eta)}{1+\exp(\eta)}$. We estimate the parameters a-posteriori.

An alternate hierarchical formulation assumes priors for the parameters. In this case we assume that the hyperparameters have prior distributions.

$$\mu_u \sim \mathcal{N}(\lambda, \tau); \quad \lambda \sim \pi_1(\lambda);\ \tau \sim \pi_2(\tau) \tag{5.11}$$

$$\psi_u \sim \left(\frac{\exp(\eta)}{1+\exp(\eta)} | \eta < \eta_0 \right); \tag{5.12}$$

$$\eta \sim \mathcal{N}(\mu_\psi, \sigma_\psi);\ \mu_\psi \sim \pi_3(\mu_\psi);\ \sigma_\psi \sim \pi_4(\sigma_\psi) \tag{5.13}$$

5.3 Bayesian Formulation with covariates

It can also be useful to incorporate observation covariates into the model. In the airline application, it is quite relevant as mentioned earlier to allow covariates into the modelling of individual passenger no-show probabilities.

In the Bayesian formulation with covariates, we adopt the earlier indexing formulation; however, for simplification we use standard indexing below. In order to ensure parameter identifiability, we assume below that all covariance matrices Σ are correlation matrices with ones on the diagonal. The joint probability of all trials is given by:

$$P(Y_1 = i_1, ..., Y_n = i_n) = \int_{A_1} ... \int_{A_{i_n}} \phi\left(Z | (X_1'W, ..., X_n'W), \Sigma\right) dZ \tag{5.14}$$

Assume all groups are (conditionally) mutually independent. Groups that belong to the same set share a common normal correlation. We then produce the following theorem:

Theorem 4. *The correlation between Bernoulli variables Y_i and Y_j, belonging to same group $g \in S_u$ is:*

$$\begin{aligned}
\rho_u[i,j] &= \qquad\qquad (5.15)\\
&= \frac{P(Y_i=1,Y_j=1)-P(Y_i=1)(P(Y_j=1))}{\sqrt{P(Y_i=1)(P(Y_j=1))(1-P(Y_i=1))(1-(P(Y_j=1)))}}\\
&= \frac{\int_{Z_i>0}\int_{Z_j>0}\phi\left((Z_i,Z_j)|\,(X'W),\begin{pmatrix}1 & \psi_u\\ \psi_u & 1\end{pmatrix}\right)-\Phi\left(X_i'W\right)\Phi\left(X_j'W\right)}{\sqrt{\Phi(X_i'W)\Phi(X_jW)*(1-\Phi(X_iW))(1-\Phi(X_jW))}}
\end{aligned}$$

All groups could share the same vector of parameters W, or the model can allow for different W per group. Again, depending on the context, it might be more useful to go for one approach or the other.

In the airline application context, we assume a same vector of parameters W for all groups (and thus all observations). The relationship between covariates and the probabilities of success do not necessarily need to vary by trials in different groups by our grouping approach. For example, the relationship between the Age of the passenger and the probability of a passenger not-showing up or not should not vary by whether the passenger is part of a group or not. This assumption can be challenged in other applications and other grouping options, as such the model can be adapted to allow for this. Furthermore, by allowing all groups and trials to share the vector of parameters W, we combine the historical data for all trials and have more information for the estimation of each parameter. Unless there is a clear reason or data evidence to support varying W by different groups, then the approach of using one shared W is preferred.

5.4 Estimating the number of no-show passengers: Explicit Groups model

Assume the Bernoulli observations count the number of booked passengers who do not show up.

1. Identify groups of passengers on a current flight using common characteristics with groups of passengers from past flights. Based on the specific context of the data, the model can easily be adapted to meet a number of criteria. For example, allowing for intergroup correlations, how to group (and create sets) of trials for the intragroup correlations, allowing intragroup correlations to vary by different group types or not, etc.

2. Construct posterior simulations for the group parameters using former data, using the simulation MCMC based inference methodology proposed by Chib and Greenberg (1998).

3. Simulate the probabilities of equation (5.1) if no covariates are present, or (5.14) if covariates are present, an appropriate number of times.

4. Compute predicted probabilities using Theorem 2 that the unobserved Bernoulli random variables are equal to 1.

5. For each particular flight use MCMC to determine estimates for the number of passengers not showing up for the flight.

5.5 Explicit Groups - Ensemble Method Combination

Another interesting idea to explore is combining the Explicit Groups model within an ensemble method framework. By doing so, we expect to be able to improve the prediction results of the Explicit Groups. A number of ways could be derived to put the Explicit Groups model in an ensemble method framework. We provide one initial idea that can serve as a starting point for future research exploring this idea even further.

Our idea explores a way to perform bagging of multiple instances of the Explicit Groups model. We perform 10 iterations where in each iteration we take a random sample with replacement representing 50% of the entire training data population and we take a random sample where only 20% of all the features are used in each iteration.

We then average the predicted probabilities provided by each iteration of the process to get the final aggregated probabilities for each trial. We note that these parameters can be adjusted to fit other needs, and other ensemble methods and approaches could also yield even better results. We will present the results of this model in the comparison sections along with the other models and algorithms.

CHAPTER 6

INFERRED GROUPS MODEL

6.1 Parametric Formulation

Marin and Robert (2007) ([33]) treat the problem of parametrically inferring the presence of groups in data. We omit the assumption that trials can be explicitly partitioned into groups. In its place we assume a univariate probit model whose parameters are assumed to be distributed according to a nonparametric mixture model. Below, we describe the algorithm we use.

1. The passengers on past flights are identified as $o_1, ..., o_m$. Their observed covariate vectors are denoted by $X_1^{(O)}, ..., X_m^{(O)}$.

2. Assume an independent probit model for the probabilities of success for the past and current flights:

$$P(Y_i = 1) \quad = \quad \int_{Z_i > 0} \phi(Z_i | X_i^{(O)} W, 1) dZ_i; \qquad i = 1, ..., m \tag{6.1}$$

3. Assume that the observed covariate vectors are distributed a priori according to multivariate normal distributions with mean vector μ_X and covariance matrix Σ_X:

$$(X_1^{(O)}, ..., X_m^{(O)}) \quad \sim \quad \mathcal{N}(\mu_X, \widehat{\Sigma_X}); \quad \widehat{\Sigma_X} = COV(X)$$

4. Assume that the parameters μ_X associated with the observed vectors $X_1^{(O)}, ..., X_m^{(O)}$ are distributed according to a Dirichlet process with base normal measure H.
$$\mu \sim \mathcal{D}P(H, \alpha)$$

5. Simulate the above parameters a posteriori using standard results from the Dirichlet process. Call the b'th simulation: $\mu_X^{(b)}$ (b=1,...,B).

6. The passengers on the current flight are identified as $c_1, ..., c_l$. Outcomes for these passengers are yet unknown. Their associated Bernoulli variables are given by $Y_1^{(C)}, ..., Y_l^{(C)}$. Denote their covariate vectors by $X_1^{(C)}, ..., X_l^{(C)}$.

7. The probability that $Y_i^{(C)} = 1$ is estimated by:

$$\begin{aligned} \widehat{P}(Y_i^{(C)} = 1 | X_i^{(C)}) &= \frac{\widehat{P}(Y_i^{(C)} = 1, X_i^{(C)})}{P(X_i^{(C)})} \qquad (6.2) \\ &= \frac{\sum_b \Phi\left(X_i^{(C)} W\right) \phi(X_i^{(C)} | \mu_X^{(b)})}{\sum_b \phi(X_i^{(C)} | \mu_X^{(b)})} \end{aligned}$$

6.2 Nonparametric Formulation

For a nonparametric formulation and approach for future predictions, see Muller et al. (2015) [37], Shahbaba and Neal (2009)[49], Neal (2000)[38], and Rossi (2014)[45].

We again omit the assumption that trials can be explicitly partitioned into groups. In its place we assume a univariate probit model whose parameters are assumed to be distributed according to a nonparametric mixture model. Below, we describe the algorithm we use in this case.

1. The passengers on past flights are identified as $o_1, ..., o_m$. Their observed covariate vectors are denoted by $X_1^{(O)}, ..., X_m^{(O)}$.

2. We assume a conditionally independent probit model for i=1,...,m:

$$Y_i = \begin{Bmatrix} 1 & if\ Z_i > 0 \\ 0 & if\ Z_i < 0 \end{Bmatrix}$$

$$Z_i \sim \mathcal{N}(X_i^{(O)} W, 1)$$

$$(X_i^{(O)} | \mu, \sigma_0) \sim G$$

$$G \sim \mathcal{DP}(H, \gamma)$$

$$\overline{X} = \frac{\sum X}{n}$$

$$COV(X) = \frac{\sum((X - \overline{X})(X - \overline{X})')}{n}$$

$$W \sim \mathcal{N}(0, H)$$

with $\mathcal{DP}$ denoting the Dirichlet Process.

3. Simulate the Z's a posteriori given the μ's and W. To do this, let $T_1, ..., T_n$ be the indices induced by the L clusters resulting from the Dirichlet Process. Define vectors $\mathbf{Z}_1, ..., \mathbf{Z}_L$ via:

$$\mathbf{Z}_j = \{Z_i : T[i] = j\}; \quad n_j = \sum_{T[i]=j} 1; \quad j = 1, ...L$$

Let $\mathbf{W}_j$ denote the matrix consisting of n_j copies of the row vector W. By assumption and standard properties, the distribution of $\mathbf{Z}_j$ is a truncated multivariate normal and individual Z variables inherit this.

$$Y_i = \begin{Bmatrix} 1 & if\ Z_i > 0 \\ 0 & if\ Z_i < 0 \end{Bmatrix} \tag{6.3}$$

$$(Z_i | \mu, W) \sim \mathcal{N}(\mu'_{T[i]} W, \sqrt{1 + W'TW})$$

4. Simulate the μ's a posteriori given cluster centers; Simulate Z's given the μ's, Y's, and cluster centers. Finally, simulate W aposteriori given the μ's, Z's and Y's using standard properties of the multivariate normal distribution.

5. Simulate the W vector a posteriori via:

$$W \sim \left(\sum X_i X_i' + (1/\kappa)\right)^{-1} \left(\sum X_i Z_i\right)$$

6. The passengers on the current flight are identified as $p_1, ..., p_l$. Their associated Bernoulli variables are given by $Y_1^{(C)}, ..., Y_l^{(C)}$. Denote their covariate vectors by $X_1^{(C)}, ..., X_l^{(C)}$.

7. The probability that $Y_i^{(C)} = 1$ is estimated by:

$$\begin{aligned} \widehat{P}(Y_i^{(C)} = 1) &= \frac{\widehat{P}(Y_i^{(C)} = 1, X_i^{(C)})}{P(X_i^{(C)})} \qquad (6.4) \\ &= \frac{\sum_T \Phi\left(X_i^{(C)} W^{(T)}\right) P(X_i^{(C)} | \mu^{(T)}) n[T]}{\sum_T \phi(X_i^{(C)} | \mu^{(T)}) n[T]} \end{aligned}$$

6.3 Estimating the number of no-show passengers: Inferred Groups model

Assume the Bernoulli observations count the number of booked passengers who do not show up.

1. Construct posterior estimates for the parameters based on results from former observations.

2. Compute probabilities that a passenger will show up using the parameter estimates (6.2 or 6.4).

3. For each particular flight add up the estimated probabilities.

4. This results in a posterior estimate for the total expected number of passengers not showing up for the flight.

CHAPTER 7

GWL2BOOST ALGORITHM

In this chapter we present our third proposed method. We make adaptations to the WL2Boost algorithm proposed by Adewale et al. (2010)[1]. The adaptations made are the following:

1. We adapt the variance-covariance matrix to allow for a grouped trial correlation structure rather than matched pair binary responses. Adewale et al. (2010)[1] discussed how their approach can be extended to more general correlation structures; we make one such extension that fits our grouped trial idea. For this reason, and for distinction purposes, we call the proposed extension on the algorithm GWL2Boost for grouped weighted least squares boosting.

2. We make a small adjustment in each step of the boosting iterations where the classifier is updated. The adjustment allows for a more direct way to get estimated probabilities and not just classifications. We show how to get these estimated probabilities and use these for comparison metrics that make use of individual trial predicted probabilities.

3. We make an adjustment that allows for a more flexible way to explicitly incorporate the correlation between predicted observations in the final classifier.

We present the WL2Boost algorithm and specify where we make the mentioned adaptations.

7.1 WL2Boost Algorithm with proposed adaptations

In Adewale et al. (2010)[1], the data structure used allows for multiple observations for a given subject or trial. In our case, observations belong to groups, as mentioned throughout our research. We note that the notation presented for this algorithm is specific to this proposed method, and is not necessarily applicable to the other proposed methods in our research.

Let $\mathbf{y} = (y_1, y_2, ..., y_n)^\top$ denote the n binary outcomes. y_i is i'th bernoulli trial/passenger outcome (i = 1,..., n). We note that in this algorithm each trial y_i is classified into either -1 or 1.

We let g ($g \in \{2, ..., G\}$) denote the group size of a given group of trials. For example, $g = 2$ would denote groups of passengers flying in groups of size 2. Passengers within a group of size g share a common intragroup correlation ρ_g. Passengers $y_{i,g}$ and $y_{j,g}$ are grouped if they belong to a group (passengers that booked together in this case) of size g. Let $\mathbf{x_i}$= $(x_{i,1}, x_{i,2}, ..., x_{i,p})$ denote the p covariates associated with passenger i. Let $\mathbf{F} = (\mathbf{F}(\mathbf{x_1}), \mathbf{F}(\mathbf{x_2}), ..., \mathbf{F}(\mathbf{x_n}))^\top$. Let $\mathbf{F_g}$ be all $\mathbf{F}$ corresponding to trials that belong to groups of size g.

Let $\mathbf{V} = \mathbf{var}(\mathbf{y})$ denote the variance-covariance matrix of $\mathbf{y}$. As Adewale et al. (2010)[1] noted, the structure of $\mathbf{V}$ is prespecified, in their case as an exchangeable variance-covariance matrix where the correlation between any pair of repeated measurements of a subject is assigned a correlation parameter (or parameters) and independence is assumed between subjects. In our case $\mathbf{V}$ is an exchangeable variance-covariance matrix where ρ_g is the correlation shared between any pair of passengers belonging to a group of size g. Independence is assumed between passengers that do not belong to a same group. This structure can be adapted to flexibly allow for other correlation structures which can be of interest in other applications. We also define $\mathbf{V_g}$ as a subset of $\mathbf{V}$ corresponding to the variance-covariance portion associated to a group of size g.

For example, for all groups of size $g = 2$, $\mathbf{V_g}$ would correspond to:

$$\mathbf{V_2} = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}$$

An adaptation to the WL2Boost algorithm is the introduction a scalar W_g, defined as:

$$W_g = \frac{1}{1 + (g-1)\rho_g}$$

We use W_g in the adapted WL2Boost algorithm below, and explain its consequences further below.

Let r denote the step-size parameter used in the WL2Boost algorithm. We change the notation used by Adewale et al. (2010)[1] for this parameter as we make use of ρ's to denote the correlation parameters.

The loss function is defined as:

$$L(\mathbf{y}, \mathbf{F}) = \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{y} - \mathbf{F})^\top \mathbf{V^{-1}}(\mathbf{y} - \mathbf{F})$$

The loss function is thus weighted by $\mathbf{V}$.

Then the gradient descent procedure begins:

1. Initialize $\hat{F}_0(\mathbf{x_i}) = \mathbf{0}$ for all subjects $i = 1, ..., n$, and initialize $\mathbf{V_0} = \mathbf{I}$.

2. Repeat for m = 1, 2, ..., M:

   a. Calculate the negative gradient vector:

$$\begin{aligned} \mathbf{u} &= -\frac{\delta L(\mathbf{y}, \mathbf{F})}{\delta \mathbf{F}} \\ &= \hat{\mathbf{V}}_{m-1}^{-1}(\mathbf{y} - \hat{\mathbf{F}}_{\mathbf{m-1}}) \end{aligned}$$

   evaluated at the previous estimate $\mathbf{F} = \hat{\mathbf{F}}_{\mathbf{m-1}}$, $\mathbf{V} = \hat{\mathbf{V}}_{\mathbf{m-1}}$.

   b. Select an index $l_m \in \{1, ..., p\}$ such that

$$l_m = \underset{1 \le l \le p}{\text{argmin}}\{||\mathbf{u} - \mathbf{g_l^{(m)}}(\mathbf{x})||^{\mathbf{2}}\}$$

   where $\mathbf{g_l^{(m)}}(\mathbf{x})$ can be any function that best approximates $\mathbf{u}$.

c. Perform numerical search for best step-size and parameters of $\mathbf{V}$

$$(r_m, \mathbf{V_m}) = \underset{\mathbf{r},\mathbf{V}}{\operatorname{argmin}} \mathbf{L}(\mathbf{y}, \hat{\mathbf{F}}_{\mathbf{(m-1)}} + \mathbf{r}\hat{\mathbf{g}}_{\mathbf{l}}^{\mathbf{(m)}}(\mathbf{x}))$$

Update

$$\hat{\mathbf{F}}_{\mathbf{(m)}} = \hat{\mathbf{F}}_{\mathbf{(m-1)}} + \mathbf{c}\hat{\mathbf{r}}_{\mathbf{m}}\hat{\mathbf{g}}_{\mathbf{l}}^{\mathbf{(m)}}(\mathbf{x})$$

where c is a small regularization constant.

We add an additional constraint in this step

$$\hat{\mathbf{F}}_{\mathbf{(m)}} = \mathbf{max}(\mathbf{min}(\hat{\mathbf{F}}_{\mathbf{(m)}}, \mathbf{1}), -\mathbf{1})$$

3. For each group of trials of size g, adjust the final classifier $\hat{\mathbf{F}}_{\mathbf{g},\mathbf{(M)}}$

$$\hat{\mathbf{F}}_{\mathbf{(g,M)}} = \hat{\mathbf{W}}_{\mathbf{g},\mathbf{(M)}}\hat{\mathbf{V}}_{\mathbf{g},\mathbf{(M)}}\hat{\mathbf{F}}_{\mathbf{g},\mathbf{(M)}}$$

Repeat for all group sizes g.

$\hat{W}_{g,(M)}$ is $\hat{W}_g$ evaluated with the optimal parameters found for $\hat{\mathbf{V}}_{(M)}$.

See subsection 7.1.1 for a detailed explanation of this step.

4. Output the final classifier sign$[\hat{\mathbf{F}}_{\mathbf{(M)}}]$ as final classifier.
5. Output the final estimated probabilities $\hat{\mathbf{p}}_{\mathbf{(M)}}$

$$\hat{\mathbf{p}}_{\mathbf{(M)}} = \frac{\hat{\mathbf{F}}_{\mathbf{(M)}} + \mathbf{1}}{\mathbf{2}}$$

We note that in Step 2(b) Adewale et al. (2010)[1] used a simple parameterized linear function, where in each iteration, $\hat{\mathbf{g}}_{\mathbf{l_m}}^{\mathbf{(m)}}(\mathbf{x}) = \hat{\beta}_0^{(m)} + \hat{\beta}_1^{(m)} x_{i,l_m}$ is the least squares fit of the most significant predictor in that Step. We tested various functions including the same linear function, decision tree regressors and regression splines. The best performing functions for our data were decision tree regressors with max depth = 2.

The regularization constant c should be a small positive value, usually between .01 and .5 that can ensure a slow learning rate of the algorithm. Typically, the smaller the learning the more iterations that are required. The constant c and the number of total iterations M can also be seen as hyperparameters to be tuned for optimization. Cross-validation is a common method for finding optimal hyperparameter values, and it is the method we use in our application.

The additional constraint in Step 2(c) ensures that the classifier updated at each step is bounded by -1 and 1. This makes intuitive sense as what we are trying to estimate shares these bounds. Adewale et al. (2010)[1] did not explicitly mention whether they perform this constraint. Since they only used the final classifier sign for classification, this is not really needed in their application. In our case, this extra constraint helps to enable a proper transformation of the final classifier into an estimated probability.

In Step 2(c) we find optimal parameters for the step-size and correlation parameters of $\mathbf{V}$ that minimize the Loss Function. Adewale et al. (2010)[1] did this by fixing all parameters except one, for example fixing the correlation parameters in $\mathbf{V}$, and minimizing the Loss Function with respect to the remaining parameter (step-size r) analytically. The minimization would then follow an iterative process, minimizing the Loss Function with respect to one parameter, given the previous values of the remaining parameters. The iterative process would continue until convergence. Software packages in common languages such as Python and R can also minimize a given function with respect to a set of parameters, and can also do this by respecting specific bounds for each parameter. We note that in our example, we set bounds to the correlation parameters of $\mathbf{V}$, meaning all ρ_g's, such that $0 \leq \rho_g < 1$ and the step-size parameter is bound by $0 \leq r \leq 1$.

The proposed algorithm can be extended to other correlation structures. In our case, the group type and group size are actually the same, and therefore notation is somewhat simplified. Imagine an example where we had a portfolio of loans with loans that originated either in California or New York. We could allow the group type to be the state of origination, and let all loans belonging to either group to share an intragroup correlation with all other loans in their group. In this case we would have two correlation parameters, say ρ_{CA} and ρ_{NY}. The group size of each group could be any group size, say g_{CA} and g_{NY}. The structures of $\mathbf{V}$ and W_g would need to be prespecified to make use of these parameters, and the same algorithm could be used.

7.1.1 Final Classifier Group Aggregation in Prediction Set

Step 3 is a modification we make to the WL2Boost algorithm that introduces sharing of information between grouped trials into their predicted final classifier/probability. The correlation structure, through $\mathbf{V}$, influences the results of the WL2Boost algorithm in two main ways. First, it impacts the resulting function $\hat{\mathbf{g}}_{\mathbf{l_m}}^{(\mathbf{m})}(\mathbf{x})$ in Step 2(b) of each iteration. Second, it impacts the step-size estimation in Step 2(c) of each iteration. Both effects occur at the training data level. The correlation structure of the prediction/test data set does not have a direct impact on the final predictions for the prediction data set. While the outcome of the prediction trials is unknown, their correlation structure is known. We know how the test trials are grouped. Therefore, modifying the final classifier/probabilities of the test data based on the known correlation structure of these trials could add value.

Adewale et al. (2010)[1] made a small adjustment that uses the known correlation structure to modify the final classifiers. For a given matched-pair that is classified they let the sign of the largest margin in absolute value dictate the classification of the pair. In other words, and adapted to our grouped data, this means taking the maximum value (in terms of absolute value) of the final classifiers amongst each pair and assigning this as the final classifier for all trials in the given group. This could also be done for aggregating the final estimated probability at a group level.

It can be argued that all trials in a group should share a final classifier and estimated probability, and that the aggregation should be done using the maximum values amongst all members of a group. In fact, in the application to match-pair data of Leukemia-Data (Adewale et al. (2010)[1]), it is clear why such aggregation method makes sense. In our analysis and application, we tested this method of modifying the final classifiers/probabilities and also other methods. We also aggregated using the mean values of the final classifier and estimated probabilities for a group. The mean value of the final classifier amongst group members would be assigned to all members of that group, this could be seen as another version of Step 3. Then Steps 4 and 5 would still be used, but using this aggregated average final classifier values for each group member. We tried aggregating using the maximum, mean and minimum values. Each option would have a different interpretation. For example, using the minimum value would say that groups of passengers either show up or not based on the likelihood of it's group member with the highest likelihood of showing up (meaning lowest probability of not showing up).

The aggregation method proposed in Step 3 is more flexible and provides other benefits over these other aggregation methods. While it is the case that in many applications, pairs or groups of trials are very highly correlated in their outcomes, it might not be the case that group members always share the same outcome. The aggregation methods of assigning the mean, minimum or maximum values, would all implicate that we always assume group members will have the exact same outcomes. The aggregation method proposed in Step 3 can be seen as taking a weighted average of the final classifiers/probabilities amongst the group members. Where more weight is given towards the individual probability of a given trial. The higher the correlation parameter of that group, the more that the classifier/probabilities of other group members are combined.

We present an example of the aggregation occurring at Step 3, using a given group of 2 passengers (g = 2). We see how the final classifier is adapted using its final classifier along with the final classifier of its corresponding paired trial. Let $\hat{F}_{(M),1}$ and $\hat{F}_{(M),2}$ represent the final classifier of members 1 and 2, respectively, of this paired group of passengers.

$$
\begin{aligned}
\hat{\mathbf{F}}_{(\mathbf{g=2,M})} &= \hat{W}_{2,(M)}\hat{\mathbf{V}}_{2,(M)}\hat{\mathbf{F}}_{(\mathbf{2,M})} \\
&= \frac{1}{1+\hat{\rho}_{2,(M)}}\begin{pmatrix} 1 & \hat{\rho}_{2,(M)} \\ \hat{\rho}_{2,(M)} & 1 \end{pmatrix}\begin{pmatrix} \hat{F}_{(M),1} \\ \hat{F}_{(M),2} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\hat{F}_{(M),1}+\hat{\rho}_{2,(M)}\hat{F}_{(M),2}}{1+\hat{\rho}_{2,(M)}} \\ \frac{\hat{F}_{(M),2}+\hat{\rho}_{2,(M)}\hat{F}_{(M),1}}{1+\hat{\rho}_{2,(M)}} \end{pmatrix}
\end{aligned}
$$

We can see that in the extreme case that if the correlation parameter is 1, then this aggregation method would be equivalent to assigning the average value of all final classifiers/probabilities to all group members. We also note that we make use of the estimated correlation parameters of the **V** matrix from the training data set, and apply it to the prediction/test data set explicitly.

The aggregation method proposed is another proposed example, or option. Depending on different situations, levels of correlation in data, intended objectives, etc. different methods of aggregation could be used and preferred. We see this as another benefit of the potential of this algorithm for predicting correlated binary data. For our data structure, where groups sizes are not very large, grouped trials are highly correlated in their outcome but do not always have concordant outcomes, this method of aggregation made sense to use.

We also note that this idea of aggregating, or sharing information of, estimated classifiers/probabilities amongst grouped/paired correlated trials can be adapted to fit other ensemble methods for classification. Where, in general, Step 3 could be applied to modify any final classifier/estimator provided by a prediction/classification algorithm. The WL2Boost algorithm can be used to get the estimates of **V** that are used per our recommended aggregation method. Of course, the parameters required by our recommended Step 3 could also be seen as hyperparameters to be tuned. In other words, the weighting parameters in our recommended Step 3 (in our case ρ_g's could be optimized in the context of another type of classification algorithm.

In the results, we compare the aggregation method used by Adewale et al. (2010)[1] and our method of aggregation and show that our proposed method performs better with the airline data.

CHAPTER 8

REAL AND SIMULATED DATA - RESULTS

In this section we introduce the real data available to fit the models proposed. We show evidence in real data that supports the motivation for the models proposed. We also generate simulated data sets meeting different criteria. Both fit the airline data to the proposed model and algorithm extensions as well as some existing competing models and algorithms. We present the results of these comparisons, with a focus on prediction capacity of different models and algorithms on the airline data. We show that the proposed extension on WL2Boost outperforms even top out-of-the-box ensemble classification algorithms that do not assume correlation of outcomes.

As the idea of correlation between trials is a key motivation for the proposed models, we also provide some frequentist and Bayesian analysis of correlation in the airline data.

8.1 Airline Data Overview

A major international airline provided access to a dataset (airline data) consisting of historical flights. Each observation in the historical dataset represents a booked passenger for a flight. Each observation has 154 unique covariates (73 at the flight level, 81 at the passenger level) as well as covariates used to determine whether the passenger showed up or not to the flight.

For the purposes of the analyses shown in this proposal we worked with a sample of 418 unique flights all corresponding to one of either two specific flight routes and times (same origin city, destination city and same departure time). We can see how these flights are distributed in Table 8.1. Both flights have a same origin city A and destination city B, but have different flight departure times.

Origin	**Destination**	**Departure Time**	**Total Flights**
A	B	W	272
A	B	X	146

Table 8.1: Number of Flights in Data per Flight Routes

One route is for a night departure time and another one is for an afternoon departure time. The cleaning and review process of these flights and the respective covariates was extensive.

We worked with these combinations of origin and destinations because we want to test for a mix of ways to combine the data when training each model. We will work with two different ways of training each model:

1. Train a separate model using only data for a specific route/departure time to predict future flights for that same specific route/departure time. For example, train a model using only the training data set from flights A-B-W (see Table 8.1) to predict the future flights of A-B-W.

2. Combine the data of routes from a same origin and destination (and differing departure time) to predict the future data corresponding to flights for that combination of origin and destination. For example, use the training data set from flights A-B-W and A-B-X to predict the data from future flights of both A-B-W and A-B-X.

By evaluating and comparing the prediction results across these methods, we could get a better understanding for how to improve the model training and selection process in practice. For example, if we see that method 2 provides better or similar prediction results as method 1, then it might be preferable for the airline to use method 2 as it does not require training, maintaining and selecting a model for each specific route/departure time combination.

8.2 Frequentist analysis of correlation

Initial basic data analysis on the airline data shows evidence of within flight passenger correlation, or at least overdispersion.

The total number of passengers not showing up for a given flight does not conform to a binomial distribution as could be expected, especially examining passenger no-show rates by different passenger group sizes. We define a passenger group as any group of passengers that are part of the same booking transaction for a given flight. If two passengers booked together, then they belong to passengers of group size 2.

Whenever we see groups of booked passengers of size 2 and up, regardless of the size, the data is very skewed towards either all passengers in the group showing up or none showing up. For every group of passengers of size 2 and up, the data shows that more than 98% (this varies by flight, route, etc.) of groups have either all or none of the passengers belonging to a group showing up. The rate of how many passengers fly in groups varies by route, time of year and other factors, but we see for our data set that on average 50% to 60% of all booked passengers belong to a group (meaning, they booked together with at least one other passenger). Meaning this grouping method applies to most of our binary trials in this application. Given the high correlation of how grouped passengers behave together, it seems intuitive to use this grouping type as the groups for our purposes of modelling group correlation.

The rate of booked passengers all behaving the same as their group members, meaning either all or no passengers that belong to a group show up, is very high (more than 98% on average). The variance seems to be higher than expected if the data conformed to a binomial distribution. We perform a binomial variance test for homogeneity on each of the specific flight routes.

The test statistic used follows a chi-square distribution with $(n-1)$ degrees of freedom, where n represents the total number of flights available for a specific route, and is given by:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Where s^2 represents the sample variance and σ^2 is the expected variance if the data followed a binomial distribution.

Given that flights vary in number of booked passengers (individual trials), we specify that the test statistic follows a chi-square distribution with $(n-1)$ degrees of freedom and is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(U_i - m_i\hat{p})^2}{m_i\hat{p}(1-\hat{p})}$$

Where n represents the total number of flights available for a specific route, U_i the observed number of no-show passengers for flight i, m_i represents the total number of booked passengers for flight i and $\hat{p}$ represents the overall estimated probability of no-show across all flights for a given route.

In Table 8.2 we can see the corresponding test statistic for each flight route corresponding to routes in Table 8.1.

Route	**Test Statistic**	**N Flights**	**p-value**
A-B-W	1092.68	272	$4.72e^{-99}$
A-B-X	585.41	146	$3.11e^{-54}$

Table 8.2: Overdispersion Test Statistics and Significance

Both routes show significant overdispersion. We continue with more specific analyses of correlation that provide further evidence of high correlation in the binary data we are working with.

As discussed in chapter 2, the Beta-Binomial model is commonly used when binary data shows overdispersion. Pairs of correlated binary trials, say Y_i and

Y_j with response probabilities p, have a common correlation coefficient δ given by

$$\delta = \frac{P(Y_i = 1, Y_j = 1) - p^2}{pq} \tag{8.1}$$

In the beta-binomial distribution (BBD), and using the parameters α and β from the BBD probability mass function (equation 2.1), the correlation coefficient δ is given by:

$$\delta = \frac{1}{\alpha + \beta + 1} \tag{8.2}$$

Tarone(1979)[53] and Prentice(1986)[43] provided a test statistic for testing that $\delta = 0$ in the beta-binomial model. They showed that for n binomial random variables U_i each with m_i Bernoulli trials, and overall estimated probability of success $\hat{p}$, the test statistic, which follows a standard normal distribution, is given by:

$$\frac{(\hat{p}(1-\hat{p}))^{-1}\sum_{i=1}^{n}(U_i - m_i\hat{p})^2 - \sum_{i=1}^{n} m_i}{(2\sum_{i=1}^{n} m_i(m_i - 1))^{1/2}} \tag{8.3}$$

We find the above test statistic with the data for each flight route. Where again n represents the total number of flights available for a specific route. U_i is the observed number of no-show passengers for flight i, m_i represents the total number of booked passengers for flight i and $\hat{p}$ represents the overall estimated probability of no show across all flights for a given route. The results are presented in Table 8.3.

Route	**Test Statistic**	**p-value**
A-B-W	35.57	$<< .00001$
A-B-X	16.36	$<< .00001$

Table 8.3: BBD Correlation Coefficient $\delta = 0$ Test per Route

As we can see in Table 8.3, every route has a very significant test statistic for testing that the BBD correlation coefficient $\delta = 0$. This provides further evidence of strong correlation between passengers.

8.3 Alternative Bayesian and Boosting approaches for estimates of correlation

The Explicit Groups model proposed allows for different normal correlations for each group of trials. We calculated the normal correlation parameters per group when fitting the proposed Explicit Group model to the airline data. Results for estimated parameters using all routes combined are presented below in Table 8.4 for the first 4 groups (2 through 5). Groups 2 through 5 represent more than 93% of all grouped passengers. Groups 6 and higher have significantly lower number of observations. For these reasons, we focus our results on correlation parameters just for groups 2 through 5.

Passengers are grouped by group size explicitly in this model. As such, Group 2, for example, represents passengers flying in pairs, Group 3 passengers flying in groups of 3 and so on. We recall that in the proposed Explicit Group model, the normal correlation between passengers of a given group type are all the same. The parameters presented are the mean values estimated over 1000 MCMC runs from the posterior density for each of the parameters. We also present the 90% Posterior Density Interval for the parameter estimates, which was obtained by taking lowest 5th percentile and highest 95th percentile values of the MCMC runs.

For grouped passengers we can see significant normal correlations are found. Serving as further evidence of correlation between grouped passengers. We can see that the highest correlation is found for groups of 3 passengers. We can also see that the normal correlation does seem to differ by group. Allowing for different correlations by group would seem to conform to the data.

We also present the estimated correlation parameters $\hat{\rho}_g$'s from the GWL2Boost

Group	**Normal Correlation Parameter**	**90% PDI**
2	.858	(.734, .948)
3	.928	(.806, .972)
4	.720	(.648, .761)
5	.776	(.639, .840)

Table 8.4: Normal Correlation Parameters for Explicit Groups Model with Airline Data

proposed algorithm, again for the first 4 groups in Table 8.5.

Group	$\hat{\rho}$
2	.882
3	.864
4	.847
5	.852

Table 8.5: Estimated Correlation Parameters for GWL2Boost Algorithm with Airline Data

We again see very high values for the estimated correlation parameters. Groups 2 and 3 have the highest estimated correlation parameters.

8.4 Simulated Data Generation

The Explicit Groups model was our initial proposed model during our research. Before we had the airline data cleaned and available, we ran simulations for 2 models to get a better understanding of this model. The first one follows the proposed Explicit Groups model, and we call it the Correlated Simulation, and is described below.

We simulate Bernoulli data respecting groups of trials (e.g., booking groups). See section (5.1) for notation. We use the notation $\mathbf{X}$ for the $n \times k$ matrix of covariates generated for all the trials. We partition the $\mathbf{X}$ matrix into:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ ... \\ \mathbf{X}_u \end{pmatrix}$$

in such a way that $\mathbf{X}_1$ corresponds to the set S_1 covariates, ..., $\mathbf{X}_u$ to the set S_u covariates. Corresponding to this notation, we partition the Bernoulli (trial) $\mathbf{Y}$ vector into:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ ... \\ \mathbf{Y}_u \end{pmatrix}$$

and the normal vector $\mathbf{Z}$ into:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ ... \\ \mathbf{Z}_u \end{pmatrix}$$

In view of our assumption of no intergroup correlation, and in conformity with the above partitions, we can partition the correlation matrix into:

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & ... & 0 \\ 0 & \Sigma_2 & 0.... & 0 \\ .... & & & \\ 0 & 0 & .... & \Sigma_u \end{pmatrix}$$

We assume that the Bernoulli vectors $\mathbf{Y}_1, ..., \mathbf{Y}_u$ are distributed according to (conditionally) independent multivariate normal probit models (see section (5.1)) with

$$\Sigma_v = I_{n_v} + \psi_v M_v$$

We assume one parametric vector W_v for each set S_v (v=1,...,u). The top-level parameters for this model are:

$$\Theta = \{(W_1, \psi_1), ..., (W_u, \psi_u)\}$$

One simulation follows the correlation structure mentioned above. The other is a simulation of independent binary data which will be used to see how models that do not assume correlation vs. the Explicit Group Model (which allows for correlation) compare:

1. Correlated Simulation (SIM1) - simulated binary data, which assumes correlation between passengers that belong to the same group.

2. Independent Simulation (SIM2) - simulated binary data, that does not contemplate any correlation between trials.

We fit the different simulated data to our Explicit Groups model and to two other statistical binary classification models (logistic and probit regression). We do this to get an initial sense of how the Explicit Groups model compares against models that assumes independence between binary trials on binary data that is either correlated or independent. We can see in Table 8.6 the AIC that we get when fitting each model to the different simulated data sets.

Model	**Simulated Data**	**AIC**
Logistic Regression	SIM1	797.51
Probit Regression	SIM1	795.55
Explicit Groups	SIM1	326.2
Logistic Regression	SIM2	707.54
Probit Regression	SIM2	707.53
Explicit Groups	SIM2	642.3

Table 8.6: AIC per Model for Simulated Data

The proposed Explicit Group model shows a somewhat better fit in terms of AIC, even in the independent binary data simulation. What is clear, and also expected, is that the Explicit Group model fits the data significantly better than traditional models that do not assume correlation, when more advanced group correlation structures are present.

8.5 Predicting Correlated Binary Data

We fit the proposed models and use them to predict the test data Bernoulli variables. We do this using the airline data, where we separate a test data set. The two available routes for the airline data consist of consecutive sequential flights for each route, from April 2019 through December 2019 (9 months). The first 8 months of data (April through November) correspond to the training sample. The remaining month (December) of data is used as the testing sample for out-of-time validation. This aligns to how an airline could actually use the models - by using historical data to predict passenger no-show for the upcoming flights for a given time period.

As mentioned earlier, it is of interest in the airline application to have accurate predictions of individual passengers that will not show up. The airline is even more interested in predictions of the total sum of passengers that will not show up to a given flight.

Depending on the application, it might be of interest to have accurate predictions at an individual trial level or for sums of trials. Our proposed methods allow for both types of predictions. We present comparisons for both types of predictions. We show that our proposed algorithm extension GWL2Boost outperforms other competing methods in terms of individual trial predictions, and significantly outperforms other methods in terms of predicting total sums of trials.

8.5.1 Predicting Individual Trial Probabilities

For each model, we estimate (predict on the test sample) the Bernoulli variables $Y_1, ..., Y_n$ to be $\widehat{Y}_1, ..., \widehat{Y}_n$ (respectively). One way to compare the models' predictions at individual trial level, when the Bernoulli variables are observed, is with ROC curves and AUC.

Other mechanisms for comparing predictions at an individual trial level when the Bernoulli variables are observed involves the MSPE (mean squared prediction error) and Mean Absolute Error (MAE) of the predicted probabilities.

$$MSPE = (1/n)\sum_{n}(\widehat{Y}_k - Y_k)^2$$

$$MAE = (1/n)\sum_{n}|\widehat{Y}_k - Y_k|$$

We also calculate Accuracy, Precision and Recall as metrics to evaluate the individual classifications predicted by the different models and algorithms.

8.5.2 Predicting Expected Total Sums of Trials

The airline is interested in being able to predict the total number of booked passengers that will not show up for a specific flight. As such, we also estimate the total number of no-show passengers for each flight in the airline test samples. We then compare this estimate to the actual total number of passengers that did now show up for each flight by both finding the mean squared difference between the expected and actual sums of no-show passengers, the mean absolute difference and the average ratio of expected and actual sums of no-show passengers.

We first find the expected total number of no-show passengers for a given flight with each of the fitted models and algorithms. We do this by first adding up the individual passenger predicted probabilities of no-show for a given flight; we call this expected sum of no-shows, $\widehat{U}_f$ for flight f. Then, given an actual sum of passengers that did not show up for flight f, U_f, the metrics evaluated are:

$$\begin{aligned}
Mean\,Squared\,Difference\,of\,Sums &= (1/n)\sum_n(\widehat{U}_f - U_f)^2 \\
Mean\,Absolute\,Difference\,of\,Sums &= (1/n)\sum_n|\widehat{U}_f - U_f| \\
Mean\,Ratio\,Expected\,To\,Actual\,Sums &= (1/n)\sum_n\frac{\widehat{U}_f}{U_f}
\end{aligned}$$

We also perform a simulation of total number of passengers that will not show up for the flights in the test data. We also incorporate the correlation between grouped trials into our simulation. We perform one simulation where we use the predicted probability at a passenger level and assume each passenger is an independent Bernoulli trial. In the second version of the simulation, we take the average predicted probability for all passengers in a given group. We use this mean probability to simulate whether all members of a given group will show up or not. We expect that this second version of the simulation will capture more extreme cases of the actual number of passengers that did not show up for a flight. This would make sense as we know what groups of passengers tend to either all show up or none show up. We perform 1000 simulations of the future flights. We then take a 90% interval, by taking the interval of the 5th percentile and 95th of the simulated results. We then see how many times we capture the actual number of passengers that did not show up for the future flights simulated.

We provide a further note regarding the data and potential operative use of the models and algorithms presented. The data we have for each historical flight provides with the information regarding all of the passengers that booked the flight. A no-show passenger is defined as such if it cancelled a booking up to 48 hours prior to departure, or if the passenger simply did not show up to the flight. Passengers book flights with varying days of anticipation before the flight. The proposed models could always estimate the expected total passengers that will not show up, from the passengers that have booked up to the point the model is run. The airline can then subtract the expected total passengers that will not show up from the remaining capacity of possible tickets to be sold at any point prior to the flight. The airline can also segment the expected total passengers that will not show up by fare, or other segments of interest in terms of tickets sold.

8.6 Competing Methods and Data Overview

We train the proposed models and competing models and algorithms to airline training data and use these trained models to predict the binary variables in the test samples. The competing models and algorithms represent a combination of different statistical models and machine learning algorithms.

8.6.1 Overview of Proposed and Competing Approaches

We present the list of proposed and competing models and algorithms that will be evaluated using the airline data. For the existing competing models and algorithms, we use the Python libraries Scikit-learn (sklearn) and xgboost with default parameters/hyperparameters except where noted.

Hyperparameter tuning was performed testing two main hyperparameters (that are common amongst some of the ensemble methods below). We tested the number of weak estimators / iterations of the ensemble going from 10 to 100 estimators by increments of 10, and from 100 to 5000 estimators by increments of 100. We also tested different learning rates, going from .01 to .05 in increments of .01 and from .05 to 1 in increments of .05. We used the results on the prediction sample to determine these optimal hyperparameters.

1. Explicit Groups Model - Proposed statistical Bayesian model allowing for within group correlation of binary data as previously discussed in chapter 5. In this specific application, we group passengers (binary trials) by common bookings. Meaning passengers that booked together belong to a same group. Correlation parameters for each group size (ex. 2, 3, 4, etc.) and other model

parameters are estimated using the training sample and then used for prediction in the test sample (see Chapter 5. Parameters are estimated using from the posterior density from 100 MCMC simulations. We use Theorem 3 from chapter 5 for the estimation of probabilities for future trials.

2. Inferred Groups Model - Proposed Bayesian model, allowing for correlation between trials as previously discussed in chapter 6. Groups of trials are not explicitly defined a priori.

3. GWL2Boost-W Algorithm - Gradient descent boosting algorithm, allowing for trial group correlation of outcomes discussed in chapter 7. In summary, it is our modified extension on the WL2Boost algorithm. This version denoted by 'W' uses our proposed Step 3 method for aggregating the final classifier/probabilities. We note that we tested multiple types of weak learner functions within this algorithm: linear regressor, decision tree regressors and spline regressors. The best results were obtained using decision tree regressors with max depth = 2. The optimal number of iterations and learning rate was the combination of 1000 iterations with a learning rate of .05. However, the results with those parameter values were almost identical to those obtained with 300 estimators and a learning rate of .2, therefore we used iterations = 300 and learning rate (c) = .2 for efficiency.

4. GWL2Boost-MAX Algorithm - Same algorithm as the one mentioned above, except that in this 'MAX' version, we perform the same aggregation as Adewale et al. (2010)[1], letting the maximum probability/classifier dictate the entire group probabilities/classifiers (see 7.1.1). We also used decision tree regressors as the weak learners with max depth = 2, iterations = 300 and learning rate (c) = .2.

5. Logistic Regression

6. Classification Decision Tree

7. Random Forest - with 1000 trees

8. AdaBoost w/RF - with base learners random forest with 100 trees, with 5 total estimators and learning rate = .5. We also tested various other base learners, and this provided the best results.

9. XGBoost Classifier - with 300 total estimators and .2 learning rate. Can be seen as a comparable version of the WL2Boost algorithm, however without any correlation between outcomes.

10. Ensemble EG - the initial idea for ensemble method approach for the Explicit Groups model (see 5.5).

We use each of the above models and algorithms with the airline data set as described below.

Each of the models and algorithms are trained using the training set from the airline data. Recall this corresponds to the first 8 months (April 2019 through November 2019) of flight data for 2 different flight routes.

Using the trained models and algorithms, and the associated parameters estimated in the training data with each method, we estimate probabilities of each passenger not showing up in the test sample (1 future month, December 2019). We use both model training and testing approaches mentioned in section 8.1. Where one method consists of training a model for each specific route and using this model to predict the future data for that same route. The second method consists of training and testing each model/algorithm with a combined training sample of routes for a same origin-destination combination.

For each model/algorithm and training/testing approach, we calculate and compare the prediction metrics described in section 8.5.

8.6.2 Data Preparation

Extensive data cleaning and checks were performed for the final data used in the comparison of models/algorithms. However, there were still cases of missing values for specific observations for some covariates.

A simple approach was taken to clean these missing value issues in order to have more complete data to run all the competing models/algorithms. For continuous numerical covariates, missing values were replaced with the mean value of non-missing observations for that covariate. An indicator variable was also created for each continuous numerical covariate, indicating (1) if the observation had a missing value for that covariate or not (0).

For categorical covariates, missing values were left as such and treated as one more possible category. Categorical covariates where also transformed through the common technique of one-hot encoding to allow the competing models/algorithms to run.

8.7 Airline Data Results - Individual Trial Prediction Metrics

We first present the results for the individual metrics (see 8.5.1).

8.7.1 Results - Method 1

The first method of training consists of training each of the competing models/algorithms for each flight route training data and using these to predict the test data for that specific flight route.

In the comparison results tables, we show the prediction metrics for individual trials: AUC, MSPE, MAE, Accuracy (ACC), Precision (PR) and Recall (RC).

In Table 8.7, we present the prediction comparative prediction metrics using the first method of model training and testing, and using the A-B-W flight route testing data.

We can see in Table 8.7 that the results attained on the prediction set across all models are very high.

We see that the ensemble methods in general perform best across most metrics, which is something typically seen in practice. The proposed algorithm GWL2Boost outperforms even some of the top out-of-the-box machine learning algorithms such as random forest and even XGBoost. This is a very promising result that further shows the value that incorporating our proposed group correlation structure has. We also see that GWL2Boost-W out performs GWL2Boost-MAX, which supports our recommendation that our rec-

Model/Algorithm	AUC	MSPE	MAE	ACC	PR	RC
Explicit Groups	0.912	0.045	0.091	0.942	0.810	0.349
Inferred Groups	0.845	0.049	0.117	0.937	0.808	0.255
GWL2Boost-W	0.924	0.040	0.083	0.952	0.903	0.341
GWL2Boost-MAX	0.920	0.041	0.089	0.951	0.828	0.381
Logistic Regression	0.895	0.048	0.095	0.939	0.805	0.288
Decision Tree	0.712	0.071	0.095	0.936	0.677	0.456
Random Forest	0.935	0.047	0.106	0.949	0.828	0.278
AdaBoost w RF	0.941	0.044	0.101	0.946	0.909	0.241
XGBoost	0.921	0.042	0.093	0.949	0.862	0.312
Ensemble EG	0.919	0.044	0.091	0.946	0.895	0.268

Table 8.7: Individual Prediction Metrics per Model - Method 1 - Route 1

ommended aggregation of final classifiers/probabilities performs better than just taking the maximum values for aggregating the classifiers/probabilities for a given group.

AdaBoost with random forests provides the highest AUC and XGBoost provides some of the highest metrics amongst competing algorithms.

It is interesting to see that the Explicit Groups model does outperform some a comparable standard statistical model such as logistic regression. This is aligned to what we had seen during the proposal with a smaller data set and less covariates. The Explicit Groups model outperforms the logistic regression in every comparative metric for this sample except for a slightly lower Precision but it does have a higher Recall. It is also very promising to see that even with an initial simple idea of bagging for the Explicit Groups model, we do see an important lift.

The inferred groups model does not show strong prediction results when compared to the other algorithms. We see opportunities for continuing to improve this model, but for now we do not see promising results when applied to the available airline data.

In Table 8.8, we present the prediction comparative metrics using the first method of model training and testing, and using the second flight route (A-B-

X) testing data.

Model/Algorithm	**AUC**	**MSPE**	**MAE**	**ACC**	**PR**	**RC**
Explicit Groups	0.724	0.048	0.105	0.953	0.782	0.373
Inferred Groups	0.682	0.049	0.127	0.949	0.738	0.241
GWL2Boost-W	0.919	0.036	0.093	0.968	0.800	0.482
GWL2Boost-MAX	0.908	0.039	0.115	0.961	0.645	0.482
Logistic Regression	0.688	0.047	0.103	0.951	0.677	0.388
Decision Tree	0.765	0.045	0.101	0.951	0.782	0.602
Random Forest	0.939	0.035	0.121	0.959	0.893	0.181
AdaBoost w RF	0.953	0.034	0.079	0.957	0.750	0.373
XGBoost	0.898	0.036	0.092	0.962	0.813	0.337

Table 8.8: Individual Prediction Metrics per Model - Method 1 - Route 2

In Table 8.8, we observe a similar tendency as for that of the first route. In general we see high levels of prediction metrics. The Explicit Groups model performs at a similar or better level in most metrics when compared to logistic regression. Again, the Implicit Groups model does not show any significant improvements.

The ensemble methods again show the best metrics. Once again, the GWL2Boost-W has the best metrics except in AUC. In precision it does not have the highest performance, but given its level of recall it is overall performing better. In terms of accuracy, the proposed model is by far the best, with XGBoost the second best. Again, we see that our proposed method for aggregation used in GWL2Boost-W outperforms that of GWL2Boost-MAX.

8.7.2 Results - Method 2

The second method of training consists of combining the routes for a same origin-destination combination. We use this combined data to train each of the competing models, and then calculate the same prediction metrics on the test data.

In Table 8.9 we present the prediction comparative metrics using this second method of model training and testing for the combined routes.

Model/Algorithm	**AUC**	**MSPE**	**MAE**	**ACC**	**PR**	**RC**
Explicit Groups	0.917	0.041	0.075	0.948	0.797	0.333
Inferred Groups	0.754	0.048	0.127	0.943	0.780	0.267
GWL2Boost-W	0.924	0.036	0.067	0.956	0.935	0.338
GWL2Boost-MAX	0.919	0.037	0.072	0.956	0.881	0.369
Logistic Regression	0.914	0.043	0.075	0.946	0.724	0.254
Decision Tree	0.733	0.065	0.075	0.943	0.774	0.421
Random Forest	0.951	0.039	0.109	0.952	0.911	0.284
AdaBoost w RF	0.957	0.038	0.091	0.951	0.959	0.254
XGBoost	0.924	0.039	0.095	0.952	0.867	0.308
Ensemble EG	0.919	0.040	0.078	0.951	0.925	0.265

Table 8.9: Individual Prediction Metrics per Model - Method 2

The results in Table 8.9 show better prediction metrics across the board when compared to the results from Method 1. This shows that combining data from the two routes that shared a same origin-destination actually resulted in having more robust prediction models and algorithms.

The Explicit Groups model again shows better metrics when compared to a statistical model that assumes independence like Logistic Regression. The Inferred Groups model again underperforms against these two statistical models.

It is very promising to see that GWL2Boost-W again provides the best metrics, except for AUC. AdaBoost with random forest provides by far the best AUC. GWL2Boost-W has the lowest MSPE, MAE, highest Accuracy. AdaBoost with Random Forests has the highest Precision, but a significantly lower Recall compared to the GWL2Boost-W, which again has one of the highest Precision metric.

GWL2Boost-W once again outperforms GWL2Boost-MAX, which provides further evidence that our proposed method for aggregation of final classifier/probabilities is preferred.

In summary, GWL2Boost-W outperforms all the competing models and algorithms in terms of individual trial predictions. It is a very good sign that supports our idea that incorporating group correlation structures would result in better predictions. AUC is the only metric where other competing ensemble methods such as AdaBoost with random forests seems to perform better. However, we believe the other performance metrics are actually more valuable for this application.

We show the top 10 features by importance for the GWL2Boost algorithm proposed, the top performing algorithm in the comparisons, in Table 8.10

Importance Rank	**Feature**
1	No show Rate Pass. 2YRS
2	Avg. Show Rate Fl.
3	Trip Fare
4	Language Settings
5	% of Max Price
6	Avg. CH Rate Fl.
7	No Show Rate 3 months Fl.
8	Sales Channel
9	Age
10	No Show Rate Day of Week

Table 8.10: Top 10 Features for GWL2Boost Algorithm

8.8 Airline Data Results - Sums of Trials Prediction Metrics

We now present the results for the sums of trials prediction metrics (see 8.5.2).

As mentioned before, while it is interesting for the airline to predict individual trials and their probabilities, it is of greater value to have good predictions on the total number of passengers that will not show up to future flights.

In Table 8.11 we present the prediction metrics on the total sums of passengers: Mean Squared Difference of Sums (MSDS), Mean Absolute Difference of Sums (MADS), Mean Ratio of Expected to Actual Sums (ExpRatio) and the percentage of coverage by our 90% intervals of the simulations of future flights.

We only use the top three performing proposed methods and top two performing competing methods from the analysis of performance on individual trial prediction metrics from the previous section. These are: Explicit Groups model, GWL2Boost-W algorithm, Ensemble Explicit Group, AdaBoost with random forests and XGBoost.

For both proposed methods that allow for trial correlation, we also performed the simulation that assumes correlation between trials. We use method 2 of combining both routes as we saw better performance in the individual trial prediction metrics.

Model/Algorithm	**MSDS**	**MADS**	**ExpRatio**	**90% CI**
Explicit Groups	19.86	3.50	1.14	.83
GWL2Boost-W	16.18	3.09	1.05	.90
AdaBoost w RF	19.89	3.43	1.14	.83
XGBoost	18.83	3.59	1.14	.83
Ensemble EG	19.22	3.47	1.14	.83

Table 8.11: Trial Sums Prediction Metrics per Competing Model

The results are even more clear of the value added from allowing for correlation of trials. GWL2Boost-W clearly outperforms the other top competing methods in every metric.

GWL2Boost-W has a Mean Absolute Difference of Sums of 3.09, meaning on average it is off by an absolute value of 3.09 passengers in its predictions. The next best method in this metric is AdaBoost with random forest with a value of 3.43, 11% higher than GWL2Boost-W.

GWL2Boost-W has an average ratio of expected to actual sums of no-show passengers of 1.05, while the rest of the methods hover around 1.14-1.15. Finally, we see perhaps the clearest evidence of the value of allowing for correlation in our binary data. In the correlated simulation using GWL2Boost-W, the 90% interval of the 1000 simulations of future flights captures the actual value in 90% of the cases, the rest of the methods only capture 83%.

This gives the airline a very significant tool that outperforms top out-of-the-box ensemble methods. Having such simulations for future flights gives the airline a tool for determining with more certainty how much can they overbook given the overdispersed, highly correlated, nature of this data.

CHAPTER 9

CONCLUSIONS AND FUTURE RESEARCH

9.1 Conclusions

We found that the proposed statistical models (Explicit Groups and Inferred Groups) indeed fit the correlated binary data well, and they also show strong potential as prediction tools, especially the Explicit Groups model. When compared to other statistical models that do not assume correlation between trials, the Explicit Groups model showed strong results despite potential for improvement. The Explicit Groups model becomes computationally burdensome when adding more data and covariates. We will look into making the code more efficient such that we can expand the number of MCMC iterations and with that, we expect to get even more robust estimated parameters from the training data which should continue improving the prediction results.

The Explicit Groups model also provides some useful insights regarding the different correlation structures of different groups of binary trials belonging to the same cluster of trials. We see strong potential for this model in practical applications. It will be interesting to see the results of the Explicit Groups model in other data and applications.

We were able to reach very high level of prediction capacity for the airline data. Some of ensemble methods such as AdaBoost and random forest showed very strong results as prediction tools. Our proposed statistical methods underperformed when compared to these methods, which is common in practice. As such, we incorporated our group correlation idea into a gradient descent boosting algorithm. This proposed algorithm extension GWL2Boost made small adaptations on the existing algorithm WL2Boost and we saw improved results with these adaptations. These adaptations can be incorporated into other ensemble methods with relative ease. GWL2Boost outperformed all other competing models and algorithms in virtually every metric when predicting individual trials and their probabilities. However, when predicting total sums and simulating intervals of predicted total sums of no show passengers, GWL2Boost significantly outperformed even the most widely used ensemble methods.

This new method gives the airline a very powerful tool. The airline currently uses ensemble methods for no-show predictions, but had never incorporated the idea of trial correlation into their algorithms. With this, the airline can determine with greater certainty how much can they overbook given the overdispersed, highly correlated nature of this data.

9.2 Future Work

A number of interesting expansions and ideas have emerged. Key future research would include:

- We want to build statistical packages (in Python and R) that enable others to use our proposed methods. The GWL2Boost algorithm can help as a practical tool for allowing correlation of binary data in a gradient descent boosting framework. We have not found an out-of-the-box package that allows for flexible correlation structures in binary data.

- Expand further on the idea of combining the proposed Explicit Groups model framework into an ensemble method. As we saw, an ensemble method that incorporated correlation outperformed existing ensemble methods. As such, we see opportunity for an even more powerful method by combining these two ideas.

- Further hyperparameter optimization and tuning of competing machine learning algorithms used for passenger no-show prediction.
- Finding other ways to group trials in the Explicit Groups model and GWL2Boost algorithm. Including potentially using unsupervised learning to cluster trials and testing the impact of these grouping approaches vs. a priori defining the exact groups.
- Using conditional random fields to model passenger information over a number of flights and over many blocks of time.

BIBLIOGRAPHY

[1] Adeniyi J Adewale, Irina Dinu, and Yutaka Yasui. Boosting for correlated binary classification. *Journal of computational and graphical statistics*, 19(1):140–153, 2010.

[2] Werner Adler, Alexander Brenning, Sergej Potapov, Matthias Schmid, and Berthold Lausen. Ensemble classification of paired data. *Computational statistics & data analysis*, 55(5):1933–1941, 2011.

[3] Werner Adler, Sergej Potapov, and Berthold Lausen. Classification of repeated measurements data using tree-based ensemble methods. *Computational Statistics*, 26(2):355–369, 2011.

[4] Patricia ME Altham. Two generalizations of the binomial distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 27(2):162–167, 1978.

[5] Takeshi Amemiya. Bivariate probit analysis: Minimum chi-square methods. *Journal of the American Statistical Association*, 69(348):940–944, 1974.

[6] JR Ashford and RR Sowden. Multi-variate probit analysis. *Biometrics*, pages 535–546, 1970.

[7] Raji Balasubramanian, E Andres Houseman, Brent A Coull, Michael H Lev, Lee H Schwamm, and Rebecca A Betensky. Variable importance in

matched case–control studies in settings of high dimensional data. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 639–655, 2014.

[8] Dale Bowman and E Olusegun George. A saturated model for analyzing exchangeable binary data: Applications to clinical and developmental toxicity studies. *Journal of the American Statistical Association*, 90(431):871–879, 1995.

[9] Peter Bühlmann and Bin Yu. Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[10] Christopher Chatfield and Gerald J Goodhardt. The beta-binomial model for consumer purchasing behaviour. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 19(3):240–250, 1970.

[11] Siddhartha Chib and Edward Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

[12] Eliana Costa e Silva, Isabel Cristina Lopes, Aldina Correia, and Susana Faria. A logistic regression model for consumer default risk. *Journal of Applied Statistics*, 47(13-15):2879–2894, 2020.

[13] Martin J Crowder. Beta-binomial anova for proportions. *Applied statistics*, pages 34–37, 1978.

[14] Sanjiv R Das, Darrell Duffie, Nikunj Kapadia, and Leandro Saita. Common failings: How corporate defaults are correlated. *The Journal of Finance*, 62(1):93–117, 2007.

[15] D. Dey, S. Ghosh, and B. Mallick. *Generalized Linear Models.* CRC Press, 2000.

[16] Carlos AR Diniz, Marcelo H Tutia, Jose G Leite, et al. Bayesian analysis of a correlated binomial model. *Brazilian Journal of Probability and Statistics*, 24(1):68–77, 2010.

[17] Darrell Duffie, Andreas Eckner, Guillaume Horel, and Leandro Saita. Frailty correlated default. *The Journal of Finance*, 64(5):2089–2123, 2009.

[18] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[19] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[20] E. George and D. Bowman. A full likelihood procedure for analysing exchangable binary data. *Biometrics*, 51:512–523, 1995.

[21] E Olusegun George and Ralph L Kodell. Tests of independence, treatment heterogeneity dose-related trend with exchangeable binary data. *Journal of the American Statistical Association*, 91(436):1602–1610, 1996.

[22] DA Griffiths. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, pages 637–648, 1973.

[23] Ramesh C Gupta and Hui Tao. A generalized correlated binomial distribution with application in multiple testing problems. *Metrika*, 71(1):59, 2010.

[24] JK Haseman and LL Kupper. Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, pages 281–293, 1979.

[25] CD Kemp and ADRIENNE W Kemp. The analysis of point quadrat data. *Australian Journal of Botany*, 4(2):167–174, 1956.

[26] A. Kuk. A litter-based approach to risk assessment in developmental toxicity studies via a power family of completely monotone functions. *JRSS Series C*, 69:369–386, 2004.

[27] Lawrence L Kupper and Joseph K Haseman. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, pages 69–76, 1978.

[28] Charles Kwofie, Caleb Owusu-Ansah, and Caleb Boadi. Predicting the probability of loan-default: An application of binary logistic regression. *Research Journal of Mathematics and Statistics*, 7(4):46–52, 2015.

[29] Richard D Lawrence, Se June Hong, and Jacques Cherrier. Passenger-based predictive modeling of airline no-show rates. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 397–406, 2003.

[30] Sen Liang, Anjun Ma, Sen Yang, Yan Wang, and Qin Ma. A review of matched-pairs feature selection methods for gene expression data analysis. *Computational and structural biotechnology journal*, 16:88–97, 2018.

[31] Alberto Luceño. A family of partially correlated poisson models for overdispersion. *Computational statistics & data analysis*, 20(5):511–520, 1995.

[32] Alberto Luceño and Federico De Ceballos. Describing extra-binomial variation with partially correlated models. *Communications in Statistics-Theory and Methods*, 24(6):1637–1653, 1995.

[33] J. Marin and C. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics.* springer, 2007.

[34] Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.

[35] Dolores Romero Morales and Jingbo Wang. Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, 202(2):554–562, 2010.

[36] Patrick Alfred Pierce Moran. *An introduction to probability theory.* Number 519.223 MOR. CIMMYT. 1968.

[37] P. Muller, F. Quintana, A. Jara, and T. Hanson. *Bayesian Nonparametric Data Analysis.* Springer Series in Statistics, 2015.

[38] R. Neal. Markov chain sampling methods for dirichlet process mixture models. *Markov Chain Sampling Methods for Dirichlet Process Mixture Models*, 9:249–265, 2000.

[39] Rainer Neuling, Silvia Riedel, and Kai-Uwe Kalka. New approaches to origin and destination and no-show forecasting: Excavating the passenger name records treasure. *Journal of Revenue and Pricing Management*, 3(1):62–72, 2004.

[40] Y e Ochi and Ross L Prentice. Likelihood inference in a correlated probit regression model. *Biometrika*, 71(3):531–543, 1984.

[41] Rubiane Maria Pires et al. Modelos de regressão binomial correlacionada. *Universidade Federal de São Carlos*, 2012.

[42] Saeedeh Pourahmad, Seyyed Mohammad Taghi Ayatollahi, S Mahmoud Taheri, and Zahra Habib Agahi. Fuzzy logistic regression based on the least squares approach with application in clinical studies. *Computers & Mathematics with Applications*, 62(9):3353–3365, 2011.

[43] RL Prentice. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81(394):321–327, 1986.

[44] Ross L Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048, 1988.

[45] P. Rossi. *Bayesian Non- and Semi-parametric Methods and Applications.* Princeton University Press, 2014.

[46] Daniel B Rowe. *Multivariate Bayesian statistics: models for source separation and signal unmixing.* CRC press, 2002.

[47] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[48] Philipp J Schönbucher. Factor models: Portfolio credit risks when defaults are correlated. *The Journal of Risk Finance*, 2001.

[49] B. Shahbaba and R. Neal. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning*, 10:1829–1850, 2009.

[50] John Gordon Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):257–261, 1948.

[51] So Young Sohn, Dong Ha Kim, and Jin Hee Yoon. Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing*, 43:150–158, 2016.

[52] A. Sveinsdottir. Passenger overbooking, using no-show probability. *University of Iceland*, 2019.

[53] Robert E Tarone. Testing the goodness of fit of the binomial distribution. *Biometrika*, 66(3):585–590, 1979.

[54] DA Williams. 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, pages 949–952, 1975.

[55] DA Williams, Lawrence L Kupper, Christopher Portier, and Michael D Hogan. Estimation bias using the beta-binomial distribution in teratology. *Biometrics*, pages 305–309, 1988.

[56] DA Williams and Kamta Rai. Dose-response models for teratological experiments. *Biometrics*, 43(4):1013–1016, 1987.

[57] David A Williams. Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2):144–148, 1982.

[58] Gary Witt. A simple distribution for the sum of correlated, exchangeable binary data. *Communications in Statistics-Theory and Methods*, 43(20):4265–4280, 2014.

[59] Chang Yu and Daniel Zelterman. Sums of dependent bernoulli random variables and disease clustering. *Statistics & probability letters*, 57(4):363–373, 2002.

[60] Cristiano Zazzara. Credit risk in the traditional banking book: a var approach under correlated default. *Research in Banking and Finance*, 2, 2000.

[61] David Zenkert. No-show forecast using passenger booking data. 2017.

[62] Lue Ping Zhao and Ross L Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648, 1990.