Combining Logistic Regression Analysis and Association Rule Mining via MLR Algorithm

Ozge Yucel Kasap Nevzat Ekmekci Utku Gorkem Ketenci

Cybersoft R&D Center
Istanbul, Turkey
Email: ozge.kasap@cs.com.tr

***Abstract*—One of the keys in marketing is to recommend the right products to the right customers. This paper proposes a solution to this problem as a part of the development of a new data mining tool PROPCA (Proximus Optimum Canistro). The aim is to use logistic regression analysis and association rule mining together to make recommendations in marketing. An innovative approach in which combination of these two algorithms provides better results than algorithms used stand-alone is presented. While association rule mining searches all rules in the data set, logistic regression predicts a purchase probability of a product for customers. The combination of these two approaches are tested on a real-life banking data set. The results of combination are shown and their suitability in general is discussed.**

Keywords–Logistic Regression; Logit; Association Rule Mining; Apriori; Ensemble Learning; Stacking; MLR.

I. INTRODUCTION

People are facing the challenge of choosing the right service or product due to wide range of choices. On the other hand in marketing, it is important to know which customers might be interested in a specific product or which customers could be annoyed by receiving the campaign mails or messages about uninteresting products.

This is a major problem in the finance and banking sector. According to a survey conducted in 2014 by IBM Silverpop [1], the unique open rate and click-through rates of e-mails are %22.4 and %3.3 respectively for the finance sector in the United States. Given these statistics, if the potential purchasers of each product are determined properly, this may increase the company profit as well as decrease unnecessary expense concerning messages, calls, mails etc. Considering the growth of wide range of products, for a marketing expert, it is inevitable to turn into statistical models or machine learning algorithms to understand and explain the customer behaviors.

In recent years, logistic regression models (logit) have been used prevalently in several domains to make predictions [2]. To be more specific, Akinci et al. examined the application domains of logistic regression models in marketing researches such as consumer behavior modelling, international marketing, branding, societal marketing, promotion, retailing, health services marketing [3].

Similarly, mining for association rules (AR) is one of the well-studied issues in data mining for marketing [4]. AR aim to discover patterns, such as two or more products that are often purchased together.

The main purpose is to build a more reliable model to predict a large number of customers who are likely to purchase specific products. Considering how big marketing data can be, it is not suitable to perform data mining algorithms in the whole data set. Instead, a sample can be selected to apply data mining algorithm and the results can be generalized for whole set of customers. However, for a better predictive force, it is inevitable to use multiple data mining algorithms together.

This research suggests and examines an approach by combining two prediction models used widely in marketing in order to obtain better results: logistic regression and association rules. These two models are implemented as a part of a new data mining tool PROPCA, the way they work is completely different from each other. Logit uses customer and product features to make predictions whereas association rule mining uses only the ownership information of product families. Combining these two complementary models by Ensemble Learning and placing a new model will be an innovative approach in consumer behavior modeling.

The outline of this paper is as follows. Section II briefly explains general concepts and related works about Logistic Regression Analysis, Association Rule Mining, Ensemble Learning. Section III describes the proposed model. Results of this research in addition to the implementations and tests are given in Section IV. Finally, Section V gives conclusions and recommendations for further researches.

II. RELATED WORK

As aforementioned, this research's aim is to combine logit and AR with Ensemble Learning methods in order to obtain better prediction (classification) results. In this section, firstly the researches about Association Rule Mining will be introduced. Then, the works on Logistic Regression models will be examined. Last but not least, the researches about Ensemble Learning will be briefly presented.

A. *Association Rule Mining*

AR mining was first proposed by Agrawal et al in 1993 [5]. After the first AR algorithm, named AIS (Agrawal, Imielinski, Swami), was discovered [5], a very known algorithm, Apriori, was introduced [6]. Although Apriori is the most used algorithm, there are many other algorithms in this field such as Eclat [7] and FP-Growth [8].

AR mining is one of the most important fields of data mining [9]. AR are generated by finding frequent patterns from the data simply by using if/then conditions and identifying the most significant relationships.

AR may be used for analyzing and predicting customer behaviors [10]. They were initially adopted for "Market Basket Analysis" to find which items purchased with which items.

 ISBN: 978-1-61208-498-5

Given a set of customers' transactions (i.e., logs of products purchased together by different customers), the aim of AR is to find frequently purchased products or items in these transactions. On the other hand, they are useful for product clustering and catalog design.

An itemset is the subset of all products in the database and transactions are product groups that a customer has purchased together. $X \Rightarrow Y$ is an example expression of an association rule, where X and Y are sets of items. The meaning of such a rule is that transactions which contain X tend to contain Y. There are two important parameters of AR algorithms: *support* and *confidence*. Support of any itemset is defined as the number of transactions in which products appear together in all transactions. Confidence is defined as the ratio of support value of $X \cup Y$ to the support value of X itemset.

This research uses AR mining algorithm to identify the mostly purchased product families in finance and banking sectors and to generate rules according to these products. Apriori is the best known algorithm to mine association rules [11], and it is adopted in this research.

B. Logistic Regression

Usage of statistical models for explaining categorical choices is a common approach [12]. These choices are generally denoted with qualitative values and logit models are appropriate for such analysis [13]. Logit models are frequently used in medical and social sciences researches because they are easily applied and easily understood. In addition to these areas, logit models have also been used in biology, psychology, economics and transportation [14]. Since 1977, logit models are being used effectively in marketing [15]. Several usages of logit models for different marketing situations are studied in [16–21].

The main aim of logit analysis is the same as other techniques used for modeling in statistics such as Least Square Regression, Curve Fitting or Generalized Linear Model. They represent the relationship between dependent and independent variables and find the best fit among them. The relationship can be represented with a simple function, $y = f(x)$, where y and x are the dependent and the independent variables respectively. Considering statistical relationships, if the value of x is known, the value of y can be estimated with a specific error term.

The following formula is used to define the logistic cumulative distribution function [12].

$$y = f(x) = \frac{1}{1 + e^{-z}} \tag{1}$$

z is called the "utility function" and can be represented as

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{2}$$

where β_0 is the intercept and $\beta_1 x_1, \beta_2 x_2, \ldots, \beta_n x_n$ are the regression coefficients multiplied by some values of the predictor, n is the number of independent variables in the data set and ε is the error term.

Logit can be classified as binomial, ordinal or multinomial. If the observed values for the dependent variable has only two types such as "yes" vs. "no" or "success" vs. "failure", binomial (or binary) logistic regression can be used. If the dependent variable has the number of types equal or more than three and types are not ordered, multinomial logistic regression can be used. If types are ordered ordinal logistic regression can be used. Because of the configuration of products and existing data in banking sector (Detailed in Section III), in this paper, product choices of customers are modeled using binomial logit.

Generally, in binomial logit, "0" or "1" is used to represent the types of dependent variable [22], e.g., 1 being purchased, 0 being not purchased. Based on one or more continuous or categorical independent variables (such as demographic or socioeconomic), binomial logit predicts the probability that an observation falls into one of two types of a dependent variable [23].

When logit is used to estimate the probability of a certain event occurring (e.g., purchasing or not), Maximum Likelihood (ML) approach provides a basis to find the smallest possible deviance between the observed and predicted values by trying different solutions through multiple iterations. ML gives the values of regression coefficient which maximizes the probability of obtaining the observed values [24]. According to the established model and learned coefficients, it will be determined whether to recommend or not a modeled product to the bank customers who do not own this product yet.

C. Ensemble Learning

Ensemble Learning aims to create a better predictive model based on the combination of multiple classification results instead of single one. Early studies about this methodology has been seen in late seventies [25]. Schapire mentioned the Boosting algorithm in 1990 [26]. Boosting tries to create a strong classifier from weak classifiers. Thus, it combines the prediction of each weak classifier using weighted average. Tumer and Ghosh studied on combining neural networks in 1996 [27]. Schapire developed a new boosting algorithm with Freund and Adaboost [28]. Bagging was introduced by Breimann [29] and Random Subspace method was discovered by Ho [30]. Skurichina compared the popular algorithms such as Bagging, Boosting and Random Subspace and mentioned optimal working conditions [31]. Zhang and Zhang proposed resampling version of Adaboost [32].

Stacking was first introduced by Wolpert [33]. According to the study of Dzeroski and Zenko, stacking is the best method for combining heterogeneous classifiers [34].

Up until today, Ensemble Learning is used by researchers from different disciplines, especially pattern recognition, statistics and machine learning [35]. Based on this method, this paper's aim is to combine results of logit and AR.

In general, combining methods in Ensemble Learning are examined in two categories [36]:

- Weighting Methods: If the classifiers are working for the same tasks and the results are comparable, it is convenient to use this approach [37].
- Meta-Learning Methods: In this approach, there are several base classifiers and one meta classifier. Firstly, base classifiers make predictions, then according to these results learning is performed and meta-classifier is created. Meta-classifier's results are the system's results at the same time. Stacking is the most popular technique used in this domain [34].

Results of classifiers cannot be evaluated directly because logit and AR give numerical values in separate scales. Given these conditions, this study was shifted to meta-learning (Stacking). Stacking has two stages, base classifiers and meta classifier, as shown in Figure 1 of Philip Chan's study [38]. Even though there are many studies on how to combine different classifier outputs, there is no specific study about the combination of logit and AR results. From the literature review of stacking, it has been seen that StackingC algorithm, which uses multi-response linear regression (MLR) method in meta learning stage [39][40], is suitable for this study since it tries to combine the probability values calculated by heterogeneous classifiers.

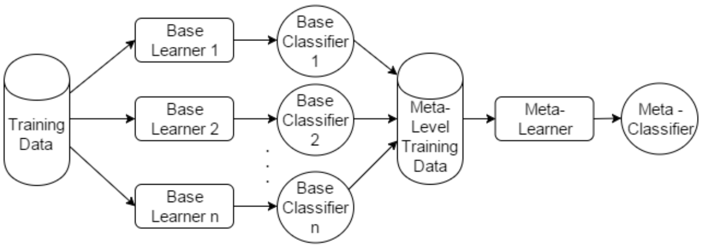

Figure 1: Meta Learning System.

Logit and AR calculate the probabilities of purchasing each product for each customer separately. MLR takes these probabilities produced by different algorithms as input. It runs linear regression and generates outputs for each product (1 if this product is bought, 0 otherwise).

Consequently, the models are formed separately for each product. Later on, when asked to predict for a new instance (new customer), all product models (AR-logit) are executed for this instance and the results of separated models are combined by MLR coefficients that are determined during the learning phase. In the end, one customer can be classified as purchaser of zero or multiple products.

The result of binomial logit model does not require any special regulation since it produces different probabilities for each product. These results can be used directly in MLR. However, determination of the probability value produced by AR is problematic. As a first approximation, probability of each product predicted by AR is identified by the confidence value of the related association rule. If the product appears in multiple rules and more than one rule is applicable for a specific customer, the confidence value of association rule with the highest confidence is selected as a probability value.

III. MODEL

As stated in Section II, logit and AR mining are widely used approaches in the marketing literature for estimating the probability of a customer purchasing an item. Considering the real-life industrial data set from a medium sized bank in Turkey, which includes features of customers and ownership information for four different retail banking product families; these approaches are applied. 1 presents owners and 0 shows non-owners of the products.

As shown in Figure 2, the training data set was prepared for the algorithms by feature selection (See Figure 2(a)) and sampling (See Figure 2(b)) steps (Detailed in Section IV). Sampling steps were taken into consideration due to decrease memory consumption and increase performance. While modeling with logit, ownership values are used as dependent and demographic informations are used as independent variables. Since the product families are not equivalent to each other and do not represent alternatives to each other, the results are calculated separately for each product family in the training data set with a separate binomial logit model. The same data set, except the demographic part, is used to mine the association rules using the Apriori algorithm. According to the ownership information of all product families for all customers, frequent itemsets are created and association rules are mined accordingly for each training data set.

While the logit model is performed separately for each product family in the training data set, it is possible to apply AR for each product family at once. Hence, the results of the binomial logit model cannot be merged directly for any customer because the results for each product are in different scale and not directly comparable. Therefore, the results from both the logit model and AR must be considered separately for each product because of the output format. This is why it is needed to combine logit and AR mining results with MLR.

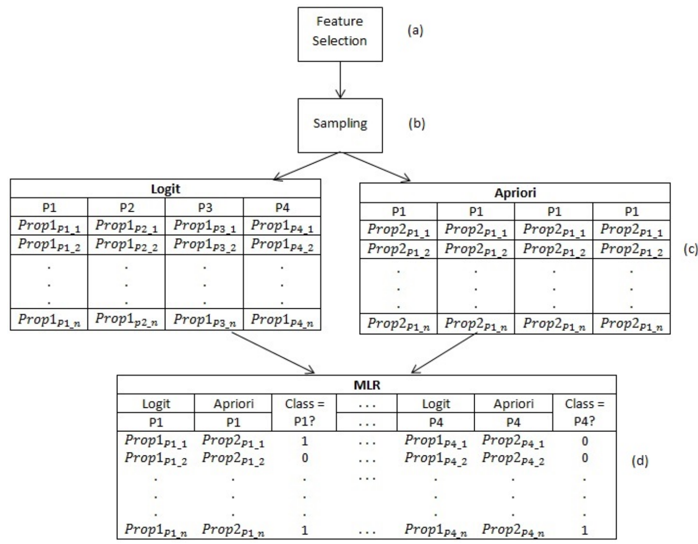

Figure 2: Illustration of MLR Algorithm with Logit and AR

In Figure 2, illustration of MLR algorithm on a data set with four classes (P1, P2, P3 and P4), n observations and two (logit and AR) base classifiers are given. $Propi_{j\_k}$ refers to the probability given by base classifier i for product j on observation number k. Figure 2(c) shows the predicted probabilities of both logit and AR separetely and Figure 2(d) shows the meta training set for MLR which will be used to learn a linear regression function to predict the ownerships of each product (class). Basically, MLR learns a linear regression function for each classifier and the calculated coefficients are used on the relevant test sets to predict the ownerships.

Coefficients for logit, association rules and coefficients for MLR are obtained at the end of this learning step. Using this learned parameters, the ownership (or purchasing) probability for each product for each customer in the test data set is calculated.

In general, MLR is used to combine two or more multiclass classifiers that have equal number of classes. Therefore, the sum of each observations's probability for each classifier is equal to one in the meta training set. In this research, there is no such restriction as logit models generate probabilities for each product independently because of the product families being not equivalent to each other.

IV. IMPLEMENTATION AND TESTS

Implementation and tests have been done via Java as part of the development of a new data mining tool PROPCA.

PROPCA includes every necessary data mining component such as feature selection, sampling, etc. Moreover, PROPCA contains two parallel offer production models: association rules and the logistic regression. These two models have different requirements in terms of data.

As aforementioned, logit models the relation between the demographical properties of customers and the purchase action. Banking data sets contain hundreds of demographical attributes. The usage of all these attributes in modeling is time consuming and it will be an error prone approach. Therefore, the identification of relevant ones in the whole set of attributes is commonly done before the modeling task. Strong predictors are selected using the Information Value (IV) algorithm implemented in PROPCA among frequently referenced attributes in banking domain. These attributes have both categorical and numerical values.

Final banking data set contains 19818 customers' data having 5 demographic features (3 numerical and 2 categorical) chosen by IV and ownerships of 4 product families. The names of the used features and products cannot be mentioned in this work because of privacy of banking customers' data. In order to apply logistic regression, categorical attributes are represented with dummy values and all of the selected attributes are standardized.

TABLE I. OWNERSHIP RATIO (≈)

	P1	*P2*	*P3*	*P4*
Ownership	32.2 %	25.3 %	62.2 %	63 %

In Table I, all product families and ratio of number of these products' owners to number of all customers are presented. Some customers own more than one of these products at the same time. Association algorithms examine the relation between these products ownership and detect customers who may purchase a new product. To perform this, AR algorithms need minimum support and minimum confidence values. These values are fixed as 0.01 and 0.20 respectivelly. While the minimum support threshold is used to find frequent itemsets, the minimum confidence threshold is used to generate rules from frequent itemsets.

The data set has splitted into three equal parts by the stratified sampling mechanism implemented in PROPCA and each part is used as training set and the rest of the data is used as test set iteratively. Therefore, effects of distinct training set on general results was observed. Each data set contains 6606 different customers' data and having slightly different ownership ratio for each product (See Table II).

TABLE II. OWNERSHIP RATIO IN DISTINCT TRAINING SETS (≈)

	P1	*P2*	*P3*	*P4*
Training Set 1	32.3 %	25.7 %	62.1 %	63.0 %
Training Set 2	32.1 %	24.8 %	61.6 %	63.2 %
Training Set 3	32.1 %	25.3 %	63.0 %	62.8 %

In each iteration, the model learns from one data set and it is tested with the remaining two. The predicted probabilities of algorithms are evaluated by a cut-off point (i.e., 0.5) to obtain "1" or "0" for the class value. The true positive rate (TPR), true negative rate (TNR) and accuracy (Acc) are calculated for each product family on distinct test sets. The aim of these tests is to show the benefit of combining different algorithms with MLR.

A. TPR and TNR

TPR and TNR are two complementary indicators. While TPR measures the ratio of correctly identified positives (e.g., purchaser of a product), TNR measures the ratio of correctly identified negatives (e.g., non-purchaser of a product). In Figure 3, the TPRs and TNRs for each product (Product 1, 2, 3 and 4) obtained from each test (Test Data 1, 2 and 3) set with different algorithms (Logit, Association and MLR) are shown.

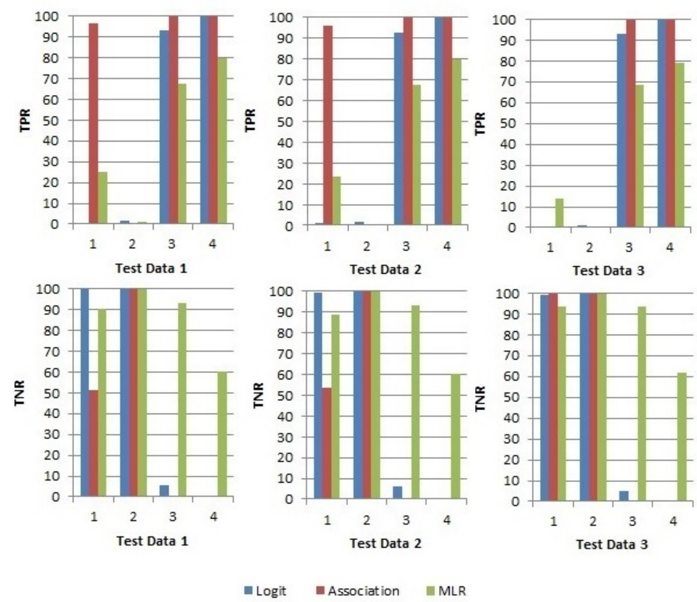

Figure 3: TPR and TNR

TPRs and TNRs of logit are respectively about 0 % and 100 % in all tests, for Product 1. Logit cannot distinguish purchaser of Product 1 from non-purchaser and it tends to classify a negligible part of customers as purchaser. One reason of these unsuccessful results may be low ownership ratio of Product 1 (less than 50 %) in whole data set (see Table I and II). In Test Data 1 and 2, association rules' TPRs and TNRs are respectively about 95 % and 50 %. However, in Test Data 3, it tends to classify none of customers as purchaser. Therefore, TPR is 0 % and TNR is 100 % in this last test. In first two tests, MLR's results are between logit's and AR's results. In the last test, MLR's TPR (≈ 14 %) is greater than and TNR (≈ 94 %) is less than both of the base classifiers. Thus, with a little loss in terms of TNR, MLR makes 14 % gain in terms of TPR.

None or a very little number of customers are classified correctly as purchaser of Product 2. Thus, poor TPRs for this product are obtained from all test data. Neither logit nor association algorithm can model the purchase action of Product 2. One reason of this fact can be the poor ownership rate of this product in whole data set (25 %). Conversely, TNRs of Product 2 equals approximately 100 % in each test. In this case, MLR does not add any gain on these results.

In all of the tests, all of the customers are classified as purchaser of Product 3 and 4 by association algorithm. Therefore, TPRs of association algorithm are always 100 % and TNRs are 0 %. Association algorithms cannot distinguish purchaser and non-purchaser. On the other hand, performance of logit is not significantly different from association algorithms. TPRs rate are greater than 90 % and TNRs are less than 10 %. MLR algorithm equalizes TPRs and TNRs. With a loss about between 10 % and 35 % on TPRs, it makes gain about between 60 % and 90 % on TNRS for Product 3 and Product 4 in different test sets.

In a nutshell, MLR takes the output of two poor classifiers' results as input and it increases the equilibrium between TNRs and TPRs. In some case, like in Test Data 3 for Product 1, MLR obtain a TPR more than both base classifiers' results. In the next section, accuracies and MLR's impact on this indicator will be examined.

B. Accuracy

Accuracy measures how well the classifiers predict positive and negative class together. As shown in Figure 4, using MLR algorithm increases the accuracy of both Logit and Association Rules. To be more specific, except for Product 2, MLR provides an increase in accuracy by approximately 2 % for Product 1 and even more (≈ 8 %) for Product 3 and 4 for each test data set.

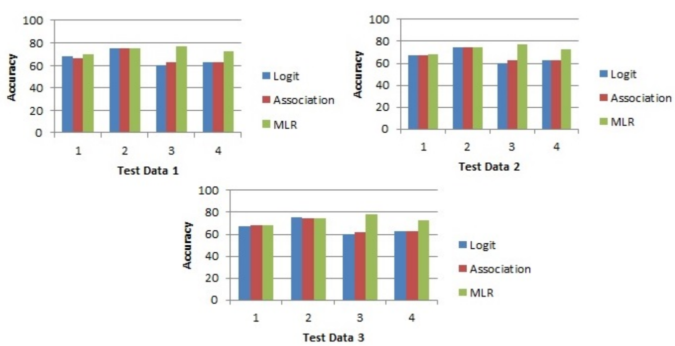

Figure 4: Accuracy

The reason that accuracy rates are around 60% for each product family can be interpreted by simply saying user behavior is complex. It is particularly more complex when the issue is recommending something to a customer. The retrieval of what is known before is not the same thing as suggesting something to a customer that he has never seen or heard before. Therefore, accuracy rates being not that high is something that can be acceptable in social sciences.

In the light of Figure 4, it can be said that MLR separates the purchaser customers from non-purchaser customers more successfully than both of Logit and Association algorithms in almost every test data set.

V. Conclusion and Future Works

In this research, it is aimed to obtain a model that combines the outputs of two seperate algorithms, logit and AR mining developed as a part of PROPCA, which have difficulties to work effectively with large data sets and also have different operation principles. The results are promising and show that this model works more efficiently and has a higher accuracy than logit or AR mining used standalone. It can be considered as a model that develops and strengthens the power of prediction of these two algorithms. However, the authors are aware of the fact that the tests have to be extended with a data containing more products and more customers and also other evaluation metrics like precision and recall can be used in order to validate the model and also camparison of the proposed approach with other ensemble techniques could also be performed.

It is known that logit works better when learning data set is balanced in terms of the number of positive and negative cases. It is possible to improve accuracy of the model by carrying attention to sampling models. If for each product, balanced training set (approximately 50 % negative, 50 % positive observations) is created, the performance of logit could be increased comparing to current sampling. Hence, this increase in performance will give better predictive probabilities as logit output and this could also positively affect the overall model's success.

In addition to balanced traning set, selection of diffrent strong attributes and various cut-off points for each product separately could also result in increasing the success of both logit and consequently the MLR model.

As a future work, to increase the efficiency of the proposed model some primarily work can be performed. Before applying association rule mining to the data set, some additional steps might be taken into consideration such as clustering (e.g., k-means or expectation maximization (EM) [41]) in order to gather customers having similar demographic features together. For each cluster of customers, AR algorithms could be executed separately.

Similarly, Latent Class (LC) approach can be used with logit. Choosing the correct number of classes in LC analysis is an important issue. A likelihood criteria such as Akaike Information Criterion (AIC) [42] and Bayesian Information Criterion (BIC) [43] could be used to compare models with different numbers of classes. Customers in the sample could be separated into segments based on customer characteristics and product preferences. Then, characterized values for each segment could be calculated using the attributes of customers in them. Using the selected attributes of each customer in the universe, the distance to each segment can be calculated and customers can be assigned to the segment having the minimum distance and the predicted probabilities can be calculated using the related segments propeties.

Acknowledgment

It is worth mentioning that this research has been carried out based on a mutual team work cooperation between the authors of this paper. The authors would like to thank to the other members of the research and development department at Cybersoft for their help throughout this work and also it should be stated that this study is conducted under the scope of PROPCA which is supported by the Scientific and Technological Research Council of Turkey (TUBITAK).

References

[1] Silverpop, "Silverpop email marketing metrics benchmark study," 2014.

[2] J. Tanguma and R. Saldivar, "Interpretation of logistic regression models in marketing journals," in The Sustainable Global Marketplace. Springer, 2015, pp. 2–2.

[3] S. Akinci, E. Kaynak, E. Atilgan, and S. Aksoy, "Where does the logistic regression analysis stand in marketing literature? a comparison of the market positioning of prominent marketing journals," European Journal of Marketing, vol. 41, no. 5/6, 2007, pp. 537–567.

[4] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in ACM SIGMOD Record, vol. 26, no. 2. ACM, 1997, pp. 265–276.
[5] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD Record, vol. 22, no. 2, 1993, pp. 207–216.
[6] R. Agrawal, R. Srikant et al., "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, 1994, pp. 487–499.
[7] M. J. Zaki, "Scalable algorithms for association mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 12, no. 3, 2000, pp. 372–390.
[8] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM Sigmod Record, vol. 29, no. 2. ACM, 2000, pp. 1–12.
[9] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," GESTS International Transactions on Computer Science and Engineering, vol. 32, no. 1, 2006, pp. 71–82.
[10] R. Srikant and R. Agrawal, Mining generalized association rules. IBM Research Division, 1995.
[11] M. Hegland, "Algorithms for association rules," in Advanced lectures on machine learning. Springer, 2003, pp. 226–234.
[12] D. Flath and E. Leonard, "A comparison of two logit models in the analysis of qualitative marketing data," Journal of Marketing Research, 1979, pp. 533–538.
[13] J. Berkson, "Maximum likelihood and minimum x 2 estimates of the logistic function," Journal of the American Statistical Association, vol. 50, no. 269, 1955, pp. 130–162.
[14] N. K. Malhotra, "The use of linear logit models in marketing research," Journal of Marketing research, 1984, pp. 20–31.
[15] P. E. Green, F. J. Carmone, and D. P. Wachspress, "On the analysis of qualitative data in marketing research," Journal of Marketing Research, 1977, pp. 52–59.
[16] J. R. Hauser and G. L. Urban, "A normative methodology for modeling consumer response to innovation," Operations Research, vol. 25, no. 4, 1977, pp. 579–619.
[17] A. J. Silk and G. L. Urban, "Pre-test-market evaluation of new packaged goods: A model and measurement methodology," Journal of marketing Research, 1978, pp. 171–191.
[18] R. R. Batsell, "Consumer resource allocation models at the individual level," Journal of Consumer Research, 1980, pp. 78–87.
[19] R. G. Chapman and R. Staelin, "Exploiting rank ordered choice set data within the stochastic utility model," Journal of marketing research, 1982, pp. 288–301.
[20] D. H. Gensch and W. W. Recker, "The multinomial, multiattribute logit choice model," Journal of Marketing Research, 1979, pp. 124–132.
[21] N. K. Malhotra, A. K. Jain, and S. W. Lagakos, "The information overload controversy: An alternative viewpoint," The Journal of Marketing, 1982, pp. 27–37.
[22] S. Hosmer, David W.; Lemeshow, Log-linear models. Springer-Verlag, 1990, ISBN:0-471-35632-8.
[23] L. De La Viña and J. Ford, "Logistic regression analysis of cruise vacation market potential: Demographic and trip attribute perception factors," Journal of Travel Research, vol. 39, no. 4, 2001, pp. 406–410.
[24] P. McCullagh and J. A. Nelder, Generalized linear models. CRC press, 1989, vol. 37.
[25] J. W. Tukey, "Exploratory data analysis," 1977.
[26] R. E. Schapire, "The strength of weak learnability," Machine learning, vol. 5, no. 2, 1990, pp. 197–227.
[27] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," Pattern Recognition, vol. 29, no. 2, 1996, pp. 341–348.
[28] Y. Freund, R. E. Schapire et al., "Experiments with a new boosting algorithm," in ICML, vol. 96, 1996, pp. 148–156.
[29] L. Breiman, "Stacked regressions," Machine learning, vol. 24, no. 1, 1996, pp. 49–64.
[30] T. K. Ho, "The random subspace method for constructing decision forests," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 8, 1998, pp. 832–844.
[31] M. Skurichina and R. P. Duin, "Bagging, boosting and the random subspace method for linear classifiers," Pattern Analysis & Applications, vol. 5, no. 2, 2002, pp. 121–135.
[32] C.-X. Zhang and J.-S. Zhang, "A local boosting algorithm for solving classification problems," Computational Statistics & Data Analysis, vol. 52, no. 4, 2008, pp. 1928–1941.
[33] D. H. Wolpert, "Stacked generalization," Neural networks, vol. 5, no. 2, 1992, pp. 241–259.
[34] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" Machine learning, vol. 54, no. 3, 2004, pp. 255–273.
[35] M. Sewell, "Ensemble learning," RN, vol. 11, no. 02, 2008.
[36] L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, no. 1-2, 2010, pp. 1–39.
[37] G. Sigletos, G. Paliouras, C. D. Spyropoulos, and M. Hatzopoulos, "Combining information extraction systems using voting and stacked generalization," The Journal of Machine Learning Research, vol. 6, 2005, pp. 1751–1782.
[38] D. W. Fan, P. K. Chan, and S. J. Stolfo, "A comparative evaluation of combiner and stacked generalization," in Proceedings of AAAI-96 workshop on Integrating Multiple Learned Models, 1996, pp. 40–46.
[39] A. K. Seewald, "How to make stacking better and faster while also taking care of an unknown weakness," in Proceedings of the nineteenth international conference on machine learning. Morgan Kaufmann Publishers Inc., 2002, pp. 554–561.
[40] K. M. Ting and I. H. Witten, "Issues in stacked generalization," J. Artif. Intell. Res.(JAIR), vol. 10, 1999, pp. 271–289.
[41] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," in Advances in Neural Information Processing Systems 7. Citeseer, 1995.
[42] H. Akaike, "Factor analysis and aic," Psychometrika, vol. 52, no. 3, 1987, pp. 317–332.
[43] G. Schwarz et al., "Estimating the dimension of a model," The annals of statistics, vol. 6, no. 2, 1978, pp. 461–464.

 ISBN: 978-1-61208-498-5