2016

Application of Isotonic Regression in Predicting Business Risk Scores

Linh T. Le
Kennesaw State University

Jennifer L. Priestley
Kennesaw State University, jpriestl@kennesaw.edu

Recommended Citation

Le, Linh T. and Priestley, Jennifer L., "Application of Isotonic Regression in Predicting Business Risk Scores" (2016). *Grey Literature from PhD Candidates*. 3.
http://digitalcommons.kennesaw.edu/dataphdgreylit/3

Application of Isotonic Regression in Predicting Business Risk Scores

Linh T. Le
Department of Statistics and Analytical Sciences
Kennesaw State University
Kennesaw, GA

Jennifer L. Priestley
Department of Statistics and Analytical Sciences
Kennesaw State University
Kennesaw, GA

Abstract*— an isotonic regression model fits an isotonic function of the explanatory variables to estimate the expectation of the response variable. In other words, as the function increases, the estimated expectation of the response must be non-decreasing. With this characteristic, isotonic regression could be a suitable option to analyze and predict business risk scores. A current challenge of isotonic regression is the decrease of performance when the model is fitted in a large data set e.g. more than four or five dimensions. This paper attempts to apply isotonic regression models into prediction of business risk scores using a large data set – approximately 50 numeric variables and 24 million observations. Evaluations are based on comparing the new models with a traditional logistic regression model built for the same data set. The primary finding is that isotonic regression using distance aggregate functions does not outperform logistic regression. The performance gap is narrow however, suggesting that isotonic regression may still be used if necessary since isotonic regression may achieve better convergence speed in massive data sets.

***Keywords*—component; isotonic regression; logistic regression; business risk score; PAVA; additive isotonic model.**

I. Introduction

Isotonic regression is a form of regression that minimizes the quadratic form

$$f(y,\theta)=\sum_{i=1}^{n} w_i(y_i-\theta_i)^2 \tag{1}$$

with y_i being the response variable, θ_i is a function of the predictors; θ must be isotonic, that is, $\theta_i < \theta_j$ for all $i < j$; and w_i is the weight for each data point. The estimation of θ can be retrieved using the Pool-Adjacent-Violators algorithm (PAVA). With this characteristic, isotonic regression seems to fit in the field of scoring risk since many attributes used in evaluating risk have this similar relationship to the risk scores. The challenge of applying isotonic regression to predicting business risk scores is the decrease in performance of the models when fitted in multidimensional data sets, especially those with more than four or five independent variables, while the risk evaluation process generally must go through many attributes.

As a result, this paper presents a number of attempts to fit isotonic regression models into a large data set – approximately 50 numeric variables and 24 million observations – to predict the business risk score for each of 3554073 companies. The models are evaluated relative to a traditional logistic regression model built for the same data set.

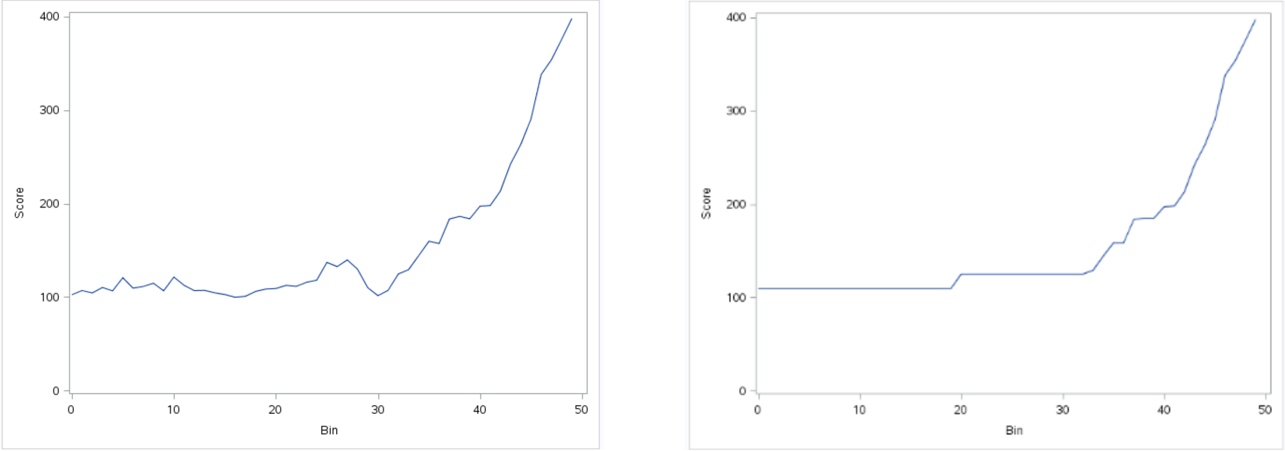
II. Literature Review

The problem of isotonic regression emerged from the 1950s [1] in the form of monotone regression. Using the least square method, the problem of isotonic regression is to find a set of functions θ_i of the explanatory variables that minimize $f(y, \theta)$ in equation (1) with respect to the assumption θ being isotonic, that is $\theta_i < \theta_j$ for all $i < j$. In a more relaxed case, it may become $\theta_i \leq \theta_j$ where w_i are nonnegative weights in the case of weighted data and in the case of unweighted data $w_i = 1$.

In 1972. Barlow et al. formalized the Pool-Adjacent-Violators Algorithm (PAVA) [2] to estimate θ_i. Briefly, PAVA estimates θ_i by splitting the list of the response values into blocks with respect to some function. The estimate expectation for each response in the same block is the average of all response values in that block and satisfies the isotonic restriction. More specifically, suppose the response y is already sorted in respect to a function of the predictors. Then θ can be estimated by iterating through the list. At point $i+1$, if $y_{i+1} \geq y_i$ then let $\theta_{i+1} = y_{i+1}$ else merge θ_{i+1} into the block before it and recalculate the mean of the block. If the condition is satisfied, then moving on to the next response value, otherwise go back one more block until the non-decreasing condition is unviolated.

An example of PAVA can be seen in figure 1. The chart on the left side is the risk score by the explanatory variable *bin* generated from the data used in this paper (more details will be provided in later sections). An overall trend of score increasing by bins can be seen although across the smaller intervals of bins the increasing constraint is violated. A PAVA process is then applied resulting in the strictly isotonic line in the chart on the right side with all the fluctuated parts replaced and becomes blocks of same values.

Besides estimation of θ_i, another important aspect of isotonic regression is the ordering of the data points by the predictors. There is no problem in the univariate case since there is only one independent variable. In the multidimensional case, sorting the response variable is overly complicated when too many predictors are introduced in the model. Currently, there are several approaches to order the data. The first is multidimensional ordering [3]. In this case, the responses are



Original Data Line *After PAVA*

Fig. 1. Example of PAVA Process

put in an array of which dimensions are the predictors; PAVA is then repeatedly applied to each dimension to generate the non-decreasing pattern until all the algorithms converge. The downside of this method is the complication of running PAVA in more than three dimensions, and moreover in such case, convergence is not ensured.

Another approach is introduced by Quentin F. Stout [4] using directed acyclic graphs. The general idea is to build a graph with partial order vertices then map the vertices into real number space. However, the author proposes that this method generates a number of unnecessary vertices which may become a serious problem in massive data sets with millions of observations and hundreds of dimensions. Additionally, optimization is guaranteed only in 2-dimensional data; whether the solution is optimal in more than three dimensional data cannot be proven.

The last method considered here is the additive isotonic model by Bacchetti [5]. The mechanism of additive isotonic model is to use the sum of multiple univariate isotonic functions of each explanatory variable. For example, if a model has three predictors x_1, x_2, x_3 then the ordering of data points will be conducted according to the sum of the three functions: $f_1(x_1) + f_2(x_2) + f_3(x_3)$ instead of the function $f(x_1,x_2,x_3)$. The weakness of this approach is that it may not be sufficient when there exists higher ordered interactions between predictors. Moreover, the author proposes this algorithm may suffer from slow convergence since the isotonic functions must be estimated separately for each variable in the model multiple times using a cycle algorithm.

III. Data

A. Data Discovery

The data used in this paper is the business risk score data set provided by Equifax that consists of 36 quarterly data subsets for the period from 2006 to 2014; snapshots were taken annually in January, April, July and October. The data records financial information of 11,787,287 companies identified by their Market Participant Identifier *(MPID)*.

The response variable is the business risk score of the companies. Table 1 provides some available information on this variable in October 2014. According to Equifax, a risk score below 450 can be considered “bad” for a company. This definition is used to define the response variable in the models built in the later section. Note that here the value 0 does not refer to a score of 0, but rather to an invalid score.

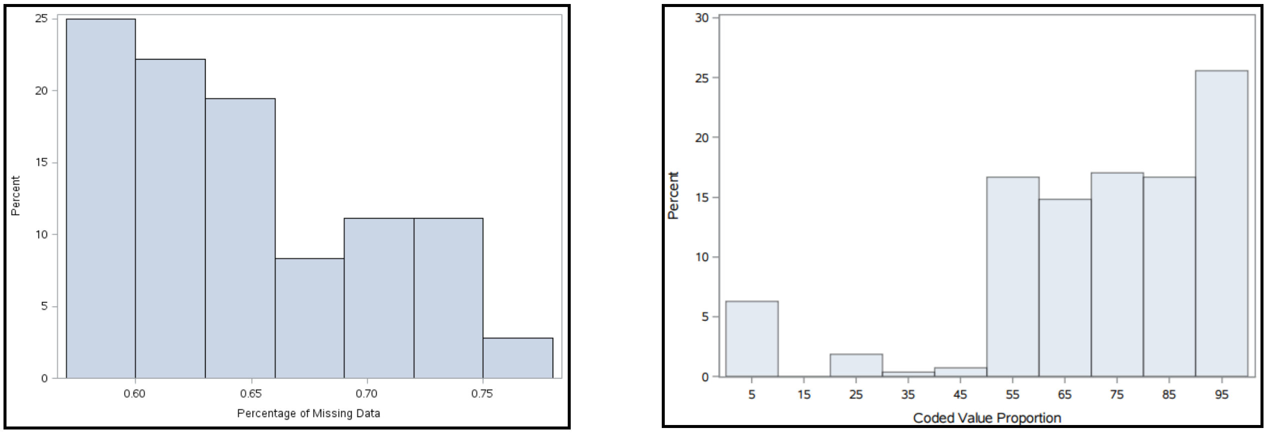
Percentile	0%	5%	25%	50%	75%	95%	100%
Quantile	0	219	262	463	497	539	619

good_score	Percent
0	47.47
1	52.53

Table. 1. Distribution of Risk Score in October 2014

Besides the risk scores, the data carries 304 other variables among which 250 are potential numeric predictors providing information of the companies’ activities in categories such as Non-Financial, Telco, etc. An important point to be aware of before using these variables in any model is the coding convention. For every variable, relevant values are given only from 0 to the variable upper bound subtracting 7. Numbers above that threshold become categorical, indicating missing or invalid data not appropriated for modelling. For example, a variable with values from 0 to 99 has a meaningful range up to 92; 93 to 99 represents categories of invalid values.

Figure 2 illustrates the proportion of missing data across the sets and coded data among the variables. It can be observed that the rates of all-missing data ranges from 55% to 80% in the 36 subsets whereas the rates of invalid data mostly lie above 50% among the 250 variables. This poses problems in any model since filtering out all missing and coded values would not only bias the data but also critically drops the number of observations. Specifically, attempts to filter the data of 2014 by particular pairs of variables may decrease the number of observations from 11 million to about 100,000 while filtering that same data by some sets of six variables results in a set of only around 6,000 data points. As a result, the



Histogram of Missing Data *Histogram of Coded Data*

Fig. 2. Distributions of Percentage of Missing Data across the Sets and Coded Data across the variables

data must be cleaned and imputed to mitigate lost information due to missing values.

B. *Data Cleansing*

First, observations with all fields missing or coded are filtered out since they would not contribute any information to analysis. Data points with partial coded values are more problematic since all variables have such a large portion of coded data that filtering them will result in a small and biased sample. On the other hand, immediate monotonic is not feasible: using a constant such as mean or median to replace more than 50% of the will nullify its variance while stratified or regression imputation among 250 variables with different coded values creates a circular reference.

Thus, a strategy using a combination of variable clustering and regression stratified imputation must be employed. The set of 250 variables are grouped into logical categories (such as group of Non-Financial activities, Telco activities etc.) followed by variable clustering procedures in order to reduce the complication of applying regression imputation. This step provides a smaller data set of 50 variables while the proportion of variation explained is still above 80%.

Next, all observations with more than 50% coded data across all variables were filtered as well since the data resulted from the previous step is still filled with invalid. This process has a disadvantage of biasing the data, however it seems to be the only available option since most of the variables still have a sufficiently high rates of coded values.

The quarterly data sets are finally compressed into half-year period data to reduce redundancy and simplify the imputation process before being joined together. This step results in a set of approximately 24 million observations with 50 predictors of which 20 have less than 10% of coded values. They were imputed using medians then used to build linear regression models to impute the rest of the variables.

The final data set consists of approximately 24 million observations with 46 numeric variables and 50 categorical variables.

C. *Variable Transformation*

Since isotonic regression is being tested in this paper, all variables are transformed to accommodate the technique. All the referenced logistic regression models will be using the same variables as the isotonic regression models to clearly contrast the performances.

Although isotonic regression models accommodate both continuous and binary responses, the current model will use a binary response and then be compared to a logistic regression model, which also takes a binary response variable. Because the methods in this paper use a system of norms to predict, complicated transformations of predictors will not be required. A binary transformation was used on about half of the variables because of the high proportion of zeros and ones in their distributions. Example of these variables can be seen in Table 2. Across the values of the variables, 95% data consists of zeros and ones, other values appear only after the 95th percentile. After transformation of these variables, all values 0 remain and all values greater than 0 become 1.

The remaining variables are normalized to accommodate aggregate functions used in the models such as Euclidian distance. The method min-max normalization was used since the range of *(0,1)* is desired:

$$normalized\ value = \frac{actual\ value - Min}{Max - Min} \qquad (2)$$

Description	Min	Med	75th P	90th P	95th P	Max
Total Non-Financial accounts in last 12 months	0	0	0	0	0	39
Total Non-Financial accounts 3-cycle past due in last 12 months	0	0	0	0	1	92
Total Non-Financial accounts 4-cycle past due in last 12 months	0	0	0	1	1	92

Table. 2. Example of Variables Transformed to Binary

Because all the variables have their minimum value at 0, equation (2) is simplified to become

$$normalized\ value = \frac{actual\ value}{Max} \quad (3)$$

A notable point is that the Max is not the "real" maximum value of the variables but rather the 90th percentile since the values after that point are generally too far from the medians and min-max normalization; using these values will make the rest of the values much closer to zeros. This results in a decrease in effects of a large portion of the variable in the output of the aggregate functions. Values greater than the 90th percentile become 1 after normalization. Examples for this type of variable can be seen in Table 3.

To ensure a consistent comparison of model performance, both modeling executions will utilize the variables in the same form.

IV. Methodology

The data is split into a training set (60%) and a validation set (40%). To build the isotonic regression model, the estimation algorithm PAVA is used. The multidimensional ordering problem is simplified into a univariate one using a weighted distance system. Multidimensional ordering and sorting using directed acyclic graphs are not used since they do not fit well into high dimensional data. An additive isotonic model used with 46 variables will result in slow convergence and complicated estimation therefore is not a good solution either.

A threshold of 450 in the business risk score is chosen to generate the binary variable: a score considered to be "good" if it is over 450.

To build the models, a non-negative relationship between the business risk score and all other 46 predictors must be guaranteed. A correlation procedure was conducted to test this assumption which shows that in reality, the risk score has a negative correlation with most of the predictors, consequently it must be transformed to satisfy the isotonic restriction. A most simple way to solve this while not changing the relationships between variables is to model a "bad score" indicator instead of a "good score" indicator:

– *bad_score = 1* if *risk_score < 450*

– *bad_score = 0* if *risk_score ≥ 450*

Here, *bad_score* is the new response variable to be predicted. Since a few explanatory variables previously have a positive correlation with the original response, they are also transformed once more: *new_value = 1 – old_value* to ensure the isotonic correlation between all variables of the model (note that since they have already been normalized, their max values are ones).

Then, the data points must be sorted in non-decreasing order of a stratified function. To begin with, the data is considered to be a 46-dimension space, and the tested aggregate functions include the L_1 norm, L_2 norm (Euclidian distance) and L_∞ norm of each data point from the origin point:

$$L_1(i) = |x_1|_i + |x_2|_i + \cdots + |x_{46}|_i \quad (4)$$

$$L_2(i) = \sqrt{x^2_{(1)i} + x^2_{(2)i} + \cdots + x^2_{(46)i}} \quad (5)$$

$$L_\infty(i) = MAX\,(|x_1|_i\,, |x_2|_i\,, \ldots, |x_{46}|_i \quad (6)$$

With $L_1(i)$, $L_2(i)$ and $L_\infty(i)$ respectively are the L_1, L_2 and L_∞ norms of observation *i* in the space, and $x_{(1)i} \ldots\ x_{(46)i}$ are the normalized predictors of the data point *i*. There is an issue with the infinity norm however: because all the variables are in the range between 0 and 1, and there are a large number of binary variables with only 0 or 1 as values, the infinity norm may become 1 for most of the observations. As a result, the weighted infinity norm will be used instead.

With the norm functions as a baseline, another assumption can be made: since explanatory variables have different correlations to the response variable, a weighted system will be used to reflex this effect. To be precise, a variable with a higher correlation should have greater effect on the response variable. Therefore, a weighted distance system is introduced and tested here along with the normal Euclidian distance system that equalizes the relationships of all predictors to the response. The simplest way to derive the weight system is to use the correlations themselves:

$$weighted\ L_1(i) = \rho_1|x_1|_i + \rho_2|x_2|_i + \cdots + \rho_{46}|x_{46}|_i \quad (7)$$

$$weighted\ L_2(i) = \sqrt{\rho_1 x^2_{(1)i} + \rho_2 x^2_{(2)i} + \cdots + \rho_{46} x^2_{(46)i}} \quad (8)$$

$$weighted\ L_\infty(i) = MAX\left(|x_{1(i)}|, |x_{2(i)}|, \ldots, |x_{46(i)}|\right) \quad (9)$$

Variable before Normalization	Min	75th P	90th P	95th P	Max
Highest Non-Financial balance in last 12 months	0	859	4944	15151	663121110
Total Cycle 1 Non-Financial past due amount in Last 3 Months	0	0	13	286	99436549
Percent of Non-Financial past due amount to total balance reported in last 12 months	0	47	96	100	999.92
Highest industry balance in last 12 months	0	574	3999	13000	657626357

Variable after Normalization	Min	50th P	75th P	90th P	Max
Highest Non-Financial balance in last 12 months	0	0.0291	0.1745	1	1
Total Cycle 1 Non-Financial past due amount in Last 3 Months	0	0	0	1	1
Percent of Non-Financial past due amount to total balance reported in last 12 months	0	0.0789	0.4845	1	1
Highest industry balance in last 12 months	0	0.0129	0.1443	1	1

Table. 3. Example of Variables before and after Normalization

Where $\rho_1 ... \rho_{46}$ are the correlation coefficients between each predictor and the bad score indicator. Since these correlations must be computed before building the second model, this model is more complicated to be built and fitted. Hence, the performances of both methods will be compared to determine whether a weighted system of distances is necessary.

After the data points are ordered, PAVA is applied to estimate the negative score of each observation. If a violation of the non-decreasing constrain is detected, the algorithm must trace back and recalculate the means of the blocks until the violation is solved; it is possible that this step would be repeated a number of times in a data set of 24 million observations. As this severely impacts the performance of the model, a simplification method is employed. During the estimation process, the data points are divided into blocks of near norm values. The response variable becomes the probability of an observation in the blocks being 1. The order of the block must satisfy the non-decreasing pattern in the norm values. PAVA is then applied on the blocks instead on all the data points. Approaching the estimation from this method lightens the burden of repeatedly iterating through the data points and recalculating the mean negative score for the blocks; the number of bins can then be fine-tuned in the implementation process to retrieve the fittest value. Implications of these different approaches are to be discussed in the next section.

In addition to the five types of aggregate functions, the models are also evaluated by using different number of variables chosen by their correlation to the response. Four correlation thresholds are used: no threshold, 0.20, 0.30 and 0.45 which results in models of 46, 29, 19 and 7 variables respectively.

To examine the performance of the isotonic regression models, all are contrasted to a logistic model with the same variables set using the C-statistic [7] which is one among the measurements of a logistic regression model. Because the chosen outcomes for isotonic regression models in this paper is the probability of the response being 1, the concepts of concordant, discordant and C-statistic similar to logistic regression can be used: a pair of observations is concordant if the predicted probability of the observation with response of 0 is less than that of the one with response of 1; the pair is tie if the two probabilities are equal, and discordant if otherwise. The C-statistic is then computed by summing the percentage of concordant pairs and half the percentage of tie pairs. Figure 3 illustrates the similarity between a logistic curve and the isotonic line estimated using the same variable L_2 norm, both represents the probability of a company having bad risk score (below 450) as the variable increases.

V. Result

With the discussed methodology, five types of isotonic regression model: using L_1, L_2, weighted L_1, weighted L_2 and weighted L_∞ are tested with four sets of 46, 29, 19 and 7 variables. A benchmark logistic regression model is also built for each set. The resulting C-statistics for all the models can be seen as in figure 4. As can be seen, the logistic regression models have higher C-statistics in all four cases, however the number gradually drops with the number of variables. The isotonic regressions models have lower C-stats but they do not seem to be effected by decreasing variables. In fact, most of the models have their C-stat increased instead.

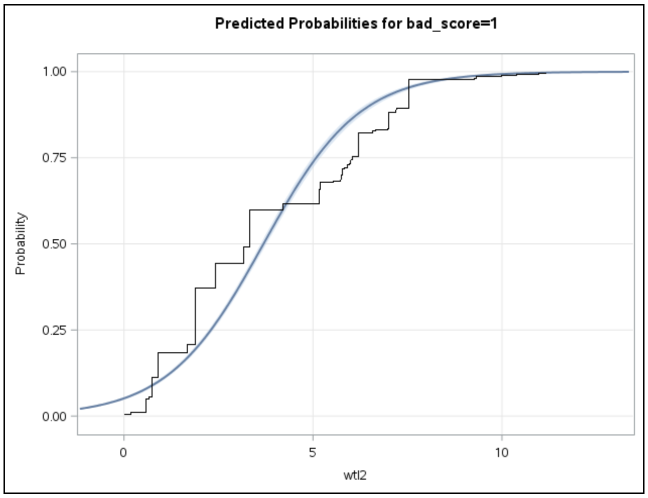


Fig. 3. The Logistic Curve and Isotonic Line Estimated by the Same Variables

Among the isotonic regression models, those with weighted norms yield better C-statistics with high number of variables, and there is virtually no differences between using weighted L_1 norm or weighted L_2 norm in the first two cases although the first model get better performance at 7 variables. The unweighted L_1 and L_2 norm models have lower C-statistics when tested with large number of variables but both raise in a smaller number of variables. With only seven variables, the L_1 norm model provides highest C-statistic among the isotonic regression models. The L_∞ models show same C-statistics with both 46 and 7 variables.

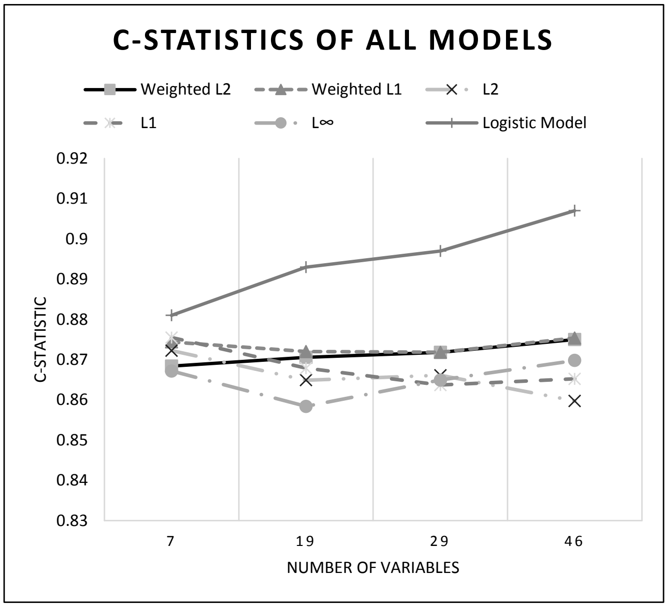


Fig. 4. C-statistics of All Models by Number of Variables

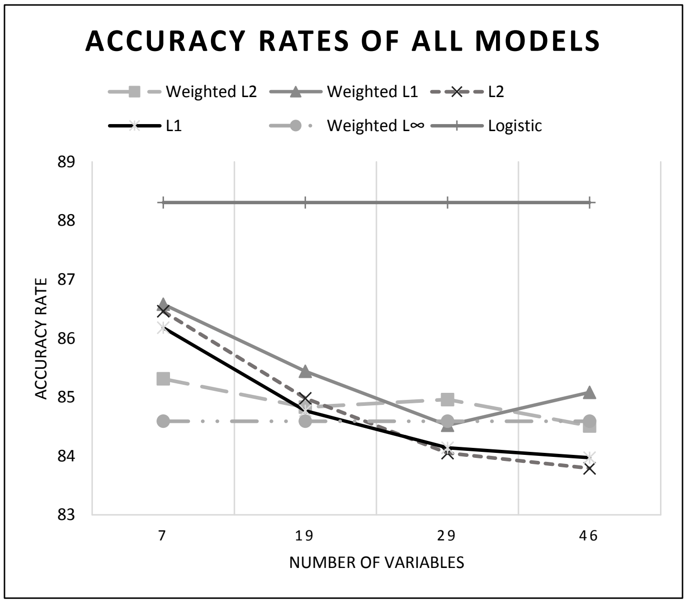


Fig. 5. Accuracy Rates of All Models by Number of Variables

Next, the prediction accuracy rates of the models can be seen in figure 5. It shows that all the isotonic regression models have a lower accuracy rate with more variables used except for the L_∞ models with a constant rate. Also, all isotonic regression models cannot outperform logistic regression in all cases. At best, the difference is still 2% between the weighted L_1 and the logistic model both using 7 variables.

VI. Discussion

The first conclusion which can be made from the test results is that this method of aggregating multiple variables using a norm system does not improve the multivariate isotonic regression problem: the performance still drops when the number of variables increases. Though the weighted L_1 and L_2 models are able to maintain the C-statistic when raising the number of variables to 46, none of the isotonic models can keep their prediction rate as high as in the case of 7 variables.

It is also revealed that among the five types of norms used, there are differences in performance but they may not be practical as the largest gap is about 0.2 in C-statistic and 3% in prediction accuracy. This indicates that a correlation weighted system may not be preferred because of the growth in complexity when being implemented.

In comparison to logistic regression models, overall none of the isotonic regression models outperforms them in both C-statistic and prediction rate. The gap in performance however is not large which suggests that isotonic regression using norm can still be used in risk scoring if necessary.

The last notable point is that whether isotonic regression has better speed in this type of data. The whole process of the method used in this paper includes computing the norms, ordering the data and estimating the isotonic function using PAVA. Computing the norm has a complexity of *O(nd)* with *n* is the total number of observations and *d* is the number of variables used; the complexity of sorting the data depends on the algorithm chosen but for algorithms such as quick sort, it can be maintain at *O(n log(n))* and lastly PAVA has a complexity of *O(n)* [6]. Overall, the whole process may have a complexity level of *O(n (d+log(n)))*. On the other hand, the complexity of logistic regression depends on the estimation method chosen and ranges from O(nd) to O(nd^2) per iteration as pointed out by Minka [8]. As a result, if the number of iterations in estimating the logistic model is high enough, the isotonic regression model may achieve better convergence speed, which is important in massive data set.

References

[1] Stout, F. Q., “Isotonic Regression for Multiple Independent Variables” Algorithmica, vol. 71, January 2013.

[2] Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), “Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression”, New York: John Wiley.

[3] Salanti, G., Ulm, K., “Multidimensional Isotonic Regression and Estimation of the Threshold Value”, Sonderforschungsbereich 386, Paper 234 (2001).

[4] Stout, F. Q., “Isotonic Regression for Multiple Independent Variables” Algorithmica, vol. 71, January 2013.

[5] Bacchetti P., Additive isotonic models, Journal of the American Statistical

[6] Best, M.J.; & Chakravarti N. (1990). "Active set algorithms for isotonic regression; a unifying framework". Mathematical Programming 47: 425–439

[7] Harrell FE Jr: Regression modeling strategies. New York, NY: Springer; 2001.

[8] Minka. T., “A Comparison of Numerical Optimizers for Logistics Regression”, Microsoft Research, October 2003.