College of Engineering: Mechanical, Industrial, and Manufacturing Engineering

# A MIXED INTEGER LINEAR PROGRAMMING MODEL FOR RISK SCORE DEVELOPMENT IN HEALTHCARE

**James McKenna; Joseph Agor, PhD**

*Oregon State University, School of Mechanical, Industrial, and Manufacturing Engineering*

## BACKGROUND

Electronic Health Record (EHR) data can be used to quantify the relationship between patient attributes and the progression of acute medical conditions. Risk scoring systems can dynamically assess the status of a patient and their risk of further degeneration leading to providers giving preventative care before the onset of a more serious condition. Panels of experts have developed risk scores to assist decision makers at the bedside.

Current methods for construction of these scoring systems include statistical modeling and machine learning. With the acceleration of computational power, mixed-integer optimization models have been shown to be an effective way to learn from large data sets through the selection of key features and point values of a risk score.

## METHODS

We have created a Mixed Integer Linear Programming (MILP) formulation for the development of a risk scoring system to identify a patient's potential of a specific adverse outcome. Given a set of patient features, our model assigns integer point values representing how much each feature contributes to the risk of the outcome.

Current approaches in score development do not attempt to decide upon cutoff values for risk stratification. Instead, cutoff values are determined after the score is developed. In our model we include a constraint to choose a cutoff value for stratifying high and low risk patients (e.g. if the score is greater and 5, then there is a "high risk"). The objective of the model is to maximize the proportion of correctly classified patients. Our objective considers data sets where outcomes are imbalanced by giving equal value to the proportion of correctly identified positive and negative outcomes.

## DATA

We test our model using three data sets:

1. Center for Medicare and Medicaid Services (CMS) from 2009-2012 with patient claims to Medicare services, to predict preventable hospital admissions due to heart failure.
2. Sylhet Diabetes Hospital in Sylhet, Bangladesh from 2018 of recently diagnosed positive and negative diabetic patients, for an early-stage diabetes detection score.
3. A/H3N2 virus set from 2004 which measures the antigenic distance between strains and the resistance to vaccinations to assess potential of mutated variants to resist immunizations.

## MODEL FORMULATION

We consider an array of data where $I$ is the number of observations and $F$ is the number of features. A vector of point values is denoted by $\boldsymbol{\lambda}$ and is constrained to be an integer with a maximum value of $P$.

For each observation $x_i$ we must bound the vector product of $\boldsymbol{\lambda} x_i$ between the upper and lower bound of a scoring interval which is decided by the optimization model.

The objective is to maximize the balanced accuracy by correctly placing each observation in a high or low risk scoring interval. Value is gained by increasing the rate of true positives and true negatives given the ratio of actual positive and negative outcomes.

A full formulation of the model is given below.

$\boldsymbol{\lambda}$ : vector of point values for each of the attributes
$l_s$ : Lower bound of score interval $s$ for $s = 1, \ldots, S$
$u_s$ : Upper bound of score interval $s$ for $s = 1, \ldots, S$
$z_{is}$ : $\begin{cases} 1 & \text{if observation } i \text{ has score } s \text{ (i.e. } \boldsymbol{\lambda}' \boldsymbol{x}_i \in [l_s, u_s]) \\ 0 & \text{otherwise} \end{cases}$
Let $\boldsymbol{Z}$ be the matrix representation of these variables.
$w_{i,s,l}$ : A linearized variable where $w_{i,s,l} \equiv l_s z_{i,s}$
$w_{i,s,u}$ : A linearized variable where $w_{i,s,u} \equiv u_s z_{i,s}$
$M$ : a very large number

$$\max_{\boldsymbol{\lambda}, \boldsymbol{Z}} \; R_1 + G_0$$

$$\text{s.t. } w_{i,s,l} \leq \boldsymbol{\lambda}' \boldsymbol{x}_i \leq w_{i,s,u} + N_i P(1 - z_{i,s}) \quad \forall i \in \Omega^{\mathrm{T}}, \; \forall s$$

$$w_{i,s,l} \leq M z_{i,s} \qquad \forall i \in \Omega^{\mathrm{T}}, \; \forall s$$

$$w_{i,s,l} \leq l_s \qquad \forall i \in \Omega^{\mathrm{T}}, \; \forall s$$

$$w_{i,s,l} \geq l_s + M(z_{i,s} - 1)$$

$$w_{i,s,u} \leq M z_{i,s} \qquad \forall i \in \Omega^{\mathrm{T}}, \; \forall s$$

$$w_{i,s,u} \leq u_s \qquad \forall i \in \Omega^{\mathrm{T}}, \; \forall s$$

$$w_{i,s,u} \geq u_s + M(z_{i,s} - 1)$$

$$\sum_s z_{i,s} = 1 \qquad \forall i$$

$$\lambda_0, \boldsymbol{\lambda} \in \{0, \ldots, P\}$$

$$z_{is} \in \{0, 1\} \qquad \forall i \in \Omega^{\mathrm{T}}, \; \forall s$$

$$l_s, u_s \geq 0 \qquad \forall s$$

$$l_s \leq u_s - 1 \qquad \forall s$$

$$l_{s+1} = u_s + 1 \qquad \forall s$$

where...

$$R_1 = \frac{\sum_i y_{i,1} z_{i,1}}{\sum_i y_{i,1}} \qquad G_0 = \frac{\sum_i y_{i,0} z_{i,0}}{\sum_i y_{i,0}}$$

## SAMPLE OUTPUT

The following is an example of output using the Sylhet Diabetes dataset. We have constrained the model to only select 5 features to provide a compact risk score.

| | |
|---|---|
| Polyuria | +4 |
| Polydipsia | +2 |
| Sudden Weight Loss | +1 |
| Genital Thrush | +3 |
| Irritability | +3 |
| **TOTAL** | ... |

| | |
|---|---|
| Score ≥ 5 | High Risk |
| Score < 5 | Low Risk |

Features that appear in the original dataset include: sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, Itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity.

## RESULTS

We compare our results to 3 popular classification algorithms including RiskSlim, LASSO, and SVM, all of which use a post hoc cutoff approach. We use statistical testing to compare results from our scoring system to the other classification methods. Our model performs better when the number of variables allowed to be selected is unconstrained. Embedding the cutoffs in the constraints of the optimization problem provides a fully optimized tool for use in numerous applications.

Compared to other classification models, we observe comparable results in the areas sensitivity, specificity, accuracy, and area under the curve (AUC). The computational time is significantly longer, however a feasible solution is still found in acceptable time. We find that there is a need to develop a heuristic solution algorithm for scalability to large datasets.

## DISCUSSION

There are some limitations to a MILP approach for risk score development. Using integers for scoring weights provides interpretability of the results, especially for use in a real-world application. However, the addition of features and observational data may increase the computational requirements on an exponential scale. As such, a pragmatic application is to use a smaller subset of data as a training set to determine scoring weights.

## SUMMARY OF OUR SOLUTION

We have developed a classification model which uses medical feature data to construct a risk score. While our investigation is ongoing, we have had success in prediction of risk for negative outcomes such as:

1. Preventable hospital admissions due to heart failure
2. Early-stage diabetes detection given a list of symptoms
3. Influenza resistance to immunizations based on comparisons of antigenic distance.

We categorize the development of our model in these 5 steps:

- **Data preparation and pre-processing** includes the selection of high-quality classification data sets. We convert our feature data to binary inputs for our model to solve.
- **Formulation and linearization** of the model to ensure that an optimal solution to our problem is found. We allow the user to specify parameters such as the total number features to be selected for a score and the largest allowable point value for any of the selected features. By linearizing the model we ensure that the formulation is a valid MILP.
- **Score computation** with mixed integer programming that produces interpretable risk scores for conditions and adverse outcomes. We use Gurobi optimization software to solve the problem to optimality.
- **Statistical hypotheses and testing** to compare our results to existing attempts at developing risk scores. We perform a 10-fold cross-validation and split our data sets into 80/20 training and testing subsets to obtain summary statistics. We compare model validation metrics such as sensitivity, specificity, accuracy, F1 score, and area under the receiver operator cure (ROC).
- **Development of a heuristic algorithm** to arrive at a solution in less time. Using integer variables in mathematical optimization is challenging from a computational perspective. Although a heuristic will provide a sub-optimal solution, the trade-off is a much quicker solution time with a minimal effect on the accuracy of the outcome.

Oregon State University