Modified Logistic Regression using the EM Algorithm for Reject Inference

Billie Anderson, Ph.D.
Research Statistician
SAS® Enterprise Miner™

J. Michael Hardin, Ph.D.
Senior Associate Dean
University of Alabama

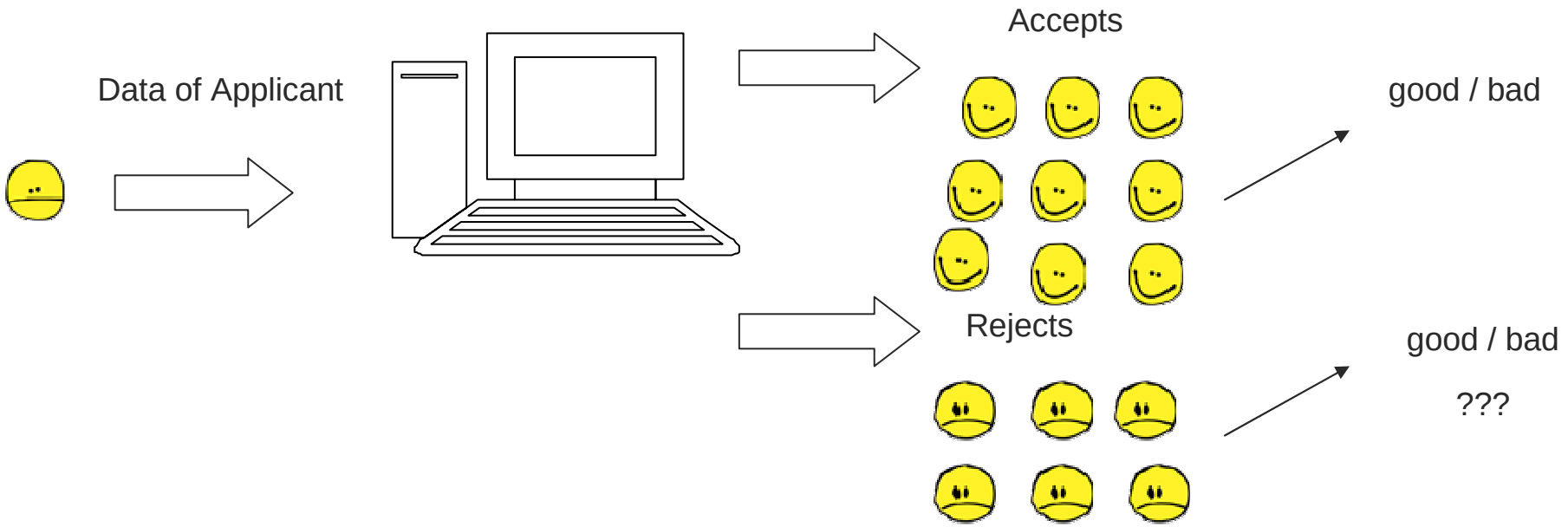
Credit Scoring

- Main aim of credit scoring is to predict whether an applicant, if accepted, will repay the loan on time or not
- Predict the probability that an applicant will default
- Compare with cutoff probability to decide whether to accept or reject applicant

Credit Scoring-continued

- In many financial institutions, the performance of only a subset of applicants is observed
- The performance of those applicants rejected in the past are not observed

Credit Scoring-Observed Performance



Issues in Credit Scoring

- Over time, credit score models degrade due to changing credit applicant populations
- Financial institutions periodically update their credit models

Issues in Credit Scoring-Reject Inference

- Most of the time the only complete available data to update the scorecard is that of the accepted applicants
- Why is this a problem?
- Sample bias
- Need statistically sound representative scorecard development sample

Modified Logistic Regression Modeling Applied to Reject Inference

- Many credit scoring models contain categorical variables for which it is not reasonable to assume an underlying distribution
- Most widely used credit scoring model is logistic regression

Modified Logistic Regression Modeling Applied to Reject Inference

- Propose a method for estimating the parameters in a logistic regression model in which the response variable for the rejected applicants will be missing
- The Expectation-Maximization (EM) algorithm will be employed to solve for the parameters for this logistic regression with missing target values
- The modified logistic regression (EM logistic model) will be compared to standard augmentation through a simulation study

Overview of Expectation Maximization (EM) Algorithm

- Basic idea of the E (for expectation) M (for maximization) algorithm is to associate the incomplete-data problem as a complete-data problem for which maximum likelihood estimation is computationally easier

Logistic Regression with Missing Values

- In multiple logistic regression, it is assumed that each observation j, j=1,…,n,n+1,…,n+m,are independent observations such that $Y_j|x \sim$ Bernoulli(p_j) where the logistic model takes the form

$$p_j = P(Y_j = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1+\exp(\mathbf{x}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}\boldsymbol{\beta})}$$

in which there are *m* missing *y* observations for the rejected applicants and

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ denotes the unknown p-vector of parameters to be estimated

EM Logistic Model

- Since each Y_j is a Bernoulli random variable, where: $P(Y_j = 1 | \mathbf{x}) = p_j$

  we can represent its probability distribution as follows:

$$f(Y_j) = p_j^{y_j}(1 - p_j^{y_i}) \quad Y_j = 0,1;\ j = 1,\ldots,n,n+1,\ldots,n+m.$$

EM Logistic Model

- The observed (incomplete) data log-likelihood is the likelihood function that contains the *n* accepted applicants. Since Y_j observations are independent, the likelihood function is:

$$g(Y_1,\ldots Y_n) = L(\boldsymbol{\beta}) = \prod_{j=1}^{n} f_j(Y_j) = \prod_{j=1}^{n} p_j{}^{y_j}(1-p_j)^{y_j}.$$

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{n} y_j \mathbf{x}\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}\boldsymbol{\beta})).$$

EM Logistic Model

- The complete data log-likelihood is the likelihood function that contains the accepted and rejected applicants. The likelihood function that contains the accepted and rejected applicants can be expressed as:

$$g(Y_1,\ldots Y_n,Y_{n+1},\ldots,Y_m)=L_C(\boldsymbol{\beta})=\prod_{j=1}^{n} p_j^{\,y_j}(1-p_j)^{y_j}+\prod_{j=n+1}^{m} p_j^{\,y_j}(1-p_j)^{y_j}.$$

$$\log L_C(\boldsymbol{\beta})=\sum_{j=1}^{n} y_j\mathbf{x}\boldsymbol{\beta}-\log(1+\exp(\mathbf{x}\boldsymbol{\beta}))+\sum_{j=n+1}^{m} y_j\mathbf{x}\boldsymbol{\beta}-\log(1+\exp(\mathbf{x}\boldsymbol{\beta}))$$

Applying the EM algorithm to the Logistic Model with Missing Values; E- step

- E-step requires the calculation of

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) = E\{\log L_c(\boldsymbol{\beta}) \mid \mathbf{x}, \boldsymbol{\beta}^{(0)}\}$$

This step is performed here by replacing each unobserved outcome variable for the rejected applicants, y_j $j=n+1,\ldots,n+m$ by its expectation conditional on $\mathbf{x}_j$, given by

$$E(Y_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(0)}) = P(Y_j = 1 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(0)}) = \frac{1}{1+\exp(-\mathbf{x}_j\boldsymbol{\beta})}$$

Applying the EM algorithm to the Logistic Model with Missing Values; M- step

- M-step maximizes the expectation of the complete log likelihood,

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) = E\{\log L_c(\boldsymbol{\beta}) \mid \mathbf{x}, \boldsymbol{\beta}^{(0)}\}$$

- Will use iteratively re-weighted least squares (IRLS), a numerical algorithm that maximizes any specified objective function using a standard least squares method (Kotz and Johnson 1983).

M-step-continued

1. Choose initial estimates of the regression coefficients, such as the estimates of $\boldsymbol{\beta}$ using the accepted applicants only

2. At each iteration, *k*, update the regression coefficients:

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} + (\mathbf{X}`\mathbf{V}^{(k-1)}\mathbf{X})^{-1}\mathbf{X}`(\mathbf{y} - \mathbf{p}^{(k-1)})$$

- **X** is the data set matrix
- **y** is the observed outcome vector

M-step-continued

- $\mathbf{p}^{(k-1)}$ is the vector of fitted response probabilities for the previous iteration, the *ith entry is*

$$p^{i,(k-1)} = \frac{1}{1+\exp(-\mathbf{x}\boldsymbol{\beta}^{(k-1)})}$$

- $\mathbf{V}^{(k-1)}$ is the diagonal matrix, with diagonal entries

$$p^{i,(k-1)}(1-p^{i,(k-1)})$$

M-step continued

3. Repeat steps 2 until $|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}|$ is close to zero

Reject Inference EM Logistic Model Simulation Study using SAS® German Credit Scoring Data Set

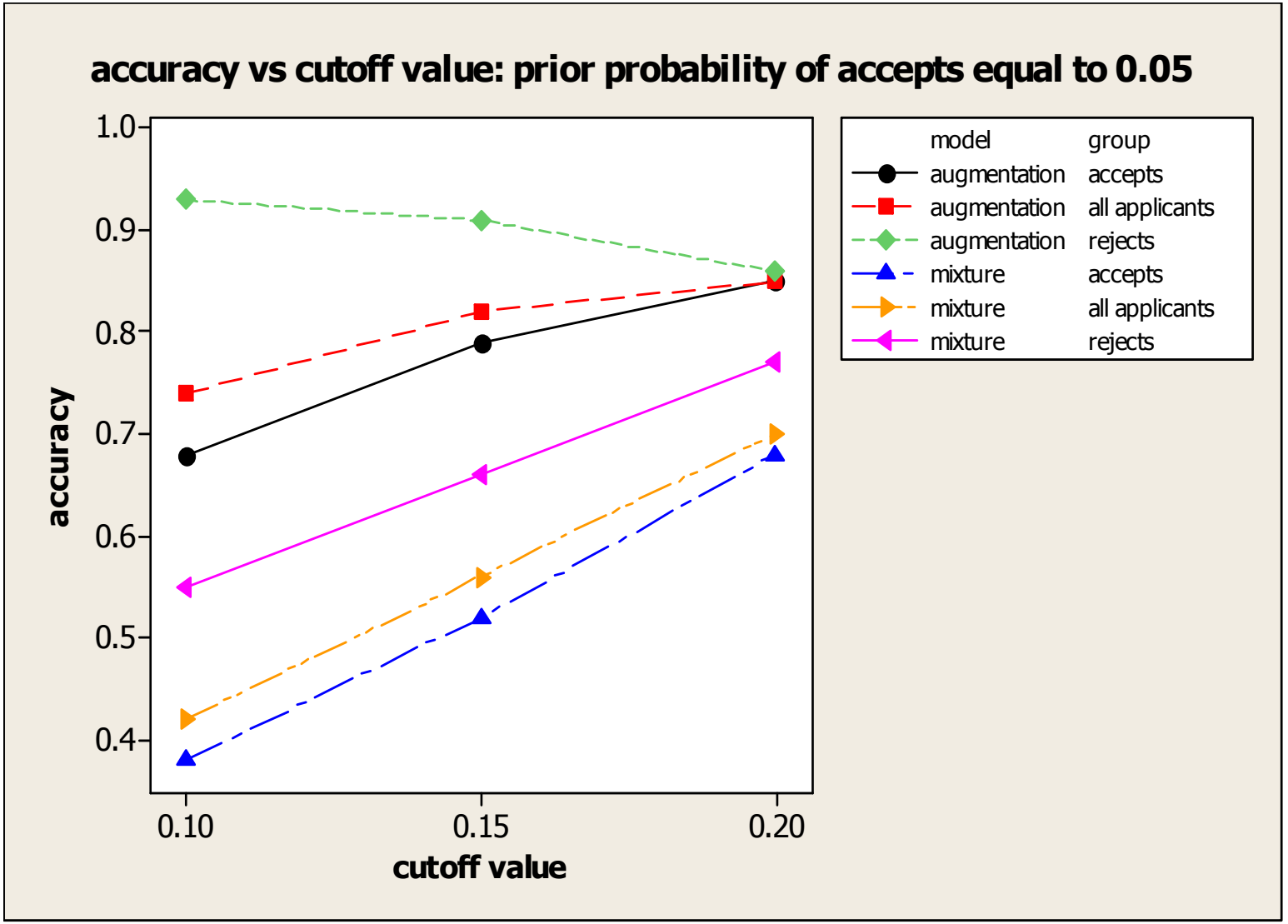
- 4500 applicants, 3,000 accepts (1500 good; 1500 bad) and 1500 rejects
- Many characteristic variables were used in this study: age, income, time at job, time at address, bank card, number of children, number of persons in household, number of loans, number of finished loans, residential status

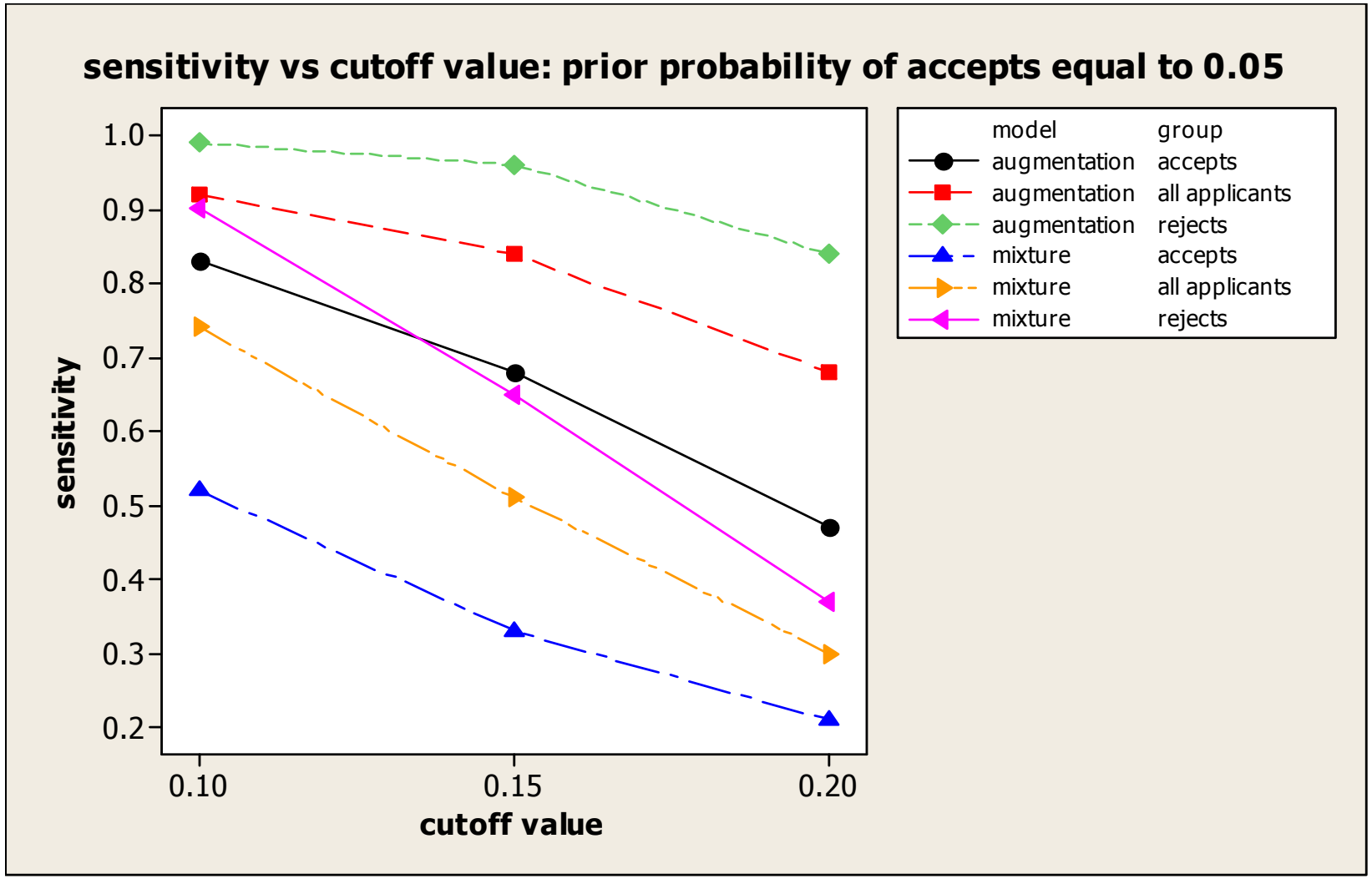
Reject Inference EM Logistic Model Simulation Study using SAS® German Credit Scoring Data Set

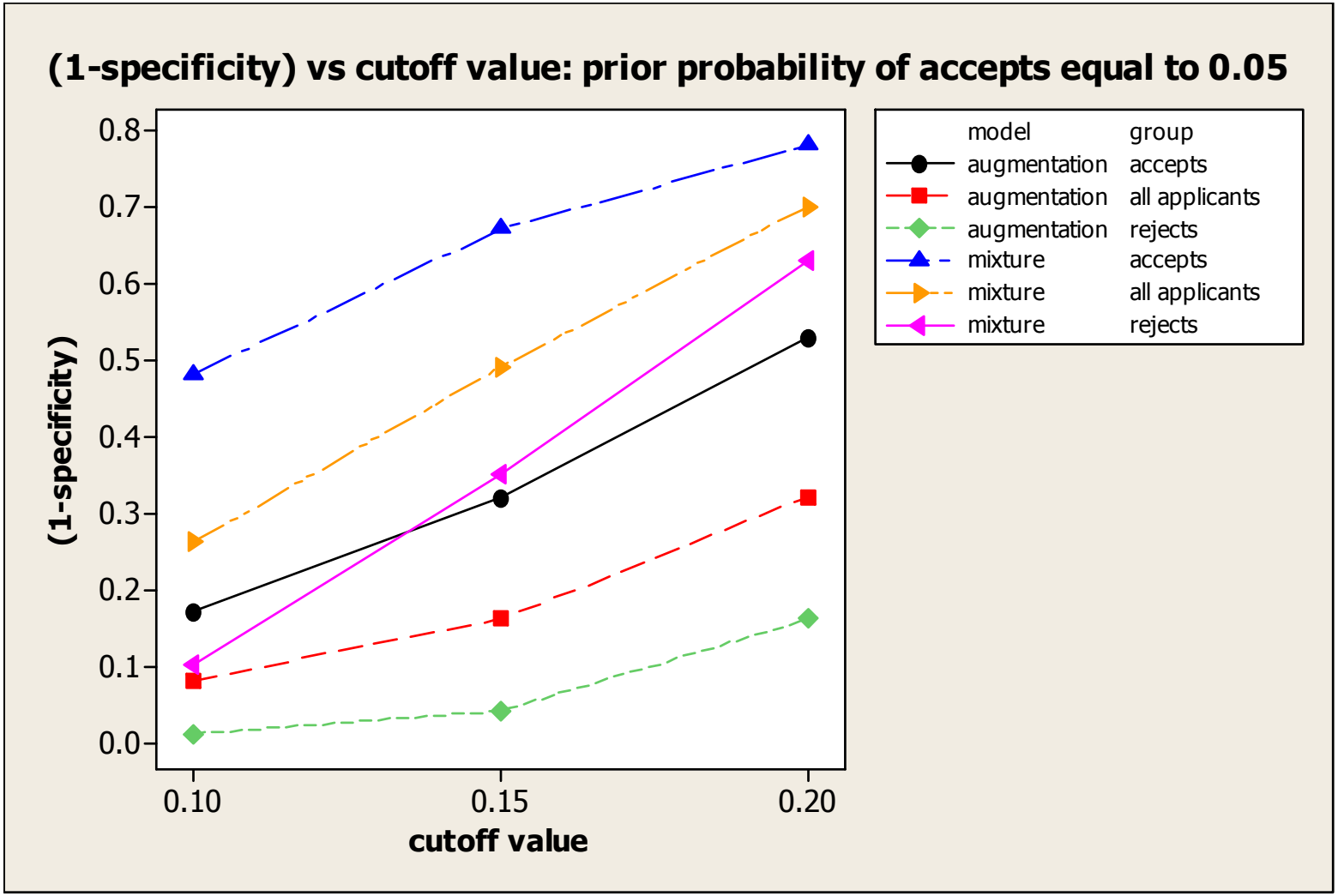
- Performed a simulation using this data set
- Compared EM Logistic model to a standard augmentation procedure using different model performance measures

EM Logistic Simulation Study Steps

1. Build logistic model using the accepted applicants with the SAS® German credit scoring data.
2. Use the same model from step 1 to determine the good/bad loan status for the rejected applicants.
3. Sample the accepted and rejected applicants in such a way as to mimic true prior probabilities in credit scoring data sets.
4. Split the sampled data into in-house and future data sets.
5. Build the reject inference models (augmentation and EM Logistic).
6. Apply the models to the future data sets.
7. Test the performance of the reject inference procedures on the future data set.
8. Accumulate the model performance measures from the future data set.







Conclusions/Questions

THE POWER TO KNOW®