Credit Card Debt Repayment Model: A Double Hurdle Approach

An applied project submitted in partial fulfillment of the requirements for the degree of
Master of Science at Virginia Commonwealth University

By

Michael A. Thurber

To

Jason R. W. Merrick, committee chair

Virginia Commonwealth University
Richmond, VA
April, 2009
Acknowledgments

The author wishes to thank his applied project committee members who have provided essential direction and support throughout this applied project. Dr. Jason Merrick and D’Arcy Mays of the Department of Statistical Sciences and Operations Research at Virginia Commonwealth University provided invaluable expertise, direction, and encouragement throughout. Kevin Busby provided both the opportunity to do this applied project and significant support to make this successful.

This is gratefully dedicated to my wife Carolyn and my children, Amy, David, and Matthew who have given me the unwavering commitment, encouragement, time, and confidence that were so much needed throughout the development of this work. Without them this would have been impossible.
# Table of Contents

CREDIT CARD DEBT REPAYMENT MODEL: A DOUBLE HURDLE APPROACH ............... I

ACKNOWLEDGMENTS ............................................................................................................... II

TABLE OF CONTENTS ................................................................................................................ III

LIST OF TABLES ........................................................................................................................ IV

LIST OF FIGURES ....................................................................................................................... V

ABSTRACT ........................................................................................................................................ VI

1 LITERATURE REVIEW AND CHOSEN APPROACH .............................................................. 1

1.1 CREDIT CARD BUSINESS BACKGROUND ...................................................................... 2

1.2 HISTORY OF METHODS ..................................................................................................... 3

1.3 ALTERNATIVE MODEL FORMULATIONS ......................................................................... 4

2 METHODOLOGY ...................................................................................................................... 13

3 DATA PREPARATION .............................................................................................................. 15

3.1 DATA GATHERING AND VARIABLE REDUCTION ........................................................... 15

3.2 TRANSFORMATIONS OF INDEPENDENT VARIABLES .................................................. 16

3.3 LIMITING THE RANGE OF VALUES ................................................................................... 18

3.4 LIMITING THE RANGE OF VALUES ................................................................................... 18

3.5 IDENTIFYING TRANSFORMATIONS FOR LOGISTIC REGRESSION ................................. 21

4 PAYER RATE MODEL DEVELOPMENT ............................................................................. 26

4.1 PAYER RATE MODEL SELECTION .................................................................................. 26

4.2 CONSIDERATION OF INTERACTION TERMS ................................................................... 27

4.3 PAYER RATE MODEL FITTING ....................................................................................... 28

4.4 PAYER RATE MODEL VALIDATION ................................................................................ 29

5 CONDITIONAL YIELD MODEL DEVELOPMENT ................................................................. 30

5.1 BOX-COX TRANSFORMATION FOR THE CONDITIONAL YIELD MODEL .................. 30

5.2 CONDITIONAL YIELD MODEL SELECTION ................................................................. 32

5.3 FITTING FINAL CONDITIONAL YIELD MODEL WITH NON-LINEAR REGRESSION ....... 36

5.4 CONDITIONAL YIELD MODEL VALIDATION ................................................................. 38

6 OVERALL YIELD MODEL EVALUATION ........................................................................... 40

6.1 OVERALL COMPARISON BY R-SQUARE ........................................................................ 40

6.2 OVERALL COMPARISON BY RANK ORDERING POWER .............................................. 41

7 MODEL INTERPRETATION ...................................................................................................... 45

8 CONCLUSIONS ..................................................................................................................... 46

BIBLIOGRAPHY ......................................................................................................................... 47
List of Tables

TABLE 1 BUSINESS CATEGORIZATION OF REGRESSORS .............................................................. 16

TABLE 2 LOGISTIC REGRESSION PARAMETER ESTIMATES OF FINAL PAYER RATE MODEL ............................................................................................................................... 28

TABLE 3 FITTED CONDITIONAL YIELD MODEL .............................................................................. 34

TABLE 4 MULTI-COLLINEARITY DIAGNOSTICS FOR THE CONDITIONAL YIELD MODEL.... 35

TABLE 5 MEANS OF LINEAR AND NON-LINEAR CONDITIONAL YIELD MODELS ............... 36

TABLE 6 FITTED NON-LINEAR REGRESSION ESTIMATES ............................................................. 38

TABLE 7 OVERALL MODEL PERFORMANCE VS. PR ONLY MODEL BASED ON R-SQUARE... 41

TABLE 8 QUANTITATIVE COMPARISON OF THE RANK-ORDERING POWER OF THE MODELS ............................................................................................................................... 44
# List of Figures

- **FIGURE 1** A TYPICAL HISTOGRAM FOR PAYMENT AMOUNT ON A SET OF PAYING ACCOUNTS  
  \[ 9 \]

- **FIGURE 2** HISTOGRAM OF DATA FROM FIGURE 1 AFTER A BOX-COX TRANSFORMATION WITH $\lambda=0.40$  
  \[ 11 \]

- **FIGURE 3** DISTRIBUTIONS BEFORE AND AFTER CONSTRAINING RANGE  
  \[ 17 \]

- **FIGURE 4** REGRESSOR DISTRIBUTION WITH DEPENDENT VARIABLE  
  \[ 20 \]

- **FIGURE 5** LOGIT VS. UNTRANSFORMED REGRESSOR  
  \[ 22 \]

- **FIGURE 6** LOGIT VS. LOG OF REGRESSOR  
  \[ 23 \]

- **FIGURE 7** LOGIT VS. SQUARE ROOT OF REGRESSOR  
  \[ 24 \]

- **FIGURE 8** LOGIT VS. SQUARE OF REGRESSOR  
  \[ 25 \]

- **FIGURE 9** SELECTION OF TWELVE MOST USEFUL REGRESSORS  
  \[ 27 \]

- **FIGURE 10** SAS PROC LOGISTIC RESULT  
  \[ 28 \]

- **FIGURE 11** LAMBDA SELECTION FOR BOX-COX TRANSFORMATION  
  \[ 31 \]

- **FIGURE 12** SELECTING MODEL SIZE, OR NUMBER OF REGRESSORS IN MODEL  
  \[ 33 \]

- **FIGURE 13** LINEAR REGRESSION ANALYSIS OF VARIANCE TABLE, WITH A BOX-COX TRANSFORMATION  
  \[ 34 \]

- **FIGURE 14** NON-LINEAR CONDITIONAL YIELD MODEL RESULT  
  \[ 37 \]

- **FIGURE 15** MEAN STATISTICS FROM CONDITIONAL YIELD MODEL VALIDATION  
  \[ 39 \]

- **FIGURE 16** THE DOUBLE HURDLE MODEL SHOWING A CLEAR LIFT OVER PAYER RATE AND CONDITIONAL YIELD MODELS ALONE  
  \[ 43 \]
Abstract

CREDIT CARD DEBT REPAYMENT MODEL: A DOUBLE HURDLE APPROACH

By Michael A. Thurber, M.S. candidate

An applied project submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University

Virginia Commonwealth University, April 2009

Committee Head: Jason Merrick, Associate Professor, Department of Statistical Sciences and Operations Research

Credit card issuers desire to objectively allocate resources to debt recovery resources to reclaim defaulted debt. Prioritization of these resources is best done by a prediction of how much a cardholder is expected to pay. A double hurdle model is developed that incorporates the probability to make at least one payment as the first hurdle, and the expected total payment amount as the second hurdle.

A literature review was conducted to find the state-of-the-art practices in cases where the binary participation response (in this case whether or not to make a payment) and extent of response (in this case amount of payment) are modeled with divergent sets of explanatory variables. The double hurdle model is well researched and proven and is used here to estimate the expected yield on defaulted accounts. The effectiveness of the double hurdle model over what a participation model or an extent of participation model alone would provide is demonstrated.
1 Literature Review and Chosen Approach

This Literature Review and Chosen Approach chapter provides the business background and discusses various published methods to simultaneously address participation and extent of participation decisions, specifically applied to credit card debt repayment. The alternative mathematical formulations are presented, and the Box-Cox double hurdle formulation is chosen.

The Data Preparation chapter describes the nuances and characteristics of the explanatory and dependent variables. Univariate analytic results are discussed as well as appropriate imputations and transformations.

The Payer Rate Model Development chapter reviews the process for specifying and selecting explanatory variables for models for the first binary hurdle, the probability of any payment. The final parameter estimates are developed and validated.

The Conditional Yield Model Development chapter reviews the process for specifying and selecting explanatory variables for models for the second continuous hurdle, the estimation of how much will be paid if any payment is made. The final parameter estimates are developed and validated.

The Overall Yield Model Evaluation chapter presents the overall double hurdle model and evaluates its performance as compared to simpler models. Performance is measured by the R-square statistic as well as model rank-ordering power.

The Model Interpretation chapter discusses the reasonableness of the models from a business perspective. Interpretability of parameters is discussed as well as the importance of the model to the business.
1.1 Credit Card Business Background

Loan repayments on defaulted loans are critical to the financial services industry, especially in the current prevailing credit crisis. The credit card sector of the financial services industry, in particular, aims to meet the revolving credit needs of its account holders with diverse circumstances and credit histories. Failure to make acceptable payments for a period of 180 consecutive days results in default and revocation of the privilege to withdraw cash or make new purchases with the card. The cardholder remains accountable for the outstanding balance, however, and the issuer of the card typically uses multiple means to recover that balance. The recovery of this debt always requires the allocation of limited resources. Efficient prioritization of these resources to recover the debt are typically based upon the likelihood of the cardholder to pay and the expected payment amount within a given finite time frame.

The aim of this applied project is to model the repayment of credit card debt based on historical behavior and present circumstances of the account holder. This is done here using a double-hurdle model that attempts to address inherent problems such as censored data and non-normal error terms.

Models of debt repayment are traditionally binary, modeling the probability of making a payment, i.e., the expected payer rate. But the realized value of an attempt to obtain a payment is the amount paid, not simply whether or not any payment was made. Thus, the objective is to go further and model the extent of payment, i.e., yield, over a given time period. While empirical evidence shows that a payer rate model slopes yield fairly well, it is not optimal. This is because the circumstances or characteristics that predict that a borrower will pay are not identical to those that estimate the amount of payment. For example, complete contact
information with address and verifiable phone numbers may indicate a higher propensity to make a payment. But access to other credit such as home equity or other credit cards in good standing may be more important in predicting how much will be paid. This leads us to explore classes of models where the propensity to pay and the expected amount of payment are treated separately.

1.2 History of Methods

In every application of the double hurdle model, discussed extensively in econometric literature, there is a binary participation decision (whether or not to make a payment, join a cause, or purchase a commodity) and another about the extent of the participation (usually, how much to pay).

Foundational work for the linear and logistic regression models has extended over a century. Sir Francis Galton (1885) introduced the concept of regression towards the mean, as it relates to heredity and human height. By the time of Watson (1967), the concept of linear least squares regression was established.

Nelder et al. (1972) outlined the generalized linear model that encompassed multiple linear regression and logistic regression. Comprehensive textbooks have been available for some time, e.g., Seber's (1977) *Linear Regression Analysis*.

The double hurdle model has been formally applied at least twice in the credit scoring literature: by Dionne et al. (1996), whose dependent variable is the number of non-payments and Moffatt (2005), who applied it to model the risk of loan default in conjunction with the amount of default, using a Box-Cox transformation. Moffatt's modeling approach is closely followed in this project to estimate the propensity to make a payment and the expected amount of payment.

The double hurdle model is intrinsically parametric in nature, where the error terms in both equations are assumed to be normally distributed. This assumption is difficult to substantiate with simple formulations. Thus, transformations of the dependent variable, such as the Box-Cox transformation, are typically used to alleviate this concern.

Another concern is extensively addressed in the literature, namely, correlation of the error terms between the participation (binary) equation and the extent (continuous) equation. If such correlation is treated as positive, a double hurdle model with dependence is required. Smith (2002) covers this approach in some detail and states that the central correlation parameter is poorly identified. He finally recommends that in practice it is most useful to assume the correlation is zero. This is also the anticipated approach with this project.

1.3 Alternative Model Formulations

The double hurdle model has been developed in stages by several authors over the last few decades. Tobin (1958) first proposed the Tobit model for censored data. Deaton and Irish (1984) proposed the more-sophisticated p-Tobit model for consumption decisions. Cragg (1971) introduced the term “double hurdle model" where the explanatory variable for the participation model can be very different than those in the consumption or extent model. Jones (2000)
reported the effectiveness of a Box-Cox transformation on the extent of participation equation and relaxes the normality assumption on the dependent variable.

1.3.1 Tobit Formulation

The Tobit formulation is based on a linear specification to treat propensity to pay and payment amount equivalently:

\[ y_i^* = x_i \beta + u_i \quad i = 1, n \]  
\[ u_i \sim N(0, \sigma^2) \]

Where \( y_i^* \) represents account i's propensity to make at least one payment, \( x_i \) is a vector of account characteristics that explain both the propensity to pay and extent of expected payment, \( \beta \) is the corresponding vector of coefficients to be estimated, and \( u_i \) is a normally distributed error term, assumed to be homoscedastic. Now, let \( y_i \) be the amount actually paid, which is never less than zero, so:

\[ y_i = \max(y_i^*, 0) \]

which is the tobit or standard censored regression equation. The corresponding log-likelihood for the joint distribution function from Tobin (1958) is:
\begin{equation}
\text{LogL} = \sum_0 \ln \left[ 1 - \Phi \left( \frac{x_i \hat{\beta}}{\sigma} \right) \right] + \sum_+ \ln \left[ \frac{1}{\sigma} \phi \left( \frac{y_i - x_i \hat{\beta}}{\sigma} \right) \right] \tag{4}
\end{equation}

where \( \Phi(.) \) and \( \phi(.) \) are the standard normal cdf and pdf, respectively. "0" indicates summation over the zero observations where no payments were made and "+" indicates summation over the positive payment amounts.

### 1.3.2 p-Tobit Formulation

The Tobit model discussed above implicitly assumes that the circumstances that drive whether or not a payment is made are exactly the same as those that determine how much will be paid. The business need for credit card debt repayment, however, is to recognize that payer rate and yield are different. In contrast to the Tobit model, the p-Tobit formulation, originally proposed by Deaton and Irish (1984), assumes that the fraction of account holders who are potential payers is \( p \), implying that the proportion \( 1-p \) are not persuadable to pay. This allows for a class of persistently non-paying account holders as a proportion of the population.

The log likelihood function for the p-Tobit joint distribution function is:

\begin{equation}
\text{LogL} = \sum_0 \ln \left[ 1 - p \Phi \left( \frac{x_i \hat{\beta}}{\sigma} \right) \right] + \sum_+ \ln \left[ p \frac{1}{\sigma} \phi \left( \frac{y_i - x_i \hat{\beta}}{\sigma} \right) \right] \tag{5}
\end{equation}

Solving this for the maximum log likelihood gives an estimate for the parameters \( p \) and \( \sigma \) as well as the vector \( \hat{\beta} \).
1.3.3 Double Hurdle Formulation

The p-Tobit model discussed above assumes \( p \) is constant. The business expects that the circumstances and characteristic that drive credit card propensity to make a payment are somewhat different than those that drive amount of payment, or yield. The double hurdle model makes this important generalization, such that the propensity to pay is a separate function of account holder characteristics, \( \Phi(z_i \alpha) \), referring to the first hurdle. The second hurdle is a separate equation, \( \Phi(z_i \beta) \), that estimates how much would be paid, given the account holder's circumstances. Thus the double hurdle model has two equations:

\[
d_i^* = z_i \alpha + \varepsilon_i \tag{6}
\]

\[
y_i^{**} = x_i \beta + u_i \tag{7}
\]

\[
\begin{pmatrix}
\varepsilon_i \\
u_i
\end{pmatrix} \sim N
\begin{bmatrix}
0 & 1 \\
0 & \sigma^2
\end{bmatrix}
\tag{8}
\]

As noted previously, the covariance between the two is assumed to be zero, hence the 0 values in the off-diagonal terms. Formally, then, the first hurdle is represented by \( d \):

\[
d_i = 1 \text{ if } d_i^* > 0 \tag{9}
\]

\[
d_i = 0 \text{ if } d_i^* \leq 0 \tag{10}
\]

and the second hurdle is:
Finally, the expected amount of payment is:

$$ y_i^* = \max(y_i^{**}, 0) $$

(11)

and the log-likelihood is:

$$ LogL = \sum \ln \left[ 1 - \Phi(z_i \cdot \alpha) \Phi \left( \frac{x_i \cdot \beta}{\sigma} \right) \right] + \sum \ln \left[ \Phi(z_i \cdot \alpha) \frac{1}{\sigma} \phi \left( \frac{y_i - x_i \cdot \beta}{\sigma} \right) \right] $$

(13)

1.3.4 Box-Cox Double Hurdle Formulation

The double hurdle model discussed above has a dependent variable $y_i$, payment amount, with a very strong positive skew. To illustrate the censoring and skewing of the dependent variable, see Figure 1.1.
The models discussed previously rely heavily on the assumption of normality of error terms and hence the dependent variable. If the error term is not distributed normally, and the ordinary least squares estimate is not necessarily the maximum likelihood estimate. This is relevant for credit card debt repayment where the amount of payment is censored at zero and is skewed far to the right. Many methods are suggested in the literature for transformation of zero-

**Figure 1** A typical histogram for payment amount on a set of paying accounts
censored data, such as zero-inflated beta distributions of Box-Cox (1964) and Ospina (2007).

The Box-Cox is commonly used for this type of problem, such as by Moffatt (2005):

\[ y^\lambda = \frac{y^\lambda - 1}{\lambda}, \text{ where } 0 < \lambda \leq 1 \]  

(14)

The Box-Cox transformation is a general one where it is linear if \( \lambda=1 \) and is logarithmic as \( \lambda \to 0 \). Realistically, we expect \( \lambda \) to lie between 0 and 1.

Figure 1.2 shows visually how the Box-Cox method can transform the skewed data in Figure 1.1 into a roughly normal distribution.
After the Box-Cox transformation, the first hurdle in equations 1.6 and 1.7 remain unchanged. But the second hurdle becomes:

\[ y_i^{*T} = \max \left( y_i^{**T}, -\frac{1}{\lambda} \right) \]  \hspace{1cm} (15) 

and

\[ y_i^T = y_i^{*T} \text{ if } d_i = 1 \]  \hspace{1cm} (16)
\( y_i^T = -\frac{1}{\lambda} \) if \( d_i=0 \). \hspace{1cm} (17)

The log-likelihood for the joint distribution function for the Box-Cox double hurdle model is:

\[
\text{LogL} = \sum \ln \left[ 1 - \Phi(z_i, \alpha) \phi \left( \frac{x_i \beta + 1/\lambda}{\sigma} \right) \right] + \sum \ln \left[ \Phi(z_i, \alpha) y_i^{\lambda-1} \frac{1}{\sigma} \phi \left( \frac{y_i^T - x_i \beta}{\sigma} \right) \right] \hspace{1cm} (18)
\]

Note that the important difference between this and the log-likelihood for the simple double hurdle model is the addition of the Jacobian term, \( y_i^{\lambda-1} \).

For the conditional expected yield, equations 1.14 and 1.15 are rearranged:

\[
y_i^* = (\lambda y_i^{*T} + 1)^{1/\lambda} \hspace{1cm} (19)
\]

Then the unconditional expected yield can now be calculated:

\[
y_i = d_i y_i^* \hspace{1cm} (20)
\]
2 Methodology

The business objective is to produce a model of expected yield on defaulted accounts using a two-stage, or double hurdle, approach. The first stage, or hurdle, is the probability to make any payment, called Payer Rate or p, modeled with logistic regression:

\[ p_i = P[payer\_makes\_payment] \]

The second stage is the expected yield if any payment is made, modeled using a Box-Cox regression technique:

\[ y_{i,cond} = y_i \mid payer\_makes\_any\_payment = E[yield \mid payer\_makes\_payment] \]

The unconditional expected yield is then the product of the scores from the two models:

\[ y_i = p_i y_{i,cond} \]

The models are built over several steps:

1. Gather data and establish the scope of the candidate regressors for both models. This includes gathering the metadata or definitions about each variable, understandable in business terms. About 254,000 observations from a limited time period are set aside for model building and 85,000 for model validation.

2. Reduce the number of regressors being considered using non-parametric techniques. In this case, decision trees are used to eliminate marginally useful variables.

3. Transform the data in preparation for modeling. This entails univariate analysis against the target variable. This step includes limiting the range of values the variable can have, imputing missing values, and performing mathematical transformations that make the independent variable more linear with the dependent variable.
4. Build each model. This involves selection of both first and second order (quadratic and interaction) terms, to maximize the predictive capability of the model. The primary objective is to maximize Somers’ D for the logistic (first hurdle) model, and R-square for the yield (second hurdle) model. Care is taken to avoid serious multi-collinearity. Particularly for the yield model, the Box-Cox technique is applied for variable selection, but the final fit of $y_{i,\text{cond}}$ is done directly using non-linear regression.

5. Validate each model. Apply the fitted model to the validation sample, and confirm that the Somers’ D and R-square are not materially different from the build sample.

6. Measure the effectiveness of the overall model with the validation sample. Compute the expected yield for each observation in the validation sample. Also, compute probability to pay and yield conditioned on whether any payment is made. Compute overall statistics such as R-square. Also, the lift of yield for the double hurdle model should be significantly better than if the payer rate model, or the conditional yield model, is used alone.

7. Interpret the model. Confirm that the model makes business sense. Do the regressors intuitively belong in the models? Are the signs of the parameters believable? Are the statistically most significant terms intuitively the most important? Describe the impact of the regressors in business terms where possible.
3 Data Preparation

3.1 Data Gathering and Variable Reduction

Prior work done by the business suggested 50 inputs that could be used for a payer rate model and 23 variables that could be used as regressors in a yield model. These had been selected from among a much larger set using advanced decision tree software from Salford Systems® called TreeNet™ version 2.0. This method was chosen because it demands much less data preparation than linear regression techniques and can quickly and recursively build hundreds of decision trees that are used in turn to detect which variables have the most information to predict a specified target variable, whether it be categorical or continuous.

In addition, a previously built payer rate model provided 19 candidate regressor variables, some of which were redundant with those from TreeNet™ data exploration. These lists were consolidated and metadata were collected on each. Some of these were dropped for business reasons, such as if it came from an outdated business strategy. The consolidated sets were then analyzed again in TreeNet™ to remove variables that had the least incremental information in the presence of the others for either payer rate or yield.

The remaining candidate variables were divided into general categories describing the account holder’s credit characteristics:

- Access to Credit—Amount of credit previously acquired
- Delinquency—Challenges experienced in repaying debt, including resolution
• Payment History—Frequency and amount of payments on debt
• Card Usage—Extent of purchases on the account
• Contact History—Past communications with individual
• Cure History—To what extent have previous delinquencies been resolved
• Willingness to pay—A history of regular payments on debt, related to cure history

Summary of inputs for consideration as regressors in models, after the first reduction:

<table>
<thead>
<tr>
<th>Category</th>
<th>PR Model</th>
<th>Yield Model</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access to credit</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Account Characteristic</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Amount of Default</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Card usage</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Cure History</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Delinquency</td>
<td>10</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Payment History</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Willingness to Pay</td>
<td>2</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Grand Total</td>
<td>20</td>
<td>22</td>
<td>42</td>
</tr>
</tbody>
</table>

Table 1 Business Categorization of Regressors

The list for the PR model includes only 2 variables in the list for the yield model, and vice versa. Note that card usage was not a category used in a payer rate model, while payment history variables are frequent in both lists. This suggests that occasionally it might prove useful to share top-ranked variables in one model in the other model as well.

3.2 Transformations of Independent Variables

Graphical techniques were used to examine and transform the data. For example, histograms of each regressor were created to help identify extreme values:
Figure 3 Distributions before and after constraining range

In this example, the variable was capped at 23000, after which the histogram was created again to see the resulting distribution.
3.3 Limiting the Range of Values

Reasons to limit the range on values of a dependent variable include:

1. Invalid values. For example, an individual cannot have -1 financial accounts.

2. Likely incorrect values. For example, a $300,000 credit limit on a bankcard is not believable.

3. Insufficient sample size in the extreme values to be reliable in the model. For example, while some individuals might have 25 open credit cards, there is not a sufficient number for statistical reliability. From another point of view, these may inappropriately become highly influential or leverage points. Something like defaulted amount of the loan in question, however, would not be truncated.

4. Statistical evidence that a linear relationship essentially stops at a certain point and becomes flat.

Histograms with univariate statistics, and histograms with empirical payer rate overlays were used to judge appropriate thresholds.

3.4 Imputing Missing Values

Depending on knowledge of the regressor, different approaches are used for missing values:

1. If a missing inherently means something equivalent to a regular value, this is imputed.

   For example, missing mortgage information could well mean that the number of mortgages is zero.
2. Missing can mean the individual failed to provide the appropriate information. In this case, the value may be imputed to the median, but an additional indicator variable can be created that indicating that the information was missing.

3. Otherwise, a common practice is to impute to the median, since this generally would not make the value influential in the regression.

   In this analysis, there were few missing values, and the general approach was to specify imputation to the median.
Following is a typical example of a graphical technique to set an appropriate range on a variable.

**Figure 4  Regressor distribution with dependent variable**

This bivariate graph in figure 4 shows the average value of the dependent variable $p$ in association with the binned values of the independent variable $X_{12}$. In this case it is clear that the dependent variable has the same value when $X_{12}$ is 1 or less than 1. So a reasonable minimum for the range is 1. Then the relationship between $p$ and $X_{12}$ is fairly linear up through
about the last bin. A reasonable maximum might be about 10. Since there are no missing values for X12 in the sample, by default any future missing would be imputed to the median, 3. Such graphs were examined for all the continuous independent variables under consideration.

3.5 Identifying Transformations for Logistic Regression

Logistic regression uses maximum likelihood to fit the $\beta$ vector in the equation:

$$p_{i,\text{hat}} = \frac{1}{1 + e^{x_i \beta + \varepsilon_i}}$$

To identify appropriate transformations, this is rearranged algebraically to solve for $X\beta$, leaving the logit as the “dependent variable:”

$$\log \left( \frac{p_i}{1 - p_i} \right) = x_i \beta + \varepsilon_i$$

Ideally, the independent regressors should be linear with this. Again, graphical techniques are used to decide which transformation of the independent variable will be most nearly linear with the logit. The independent variable transformations considered are:

1. None. The untransformed version is roughly linear.
2. Log. This is quite common.
4. Square. This is rarely the best for the logit.

A logit graph is produced for each in this X5 example:
Figure 5 Logit vs. untransformed regressor
Figure 6  Logit vs. log of regressor
Figure 7  Logit vs. square root of regressor
Even though none of the four transformations is perfect, the log transformation is the best of the four, showing the best linear relationship. Such graphs were produced for all the candidate regressors for the logistical model, and a transformation was chosen in a similar fashion for each.
4 Payer Rate Model Development

4.1 Payer Rate Model Selection

A binary variable such as whether a person pays or not within a specified time period is modeled with logistic regression. Forward and backward selection methods are applied to all the transformed inputs in SAS PROC LOGISTIC to select a model. None of the regressors is rejected at the 95% confidence level. This is because the list was already culled down to only valuable variables, and the number of observations in the build sample was so large (254,000) that the Wald Chi-square test statistic was consistently above 4.

The business has an interest in keeping the number of inputs to its model low. First, it is important to interpret the model, and this becomes increasingly important as the number of regressors increases. Second, business models must be monitored carefully and regularly, along with their inputs, and this becomes more tedious and time consuming as the complexity of the models increases.
To minimize the number of variables in the model without significantly reducing its value, a forward selection technique was applied, tracking the Somers’ D statistic with the addition of each new variable. The result in Figure 9 shows that after 12 regressors, the Somers’ D statistic is 0.487, and there is relatively little advantage in adding any of the other variables:

![Somers' D Chart]

4.2 Consideration of Interaction Terms

For simplicity and interpretability, the only interactions considered were second order, where at least one of the two was a binary or class variable. The three most important class variables were interacted with the three most important continuous variables. Backward selection was applied, resulting in a model whose Somers’ D statistic was 0.501 with 35 terms. This constitutes a great cost in terms of complexity with marginal improvement over simpler models. Hence, no interaction terms are included in the final logistic model.
4.3 Payer Rate Model Fitting

Fitting the 12 regressors with PROC LOGISTIC yields:

Figure 10 SAS PROC LOGISTIC result

<table>
<thead>
<tr>
<th>Category</th>
<th>Generic Name</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cure History</td>
<td>XPR1</td>
<td>5447.50</td>
<td>&lt;.0001</td>
<td>0.1008</td>
</tr>
<tr>
<td>Delinquency</td>
<td>XPR2</td>
<td>2260.95</td>
<td>&lt;.0001</td>
<td>-0.2064</td>
</tr>
<tr>
<td>Payment History</td>
<td>XPR3</td>
<td>1488.45</td>
<td>&lt;.0001</td>
<td>0.1572</td>
</tr>
<tr>
<td>Delinquency</td>
<td>XPR4</td>
<td>1419.67</td>
<td>&lt;.0001</td>
<td>-0.0115</td>
</tr>
<tr>
<td>Cure History</td>
<td>XPR5</td>
<td>1069.01</td>
<td>&lt;.0001</td>
<td>0.32</td>
</tr>
<tr>
<td>Willingness to pay</td>
<td>XPR6</td>
<td>1054.59</td>
<td>&lt;.0001</td>
<td>0.0816</td>
</tr>
<tr>
<td>Delinquency</td>
<td>XPR7</td>
<td>904.72</td>
<td>&lt;.0001</td>
<td>0.355</td>
</tr>
<tr>
<td>Cure History</td>
<td>XPR8</td>
<td>698.05</td>
<td>&lt;.0001</td>
<td>0.3544</td>
</tr>
<tr>
<td>Access to Credit</td>
<td>XPR9</td>
<td>635.15</td>
<td>&lt;.0001</td>
<td>0.00118</td>
</tr>
<tr>
<td>Account Characteristic</td>
<td>XPR10</td>
<td>461.42</td>
<td>&lt;.0001</td>
<td>class variable</td>
</tr>
<tr>
<td>Delinquency</td>
<td>XPR11</td>
<td>427.58</td>
<td>&lt;.0001</td>
<td>-0.0829</td>
</tr>
<tr>
<td>Delinquency</td>
<td>XPR12</td>
<td>378.84</td>
<td>&lt;.0001</td>
<td>-0.094</td>
</tr>
</tbody>
</table>

Table 2 Logistic regression parameter estimates of final payer rate model

Note that the signs on the coefficients are intuitive, e.g., an individual with a past history of resolving prior delinquencies (cure) will be more likely to cure again.
4.4 Payer Rate Model Validation

The performance of model is compared between the build and validation sample. The model is applied to each and SAS PROC FREQ with the MEASURES option was used to calculate the Somers’ D and other statistics based on the actual and fitted (scored) values. The build sample had a Somers’ D statistic of 0.4875, while the validation sample had a Somers’ D statistic of 0.4789, 1.8% lower than on the build sample. This is an acceptable difference, and we conclude that over-fitting is not a significant problem for the payer rate model.
5 Conditional Yield Model Development

The model for the second hurdle of the overall yield model is the amount paid by individuals who make a payment, i.e., the yield conditioned upon whether or not a payment was made. The dependent variable is continuous and zero-censored.

5.1 Box-Cox Transformation for the Conditional Yield Model

To accommodate the zero-censored distribution for yield on defaulted accounts, a simplified Box-Cox transformation is applied on the target variable. This allows us to perform ordinary least squares multiple linear regression to make a final selection of regressors in the yield model with fewer concerns about the assumption of homogenous variance.

\[ y_{i, \text{box-cox}} = y_i^\lambda \mid y_i > 0 = x_i \beta + \varepsilon_i \]

The technique for choosing \( \lambda \) is done by iteratively applying ordinary least squares regression using different values of \( \lambda \) and maximizing the log likelihood each time, where

\[ \log L = -\frac{n}{2} \ln(\sigma_{\text{hat}, \lambda}^2), \text{ where } \sigma_{\text{hat}}^2 = \frac{SS_{\text{ERROR}}}{n}. \]
A SAS macro is used to estimate an appropriate value for $\lambda$ of 0.23:

Figure 11  Lambda selection for Box-Cox transformation
With the number of candidate regressors, including interaction terms, reduced to 27, the business objective of having a relatively small model must be met. SAS PROC RSQUARE allows every combination of regressors to be considered for modeling the transformed yield target, $y_{i,box\_cos}$, reporting R-square and Mallow’s CP statistics for each combination. Model size selection is done by plotting the best R-square statistic and best CP statistic against the number of regressors as shown in Figure 12 below. In this case, 13 regressors appear to be sufficient. This is can be seen since 27 regressors produce an R-Square value of 0.4213, only 1% higher than 0.4177, achieved with less than half as many regressors.
Figure 12  Selecting model size, or number of regressors in model

Note that a Cp value of 375 would typically be considered high, but it is satisfactory here, given 254,000 observations.

Multiple linear regression on the y_box_cox target variable fits the 13 regressors. Six of these are two-way interactions among the transformed variables. Twelve separate inputs are required to produce these 13 regressors.
Standard multiple linear regression yields:

Figure 13 Linear regression analysis of variance table, with a Box-Cox transformation

Table 3  Fitted conditional yield model
<table>
<thead>
<tr>
<th>Number</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.28689</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1.33049</td>
<td>2.17376</td>
</tr>
<tr>
<td>3</td>
<td>1.06343</td>
<td>2.43144</td>
</tr>
<tr>
<td>4</td>
<td>0.9259</td>
<td>2.60577</td>
</tr>
<tr>
<td>5</td>
<td>0.8212</td>
<td>2.76689</td>
</tr>
<tr>
<td>6</td>
<td>0.63218</td>
<td>3.15354</td>
</tr>
<tr>
<td>7</td>
<td>0.55613</td>
<td>3.36225</td>
</tr>
<tr>
<td>8</td>
<td>0.44814</td>
<td>3.74552</td>
</tr>
<tr>
<td>9</td>
<td>0.30008</td>
<td>4.57716</td>
</tr>
<tr>
<td>10</td>
<td>0.26219</td>
<td>4.89674</td>
</tr>
<tr>
<td>11</td>
<td>0.21814</td>
<td>5.36844</td>
</tr>
<tr>
<td>12</td>
<td>0.11578</td>
<td>7.36883</td>
</tr>
<tr>
<td>13</td>
<td>0.03945</td>
<td>12.62381</td>
</tr>
</tbody>
</table>

Table 4  Multi-collinearity diagnostics for the conditional yield model

The parameter estimates are all significant. Multi-collinearity is assessed and shown to exist, given the coded interaction terms, as evidenced by some variance inflation factors above 10. However no condition index is greater than 13, which is acceptable since it is well below 30. These 13 regressors are therefore selected for the final model.
5.3 Fitting Final Conditional Yield Model with Non-Linear Regression

Now that the most appropriate regressors have been identified using a Box-Cox transformation of the target variable, and major collinearity problems are ruled out, steps are taken to assure that the expected mean of the true target variable, yield, matches the expected predicted value, i.e., eliminate the bias intrinsic in the yield estimate. Note that without this, the expected values are considerably different than actuals:

![Table 5 Means of linear and non-linear conditional yield models](image)

The concern above is that after pred_y_box_cox is transformed back into normal yield (normal_y_hat), its expected value (739) is not the same as the mean of yield (872). This can be resolved by modeling yield directly, with non-linear regression:

\[ y_i | y_i > 0 = (x_i \beta)^{1/\lambda} + \varepsilon_i \]
Initial values for $\lambda$ and the $\beta$ parameters are required for non-linear regression. The initial value for $\lambda$ is .23, from the Box-Cox procedure. The $\beta$ vector is initialized with the values from the above regression for variable selection.

SAS PROC NLIN is used to perform the nonlinear regression itself, using the Marquardt method with the NOHALVE option, to find the direction of steepest descent at each iteration. Marquardt's method is equivalent to performing a series of ridge regressions and is especially helpful when the parameter estimates are highly correlated.

![The NLIN Procedure](image)

Figure 14  Non-linear conditional yield model result
Table 6  Fitted non-linear regression estimates

Note that with this transition to a direct (non-linear) model, no parameter changes sign and the order of magnitude remains similar, thus any interpretation remains the same. However, the parameters have significantly changed. Notably, $\lambda$ has increased from 0.23 to 0.59. This change happens because this is a model of yield directly, rather than yield raised to the $\lambda$ power. Also of note is that the mean square error of the normal yield has dropped from 1016427 down to 954062, so the nonlinear model is a better fit.

5.4 Conditional Yield Model Validation

As a check to validate that the model has not been over-fit, a validation (or hold-out) sample has been reserved for this. The observations in the sample have not been used for any of the variable selection or parameter estimation steps. It is about one-third the size of the build sample. The model fitted above is now applied to the validation sample, and the mean values are reported:
From this an R-square value is computed on the validation sample, yielding an R-square of 0.6382. Note that the average residual is not zero but is just 1.3% of the average yield, $y$, in the validation sample. This is acceptable, and we conclude that this yield model is adequate.
6 Overall Yield Model Evaluation

The double hurdle model presents another practical challenge: how to effectively measure the performance of the resulting model. As indicated previously, the unconditional expected yield from the double hurdle model $y_{i, dh, hat}$ is the product of the scores from the payer rate and conditional yield models:

$$y_{i, dh, hat} = p_{i, hat} y_{i, cond, hat}$$

6.1 Overall Comparison by R-square

The effectiveness of a yield model such as this can be measured by manually computing an R-square statistic:

$$R_{square, double, hurdle} = 1 - \frac{\sum_{i=1}^{n} (y_i - y_{i, hat})^2}{\sum_{i=1}^{n} (y_i)^2}$$

The payer rate model alone may be used as a point of comparison. The simplest form assumes a known mean payment amount for anyone who pays, treated here as a constant, $\bar{y}_{cond}$:

$$y_{i, PR, hat} = p_{i, hat} \bar{y}_{cond}$$

Then,

$$R_{square, PR} = 1 - \frac{\sum_{i=1}^{n} (y_i - y_{i, PR, hat})^2}{\sum_{i=1}^{n} (y_i)^2}$$
These statistics, as well as the mean residuals, are computed on both the build and the validation sample:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Model</th>
<th>Build Sample</th>
<th>Validation Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Square</td>
<td>Double Hurdle</td>
<td>0.2709</td>
<td>0.2703</td>
</tr>
<tr>
<td>R-Square</td>
<td>PR with pm per payer</td>
<td>0.1224</td>
<td>0.1224</td>
</tr>
<tr>
<td>Mean Residual (Bias) Double Hurdle</td>
<td>9.6991</td>
<td>7.2367</td>
<td></td>
</tr>
<tr>
<td>Mean Residual (Bias) PR with pm per payer</td>
<td>11.6899</td>
<td>10.6333</td>
<td></td>
</tr>
</tbody>
</table>

Table 7  Overall model performance vs. PR only model based on R-square

This shows that the double hurdle model validates very well with similar R-square statistics for the build and validation samples. Further, the R-square for the double hurdle model has an R-square value 2.21 times the R-square value than the payer rate model.

Note that both models are biased, though the double hurdle model is less so. The bias results from the simple multiplication of the two models. In order to remedy this, the two parts of the overall yield model might be modeled directly, in the double hurdle structure using non-linear regression:

$$y_{i, complete} = p_i y_{i, cond} + \varepsilon_i = \left(\frac{1}{1 + e^{x_i \beta}}\right) \left(z_{i, cond \_yield} \gamma_{cond \_yield}\right)^{1/\lambda} + \varepsilon_i$$

Fitting this combined overall model is not in scope for this effort, though it is a potential next step.

6.2 Overall Comparison by Rank Ordering power

Since models such as this might be used to prioritize work in order of decreasing expected yield, it is helpful to measure the performance of the model from that perspective. More resources could be allocated to those accounts with the highest expected yield. This suggests rank ordering by a model score, such as the estimated expected yield, \(y\), or the payer...
rate, or the conditional liquidation rate. To use this as a measure of performance on the validation sample, order the observations by the score in descending order. Then measure the cumulative actual yield, starting from the highest score and going to the lowest. The steeper the rise of this curve, the better the model. For purposes of comparison, calculate the optimal, or perfect model, then observe how close each individual model is to it, relative to a random model. (A perfect model would be one that predicts \textit{a priori} the exact yield of every account.)

This can best be illustrated graphically with lift curves:
Figure 16  The Double Hurdle model showing a clear lift over payer rate and conditional yield models alone

An excellent numerical measure of the usefulness of a model is its lift above the random model, or the area between the curve and the diagonal line in figure 16, as compared to an optimal, or perfect, model.
<table>
<thead>
<tr>
<th>Model type</th>
<th>Lift above random model</th>
<th>% of optimal model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal, or perfect</td>
<td>0.4474</td>
<td>100%</td>
</tr>
<tr>
<td>Double Hurdle</td>
<td>0.2898</td>
<td>65%</td>
</tr>
<tr>
<td>Conditional Liquidation Rate only</td>
<td>0.2464</td>
<td>55%</td>
</tr>
<tr>
<td>Payer Rate only</td>
<td>0.2331</td>
<td>52%</td>
</tr>
<tr>
<td>Random</td>
<td>0.0000</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 8  Quantitative comparison of the rank-ordering power of the models

In summary, the double hurdle model is clearly better than each model separately, in terms of rank ordering by expected yield.
Each model should be consistent with well grounded business knowledge. In the case of the payer rate model, the relative significance of parameter estimates show that past delinquencies that were resolved, or cured, in the past is the greatest predictor of probability to make a payment. Other positive contributing factors are willingness to pay, payment history, and access to credit. The negatively contributing factors are existing delinquencies.

In the conditional yield model, the most important factor is the amount owed, which is represented in the model as both a first and second order term, and with some interactions. The extent of prior card usage also comes into play as a singular term and in interactions. Some regressors, such as cured delinquencies, were also in the payer rate model. Finally, prior payment behavior and overall indebtedness play significant roles in the model. While it is obvious that such terms should be in the model, the complexity of the second order terms makes it difficult to explain the coefficients. Attempts were made to include fewer or more intuitive interaction terms, but this did not prove fruitful in the case of the conditional yield model.
8 Conclusions

Modeling yield on defaulted credit card debt is a challenge to model effectively. While a payer rate model is simple and generally interpretable, it does not predict yield reliably.

The double hurdle approach implemented here uses a multiplicative structure between a payer rate logistic model and a conditional yield Box-Cox model. For manageability, the number of regressors was kept relatively low without sacrificing significant predictive power.

Evaluation of the overall model demonstrated the superiority of the double hurdle model over a simple extension of the payer rate model alone or the conditional yield model alone. While additional steps could be taken to remove bias from the model used here, it is clear that this model significantly improves the prediction of yield on defaulted credit card accounts.
Bibliography


