Credit Case Study

Stephen Merrill and William White
STAT 536

February 11, 2017

1 Introduction

Credit card companies are interested in predicting outstanding balances of potential customers. This is due to the potential profit to be made from interest on outstanding balances and the potential loss from customers that declare bankruptcy. A company has collected data about their current customers and hope this information can be used to predict potential customer behavior. An accurate prediction method will allow the company to identify potential customers that are more likely to maintain a moderate monthly balance, who have the greatest potential for profit.

The dataset used in this analysis includes information on the outstanding credit card balance and several quantitative and qualitative characteristics of each cardholder. The main problem is that there are 90 cardholders out of the 400 in the dataset that have zero outstanding balance. This disrupts the assumption that the data are normally distributed. Also it negatively affects the linear relationship between balance and other quantitative variables.

This analysis is meant to produce a predictive model to be used for potential customers. The most important characteristics to predict balance are identified and a model is derived. Assumptions that are made for the model are checked to see whether or not they are reasonable. This model is then tested for prediction accuracy. All of these steps are meant to provide the most predictive model possible using available methods.

2 Model Formulation & Methodology

2.1 Multiple Linear Regression (MLR) Model

$$Y = \beta_0 + X_I\beta_I + X_R\beta_R + X_S\beta_S + X_A\beta_A + \epsilon$$

- Y : Credit Balance
- β_0 : Intercept - balance for a non-student subject, aged zero, with no income or rating.
- β_I : Effect for Income - for every \$1000 change in income, balance will change by β_I dollars.
- β_R : Effect for Rating - for every one unit change in rating, balance will change by β_R dollars.
- β_S : Effect for Student status - a status change from student to non-student or vice versa will change balance by β_S dollars.
- β_A : Effect for Age - for every one year change in age, balance will change by β_A dollars.
- ϵ : The random error in the model.

A linear regression model allows for both inference and prediction to be made. For this problem, interest lies in making predictions of future customers' credit card balances. The model accomplishes this goal by determining the effect sizes (the β values) of the significant variables: income, rating, student status and age. Once determined, prediction for balance can be made by gathering data on the subjects and making calculations according to the model.

2.2 MLR Assumptions

In order for multiple linear regression to be a valid model, the following assumptions about the data must be met.

Linearity Each variable must have a linear relationship with the response. If this is not the case, the entire model is invalid since it would be fitting a line to non-linear data. Transformations of the data are often used to solve this problem.

Independence The data must be independent. If this assumption is violated, measures of variability will typically be too small. This is a difficult assumption to verify. Usually prior knowledge of the data is required.

Normality The errors must be normally distributed. Otherwise, confidence and prediction intervals that depend on t distributions are incorrect.

Equal Variance The errors also must have equal variances. Without this, measures of variability will once again be invalid.

3 Model Justification & Performance

3.1 Model Selection

Adjusting for Zero Balance Prior to finalizing the model selection, data analysis revealed serious violations of the normality assumption. This problem was a consequence of the 90 data points with zero balance. Since no transformation created normality, those 90 data points were removed. After doing so, all the assumptions were met.

There are consequences of removing those data points. It is probable that there is some kind of relationship between the subjects with zero balance, and by removing them, the data can no longer be considered a random sample. This weakens the model's ability to make predictions since it is now built on the assumption that the subject has a non-zero balance. Therefore the credit card companies will not be able to accurately identify customers that will have zero balance using this model.

Best Subset Selection The raw sample data contained 10 different explanatory X variables. In order to determine which variables were most significant and find the model that yielded the best prediction of balance, the best subset selection method was used. This method looks at all possible combinations of variables in the model and selects the "best" one. However, there are many different criteria that define which model is best.

A k-fold cross validation algorithm was selected as the criteria in selecting the best subset of variables to include in the model. This algorithm selects k subsets (called training sets) and determines the Mean Square Error (MSE) for each possible number of variables included in the model by making a prediction and comparing to the data not included in the training set. That data

are known as the testing set. Once the number of variables to include is determined, best subset selection is done on the entire data set and the best model of the specified size is selected. Here, best is quantified using Residual Sums of Squares, the default criteria for best subset selection.

This algorithm, and use of cross validation error as the acceptance criteria, is appropriate for the data since the question of interest lies in making prediction and cross validation depends upon the subset prediction method previously outlined.

Interaction A possible interaction between income and student status was considered. However, including this interaction in the model did not produce a significant effect, so it was not included in the final model.

3.2 Model Assumptions Verification

Linearity This Added-Variable plot matrix shows a clear linear relationship between each of the four explanatory variables and the response.

Independence Independence is difficult to determine. Because there is no prior knowledge that would suggest a violation of this assumption, it is assumed to be met.

Normality This histogram and quantile plot of the residuals show that the errors are normally distributed.

Equal Variance This plot of the residuals offers no evidence of an unequal variance.

3.3 Model Fit

The model fits the data very well. Every β has a highly significant t value, which suggests that each individual effect is significant, and the entire model has a significant F-statistic, which means there is a significant effect from at least one β.

3.4 Prediction Results

In order to assess the ability of the model to predict accurately, another cross validation algorithm was used. The data were randomly subset into a training

and testing set, with 90% of the data in the training set and 10% in the testing set. The model was fit to the training set, and then used with the data in the testing set to calculate prediction intervals.
For this subset of the data, the prediction intervals contained all 31 points in the testing set.

4 Results

The model was able to predict the balance of potential customers. In order to use the regression model to make a prediction the subject's income, credit rating, age and student status are needed. The table below gives the estimates of the effect size of each variable in the model along with 95% confidence intervals for those estimates.

Income, age and rating can all be interpreted in the following way. The estimate for income effect size is about -9.72, this means balance would be expected to decrease by $9.72 on average as income is increased by $1000. The student variable is interpreted a little differently. The estimate of student status effect size is roughly 479.51, this means balance would be expected to be $479.51 more on average if a customer is a student.

5 Conclusions

This analysis was done to provide banks the most predictive model for potential customer balance. This was accomplished by finding the best number of variables to use in the model and then carefully selecting that number of variables that were most predictive of balance. Underlying assumptions for the model were then verified. The model was then used to predict the values of a small subset of the data and accuracy of the model was assessed.

There were problems that had to be addressed. Over 20% of the customers in the dataset had zero balance. This complicated the assumptions of normality and linearity. It was decided to exclude these customers because no transformation or combination of transformations could be found to deal with these zero balance customers. For future analysis a logistic regression equation could be used to deal with customers that potentially could have zero balance. A logistic regression would predict whether a potential customer would have zero balance or not. Then those potential customers predicted to have a non-zero balance would be run through the model described in this paper to predict their balance. This two-step analysis would allow for the use of all of the data and to take into account potential customers that could have zero balance.