James Taylor

Assignment 1

University of Maryland University College

DATA640

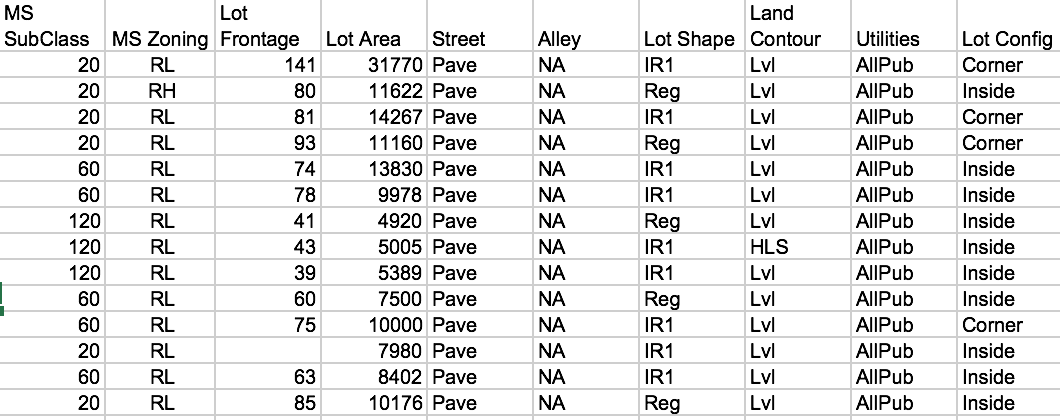
Dr. Knode

james.taylor.ps3@gmail.com

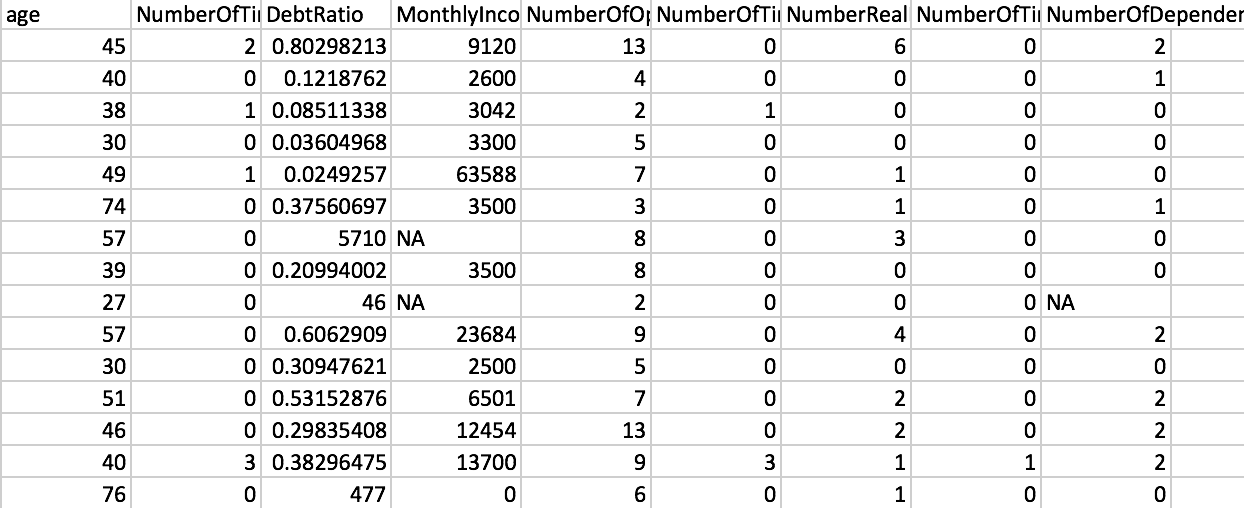
**Introduction**

The goal of this paper is to use SAS Enterprise Miner regression techniques to understand what independent variables explain for the variation seen in the target variable. There will be a linear and logistic regression used on two different datasets.

The linear model will use the Ames Housing dataset. It has 2,930 observations of homes sale in Ames, Iowa from 2006 to 2010 on 82 features. It has 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables along with two unique identifiers. The target variable will be the sale price which is continuous. The variables are specific features of the house and the land it sits on. Some examples are the dwelling type (one or two story), square footage, number of bathrooms and kitchen quality. Below is an example showing some features and observations.



The purpose of the model will be to see how certain features are correlated with increased/decreased sale prices keep all else equal, not for predicting on unseen data. Because regressions have variable estimates, we can inspect the estimated effect each one has on the dependent/target variable.

The logistic model will use the Credit Loan Default dataset. There are 150,000 observations on bank customers who took credit loans out on 12 features. The target variable will be the binary variable indicating if the customer was in serious delinquency of their loan within two years. The variables deal with age, measures of other types of debt, income, dependents and borrowing history. Below is a sample of data.

The purpose of the logistic model is to see what features contribute most significantly to a person’s credit default risk. It obviously shows missing values in some categories which will be handled later in this paper. The variable estimates will indicate which variables are correlated with increased or decreased odds of being delinquent.

**Data Preparation**

Exploring a dataset with 80 features creates an immense amount of potential relationships to investigate, so this exploratory analysis will be limited. The first relationship looked at was average home price given the year the home was built. It shows there were not homes sold in this data that were built each year, like 2006 for example. Houses appear to be more expensive the newer they are unsurprisingly. Interestingly homes built in 1950 and earlier generally have the same value and do not less valuable even if they are older. This measure does not capture things like renovations.

The house price varies depending on proximity to various conditions as well. The two most expensive groups are near positive off-site feature (park, greenbelt, etc.) and adjacent to positive off-site feature. The two lowest average groups are Adjacent to arterial street and Adjacent to East-West Railroad. It seems the quality of the home’s area drives desirability.

There are some missing values in certain variables in the Ames dataset, but because creating a model with 80 features would take too long to analyze only some were used. None of the features selected will have missing values or be transformed for the purposes of the model.

The Credit Loan data set needed preparation on two variables, monthly income and number of dependents. Those variables were imported as nominal variables, so the NA values were imported as levels and not detected. I specified the ‘NA’ values in those variables to be detected as missing using the ‘Replacement’ node. This allowed those variables to then be recognized as interval and used more easily for the model’s computation. There were no transformations done to any of the variables used to train the data.

**Predictive Models Developed**

The first linear model on Ames data had six features in it (Fireplaces, Garage\_Area, Lot\_Area, Overall\_Cond, Overall\_Qual and Year\_Built). They are all ordinal or continuous variables. These were chosen to attempt to measure the overall quality of the home. None of them were nominal, so there was no risk of the model automatically applying numbers to categories in a non-logical way. The ordinal variables make sense because they’re numerical and have a triage to them. Many of the continuous features in the set have data issues with them, for example if a house had no basement, the basement square-footage value would be zero. This throws off the normal distribution of the variable, making it difficult to fit a linear model to it.

The second linear model had 7 features as well (Above ground living area sqft., basement full baths, basement half baths, full baths above grade, half baths above grade, bedrooms above grade and kitchens above grade). They were all discrete variables. The goal is to see how the rooms in a house can influence the sale price. For example, this model will be able to determine if homes with one full and two half baths sells for more on average than a home with two full baths. A hypothesis may be that the large houses are more expensive, the more rooms of any type the more expensive.

The credit default is attempting to find which variables have predictive validity for a person credit worthiness. The first model used all the variables in the model. This will allow for the model to have all the features at its disposal to help make predictions. The second logistic model will have the Backward selection process for the features, where it starts with all the features and removes the least significant variable until all the remaining ones are very significant. This helps reduce dimensionality and simplify a model without sacrificing a lot of accuracy. The ROC curve will be analyzed to see what is the optimal value to round the models output. This is because the model outputs a probability between zero and one. This number has to be rounded down or up because there are only two values in the data, being zero and one. If one chooses to round a model up at the .9 and above, that means it will mark predictions with a 90% likelihood of being a true value (or one) as a one and the rest false. The consequences of this rounding are mentioned later in the paper.

**Results**

*Figure 1-Linear Model #1 Estimators* shows the estimated coefficients of the features. The R-squared value was .73, or these variables explain for 73% of the variation seen in the data. Also notable is that all the coefficient estimators were positive. This means for this model there is no feature that decreases the predicted sale price. For each variable, an increased value in its measure is increasing the predicted sales price. For example, for every increase of one square footage of lot area, the model predicts the sale price to be $1.31 higher. Of course lot sizes are large values (the mean value is about 10,000), so this would make up a significant part of the homes predicted price even if the estimator’s value doesn’t make that obvious.

The largest estimate value was for the Overall Quality variable. This was an ordinal rating from one to ten of the quality of the material and finish of the home. There is no surprise this is a useful predictor, a one unit increase corresponds to $30,000 more dollars on the final sale price. It was also the most statistically significant variable of the six with a t-value of 39, making it very useful in explaining the observed variation. Interestingly the Overall Condition of the house was less influential. It is another ordinal rating from one to 10, and a one unit increase only increases the predicted sale price by $2,500. One theory on why there was such a difference between overall quality and condition is the condition is easily improvable and quality is not. If the roof is dilapidated or parts of the house are sagging or crumbling, that can be an expensive fix that may take serious intrusive work. In contrast, to improve the condition of a house may just take some exterior power-washing and new paint.

The estimate of the number of fireplaces was surprisingly high. Each fireplace a home had increased the sale price by $18,000 on average. There is a possibility this feature is capturing a confounding variable, which is house age. Older homes needed fireplaces to warm them before the invention of centralized heating systems. Modern homes now have fire places as a luxury item, not as a necessity, which likely means the house is more luxurious overall.

The second linear model’s estimates are displayed in *Figure 2- Linear Model #2 Estimators*. The model only has estimates for six features when there were seven included. This is because Half Bathroom feature was not significant to the 5% level so it was not included. This model had an R-squared value of .65.

The values of the estimates are interesting. Bedrooms and Kitchens above ground have negative estimates, which really in unintuitive. Having a kitchen and bedroom seems like the standard, and having more would probably mean having larger, more expensive house. In fact, the model says that if a house has one kitchen above grade it decreases the value by $68,000. One of the ways the model is still able to make accurate predictions even with these estimators is adjusting the intercept. The intercept of the first linear model was -693,094, while the second one was 82,544 which is significantly different. It’s not obvious why kitchens and bedrooms have the negative estimates, but one theory is they have diminishing returns. The more beds and kitchens a house has, the less room for others like living rooms. Also, having two kitchens likely doesn’t make sense unless there’s lots of people living in the house. The same is true with bedrooms since they will go unused unless people live there to fill them. Ultimately this model doesn’t fit the data well, with 65% of the variation in sales price explained by these features.

The first logistical model on the credit data used all the features to make its predictions. The estimates are in *Figure 3- Logit Model #1 Estimators*. The estimates are in log-odds, but it is a monotonic transformation. That means positive estimates mean it increases the probability of the value being one (delinquent on credit) and negative estimates less likely. For example, the older someone is the less likely they are to be delinquent because the estimator is negative. Conversely, the more dependents one has the more likely they are going to be delinquent. One unexpected result was the Debt Ratio, were it was monthly debt payments divided by monthly income. The larger the value the more their livelihood is stressed by debt. Yet the estimator is negative, meaning the higher the debt ratio the less likely they are to be delinquent. The estimate does not have major influence on the prediction only having a -.00015 value.

Another interesting result was the estimate on times late in 60-89 days. Its estimate was -.906, which is a significantly negative value. A reasonable assumption would be that people who are 60-89 days late once are in a worse position than someone who was 30-59 days late once. The estimates don’t support this.

The rounding threshold used on the model’s predictions have important consequences. *Figure 4- Logistic regression #1 Classification* shows the classifications on the .5 rounding threshold. It shows an accuracy of 93% (correct classifications / all classifications). The sensitivity (ability to predict who will default) was low at 3%. The .5 rounding threshold had 9,705 false negatives and 321 true positives. *Figure 5- Logistic Regression #1 Cutoff* shows the tradeoff between sensitivity and specificity. The more sensitive a model has to be the less accurate it be at finding true negatives. The threshold is very important when the cost of a false positive is different than the cost of a false negative. In this credit loan scenario, a false positive means a lost customer that would have turned a profit, and a false negative means a customer becomes delinquent and the loaner may not get their money back which would be a major loss.

The second logistic model had a backward feature selection, which means it started with all available features and removes insignificant ones until the remaining are all significant. This selection removed two variables that did not meet the 0.05 level of significance, number of open credit lines and revolving utilization of unsecured loans. Removing two features does reduce dimensionality but not dramatically. The estimates are in *Figure 6- Logistic Regression #2 Estimates*. There are no surprises in terms of positive/negative estimates compared to the first model, although some values did change some.

The classification accuracy did slightly reduce as expected. *Figure 7- Logistic Regression #2 Classification* shows the accuracy at 93% like the previous model, but only four less correct predictions. This makes it quite advantageous to use this reduced model because nearly all the accuracy is retained. Like the previous model, the rounding threshold doesn’t make dramatic classification changes until the .1 value. The user of this credit model, like mentioned above, will have to make the decision on where this threshold should lie.

**Conclusions and Takeaways**

The regressions revealed some interesting insights. Houses have lots of attributes to them that drive the price. The regressions had separate goals, the first was using variables for the overall quality of the home and the second was the room structure. The quality model explained 8% more of the variation seen in the sales price than the room model. The room model had unexpected estimates, particularly having negative coefficient estimates on above-grade kitchens and bathrooms. This seems unusual since the model would lead one to believe a house would sell for a higher price if they didn’t have an above-grade kitchen/bathroom. The difference between quality and condition was surprising. It was difficult to tell the difference between the two by reading the variable description, but the estimates made it clearer.

The room model (model #2) made it clear that interactions between variables may be very important. This should be an area of further research. For example, if a home has a large outdoor pool and a small yard may be less desirable because of the lack of yard. Other associations should be looked at, like luxuriousness and location. It’s possible luxury homes are regional, in nice parts of town without interstates and highways.

The credit default models achieved high levels of accuracy, however this is because the number of delinquent clients was low, so even if a model predicted all of them to be non-delinquent it would be more than 90% accurate. The model struggled to achieve reasonable specificity, or to accurately recognize a future delinquent customer. Achieving better specificity will cost the overall classification accuracy, but as mentioned previously it may be necessary for the user’s goal. The coefficient estimates had some surprising results, notably 60-89 days was associated with decreased delinquency when 30-59 and 90+ days had increased delinquency.

Appendix A

Figure 1. Linear Model #1 Estimators

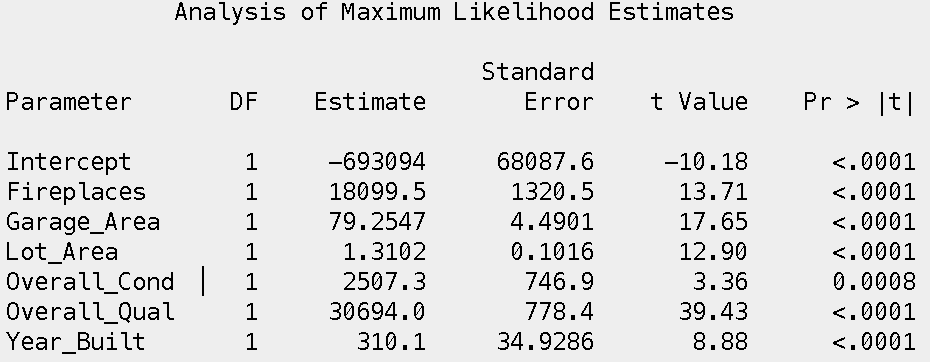


Figure 2. Linear Model #2 Estimators

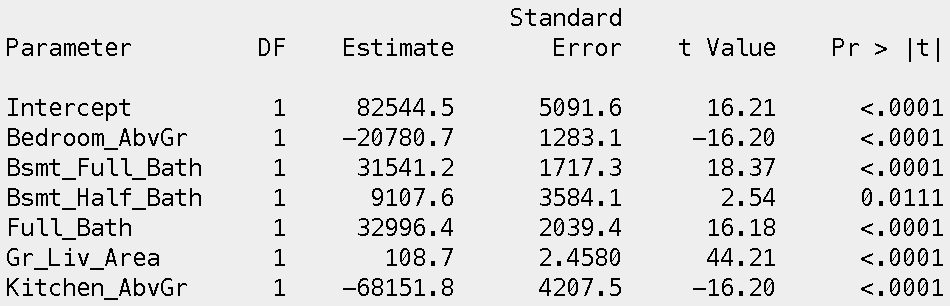


Figure 3- Logit Model #1 Estimators

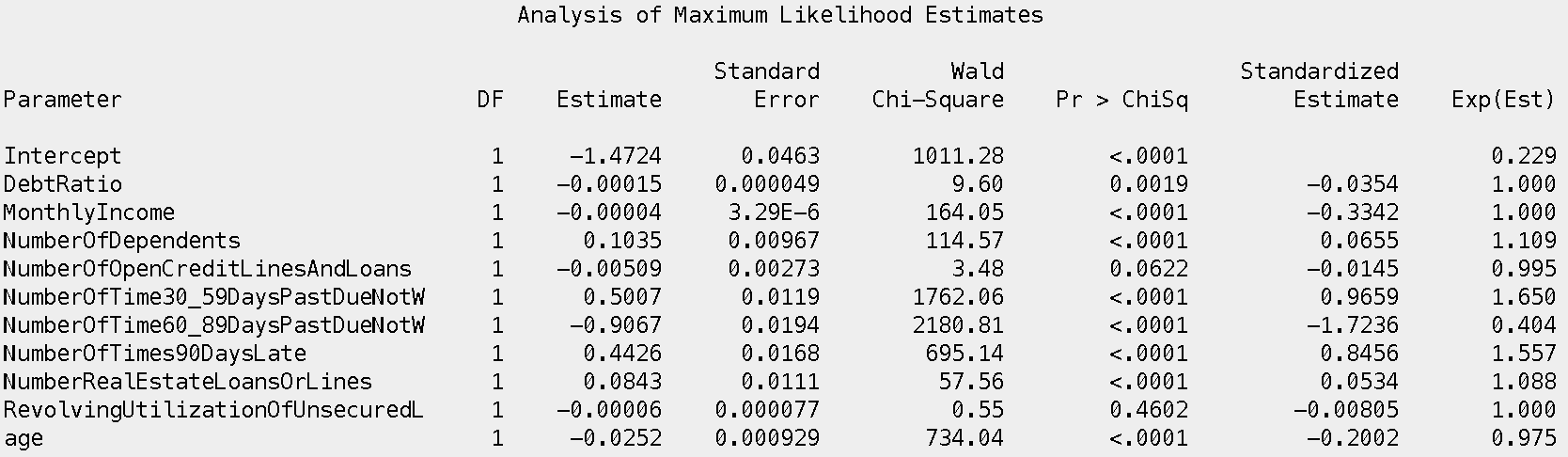


Figure 4- Logistic regression #1 Classification

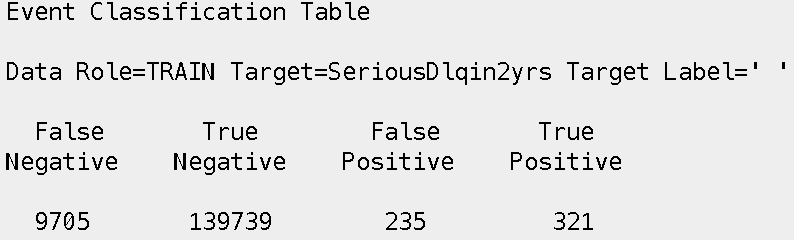


Figure 5- Logistic Regression #1 Cutoff

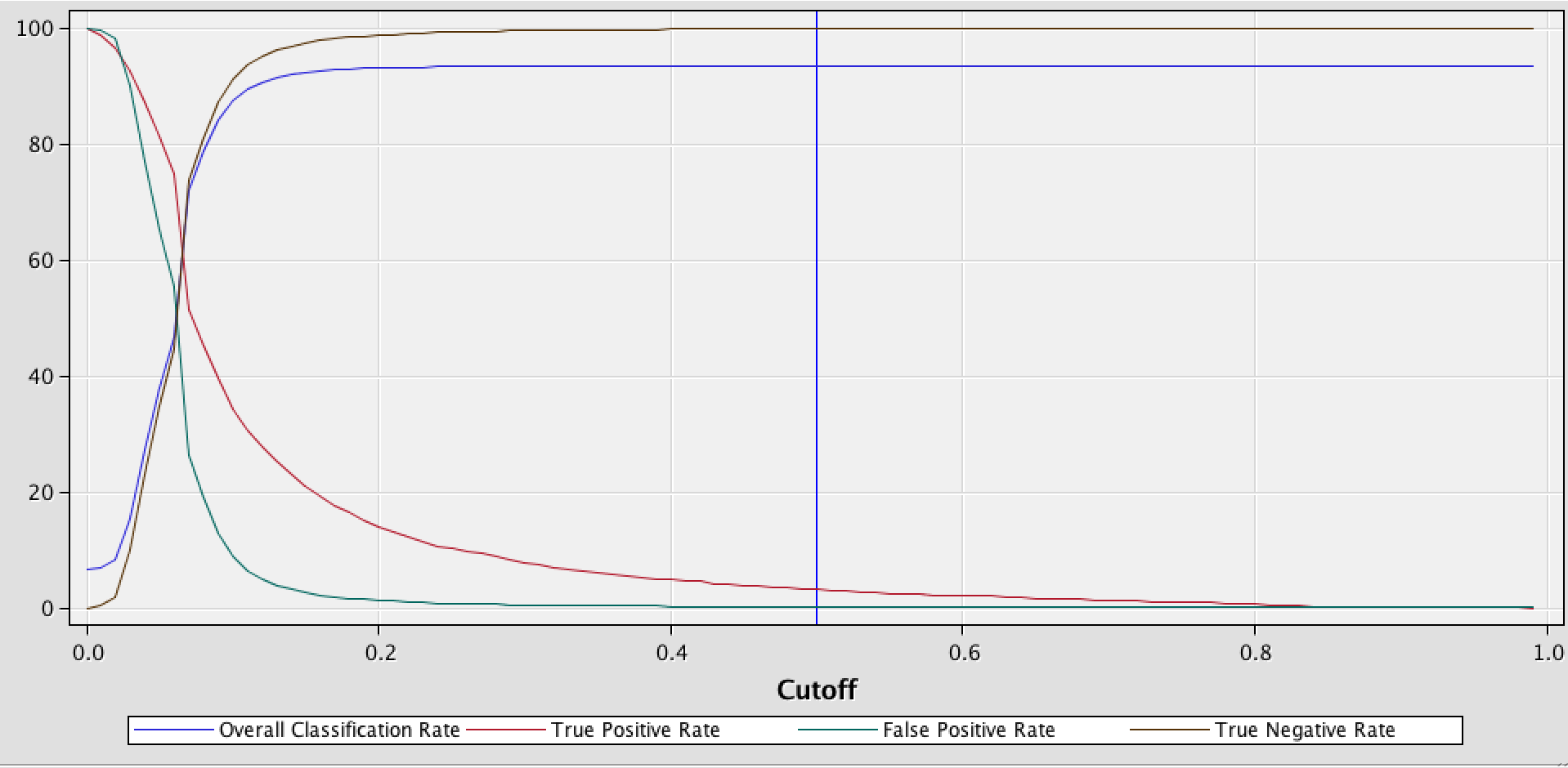


Figure 6- Logistic Regression #2 Estimates

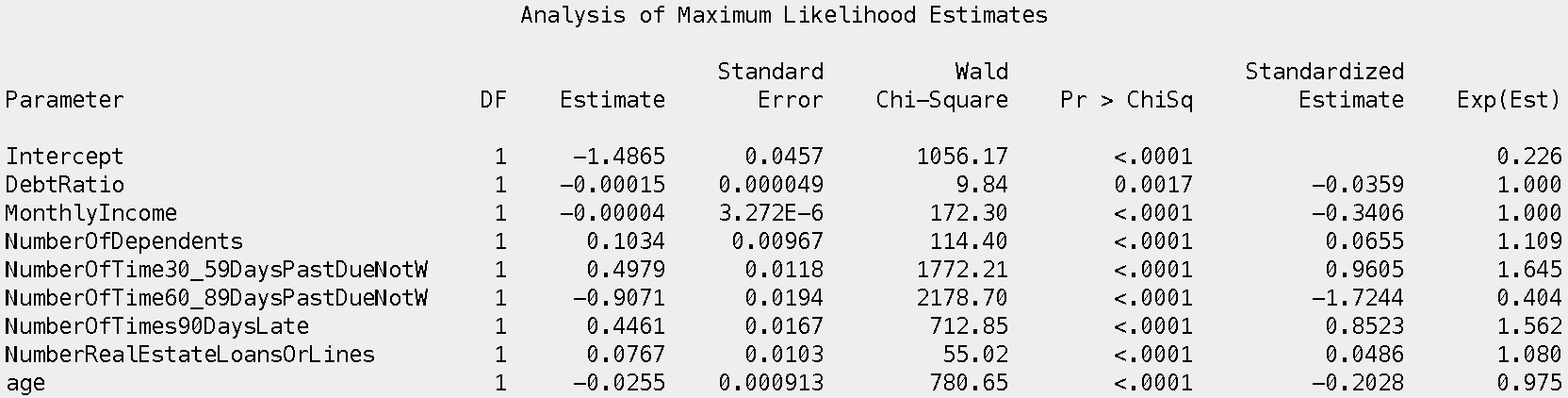


Figure 7- Logistic Regression #2 Classification

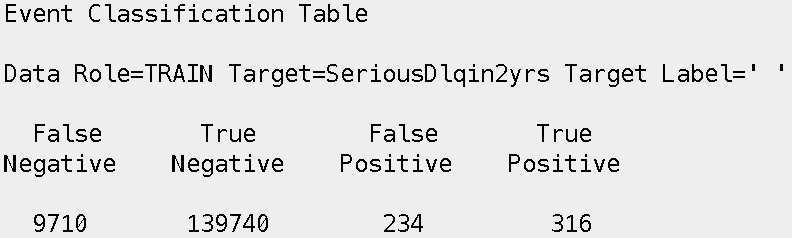


Figure 8- Logistic Regression #2 Cutoff

