Sensitivity = recall = TP / P

Specificity = TN / (TN + FP)

Precision = TP / (TP + FP)

Initial Model with all the variables:

Reduced Model with reduced number of variables suggests that although the Chi-squared test is significant, prediction and a test of ROC curves is not significant. So the reduced model is good enough as the full model.

Variable Selection

Logistic Regression

In the test data if a model predicts NO default 100% of the time then misclassification rate will be 22%. So, we expect the model to do better and have a lower misclassification rate.

Test Data had 6000 observations of which 4675 were NOT Default and 1325 were Default

Table has values for test data that the model has never seen before. Train:Test ratio = 80:20

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Misclassification Error | False Positives | True Positive | ROC (AUC) | Hyper Parameters | HL – GOF |
| Lasso (R) | 18.95% | 1.07% | 15% | 0.721 | Lambda = 0.014  Alpha = 1  Parameters = 12 | Significant |
| Elastic-Net (R) | 19.43% | 2.09% | 19.39% | 0.723 | Alpha = 0.2105  Lambda = 0.0193  Parameters = 11 | Significant |
| Logistic Regression (R) | 19.01% | 2.06% | 23.16% | 0.722 | AIC = 22462  Parameters = 12 | Significant |
| Logistic Regression (F) | 19.0% | 2.52% | 22.71% | 0.721 | AIC = 22457  Parameters = 16 | Significant |
| QDA (R) | 23.02% | 17.95% | 59.09% | 0.728 | Parameters = 182 | Significant |
| Random Forest | 17.75% | 4.46% | 36.15% | 0.657 | mtry = 2  trees = 500 | Significant |
| Neural Network | 22.10% | 0.021% | 0% | 0.711 | Hidden = 5  Decay Rate = 0.01  Bags = 50 | Significant |

Presentation Outline:

1. First page: My Introduction
2. Second page:
   1. Data source
   2. Motivation of Original data gatherers.
3. Problem Description:
   1. What will happen if a good model is established?
   2. What are the direct consequences of a good model in terms of service to customers and overall use of resources in a better way?
   3. The model may provide a better way to manage customers credit limits thus reducing overall risk from credit card defaults?
4. Data Description:
   1. What date range is the data from?
   2. What are some issues with the data? (Obvious outliers, highly structured, giving false confidence, undocumented levels of Marriage, Gender and Education)
   3. Conversion from New Taiwan Dollar to US Dollars
   4. Conversion from “numeric” to “factor” for dummy variable regression because Marriage, Gender and Education are nominal variables along with the response.
5. A couple of slides with charts to advance the story.
6. Talk about a naïve approach and get good results by looking at the distribution of “default” cases.
7. Model fitting motivation:
   1. Motivation for 80:20 split because, 80% of the data most probably contains underlying pattern.
   2. Classic fit using Logistic regression and all the available predictors.
   3. Common theme across these two models was the insignificance of the farthest 3 months of billed amount and paid amount.
   4. Why did I choose to use only the last 3 month information?
   5. Construction of new variable “trailing balance.” This new variable is more complex and has combined information from the last 3 months which also is logical and this also mitigates any possible multicollinearity problem.
   6. Selecting a reduced model and various ways to compare the two models, deviance test, roc test, inspecting prediction results and AIC.
8. Describe anomalies of predictors from different models.
   1. On average odds increase with AGE?? For GLM type models.
   2. Balance limit and trailing balance have almost no effect on the odds of default on average?
   3. Partial Plots from random forest for important variables and comparison with logistic regression estimates.
9. Conclusion:
   1. Ensemble methods perform poorly for this dataset because most of the variables are important towards a good model.
   2. Ensemble models are good when highly interactive behavior exist and to understand how to get better predictions and explanations using less important variables at the root i.e. giving the small guys a chance to add to the big picture.