Predict Churning Customers

DSC 424, Advanced Data Analysis and Regression
Bode Faleti, Eddie Cojulun, Sameer Ishaq

Introduction

Our project features a dataset of credit card customers that are categorized based on age, gender, education, and account activity among other variables. The goal of this analysis is to be able to predict credit card customers are likely to be churned so that the company can proactively provide these high-risk customers with better services to keep their business. rather than retain their credit accounts. We applied a multivariate regression technique to manage missing data and preformed exploratory analysis on the processed data to inform the analysis techniques we would incorporate. Ultimately, we go on to use principal components analysis, linear discriminant analysis, and regularized logistic regression techniques to identify patterns in the data, reduce the dimensionality of the dataset, and maximize the identification of these high-risk customers. We found that transaction activity rate by the customers contributed the greatest amount of variance in the numeric portion of the data. In addition, in comparison to LDA, regularized logistic regressions performed better at identifying the type of customers that require more attention and services in order to be retained.

Exploratory Analysis of the Data

The original dataset had 10,127 observations and 23 variables. We were able to immediately eliminate two variables 'Naïve Bayes' variables because they served as leftover results from a previous Naïve Bayes classification attempt. In addition, we were also able to eliminate the 'CLIENTNUM' variable because identification numbers would not serve a purpose in our analysis. Interestingly, 30.08% of the observations in the dataset (3046 observations) had at least one missing value and all of the missing fields occurred within the following categories: education level, marital status, and income category. Figure 1 features the distributions of variables with missing fields based on whether observations had one, two, or three missing fields within a particular observation. The observations with three missing values were immediately discarded because they only made up 0.069% of the dataset. In addition, the observations with two missing values were also purged for multiple reasons: firstly, they only make up 3.16% percent of the dataset but also because there's more room for variance when having to classify two different values. Finally, we retained the observations with just a single missing value because they made up 26.84% of the dataset.

Figure 1

Distribution of Observations with Unknown Values			
Single	Education	1267	Total = 2719
	Marital Status	560	% of dataset = 0.069
	Income	892	
Double	Education & Marital Status	107	Total = 320
	Marital Status & Income	75	% of dataset = 3.16%
	Education & Income	138	
Triple	Education, Marital Status & Income	7	Total = 7
			% of dataset = 26.85%

Figure 2

Absolute % Difference in Distribution of Classes (completed observations vs classified observations)					
Education Level		**Marital Status**		**Income Bracket**	
Class	Absolute % difference	Class	Absolute % difference	Class	Absolute % difference
Uneducated	.02%	Single	0.81%	Less than 40K	14.29%
High School	.06%			40K – 60K	17.58%
College	.08%	Divorced	0.2%	60K – 80K	13.46%
Graduate	.01%			80K – 120K	13.52%
Post-Graduate	.04%	Married	1.02%	120K+	4.89%
Doctorate	.05%				

Ultimately, we decided to use multinomial logistic regression to classify the remaining observations with just one missing categorical value. Firstly, we created 3 different regression models that were trained by the remaining, complete 7,081 observations; each one with the goal of producing probabilities of a certain class within the education, marital, and income variables. Next, we applied the models to the appropriate 2,719 incomplete observations and produced probabilities of each class occurring for every aforementioned observation. Finally, we used a Python script to classify each unknown observation based on the provided probabilities. After classification, we compared the class distribution of the complete observations to the recently classified observations for each variable. Figure 2 shows the difference in distribution among the classes between the complete observations and the recently classified ones. Our multinomial models produced similar distributions for education and marital status but somewhat different for income bracket. It was decided that the distribution for the unknown observations is simply different than the rest of the complete dataset.

Figure 3

Variable	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Range
Customer_Age	26.000	41.000	46.000	46.356	52.000	73.000	47.000
Dependent_count	0.000	1.000	2.000	2.343	3.000	5.000	5.000
Months_on_book	13.000	32.000	36.000	35.976	40.000	56.000	43.000
Total_Relationship_Count	1.000	3.000	4.000	3.814	5.000	6.000	5.000
Months_Inactive_12_mon	0.000	2.000	2.000	2.342	3.000	6.000	6.000
Contacts_Count_12_mon	0.000	2.000	2.000	2.458	3.000	6.000	6.000
Credit_Limit	1438.300	2543.500	4504.000	8591.435	10994.500	34516.000	33077.700
Total_Revolving_Bal	0.000	430.000	1279.000	1165.736	1785.000	2517.000	2517.000
Avg_Open_To_Buy	3.000	1299.500	3423.000	7425.699	9815.000	34516.000	34513.000
Total_Amt_Chng_Q4_Q1	0.000	0.630	0.736	0.760	0.858	3.397	3.397
Total_Trans_Amt	510.000	2143.000	3891.000	4398.440	4740.000	18484.000	17974.000
Total_Trans_Ct	10.000	45.000	67.000	64.798	81.000	139.000	129.000
Total_Ct_Chng_Q4_Q1	0.000	0.581	0.700	0.712	0.818	3.714	3.714
Avg_Utilization_Ratio	0.000	0.025	0.178	0.277	0.506	0.999	0.999

In an effort to identify potential outliers, we produced a 5-number summary (including mean and range) for the numeric variables in the dataset [Figure 3]. The means and medians appear not to deviate significantly for most of the continuous variables except for credit limit, average open to buy, and total transaction amount. Furthermore, each of these variables have maximum values that are three times greater than their corresponding upper quartile values. However, these outliers rightfully exist in real life as credit customers with larger credit limits and spending habits are in the minority. Lastly, the aforementioned monetary variables (as well as total revolving balance) hint at possible scaling issues that may have to be addressed later as our analysis evolves. The 5-number summary supports this claim as these variables have significantly larger ranges than the other numerical variables.

Our exploratory analysis continues with the graphical distributions of the numerical variables included in this dataset. Most of them were normally distributed or not too far off. However, the following four variables were noticeably non-normal and a couple were more easily contextualized than others. Credit limit and average utilization ratio were both significantly skewed right for related reasons. Credit limit is ultimately a measure of one's ability to pay off debts while average utilization ratio is a function of credit limit and distributions of the aforementioned ability and ratio are skewed right in real life. Total Revolving Balance [Figure 4] is primarily multi-modal in nature. One mode occurs at closer $0 while the other mode occurs at $1500. We believe that the first peak represents credit customers that pay their bills on time for the most part while the 2nd peak represents a more modest distribution of credit customers that have become accustomed to accruing interest. Finally, total months of inactivity within a 12-month period steadily rises until the 3-month peak and then drops dramatically.

Figure 4

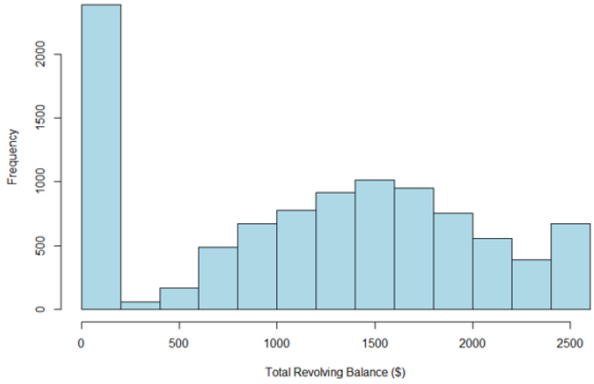
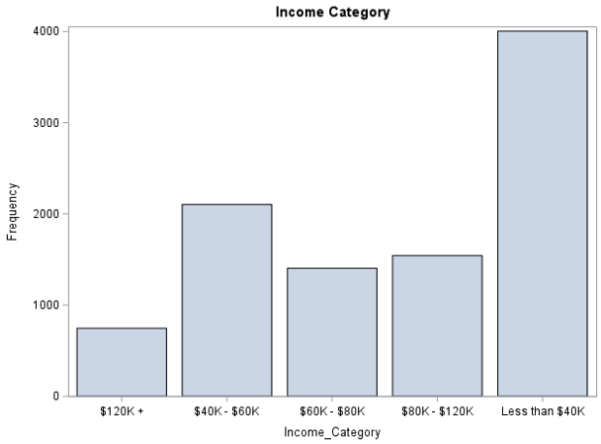
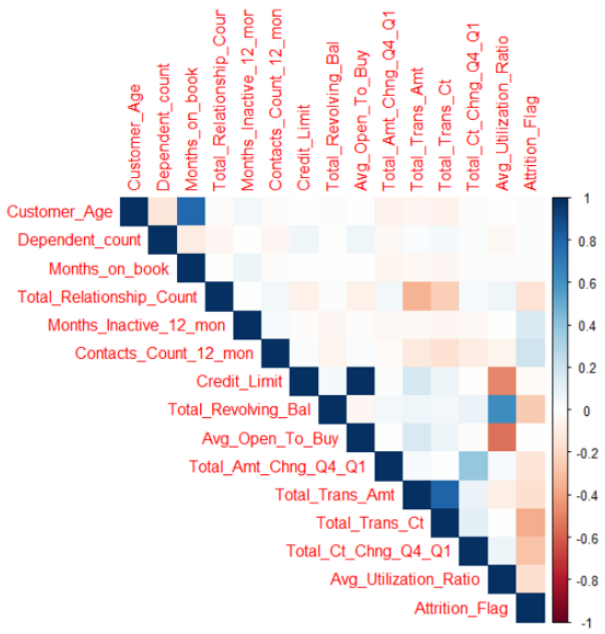


Figure 5



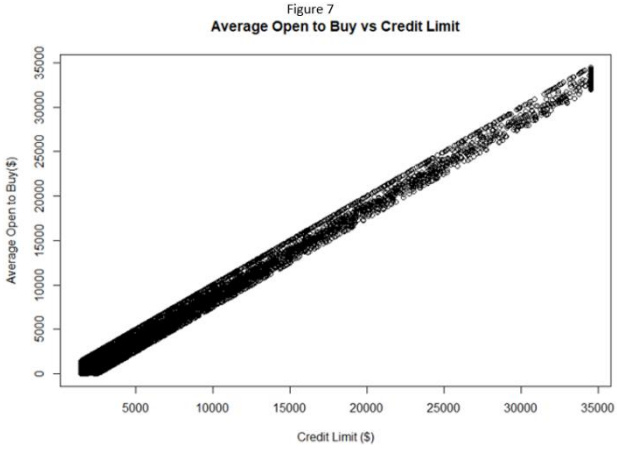
Investigation of the categorical variables did not yield as much engaging information. University graduates, high school graduates, and uneducated participants made up most of the dataset. There are more women than men in the dataset although men were 5% more likely to be an attrited customer than an existing one. The overwhelming majority of observations in the dataset were either never-married singles or married at the time of the survey. As expected, the majority of credit customers would be those of lower income [Figure 5] due to the inferred gap between average income and income required to live an adequate life (forcing people to live above their means via credit). Furthermore, lower income observations made up a greater proportion of attrited customers. Finally, most customers in the dataset were in possession of blue cards as expected. Only greater combinations of high income and credit history can permit the possession of credit cards with more flexibility. The distribution was not affected by attrition status.

Figure 6



As customary with statistical analysis, a categorical dependent variable cannot reliably be involved in an investigation of correlation or covariance with its independent variables. That being said, we formed a colorful correlation matrix [Figure 6] that consisted of the numeric independent variables involved in this dataset. Figure 6's colorful correlation matrix revealed one key indicator of multicollinearity: credit limit vs average open to buy with a Pearson correlation coefficient of 0.996. A closer look at this particular scatterplot [Figure 7] confirms our suspicion. The points in this plot form a thick, straight diagonal line. The other independent variables with high correlation coefficients do not have accompanying scatter plots that exhibit as strong of a linear relationship.

Figure 7



Prior to a deep dive into model selection, an initial full OLS model was conducted and studied. Our analysis features a categorical dependent variable along with numeric independent variables and various other categorical independent variables that were converted to dummy variables for a reliable regression to occur. The following independent variables were deemed categorical: gender, card category, education level, marital status, and income category. The dependent variable, attrition flag, seeks to determine the odds of customer churn as opposed to being retained. Figure 8 displays the result of this initial binomial logistic regression. The model resulted in an AIC value of 4653.5 and a net deviance (null deviance - residual deviance) of 4028.7. It is said that a positive net deviance indicates improved fit of a model with independent variables involved. In addition, the p-values of the beta coefficients in this model indicate that there are a lot of statistically insignificant predictors included. The following independent variables were found to have p-values greater than 0.05: months on book, customer age, average utilization ratio, all the education dummy variables, the divorce dummy variable of marital status, all the income category dummy variables (except for the $120K+ one) and the platinum dummy variable for the card category. It is clear that there's still room for parsimony even though we may lose a marginal portion of predictive power in the long run. Most importantly, statistical metrics were not defined for the "average open to buy" variable because of its strong multicollinearity with credit limit (as anticipated). As a result, we decided to remove that variable from any future analysis of the data.

Figure 8

```
Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              6.058e+00  3.892e-01  15.564  < 2e-16 ***
Customer_Age            -6.902e-03  7.810e-03  -0.884  0.37687
Dependent_count          1.413e-01  3.032e-02   4.661 3.15e-06 ***
Months_on_book          -3.988e-03  7.779e-03  -0.513  0.60819
Total_Relationship_Count -4.425e-01  2.785e-02 -15.886  < 2e-16 ***
Months_Inactive_12_mon   5.107e-01  3.836e-02  13.315  < 2e-16 ***
Contacts_Count_12_mon    5.053e-01  3.677e-02  13.740  < 2e-16 ***
Credit_Limit            -1.778e-05  6.949e-06  -2.559  0.01050 *
Total_Revolving_Bal     -9.020e-04  7.244e-05 -12.453  < 2e-16 ***
Avg_Open_To_Buy                 NA         NA      NA       NA
Total_Amt_Chng_Q4_Q1    -4.236e-01  1.891e-01  -2.240  0.02512 *
Total_Trans_Amt          4.817e-04  2.326e-05  20.707  < 2e-16 ***
Total_Trans_Ct          -1.181e-01  3.778e-03 -31.246  < 2e-16 ***
Total_Ct_Chng_Q4_Q1     -2.774e+00  1.913e-01 -14.496  < 2e-16 ***
Avg_Utilization_Ratio   -2.141e-01  2.484e-01  -0.862  0.38864
ifMale                  -8.001e-01  1.387e-01  -5.768 8.00e-09 ***
HighSchool              -2.309e-02  1.223e-01  -0.189  0.85025
College                 -1.361e-01  1.479e-01  -0.920  0.35764
Graduate                -6.212e-02  1.126e-01  -0.552  0.58108
PostGraduate             2.153e-01  1.806e-01   1.192  0.23323
Doctorate                2.005e-01  1.837e-01   1.092  0.27488
Divorced                -8.866e-02  1.499e-01  -0.592  0.55414
Married                 -5.768e-01  8.226e-02  -7.012 2.36e-12 ***
Forty                   -1.956e-01  1.125e-01  -1.739  0.08198 .
Sixty                    1.849e-02  1.817e-01   0.102  0.91898
Eighty                   3.421e-01  1.851e-01   1.848  0.06461 .
One_Twenty               6.019e-01  2.127e-01   2.830  0.00465 **
Silver                   4.168e-01  1.975e-01   2.111  0.03481 *
Gold                     1.001e+00  3.553e-01   2.818  0.00483 **
Platinum                 9.695e-01  7.046e-01   1.376  0.16884
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8624.2  on 9799  degrees of freedom
Residual deviance: 4595.5  on 9771  degrees of freedom
AIC: 4653.5

Number of Fisher Scoring iterations: 6
```

Application of Analysis Techniques

The analysis is split into transformations of the numeric variables (PCA), construction of linear combinations of numeric features that separate our dependent classes of interest (LDA) and classification of our observations via regularized regressions of all remaining variables.

Principal Components Analysis

Principal Component Analysis (PCA) is a technique used for dimensionality reduction as well as revealing latent factors (latent variables). The technique consists of transforming a larger dataset into a more refined smaller dataset at the expense of reducing the number of variables. It is

often used when simplifying data, reducing noise and revealing latent factors. The typical intention of applying PCA is to ideally deduce and establish trends in the data from the newly formed components. We decided to run a scaled PCA because the numeric variables involved in our dataset had different units and ranges for their values. Using the results discovered in PCA, we then performed a portion of Principal Factor Analysis.

The first PC was found to account for 16.44% of the variance. Figure 9 provides the particular and cumulative variances for each following PC. It took 4 PCs to obtain a cumulative proportion of greater than 50% (54.76% specifically). The scree plot in Figure 10 was more constructive in determining a good number of PCs to use in our latent factory discovery. The first four PCs in the scree plot clearly have average variances of greater than 1. The fifth PC barely clears that baseline and PC6 has a very similar variance value. However, that decrease in marginal variance increase from PC4 to PC5 encouraged us to primarily consider 4 factors for our latent factor discovery. Even though they collectively account for only 54.76% of the variance, we wanted fewer factors with the intent of model parsimony. Figure 11 shows the loadings for all the PCs.

Figure 9

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.4619	1.3580	1.3375	1.1615	1.01750	0.99944	0.99248	0.93936	0.89753	0.77438	0.45824	0.44639	0.40561
Proportion of Variance	0.1644	0.1419	0.1376	0.1038	0.07964	0.07684	0.07577	0.06788	0.06197	0.04613	0.01615	0.01533	0.01266
Cumulative Proportion	0.1644	0.3063	0.4439	0.5476	0.62729	0.70412	0.77989	0.84777	0.90974	0.95586	0.97202	0.98734	1.00000

Figure 10

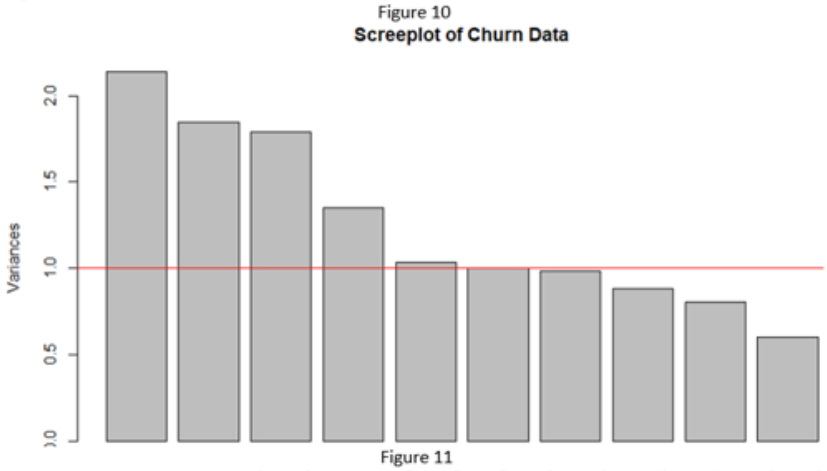


Figure 11

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Customer_Age	0.221	-0.376	0.530	-0.078	0.0648	0.051	-0.040	0.100	-0.0400	2.5e-02	-0.7005	-0.0463	0.096
Dependent_count	-0.113	0.040	-0.149	0.061	0.4968	0.575	0.277	0.520	-0.1526	1.1e-01	-0.0190	-0.0209	-0.016
Months_on_book	0.212	-0.383	0.526	-0.085	0.0631	0.056	-0.010	0.115	-0.0600	5.6e-02	0.6965	0.0557	-0.092
Total_Relationship_Count	0.322	0.138	-0.110	-0.214	0.0808	0.020	-0.067	-0.283	-0.8411	9.9e-02	-0.0143	-0.0643	-0.078
Months_Inactive_12_mon	0.082	-0.092	0.035	0.069	-0.3045	0.086	0.900	-0.261	-0.0188	-1.2e-02	-0.0143	-0.0092	0.013
Contacts_Count_12_mon	0.154	-0.073	-0.162	0.048	0.0198	-0.705	0.221	0.593	-0.1893	-8.5e-02	-0.0070	0.0140	0.024
Credit_Limit	-0.207	-0.325	-0.140	-0.208	0.6027	-0.252	0.137	-0.363	0.0823	-4.8e-02	-0.0533	0.4107	-0.176
Total_Revolving_Bal	-0.013	0.422	0.352	0.118	0.4881	-0.249	0.147	-0.198	0.0886	9.0e-06	0.0811	-0.4601	0.310
Total_Amt_Chng_Q4_Q1	-0.076	0.203	0.043	-0.651	-0.1026	-0.091	0.097	0.118	0.1778	6.7e-01	-0.0155	0.0317	0.039
Total_Trans_Amt	-0.594	-0.077	0.199	0.079	-0.0970	-0.121	0.018	0.025	-0.2104	9.0e-02	-0.0622	-0.3835	-0.606
Total_Trans_Ct	-0.570	-0.015	0.210	0.098	-0.1416	-0.034	-0.014	0.022	-0.3704	4.2e-02	0.0417	0.3640	0.572
Total_Ct_Chng_Q4_Q1	-0.131	0.209	0.138	-0.617	-0.0611	0.092	0.074	0.128	-0.0095	-7.1e-01	0.0082	-0.0208	-0.024
Avg_Utilization_Ratio	0.128	0.550	0.364	0.228	0.0092	-0.030	0.048	0.067	0.0034	1.8e-02	-0.0902	0.5732	-0.389

For the first PC, total transaction count and total transaction amount have the greatest absolute loading values of 0.5938 and 0.5695, respectively. However, both variables are actually negatively correlated with this component. We conclude that PC1 likely represents a customer that has been somewhat inactive users of this bank and credit account (at least within the last 12 months). Average utilization ratio and total revolving balance have the greatest absolute value loadings for the second PC at 0.55 and 0.422, respectively. This likely represents an underlying trend of users who use higher levels of their credit limit on a monthly basis. Customer age and 'months on book' (period of relationship with the bank) are the strongest variables for the third principal component given how their loadings only have a difference of about 0.004. This component likely represents long-term and loyal customers of the bank because older people are more likely to stick to what they know. If the loadings for both variables are found to be low, it would imply that the customer has not had a long enough relationship with that current bank and or does not appear to be a loyal customer. PC4 describes and underlying trend of credit customers that spent less and

generated less transactions ins quarter 1 relative to their quarter 4 activity due to 'Total_Ct_Chng_Q4_Q1' and 'Total_Amt_Chng_Q4_Q1' being the most dominant variables here.

Moreover, a varimax rotation was applied to the principal components to allow for further emphasis of certain variables within each PC. To confirm that the correct number of components were chosen, we generated loadings for 5 factors as well and found that the following two variables were suppressed (Contacts_Count_12_mon and Months_Inactive_12_mon). We decided that these two variables contribute mostly noise in regard to the rest of the data and do not belong to any factors or any underlying patterns in our data. There are some similarities between the results of the varimax rotation of our principal components shown in Figure 11 and the results of our principal factor analysis. The variables with the greatest loading values in the PCs matched the dominant variables across the four factors except for the 2nd PC. In PC2, customer age and months on book had greater loading values and were excluded from the 3rd factor while credit limit had the 5th highest loading value within that component and was included in the corresponding factor. This reinforces the pattern of long-term loyal customers of the bank that we see in our dataset. Other than PC2, the factors have the same significant variables as their corresponding principal components. Figure 12 summarizes this comparison.

Figure 12

PCA

	PC1	PC2	PC3	PC4
Customer_Age	0.221	-0.376	0.530	-0.078
Dependent_count	-0.113	0.040	-0.149	0.061
Months_on_book	0.212	-0.383	0.526	-0.085
Total_Relationship_Count	0.322	0.138	-0.110	-0.214
Months_Inactive_12_mon	0.082	-0.092	0.035	0.069
Contacts_Count_12_mon	0.154	-0.073	-0.162	0.048
Credit_Limit	-0.207	-0.325	-0.140	-0.208
Total_Revolving_Bal	-0.013	0.422	0.352	0.118
Total_Amt_Chng_Q4_Q1	-0.076	0.203	0.043	-0.651
Total_Trans_Amt	-0.594	-0.077	0.199	0.079
Total_Trans_Ct	-0.570	-0.015	0.210	0.098
Total_Ct_Chng_Q4_Q1	-0.131	0.209	0.138	-0.617
Avg_Utilization_Ratio	0.128	0.550	0.364	0.228

PFA

Loadings:

	RC1	RC3	RC2	RC4
Total_Relationship_Count	-0.548			
Total_Trans_Amt	0.906			
Total_Trans_Ct	0.875			
Customer_Age		0.933		
Months_on_book		0.931		
Credit_Limit			-0.575	
Total_Revolving_Bal			0.722	
Avg_Utilization_Ratio			0.947	
Total_Amt_Chng_Q4_Q1				0.811
Total_Ct_Chng_Q4_Q1				0.812
Dependent_count				
Months_Inactive_12_mon				
Contacts_Count_12_mon				

Linear Discriminant Analysis

Our dataset features a mixture of both numeric and categorical variables while our dependent variable is binary in nature. As a result, we wanted to model a way to classify our observations while considering as many variables as possible. This, of course, takes us to a higher dimensional space because we have 19 independent variables remaining in the dataset. We applied a linear discriminant analysis because it'll allow us to create a linear combination that can separate a projected version of the data into (in this case) two predetermined classes since this is a supervised dataset. The dependent variable in this technique will be 'Attrition Flag' which indicates whether or not a customer's account is churned (existing or attrited).

Prior to creating the model, the dataset was partitioned 80/20 into training and test observations so that we could observe the model's performance both on training and test data. Figure 13 represents the loadings of the single discriminant produced which essentially serve as coefficients for an equation that, when computed fully, leads to a score for each observation in the set. 'Total Count Change (QA-Q1)' has the strongest overall loading.

Figure 13

```
Coefficients of linear discriminants:
                                LD1
Customer_Age            -2.293276e-03
Dependent_count          6.151853e-02
Months_on_book           3.790687e-04
Total_Relationship_Count -2.517408e-01
Months_Inactive_12_mon   2.414033e-01
Contacts_Count_12_mon    2.312302e-01
Credit_Limit            -7.610794e-06
Total_Revolving_Bal     -5.640533e-04
Total_Amt_Chng_Q4_Q1    -3.341622e-01
Total_Trans_Amt          1.920963e-04
Total_Trans_Ct          -5.414357e-02
Total_Ct_Chng_Q4_Q1     -1.703515e+00
Avg_Utilization_Ratio    9.718315e-02
```

Figure 14

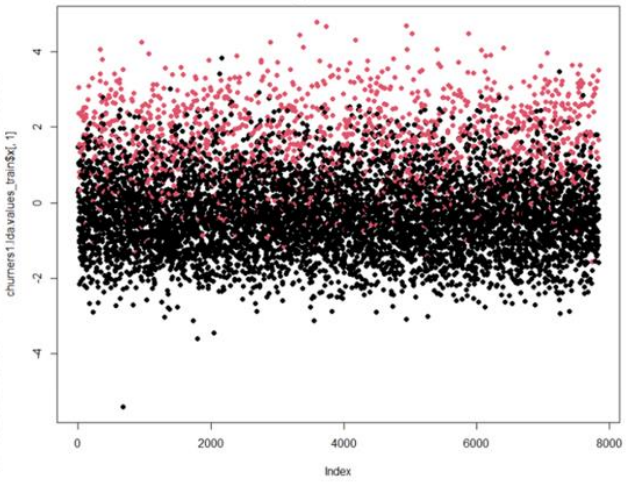


Figure 14 is a plot of the model's predictions of the training observations. There is a significant degree of overlap, but separation does exist here. As useful as graphical representations are, LDA often yields a confusion matrix to summarize classification results. Figure 15 is a hybrid confusion matrix that features raw values in addition to the accuracy values for each cell. The overall accuracy of the model on the training data was 89.9% but most importantly, the true negative rate was 73.64%. Given that this dataset has a class imbalance, and we want to limit the number of false positives (assuming existing customers are the positive class), true negative rate would be the best evaluator of our model's viability. Furthermore, we applied the model to the test set to get confirmation of the model's ability to maximize specificity (true negative rate). Once again, the scatterplot exhibited a similar performance to that of the training set. LDA appears to have done a solid job of separating our classes. Figure 16, once again, is a hybrid confusion matrix that confirms our priors. Overall accuracy was 91.33% (an improvement) and the true negative was 74.89% (also an improvement). Ultimately, LDA proved to be a useful classifier for our numeric data but it isn't the most useful technique to consider when it comes to categorical data. The assumption of multivariate normality prevents categorical variables from being accurately considered. Logistic regression would probably be better because it's based on a maximum likelihood model that doesn't require the assumption of a normal distribution.

Figure 15

LDA Training Set Confusion Matrix		
	Existing Customer	Attrited Customer
Existing Customer	6277	276
	92.4%	7.6%
Attrited Customer	516	771
	26.36%	73.64%

Figure 16

LDA Test Set Confusion Matrix		
	Existing Customers	Attrited Customers
Existing Customers	1620	57
	93.48%	6.52%
Attrited Customers	113	170
	25.11%	74.89%

Figure 17

	Training	Test
Overall Accuracy	90.2	90.10714
Specificity	74.1068	73.709

Although we did not originally run LDA predictions on multiple training and test sets, we revisited this technique and used a loop to iterate the process one hundred times, multiple times. Data points were randomly assigned to the training and test sets and the average accuracy value from the contingency tables was calculated after the loop as well as the average specificity (true negatives) from the confusion tables of every iteration. The table in Figure 17 shows the values for these metrics with an overall accuracy percentage in test sets of 90.1% and a Specificity accuracy percentage of 73.7% in the test sets.

Regularized Logistic Regression

Earlier, we used logistic regression to model the probability of a certain class occurring. Specifically, we tried to determine the likelihood that a customer will either continue to exist or soon become attrited. Furthermore, we wanted to maximize our ability to identify future churn so that we can offer promotions or programs to retain those credit card customers. Our final technique involves regularization of a logistic model in order to decrease the variance in results that can come about from an ordinary logistic regression. In an earlier preliminary analysis, we observed that a logistic regression can produce a very high overall accuracy and acceptably high accuracy in identifying those potential attrited customers. We believe that either suppressing our beta coefficients or outright eliminating certain variables from our computations can decrease the model's misclassification error and maximize the model's specificity (true negative rate).

The process can essentially be broken down into the following four analyses: ordinary logistic regression, ridge logistic regression, LASSO logistic regression, and elastic net logistic regression. Essentially, we intended to compare the outcomes of each regression and measure their performance in terms of overall accuracy, true negative rate (given that existing customers are our positive class), and deviance ratio. Overall accuracy, for the rest of this analysis, is defined as the model's ability to correctly identify both existing customers and attrited customers based on various independent variables (some continuous and other categorical). Specificity, or true negative rate, is defined as the ratio of correctly identified attrited customers to those in addition to incorrectly predicted existing customers (also known as false positives). In our analysis, false positives are worse than false negatives because it is more costly to misidentify a churn and lose out on that customer than it is overinvest in a customer that was going to retain their credit account. Deviance ratio, is a measure of the model's ability to predict the class of the dependent variable of the model, given certain predictors. Moreover, deviance ratio is explicitly defined as $1 - \frac{residual\ deviance}{null\ deviance}$. Null deviance is the measure of the model's ability to predict classes when only the intercept is involved in the logistic regression while residual deviance is similar but all the relevant independent variables and their corresponding beta coefficients are involved in the logistic regression. We always want the residual deviance to be lower than the null deviance and for the ratio between the two to be as small as possible. Finally, we subtract this ratio from 1 so that the overall deviance ratio can aimed to be maximized and approach 1. Obviously, we would like to maximize overall accuracy but for the practical sake of this analysis, specificity is the most important metric because the business cost of a false positive is much greater than that of a false negative. For each of the regressions (logistic, ridge logistic, and LASSO logistic), 100 iterations of each of them were ran on different distributions of training and test sets. The average values for the aforementioned model quality metrics were recorded for comparison. In addition, we used the largest value of lambda such that error is within 1 standard error of the minimum binomial deviance for all four regression types. The elastic net regression was a bit more involved because we had to find an optimal alpha value which would represent the weight of LASSO and ridge regularization in the regression. We ran 100 iterations for each alpha value from 0 to 1 (inclusive) with 0.1 steps (aka 0.1, 0.2, 0.3, ... 1). The optimal alpha value would then be accompanied with its corresponding average values for specificity, overall accuracy, and deviance ratio. Once again, Figure 8 (on page 4) features the output of a preliminary, ordinary logistic

model of the data. We first noticed that every beta coefficient had an absolute value of less than 2 with most of them below 1. These values indicate that the data is going to be very sensitive to any sort of regularization given that the values are already pretty small. The deviance ratio was just 46.71% and while that isn't particularly high, we'll mainly be using deviance ratio for comparison as opposed to outright maximization. While this iteration has a nice overall accuracy of 90.82%, the specificity lags a bit behind at 75.11%. It seems that this particular model was significantly better at identifying existing customers than attrited ones.

Figure 18

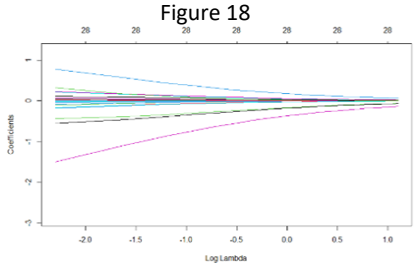


Figure 19

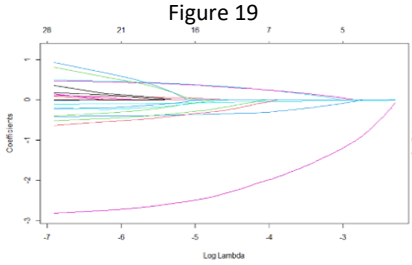
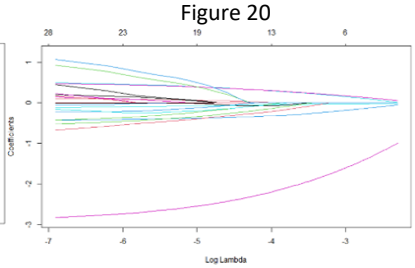


Figure 20



Preliminary models for each of the regularized regressions were executed as well. We produced some log-lambda plots comparing its variance to the number of coefficients in certain models and binomial deviance. Figures 18 and 19 are the log-lambda coefficient plots for ridge and lasso regressions, respectively and they are shaped as expected. The variables in the ridge plot all converge towards 0 at the same time but never quite get there meanwhile, some variables in the LASSO plot are eliminated earlier than others as log-lambda increases over time. Figures 20 and 21 plot log-lambda values against binomial deviance for the identification of ideal lambda values that either produce minimum binomial deviance or are within 1 standard error of the minimum. For the rest of the analysis, we'll be referring to them as lambda.min and lambda.1se, respectively. The shapes of each of these graphs provide an early indication that the elastic net version of this model may prefer a heavier weight for LASSO regression because the model can become simpler with a smaller lambda value than that of the ridge regression. Specifically, the LASSO model appears to eliminate 11 variables at its lambda.1se which is already smaller than the lambda.min for the ridge model. Figures 22 and 23 feature the log-lambda plots for the initial analysis of elastic net regression for our data. The alpha value for this initial analysis was 0.5 so that we could apply equivalent weights to ridge and LASSO techniques. Interestingly, both plots looked very similar to their LASSO equivalents, all in terms of the behavior of the coefficients, the shape of the graph log-lambda graph plotted against binomial deviance, and even the relative values for lambda.min and lambda.1se. We can say, however, that the key difference between the LASSO and elastic net coefficient charts is that the variables that do become eliminated do so at relatively greater log-lambda values which is probably a result of the ridge portion of the technique delaying the inevitable.

Figure 20

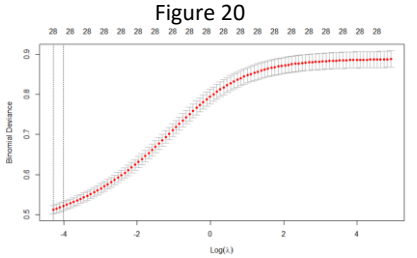


Figure 21

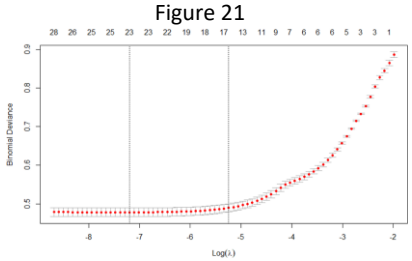


Figure 23

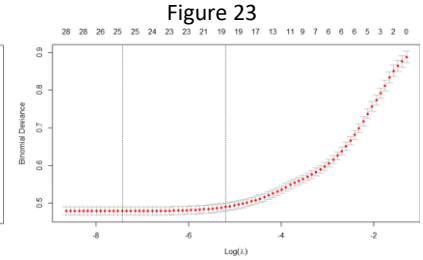


Figure 24 summarizes our metric results across the four preliminary regressions for deviance ratio, overall accuracy, and specificity. Based on this first distribution of training and test observations,

Figure 24

	Ordinary Logistic	Ridge	LASSO	Elastic Net (a = 0.5)
Deviance Ratio	46.87	41.89	45.25	45.27
Overall Accuracy	90.82	90.36	90.51	90.41
Specificity	75.11	81.55	77.95	77.72

the ordinary logistic model had the highest deviance ratio and overall accuracy while the ridge model had the highest specificity. In addition, LASSO performed similarly to elastic net which matched expectations based on the log-lambda charts. Of course, because this is based on just one distribution of test and training observations, many more variations must be analyzed before a reliable conclusion can be drawn.

In order to produce data from more iterations, we installed 4 different loops to run through ordinary logistic, ridge, and LASSO regressions 100 times each and had the elastic net complete 1100 iterations (100 for each alpha value such as 0, 0.1, 0.2….1). Figure 25 summarizes the results of those loops. The ordinary logistic model maximized both the deviance ratio and the overall accuracy although the accuracy didn't vary too greatly across the three models. However, the ridge model performed noticeably better in maximizing specificity and worse in maximizing deviance ratio. The LASSO model performed right in-between the logistic and ridge models in terms of specificity. A summary of the result of the intensive elastic net model loops are displayed in Figure 26 and the results were fascinating. The alpha value of 0 produced the greatest specificity but once any sort of LASSO weight was introduced, specificity fell by about 4% while the deviance ratio saw a similar change in the positive direction. From there on, the metrics stayed relatively stable across all different alpha values up to 1.

Figure 25

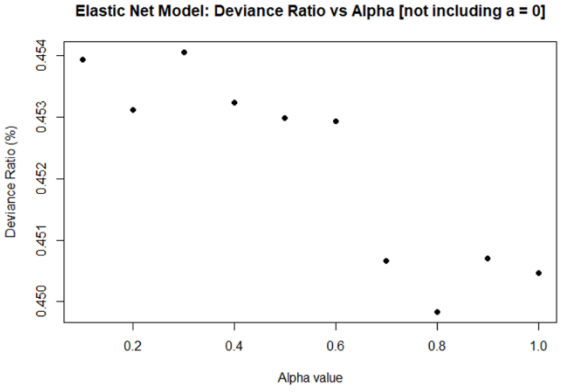
	Ordinary Logistic	Ridge	LASSO	Elastic Net (best a = 0)
Deviance Ratio	46.74	41.75	45.16	41.65
Overall Accuracy	90.20	89.87	90.03	89.79
Specificity	75.80	82.30	78.85	82.32

Figure 26

Alpha	Deviance Ratio	Accuracy	Specificity
0	41.65	89.79	82.32
0.1	45.39	90.04	78.43
0.2	45.31	89.90	78.38
0.3	45.41	90.06	78.06
0.4	45.32	89.98	78.21
0.5	45.30	90.00	78.23
0.6	45.29	90.02	78.46
0.7	45.07	90.05	78.66
0.8	44.98	90.07	78.35
0.9	45.07	90.02	78.39
1	45.05	90.13	78.36

Certain correlations and relationships that can be gathered from the table in Figure 26. Generally speaking, LASSO appeared to immediately affect all four metrics as soon as any weight was applied to it in the elastic net model while the single fully ridge data point was significantly different. We also investigated some of these relationships without the ridge model data point and Figure 27 highlights the strongest correlation. There's a clear negative correlation between deviance ratio and alpha value (when a >0) and the Pearson correlation coefficient was found to be -0.8895.

Figure 27



The results of this logistic analysis were fairly distinct. Ridge regression is the pathway if the goal is to maximize specificity (and sacrifice a bit of deviance ratio in the process). LASSO is the pathway if the goal is to break even and improve specificity while retaining a high overall accuracy and deviance ratio, relative to the ordinary logistic model. Regarding the elastic net

model, it started to behave primarily like its LASSO equivalent as soon as it wasn't fully weighted by the ridge model. We believe that's because the preferred lambda.min and lambda.1se of the LASSO model are significantly lower than that of the ridge model so any sort of weight of LASSO pushes the overall preferred lambda values in that direction. We believe that ridge was better at maximizing specificity than LASSO because identifying a minority class (attrited customers) becomes harder to do if there are less variables to work with, relative to the ridge model that retains all possible variables. However, LASSO was still better than the ordinary logistic model at maximizing specificity because it was still suppressing beta coefficients before eliminating the variables those coefficients represent. Ultimately, in a practical context, it is up to the business to prioritize certain metrics but in our opinion, the cost of false positives is much greater than that of false negatives. For that reason, we believe the ridge model is the best option.

Conclusion and Practical Significance

The resulting ridge logistic model can easily be applied to future data produced from the same firm for the sake of identifying high-risk customers in danger of churn. A limitation of that model, however, is that it retains all independent variables which makes for a more complicated model. The bank may also find it computationally and time-expensive to analyze all these metrics at recurring points throughout the year to limit customers from slipping through the cracks. The LASSO-weighted models and even PCA counter that disadvantage by producing models and datasets, respectively, that would contrast the ridge model in terms of technical expenses and opportunity cost. The LDA product is not as practical since it did not include the categorical variables in the training of its discriminant. Other supervised machine learning algorithms such as K-nearest neighbors, random forests, naïve bayes, and even support vector machines may offer better metrics for this purpose, albeit at the expense of interpretability in some cases.