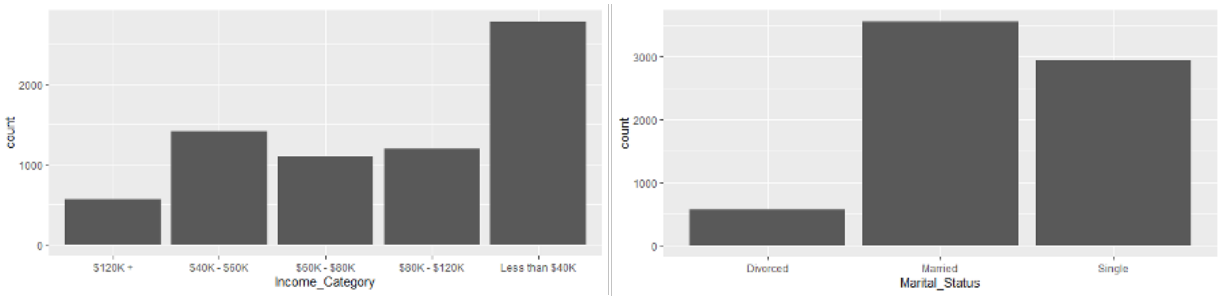
Group 17: John Sipala, Xuxi Zhu, Zhihao He
Professor Kuan
AMS 572
December 1, 2021

Credit Card Customer Attrition Data Analysis

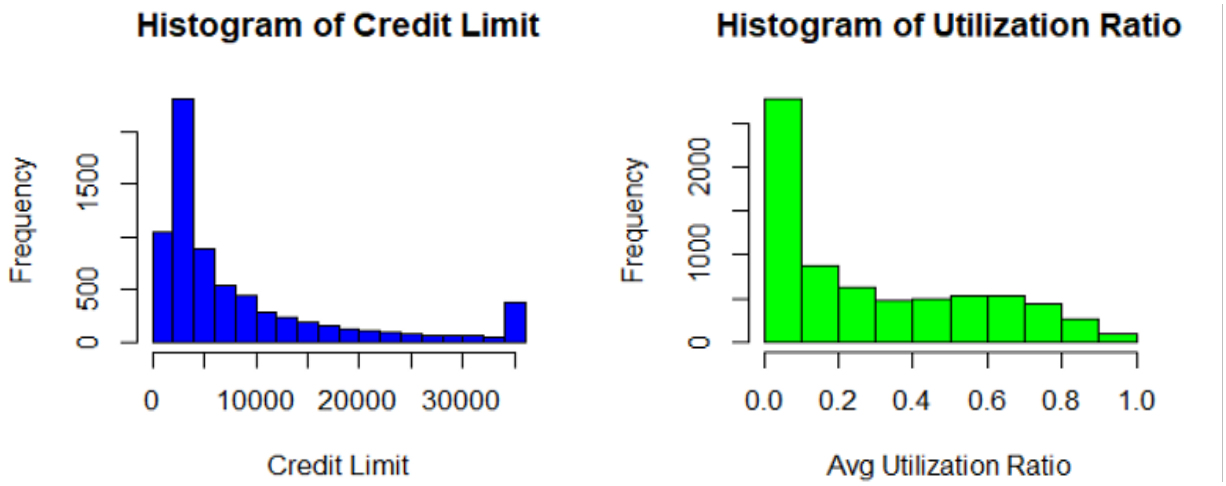
Introductory Info

The dataset we chose is based on information regarding credit card customers. The dataset came from the website *Kaggle.com*, which is traditionally known for having datasets that are conducive to Machine Learning oriented problems, but the one we chose works very well for this project as well. The original dataset, before we did any cleaning, contained over 10000 observations, each with 24 variables. Each observation (row) represents a customer and the variables consist of information about the customer. The first variable in the dataset is



"Attrition_Flag", which represents whether or not a customer is still an existing customer. Naturally, the "Attrition_Flag" variable brings up a typical classification question. Other variables include categorical variables, such as: gender, marital status, income category, and credit card type. The above bar graphs show the count of the distinct categories within the categorical variables "Income Category" and "Marital Status". There were also many numerical variables, a lot of which were relevant credit card information/statistics. For example, "months_on_book", which is the length of time in months that a customer has been doing business wih the credit card company. There was also information on total number of credit cards held by a customer, count of months where they were inactive in using the credit card as well as many others. From researching the topic of credit cards, we found out that "Credit Limit" and "Average Utilization Ratio" are both

significant attributes of a typical credit card customer. Above you see the histogram plot of the Credit Limit and Avg Utilization Ratio variables, which both seem to have a right skew.



Data Cleaning Note

As I mentioned earlier, the original data contained 10000+ observations. However, some of the categorical variables, including Education, Income Category and Marital Status contained had observations that were listed as "Unknown". So, we remove any rows that contained an unknown variable. We decided that reducing the data by removing those rows was the most practical option, leaving us with 7081 observations, which is still more than enough to work with.

It is also important to note that when dealing with some of our categorical variables, for example, income category, which has five unique sub-categories, we decided to handle those categories by assigning ordinal values from 1-5, corresponding to the lowest and highest income categories in ascending order. We used the same approach for Credit Card Type (Blue, Silver, Gold, Platinum) and Education level, which included 6 different levels of education. Some binary categories have been dealt with by substituting 1s and 0s.

The First Hypothesis

With certain categorical variables such as income category, without industry knowledge, we have no reason to believe that there would be a misleading bias in our data due to the high proportion of observations in the "Less than 40k" for example. However, with the gender category,

we believe it is logically fair to say that the age distribution of males and females should be relatively equal in our sample of 7081 observations. That is,

$$H_0: \overline{\text{Age}}_{\text{male}} = \overline{\text{Age}}_{\text{female}} \quad vs \quad H_a: \overline{\text{Age}}_{\text{male}} \neq \overline{\text{Age}}_{\text{female}}$$

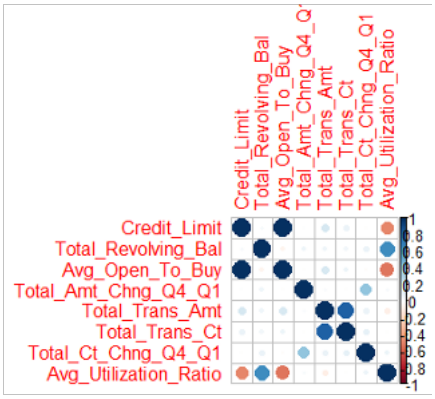
We tested this hypothesis with a two-sided t-test and revealed the results of a p value of about .37. Since this p-value is greater than .05, we fail to reject the null hypothesis and conclude that the age of Males and Females in our dataset are relatively equal. Of course, there are many combinations of variables that we could have tested, but we decided that inspecting the male and female category is important because there is a natural expecation for the two classes of male and female to be equal in various regards. If our hypothesis testing showed that they were not equal we would potentially have to consider removing some observations to avoid bias in our data.

The Second Hypothesis

As stated earlier, the variables in this dataset are oriented towards the "Attrition_Flag" category, which signifies whether or not a customer is still an existing customer. Our second hypothesis revolves around the significance of the coefficients in our model. That is,

$$H_0: \text{All } \beta_i = 0, i = 1, \ldots K \quad vs \quad H_a: \text{At least one } \beta_i \neq 0$$

We decided that a logistic regression model would be appropriate for this classification question because the result is binary ("Existing" or "Attrited" customer). Previously, in our EDA, we edited the "Attrition_Flag" variable to have 1s and 0s instead of "Existing" or "Attrited". Our process for selecting which features would go into the logistic model started by looking at the linear independence of certain features. By intuition, we decided that age and number of dependents were likely to be closely dependent. Using the "cor.test" function in R, we showed that the two features were highly correlated and linearly dependent so we dropped one of them and used this same reasoning to compare "income category" and "credit card type". The next step in our process was to evaluate the linear independence of a handful of continuous variables. Our code shows a correlation matrix that is created by a couple of nested "for" loops, which shows the statistics that we used to make our inferences on which variables to choose. The correlation matrix on the right



serves as a visual representation of this process. As a result, we removed x_9~x_{11} (i.e., Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon), x_{13}~x_{15} (i.e., Total_Revolving_Bal, Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1) and x_{17}~x_{18} (i.e., Total_Trans_Ct, Total_Ct_Chng_Q4_Q1).

Finally, using our selected features we created our first model using this line of code (fit1<-glm(Y~ x[,1]+ x[,2]+ x[,4]+ x[,6]+ x[,8]+ x[,12]+ x[,16]+ x[,19], family=binomial(link="logit")).

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.240e-01  2.632e-01  -2.371   0.0177 *
impData2[, 1] -2.856e-03  6.768e-03  -0.422   0.6731
x[, 2]         5.914e-01  1.153e-01   5.130 2.89e-07 ***
x[, 4]         4.994e-02  2.432e-02   2.054   0.0400 *
x[, 6]         9.188e-02  4.588e-02   2.002   0.0452 *
x[, 8]         2.182e-03  6.662e-03   0.328   0.7432
x[, 12]       -2.316e-05  5.151e-06  -4.496 6.94e-06 ***
x[, 16]       -2.228e-04  1.752e-05 -12.712  < 2e-16 ***
x[, 19]       -2.771e+00  1.581e-01 -17.533  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

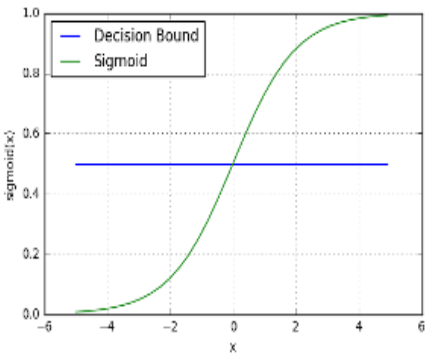
Summarizing the results of this model and looking at the p-values for each variable leads to further reduction in the number of features in our model. One of the benefits of reducing the number of features in our model is that it reduces the multicollinearity of the data. We can prove that our feature selection process was successful in regards to reducing the multicollinearity by looking at the VIF score of the variables in our model. The vif() function in R shows our results:

Variable	**x[, 2]**	**x[, 4]**	**x[, 6]**	**x[, 12]**	**x[, 16]**	**x[, 19]**
VIF	2.826574	1.000364	3.353696	1.693754	1.014766	1.126205

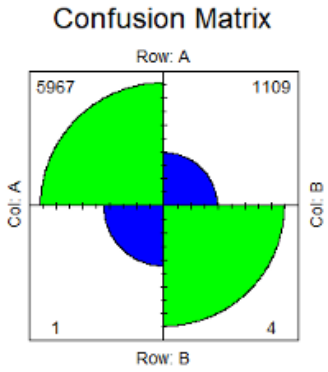
As you can see, none of the corresponding VIF scores are higher than 5, which according to our research is an important threshold when it comes to measuring the impact that a certain variable will have on the multicollinearity of the data. So, the final model is:

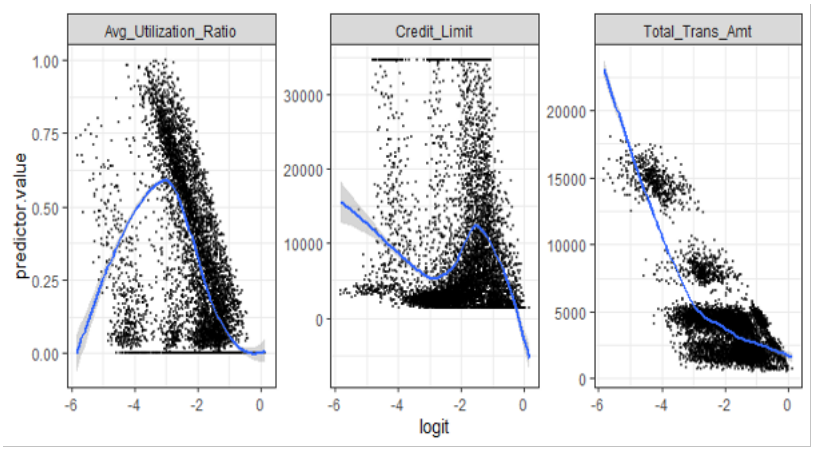
$$\text{Logit}(Y) = -0.673 + 0.59x_2 + 0.05x_4 + 0.091x_6 - 0.000023x_{12} - 0.000223x_{16} - 2.77x_{19}$$

So, the result of our hypothesis is that the coefficients in our model are significant. In addition to our hypothesis, we decided to look at the results of the prediction accuracy of our model. Unfortunately, the results of the model were not as successful as we hoped. On the right is a visual representation of how the prediction function returns a probability score between 0 and 1, and then maps that result to a discrete class (in this case, 0 or 1).



Also, on the right side here, you can see a visual representation of the confusion matrix. The results show that our model, with the exception of 5 observations, predicted the customer to be still an existing customer. Given that the dataset that we were given had about 84% existing customers, our model predicted near 100% of the customers to be existing, and thus, our model was about 84% accurate. With any logistic model, it is important to check the assumption of linearity among the continuous variables. As you can see on the left, we checked the linearity of the continuous variables in our model, and unfortunately, they are not linear, which could be a reason that our model didn't yield the results we hoped for. However, in our research we found that getting the continuous variables in our dataset to pass this linearity assumption requires processes that we would hope to include into this dataset in the future.





MCAR and MNAR

The third part of our analysis is in regards to missing data. The two types of missing data that we will focus on is MCAR (Missing Completely at Random) and MNAR (Missing Not at Random). As we mentioned earlier, the original dataset contained over 10000 observations and we got rid of any rows that contained a value labeled as "unknown". This method of filtration brought our dataset to 7081 observations. Now, one of the main reasons we chose to handle the data this way was for simplicity and practicality, especially because the unknown values were spread across multiple columns in the dataset. Taking into consideration that each column may have had to be handled differently, potentially using a separate model (such as KNN, linear regression, or random forest) to impute values as a solution to dealing with the "unknown" values would've created a lot of extra work. Also, the fact that we still had a large number of 7081 observations left to work with

was sufficient for us to move forward. Generally, a column or feature with more than 5% missing, is one that you should consider leaving out, which is another reason why our decision to reduce the data was a good idea. So, using this reduced dataset we generated our own version of missing data by using the sample function in R to create a random vector of index numbers that correspond to rows in our dataset. We chose for this vector to be length of 71 because 71 is approximately 1% of our dataset, which is less than our threshold of 5%. As shown in our code, we use those 71 indices to replace values in our dataset with NA. These randomly selected NAs were put into the "age" column. Now, one of the other columns from the dataset, which is highly correlated with the age column, is the "number of dependents" column. Using the MICE package in R which uses a distribution that is "specifically designed for each datapoint". The purpose of the MICE method is to impute values that keep the mean relatively the same without decreasing the variance. For our MNAR segment, we chose to remove only observations where the age of the customer is greater than 60. Hence, missing not at random. We purposely chose to remove older customers, which creates a bias in the data. Again, using the MICE package, we imputed the data. Here are the summary statistics of the age column (original, MCAR and MNAR).

Summary of original data

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
26.00	41.00	46.00	46.35	52.00	73.00	71

Summary of MCAR

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
21.40	41.00	46.00	46.36	52.00	73.00	0

Summary of MNAR

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
26.00	41.00	46.00	46.2	52.00	73.00	0

Of course, we also retested our first and second hypothesis using the both forms of the imputed data. Our first hypothesis had slightly higher p values for both the MCAR and MNAR data, but the result of failing to reject the null hypothesis was the same in all three cases. In regard to our second hypothesis, the p values for some of the variables were different, but again in all three cases (original, MCAR and MNAR) the result was very similar to our original findings. The results of the VIF and StepAIC functions in R were very similar as well.

Summary/ Final Notes

Our first hypothesis tests the mean of classes within the gender variable with respect to the age variable in our datset. The second hypothesis tests the significance of the coefficients in our logistic regression model. In order to get our model to produce robust results in the confusion matrix it would require us to use advanced feature engineering techniques, a lot of trial and error testing and is possibly considered to be beyond the scope of this class. The logistic regression model that we made was successful in showing that all of our features that we chose after going through our process of feature selection, were significant features. Also, in our code and in this report, shows the process of checking the linearity assumptions of the continuous variables in the model as well as checking the multicollinearity of the features that we chose. Finally, we manipulated our data to have one of the few features that is used in our model to have data that is MCAR as well as a separate datset where that same feature is MNAR. Within this report is a few notes on things that we could've done if we had more time, but overall, we have clearly mapped out the many steps of our data analysis process.