Credit Card Attrition Project

Project Background

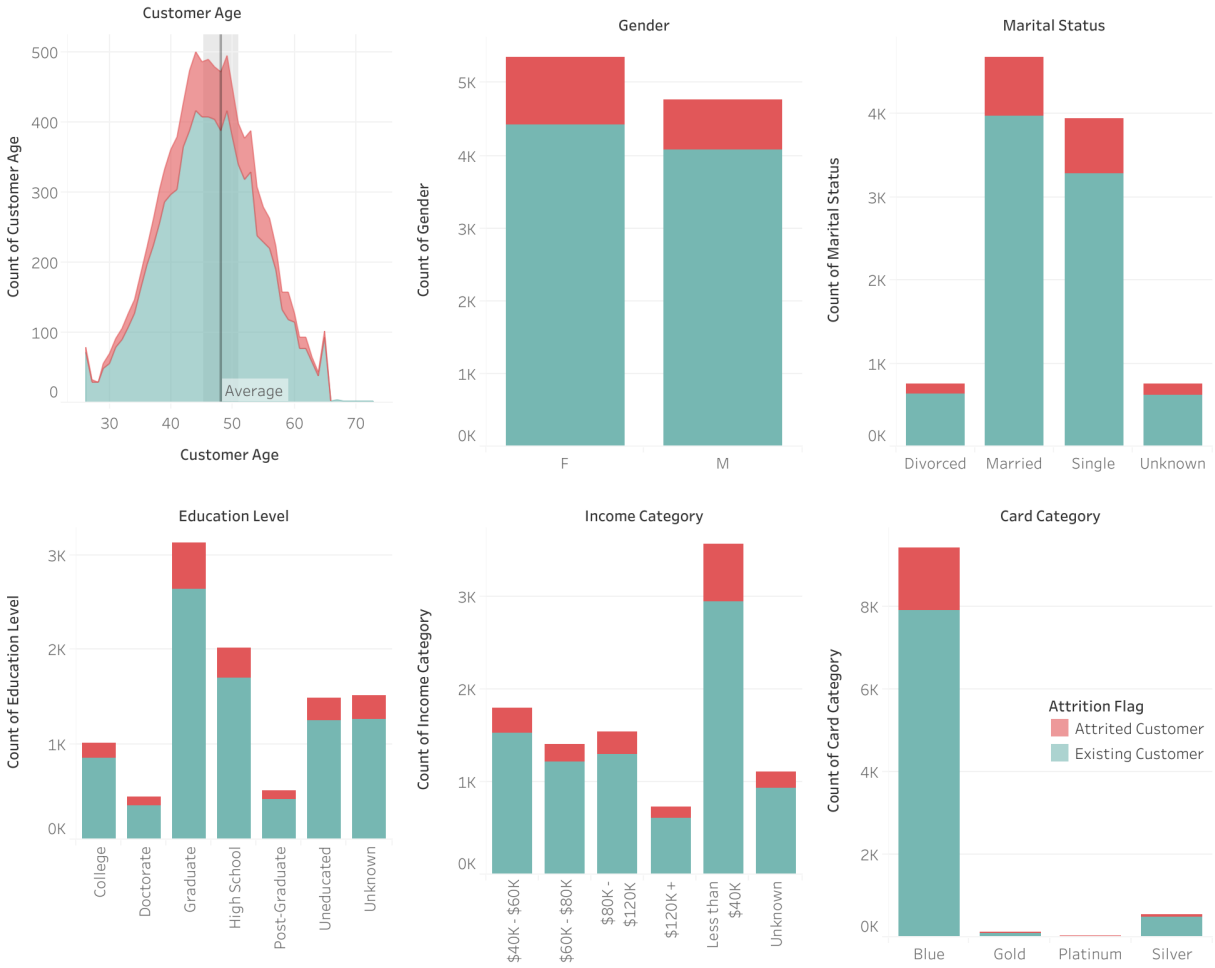
Customer attrition is defined as loss of customers, in our case, is when a customer closes their credit card services. Since the bank has an increased number of customers leaving their credit card service, our goal with this dataset is to investigate relationships between demographic, educational, and membership categories and customer attrition. Another goal this project has is for bank managers to predict possible attrited customers and allow them to build a stronger relationship with those customers to prevent eventual attrition. We will be creating and analyzing multiple linear regression models in R-studio to find the relationship of various variables on the length of a customer relationship and will also try to predict customer attrition using logistic regression models.

Dataset Background

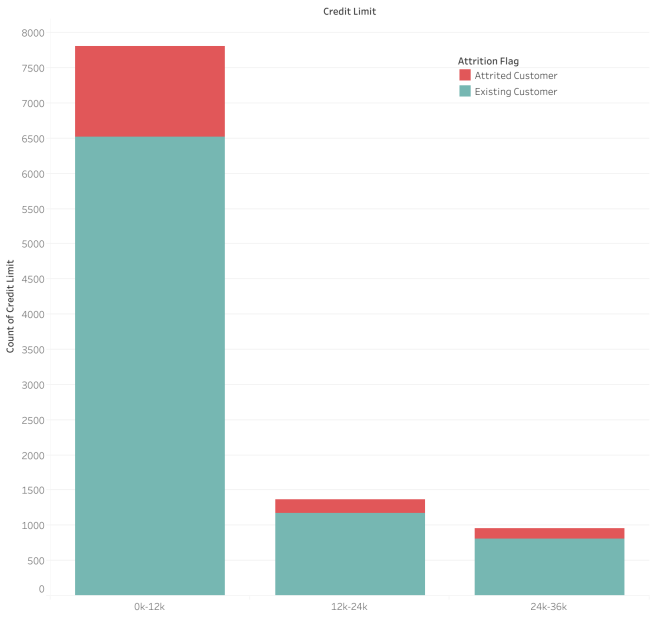
The dataset we are using is from a public database website called Kaggle.com. The dataset consists of 10,000 customers from existing to attrited customers. There are a total of 22 different variables for each customer. For our multiple linear regression models, we will use the 'Attrition_Flag' variable to subset the dataset into two data tables, existing and attrited. Our response variable will be the 'Months_on_book', which represents a customer's period of relationship with the bank. For our logistic regression model, the categorical variable 'Attrition_Flag' will be our response variable.

Predictor Variable	**Explanation**
Customer_Age	Demographic variable - Customer's Age in Years
Gender	Demographic variable - M=Male, F=Female
Marital_Status	Demographic variable - Married, Single, Divorced, Unknown
Education_Level	Educational variable - Educational Qualification of the account holder (Uneducated, Unknown, High School, College, Graduate, Post-Graduate, Doctorate)
Income_Category	Educational variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, >$120K)
Credit_Limit	Membership variable - Credit Limit on the Credit Card
Card_Category	Membership variable - Type of Card (Blue, Silver, Gold, Platinum)

We performed some data exploration on the predictor variables using Tableau.



For the `Credit_Limit` variable, the range was too wide to be shown clearly on a histogram so we transformed the continuous variable into a categorical variable, only for the data visualization, with 3 bins.



The main takeaway from the data visualizations are that there are a larger number of existing customers compared to attrited customers. Some main observations of each variable are that the average customer age of the bank is 48 years old, there are more female customers, majority of existing and attrited customers are married, hold a graduate degree, and have an annual income of less than $40,000. Most customers hold a blue card category and all five of the platinum card customers have attrited. Majority of customers have a credit limit of between $0 to $12,000.

Methodology

The primary data analysis approaches we will be using in this project are fitting linear regression models and logistic regression models on our given data. We will be answering 2 questions in this project:

1. What factors within attrited and existing customers will affect customers' months-on-book?
2. What kind of model can a bank implement to help predict what factors will contribute to current existing customers becoming attrited customers?

For the first question, we are looking at which of the categories, within the attrited customers and, separately, within the existing customers, will have an effect on the `Months_on_book` variable. We will fit the variables separately based on their category against the response variable. We will then analyze the summary statistics of the individual models, create residual plots, and measure multicollinearity.

For the second question, we hope to predict the relationship between different categories on the `Attrtion_Flag` variable with our dataset with a pseudo dataset. Since there are only two outputs in this variable, we will use logistic regression. We expect the methodology to be useful in that we can see where the bank is losing its customers in terms of the variables and what effect each category has on how long the customers stay with the bank.

```
set.seed(10)
library(readr)
library(GGally)
library(car)
library(knitr)
library(ROCR)
library(ggplot2)

path = "~/Desktop/uc davis/Classes/STA 141A/"
opts_knit$set(root.dir = path)
setwd(path)
```

```
# Credit card Dataset
bank = read.csv("bankchurners.csv", sep = ",")
bank = subset(bank, select = c(Attrition_Flag, Months_on_book, Customer_Age, Gender,
Marital_Status, Education_Level,Income_Category, Credit_Limit, Card_Category))
existing =  subset(bank, Attrition_Flag == "Existing Customer")
attrited = subset(bank, Attrition_Flag == "Attrited Customer")

# Demographic Linear Model Code
demographic_existing = subset(existing, select=c("Customer_Age", "Gender","Marital_Status","Months_on_book"))
demographic_attrited = subset(attrited, select=c("Customer_Age", "Gender","Marital_Status","Months_on_book"))

# Educational Linear Model Code
existing_educational = subset(existing, select = c(Months_on_book, Education_Level,
Income_Category))
attrited_educational = subset(attrited, select = c(Months_on_book, Education_Level,
Income_Category))

# Membership Linear Model Code
membership_existing = subset(existing, select = c(Credit_Limit, Card_Category,
Months_on_book))
membership_attrited = subset(attrited, select = c(Credit_Limit, Card_Category,
Months_on_book))
```

Part I: Multiple Linear Regression

To identify the largest relationship each of our categories has with the `Months_on_book` variable, we first created multiple linear regression models on the dependent variable `Months_on_book`.

```
# Demographic Linear Model
demographic_existing_model = lm(Months_on_book~., data=demographic_existing)
summary(demographic_existing_model)
```

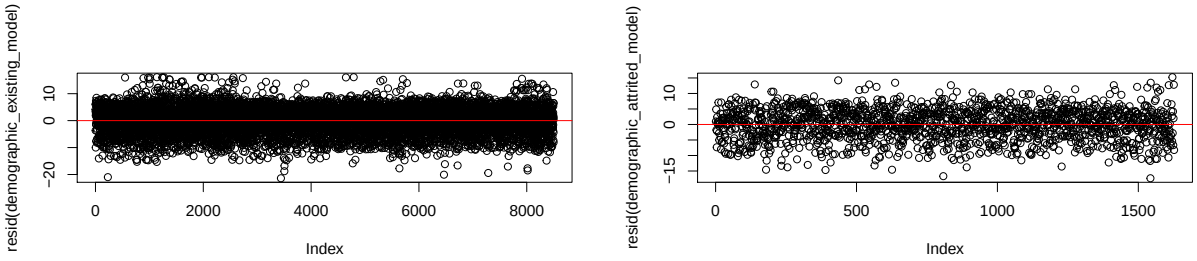
```
##
## Call:
## lm(formula = Months_on_book ~ ., data = demographic_existing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3929  -2.9601   0.2917   3.4316  16.1187
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.39321    0.35855  -1.097    0.273
## Customer_Age           0.78752    0.00658 119.688   <2e-16 ***
## GenderM                0.13518    0.10620   1.273    0.203
## Marital_StatusMarried -0.24944    0.21043  -1.185    0.236
## Marital_StatusSingle  -0.20106    0.21324  -0.943    0.346
## Marital_StatusUnknown -0.39916    0.27694  -1.441    0.150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.889 on 8494 degrees of freedom
## Multiple R-squared:  0.6288, Adjusted R-squared:  0.6286
## F-statistic:  2878 on 5 and 8494 DF,  p-value: < 2.2e-16
```

```
demographic_attrited_model = lm(Months_on_book~., data=demographic_attrited)
summary(demographic_attrited_model)
```

```
##
## Call:
## lm(formula = Months_on_book ~ ., data = demographic_attrited)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3372  -3.2259   0.3211   3.5532  15.2419
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.270319   0.883521  -0.306    0.760
## Customer_Age           0.779004   0.016239  47.970   <2e-16 ***
## GenderM               -0.004740   0.251306  -0.019    0.985
## Marital_StatusMarried -0.001012   0.493285  -0.002    0.998
## Marital_StatusSingle   0.231135   0.495641   0.466    0.641
## Marital_StatusUnknown  0.103281   0.634649   0.163    0.871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.014 on 1621 degrees of freedom
## Multiple R-squared:  0.5876, Adjusted R-squared:  0.5863
## F-statistic:   462 on 5 and 1621 DF,  p-value: < 2.2e-16
```

When analyzing the linear model results on attrited and existing customers. Demographic variables were used as predictor variables while `Months_on_book` are response variables. Our outcome results have shown that all factors of `Marital_Status` have a negative relationship with existing customers' length of relationship with the bank. For attrited customers, the `Gender` Male and `Marital_Status` Married also show a negative impact on our response variable `Months_on_book`. When analyzing demographic variables that do have a

big impact on existing and attrited customers, only age played a big impact. We found this out through the Age P value outcome of less than .05 hence the variable is statistically significant. We get R^2 is 0.628 roughly 63% of the variance found in the response variable can be explained by the predictor variables (demographics).



When looking at the residual plots for the demographic model for both existing and attrited customer data, we can see that both plots give a homoscedastic shape. This means that the sequence of the random variables has equal variances, inferring linearity to the model.

```
# Educational Linear Model
educational_existing_model = lm(Months_on_book~., data=existing_educational)
summary(educational_existing_model)
```

```
##
## Call:
## lm(formula = Months_on_book ~ ., data = existing_educational)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1783  -4.5630   0.1382   4.3872  20.7998
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    36.5470     0.4181  87.413  < 2e-16 ***
## Education_LevelDoctorate        1.6702     0.5054   3.305 0.000953 ***
## Education_LevelGraduate         0.4309     0.3148   1.369 0.171030
## Education_LevelHigh School      0.2198     0.3353   0.655 0.512182
## Education_LevelPost-Graduate   -0.2774     0.4757  -0.583 0.559803
## Education_LevelUneducated       0.3651     0.3552   1.028 0.304021
## Education_LevelUnknown          0.3537     0.3545   0.998 0.318412
## Income_Category$40K - $60K     -0.9840     0.3864  -2.547 0.010889 *
## Income_Category$60K - $80K     -1.3651     0.3998  -3.414 0.000643 ***
## Income_Category$80K - $120K    -0.8489     0.3958  -2.145 0.031991 *
## Income_CategoryLess than $40K  -1.0389     0.3587  -2.896 0.003787 **
## Income_CategoryUnknown         -1.3468     0.4200  -3.207 0.001348 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.013 on 8488 degrees of freedom
## Multiple R-squared:  0.003365,   Adjusted R-squared:  0.002074
## F-statistic: 2.606 on 11 and 8488 DF,  p-value: 0.002592
```

```
educational_attrited_model = lm(Months_on_book~., data=attrited_educational)
summary(educational_attrited_model)
```

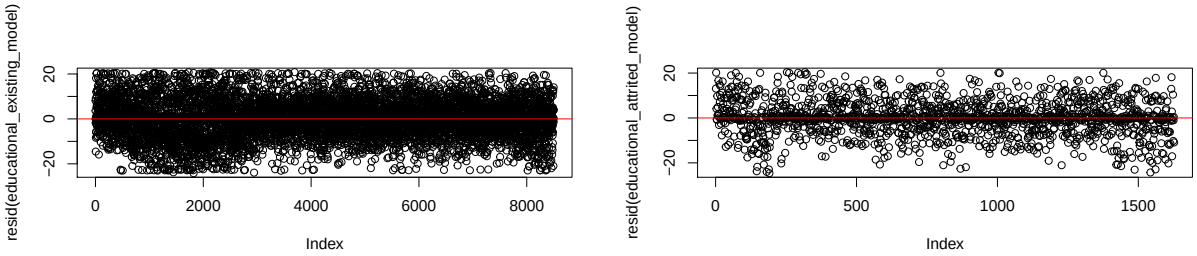
```
##
## Call:
## lm(formula = Months_on_book ~ ., data = attrited_educational)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -24.613  -3.906   0.094   4.229  20.399
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     37.0681     0.9183  40.368   <2e-16 ***
## Education_LevelDoctorate        -1.1001     1.0209  -1.078    0.281
## Education_LevelGraduate         -0.4406     0.7235  -0.609    0.543
## Education_LevelHigh School       0.7021     0.7726   0.909    0.364
## Education_LevelPost-Graduate    -0.5857     1.0296  -0.569    0.570
## Education_LevelUneducated       -0.2972     0.8085  -0.368    0.713
## Education_LevelUnknown          -0.7982     0.7975  -1.001    0.317
## Income_Category$40K - $60K      -1.0265     0.8420  -1.219    0.223
## Income_Category$60K - $80K      -0.7699     0.8987  -0.857    0.392
## Income_Category$80K - $120K     -0.4974     0.8586  -0.579    0.562
## Income_CategoryLess than $40K   -0.7215     0.7648  -0.943    0.346
## Income_CategoryUnknown          -0.1569     0.8991  -0.174    0.862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.798 on 1615 degrees of freedom
## Multiple R-squared:  0.006338,   Adjusted R-squared:  -0.0004304
## F-statistic: 0.9364 on 11 and 1615 DF,  p-value: 0.5039
```

From the summary table of the educational model of existing customers, customers with a doctorate educational level have an average of 1.6702 months more in the length of relationship with the bank than other educational levels. Since the p-value is less than 0.05, this difference is statistically significant. In the income category, $60K - $80K of annual income has an average of -1.3651 months less in Months_on_Books than other income categories Since the p-value is less than 0.05, this difference is statistically significant.

All the estimated coefficients have a p-value less than 0.05 except for Education_LevelGraduate, `Education_LevelHigh School`, `Education_LevelPost-Graduate`, `Education_LevelUneducated`, `Education_LevelUnknown`, meaning their difference is not statistically significant. We are able to see that only customers with a doctorate educational level seem to have a slightly strong positive effect on `Months_of_Books`. All of the income category variables seem to have negative effects on months on `Months_of_Books`.

The Adjusted R-squared of the educational model of existing customers is 0.002074 meaning that roughly 0.2074% of variance found in the response variable can be explained by the educational predictor variables.



Looking at the residual plots for the educational model for existing and attrited customers, we can see that both plots give a homoscedastic shape. This means that the sequence of the random variables has equal variances. There is also no pattern in the residual plots, meaning that both models follow linearity.

```
# Membership Linear Model
membership_existing_model = lm(Months_on_book~., data = membership_existing)
summary(membership_existing_model)
```

```
##
```

```
## Call:
## lm(formula = Months_on_book ~ ., data = membership_existing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0949  -4.8411   0.1324   4.1596  20.8709
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.582e+01  1.227e-01 291.840   <2e-16 ***
## Credit_Limit           1.267e-05  1.115e-05   1.136   0.2559
## Card_CategoryGold     -2.918e-01  8.593e-01  -0.340   0.7341
## Card_CategoryPlatinum  1.072e-01  2.091e+00   0.051   0.9591
## Card_CategorySilver   -8.450e-01  4.279e-01  -1.975   0.0483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.022 on 8495 degrees of freedom
## Multiple R-squared:  0.0004721,  Adjusted R-squared:  1.457e-06
## F-statistic: 1.003 on 4 and 8495 DF,  p-value: 0.4044
```

```
membership_attrited_model = lm(Months_on_book ~., data = membership_attrited)
summary(membership_attrited_model)
```

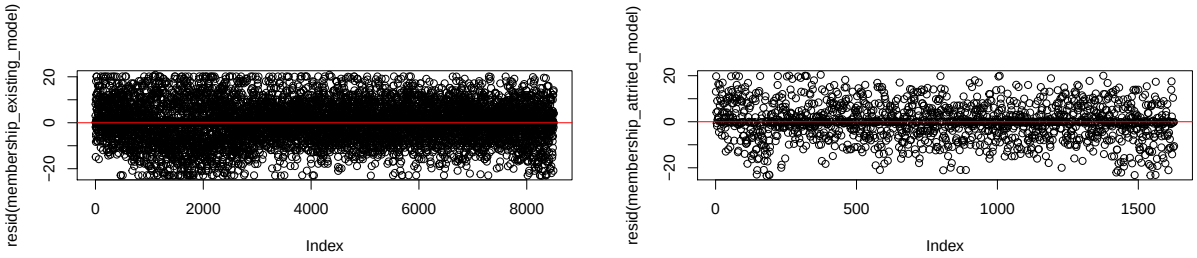
```
##
## Call:
## lm(formula = Months_on_book ~ ., data = membership_attrited)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4923  -3.9494  -0.0069   4.0260  20.3938
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.586e+01  2.643e-01 135.672   <2e-16 ***
## Credit_Limit           5.039e-05  2.535e-05   1.988   0.0470 *
## Card_CategoryGold     -3.457e+00  1.809e+00  -1.911   0.0562 .
## Card_CategoryPlatinum -1.116e+00  3.521e+00  -0.317   0.7513
## Card_CategorySilver   -7.989e-01  1.008e+00  -0.793   0.4281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.792 on 1622 degrees of freedom
## Multiple R-squared:  0.003563,   Adjusted R-squared:  0.001106
## F-statistic:  1.45 on 4 and 1622 DF,  p-value: 0.2151
```

When looking at just the existing data for the membership model, we can see that both `Credit_Limit` and Platinum `Card Category` members seem to have a positive effect on the length of relationship with the bank and all other variables have a negative impact. This tells us that the bank seems to be doing well with credit limit and Platinum card members. As the customer's credit limit increases, they tend to stay longer with the bank; and platinum members tend to remain long-time customers of the bank. From the attrited data, we can see that only the `Credit_Limit` seems to have a positive effect on `Months_on_book`.

From the summary table of the membership model of just existing customers, we can see that customers with Platinum Card Category Membership have an average of -0.845 months more in `Months_of_Books` than other card category levels. Since the p-value is less than 0.05, this difference is statistically significant in estimating `Months_of_Books`. From the summary table of the membership model of just attrited customers, we can see that customers' credit limit seems to have an average effect of 0.00005039 on the months on book variable. Since the p-value is less than 0.05, this coefficient is statistically significant. All other estimated coefficients seem not to be statistically significant.

With just existing customer data, we get an adjusted R^2 of 0.000001457, meaning 0.00001456% of variance found in the response variable can be explained by the predictor variables. With just attrited customer

data, we see that 0.1106% of the variance found in the response variable, can be explained by the predictor variables.



We can see that the residual plot gives a homoscedastic shape, meaning that the sequence of the random variables has equal variances. There is no pattern, which implies the model follows linearity.

```
# VIF Demographic
vif(demographic_existing_model)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Customer_Age    1.005368  1        1.002680
## Gender          1.000948  1        1.000474
## Marital_Status 1.005432  3        1.000903
```

```
vif(demographic_attrited_model)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Customer_Age    1.002108  1        1.001053
## Gender          1.000662  1        1.000331
## Marital_Status 1.002708  3        1.000451
```

```
# VIF Educational
vif(educational_existing_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Education_Level 1.003702  6        1.000308
## Income_Category 1.003702  5        1.000370
```

```
vif(educational_attrited_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Education_Level 1.027367  6        1.002252
## Income_Category 1.027367  5        1.002704
```

```
# VIF Membership
vif(membership_existing_model)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Credit_Limit    1.355694  1        1.164343
## Card_Category 1.355694  3        1.052027
```

```
vif(membership_attrited_model)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Credit_Limit    1.423841  1        1.193248
## Card_Category 1.423841  3        1.060662
```

For categorical variables, we will analyze its GVIF^(1/(2*Df)), the proportional change of the standard error and confidence interval of their coefficients due to the level of collinearity, to quantify the severity of multicollinearity in our multiple regression models. We follow the rule of $\text{GVIF}(1/(2\times\text{Df})) < 2$, where if the GVIF value follows, there is a low reduction in the precision of the coefficients' estimation due to collinearity.

The GVIF of the predictor variables in both attrited and existing customers for the demographic model all have a value of less than 1, meaning that there is a low reduction in the precision of the coefficients' estimation due to collinearity. The GVIF of the predictor variables in both attrited and existing customers for the educational model all have a value of less than 1, meaning that there is low collinearity. The GVIF of the predictor variables in both attrited and existing customers for the membership model all have a value greater than 1, meaning that there is low collinearity.

Part II: Logistic Regression

To predict the status of existing customers in the bank to see if the customers have a likelihood of attrition from the bank, we will first assign 0 to "Attrited Customer" and 1 to "Existing Customer" to our binary outcome variable, `Attrition_Flag`. Then, we will split our subsetted dataset for each category into training and testing data. Then, we will build a logistic regression model for each category using the training dataset with the glm() model.

```
# Demographic Logistic Model Code
demographics = subset(bank, select = c(Attrition_Flag, Customer_Age, Gender, Marital_Status))
demographics$Attrition_Flag = ifelse(demographics$Attrition_Flag == "Attrited Customer", 0, 1)
train_demographics <- demographics[1:5063,]
test_demographics <- demographics[5064:10127,]

# Educational Logistic Model Code
educational = subset(bank, select = c(Attrition_Flag, Education_Level, Income_Category))
educational$Attrition_Flag = ifelse(educational$Attrition_Flag == "Attrited Customer", 0, 1)
train_educational <- educational[1:5063,]
test_educational <- educational[5064:10127,]

# Membership Logistic Model Code
membership = subset(bank, select = c(Attrition_Flag, Credit_Limit, Card_Category))
membership$Attrition_Flag = ifelse(membership$Attrition_Flag == "Attrited Customer", 0, 1)
train_membership <- membership[1:5063,]
test_membership <- membership[5064:10127,]
```

```
# Logistic regression model for demographic category
demographics_logistic_model = glm(Attrition_Flag~., data=train_demographics, family=binomial)
summary(demographics_logistic_model)
```

```
##
## Call:
## glm(formula = Attrition_Flag ~ ., family = binomial, data = train_demographics)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3483   0.4345   0.4794   0.5221   0.6630
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              3.120594   0.296050  10.541  < 2e-16 ***
## Customer_Age            -0.021587   0.005072  -4.256 2.08e-05 ***
## GenderM                  0.218717   0.088300   2.477   0.0133 *
## Marital_StatusMarried   -0.120153   0.189726  -0.633   0.5265
## Marital_StatusSingle    -0.314312   0.191198  -1.644   0.1002
## Marital_StatusUnknown   -0.176011   0.249514  -0.705   0.4806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

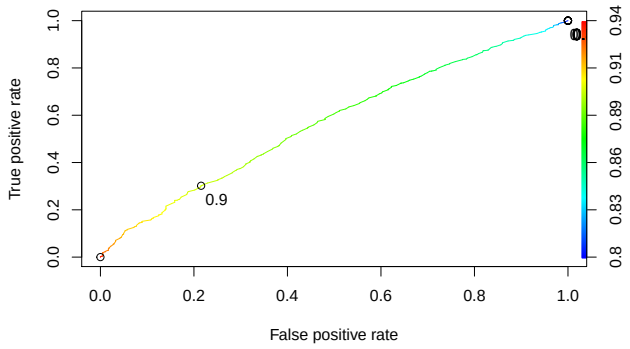
```
##     Null deviance: 3624.6  on 5062  degrees of freedom
## Residual deviance: 3594.7  on 5057  degrees of freedom
## AIC: 3606.7
##
## Number of Fisher Scoring iterations: 5
```

For the demographic logistic model, a decrease of 1 year of customer age multiplies the odds of being existing customers by 1.022. This is proven significant with a p-value of less than 0.05.

Using the predict() function, we will make predictions with the logistic models on the training dataset first with the argument type= "response", this will give us the probabilities of each customer being 1 - Existing.

```
# ROC For demographics
predictTrain = predict(demographics_logistic_model, type="response")
```

We will then convert the probabilities to predictions using a threshold value, t. If the probability of Attrition_Flag is greater than t, we predict that the customer is existing. If the probability of Attrition_Flag is less than t, then we predict that the customer is attrited. We will pick an optimum threshold value using a Receiver Operator Characteristic curve (ROC curve) by finding the best value for the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). We created the logistic regression model using our training dataset and then applied the model to our testing data set to predict its Attrition_Flag variable (attrition or existing) and then compared it with the actual test output to find its accuracy rate.



For the demographic model, we chose `t = 0.85` as the TPR is around 0.4 and the FPR is around 0.25.

```
predictTest = predict(demographics_logistic_model, type = "response", newdata = test_demographics)
table(test_demographics$Attrition_Flag, predictTest> 0.85)
```

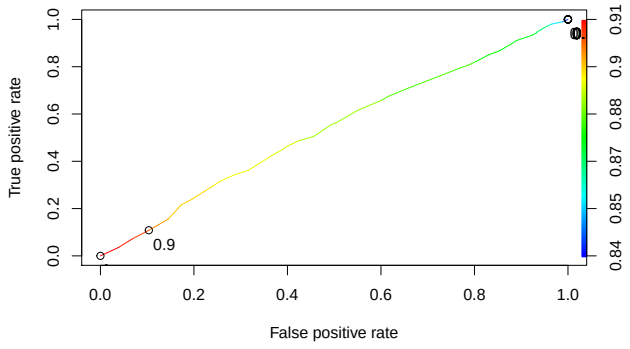
```
##
##      FALSE TRUE
##   0    114  928
##   1    468 3554
```

The model with a threshold value of 0.85 can accurately predict attrited customers with a training set accuracy of 81.8%. Accuracy = `(82+4048)/5063` = `0.816`. The model with a threshold value of 0.85 can accurately predict attrited customers with a testing set accuracy of 74.7%. Accuracy = `(114+3554)/5063`= `0.724`

```
# Logistic regression model for educational category
educational_logistic_model = glm(Attrition_Flag~.,data=train_educational, family=binomial)
summary(educational_logistic_model)
```

```
##
## Call:
## glm(formula = Attrition_Flag ~ ., family = binomial, data = train_educational)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1843   0.4708   0.4860   0.5105   0.5982
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.85552    0.19560   9.486   <2e-16 ***
## Education_LevelDoctorate      -0.22570    0.23397  -0.965    0.335
## Education_LevelGraduate        0.04148    0.15836   0.262    0.793
## Education_LevelHigh School     0.21263    0.17272   1.231    0.218
## Education_LevelPost-Graduate  -0.00947    0.23637  -0.040    0.968
## Education_LevelUneducated      0.10650    0.18014   0.591    0.554
## Education_LevelUnknown         0.05724    0.17900   0.320    0.749
## Income_Category$40K - $60K     0.12805    0.18113   0.707    0.480
## Income_Category$60K - $80K     0.22088    0.18432   1.198    0.231
## Income_Category$80K - $120K    0.18169    0.18099   1.004    0.315
## Income_CategoryLess than $40K  0.07517    0.16568   0.454    0.650
## Income_CategoryUnknown         0.02142    0.19937   0.107    0.914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3624.6  on 5062  degrees of freedom
## Residual deviance: 3617.0  on 5051  degrees of freedom
## AIC: 3641
##
## Number of Fisher Scoring iterations: 4
```

For the educational logistic model, none of the variables are associated with a significant change in the odds of customer attrition due to a high p-value.



For the educational model, we chose `t = 0.85` as the TPR is 20% and the FPR is slightly lower.

```
table(train_educational$Attrition_Flag, predictTrain > 0.85)
```

```
##
##      FALSE TRUE
##   0     21  564
##   1     91 4387
```

```
predictTest = predict(educational_logistic_model, type = "response", newdata =
test_educational)
table(test_educational$Attrition_Flag, predictTest > 0.85)
```

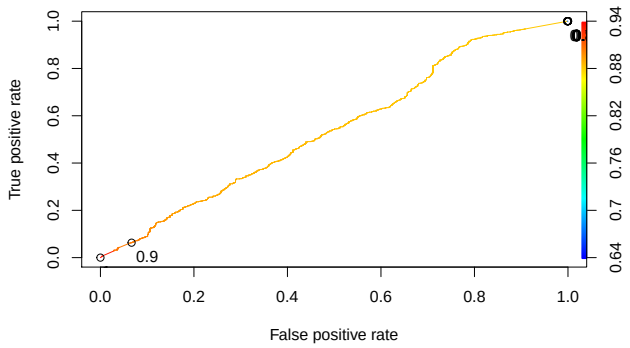
```
##
##      FALSE TRUE
##   0     45  997
##   1    108 3914
```

The model with a threshold value of 0.85 can accurately predict attrited customers with a train set accuracy of 87.06%. Accuracy = `(21+4387)/5063` = `0.8706`. Applying the training model to the test dataset to predict its Attrition_Flag variable, we found that the model with a threshold value of 0.85 can accurately predict attrited customers with a test set accuracy of 78.18%. Accuracy = `(45+3914)/5064` = `0.7818`.

```
# Logistic regression model for membership category
membership_logistic_model = glm(Attrition_Flag~., data = train_membership, family = binomial)
summary(membership_logistic_model)
```

```
##
## Call:
## glm(formula = Attrition_Flag ~ ., family = binomial, data = train_membership)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3430   0.4776   0.4997   0.5065   0.8822
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.965e+00  6.339e-02  30.997   <2e-16 ***
## Credit_Limit           7.948e-06  5.632e-06   1.411    0.158
## Card_CategoryGold      4.391e-01  7.420e-01   0.592    0.554
## Card_CategoryPlatinum -1.496e+00  1.232e+00  -1.215    0.224
## Card_CategorySilver   -3.698e-02  2.325e-01  -0.159    0.874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3624.6  on 5062  degrees of freedom
## Residual deviance: 3620.6  on 5058  degrees of freedom
## AIC: 3630.6
##
## Number of Fisher Scoring iterations: 4
```

In the membership logistic model, the variables do not seem to contribute to a significant change in the odds of customer attrition due to a high p-value for all the estimated coefficients. This is shown because using our pseudodata of 200 customers, there are only two predicted attrited customers. The membership category seems to not have an implication on whether or not the customers are leaving.



For the membership model, we chose the `t = 0.65` as the TPR is approximately 10% and the FPR is slightly lower than 10%.

```
table(train_membership$Attrition_Flag, predictTrain > 0.65)
```

```
##
##      FALSE TRUE
##   0      1  584
##   1      0 4478
```

```
predictTest = predict(membership_logistic_model, type = 'response', newdata = test_membership)
table(test_membership$Attrition_Flag, predictTest > 0.65)
```

```
##
##      FALSE TRUE
##   0      1 1041
##   1      2 4020
```

The model with a threshold value of 0.65 can accurately predict attrited customers with a train set accuracy of 88.47%. Accuracy = `(1+4478)/5063` = `0.8847`. Applying the training model to the test dataset to predict whether the customers are existing or attrited, we found that the model with a threshold value of 0.65 can accurately predict attrited customers with a test set accuracy of 79.40%. Accuracy = `(1+4020)/5064` = `0.7940`.

Psuedodata predictions for eventual attrition

Then, to apply concepts of supervised learning, we will create a pseudodata of 200 clients for each of our three models using the sample() function from the bank dataset to try to predict if the customer belongs to the attrition or existing category using our if we were hypothetically given another dataset. This data can then be used by the bank to identify possibly attriting customers and build stronger relationships with those customers to prevent eventual attrition. Our pseudodata will be separated by our three categories and we will apply the individual logistic model to predict customer IDs with 0, which means that the models predict the customer to be an "Attrited Customer".

```
# Pseudodata prediction for Demographics
pseudodata = data.frame(Customer_Age = sample(bank$Customer_Age, size=200, replace = TRUE),
                        Gender = sample(bank$Gender, size=200, replace = TRUE),
                        Marital_Status = sample(bank$Marital_Status, size=200, replace =
TRUE))
demographicsSample <- predict(demographics_logistic_model, newdata=pseudodata,type='response')
fittedDemographicsSample <-ifelse(demographicsSample > 0.85, 1, 0)
```

When looking at the Attition_flag output of the pseudo data using the demographic logistic regression model, it predicted that customers with the ID of 14, 23, 76, 97, 90, and 121 have a value of 0, meaning that they are predicted to be attrited customers. The customer IDs that are predicted to be attrited are customers whose ages are ranges between 40s more on the 60's range. With that being said, the bank can draw a conclusion that an increase in age plays a significant effect on customer attrition, which is also supported by the summary statistics seen above.

```
# Pseduodata prediction for educational category
pseudodata = data.frame(Education_Level = sample(bank$Education_Level, size=200, replace =
TRUE),Income_Category = sample(bank$Income_Category, size=200, replace = TRUE))

educationalSample <- predict(educational_logistic_model, newdata=pseudodata, type='response')
classifiedEducationalSample <- ifelse(educationalSample > 0.85, 1, 0)
```

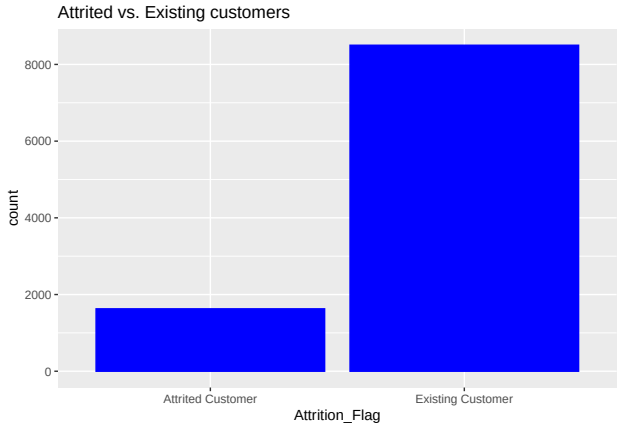
For the educational logistic regression model, customers with the ID of 45, 65, 71, 115, 174, and 198 have a value of 0. The customer IDs predicted to be attrited customers all of the predicted attrited customers have a Doctorate educational level, this may seem to play an effect in customer attrition but more analysis is required as all the estimated coefficients in the logistic model seemed to be statistically insignificant.

```
# Pseduodata prediction for membership category
pseudodata = data.frame(Credit_Limit = sample(bank$Credit_Limit, size = 200, replace = TRUE),
                        Card_Category = sample(bank$Card_Category, size = 200, replace =
TRUE))
membershipSample = predict(membership_logistic_model, newdata = pseudodata, type
= 'response')
fittedMembershipSample = ifelse(membershipSample > 0.65, 1, 0)
```

For the membership logistic regression model, customers with the ID of 48 and 153 have a value of 0. Since there are only two predicted attrited customers, the membership category seems to not have an implication on whether or not the customers are leaving, which is supported by the high p-value of all the estimated coefficients in its logistic model.

This model will help bank managers analyze the customers by looking at root cause demographic, educational, and membership variable patterns among those customers. Through analyzing and looking at indications, managers can act on what can be done to strengthen the relationships with customers to prevent eventual attrition. A suggestion for banks to build a better relationship with customers of older age or customers of doctorate educational level to prevent eventual attrition is to create a satisfaction survey. Through the survey, the bank can ask customers what they like about their current credit card and what benefits customers would desire from the bank through either promotion, lower interest rate or high cashback.

Discussion



In the bank dataset given, only 16.07% of customers are attrited customers. Thus, it's a bit difficult to train our model to predict customer attrition if the representation from "Attrited Customers" is so low compared to "Existing Customers". This explains a considerably low accuracy when predicting our testing dataset compared to our training dataset for each of our logistic models. Also, we have to consider the bias created by visually looking for patterns in the data visualizations.

To improve the models, we can focus mainly on factors that play a big contribution to why users are leaving, for example, the customer's age, educational level, etc. In conclusion, to help improve relationships with customers, we recommend the bank to implement a satisfaction survey. Through that survey, we can further analyze what factors the customers desire and what the bank can improve on.