Causes for Customer Attrition

By: Semeen Hajira, Eunice Kim, Victor Lai

Presentation Outline

- Research Question
- Dataset Background
- Data Cleaning
- Linear Regression
- Logistic Regression
- Sensitivity and Specificity
- Conclusion

Research Question

According to the bank's data, what variables are most influential in customer attrition?

Potential variables:

- Average utilization ratio
- Dependent_count
- Marital status
- Credit Limit
- Customer Age
- Education Level

About the dataset

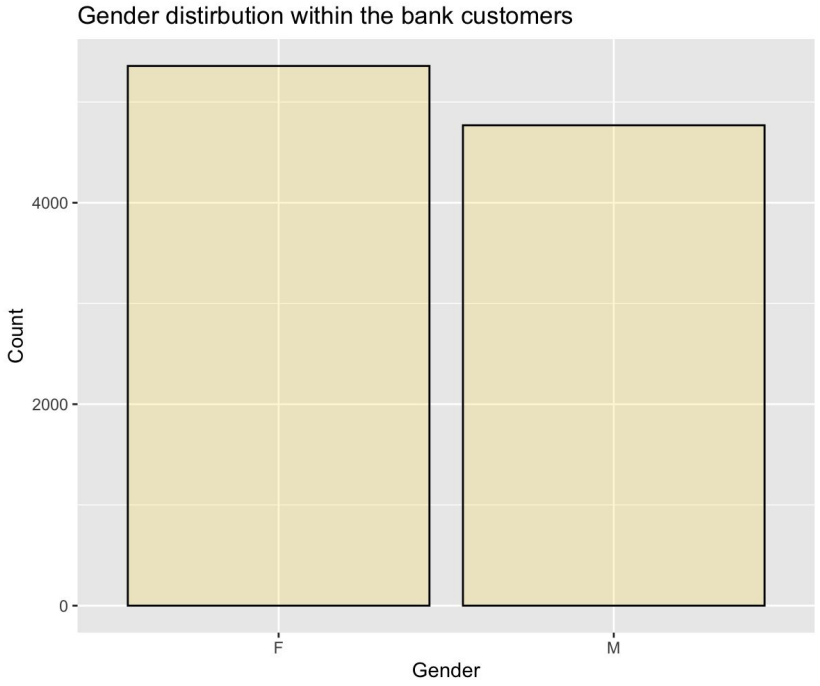
- Developed for the purpose of predicting what kind of credit card customers are leaving the bank
  - Once the variables are understood, the bank will adjust their services to better retain customers

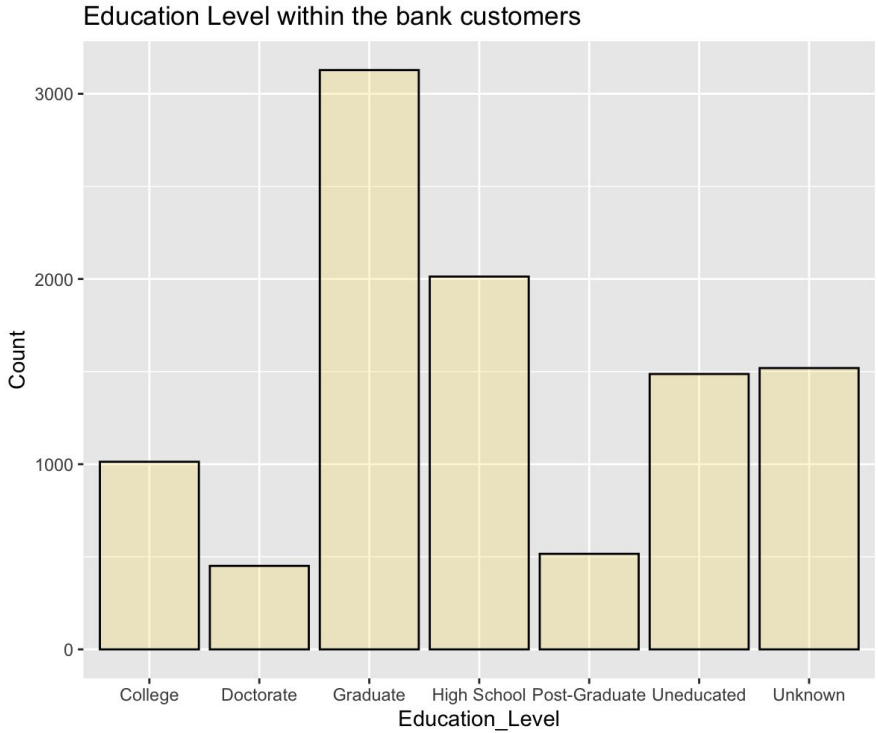
CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_
768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	5	1	3	12691	777	11914	
818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	6	1	2	8256	864	7392	
713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4	1	0	3418	0	3418	
769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3	4	1	3313	2517	796	
709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	5	1	0	4716	0	4716	
713061558	Existing Customer	44	M	2	Graduate	Married	$40K - $60K	Blue	36	3	1	2	4010	1247	2763	
810347208	Existing Customer	51	M	4	Unknown	Married	$120K +	Gold	46	6	1	3	34516	2264	32252	
818906208	Existing Customer	32	M	0	High School	Unknown	$60K - $80K	Silver	27	2	2	2	29081	1396	27685	
710930508	Existing Customer	37	M	3	Uneducated	Single	$60K - $80K	Blue	36	5	2	0	22352	2517	19835	
719661558	Existing Customer	48	M	2	Graduate	Single	$80K - $120K	Blue	36	6	3	3	11656	1677	9979	
708790833	Existing Customer	42	M	5	Uneducated	Unknown	$120K +	Blue	31	5	3	2	6748	1467	5281	
710821833	Existing Customer	65	M	1	Unknown	Married	$40K - $60K	Blue	54	6	2	3	9095	1587	7508	
710599683	Existing Customer	56	M	1	College	Single	$80K - $120K	Blue	36	3	6	0	11751	0	11751	
816082233	Existing Customer	35	M	3	Graduate	Unknown	$60K - $80K	Blue	30	5	1	3	8547	1666	6881	
712396908	Existing Customer	57	F	2	Graduate	Married	Less than $40K	Blue	48	5	2	2	2436	680	1756	
714885258	Existing Customer	44	M	4	Unknown	Unknown	$80K - $120K	Blue	37	5	1	2	4234	972	3262	
709967358	Existing Customer	48	M	4	Post-Graduate	Single	$80K - $120K	Blue	36	6	2	3	30367	2362	28005	
753327333	Existing Customer	41	M	3	Unknown	Married	$80K - $120K	Blue	34	4	4	1	13535	1291	12244	
806160108	Existing Customer	61	M	1	High School	Married	$40K - $60K	Blue	56	2	2	3	3193	2517	676	
709327383	Existing Customer	45	F	2	Graduate	Married	Unknown	Blue	37	6	1	2	14470	1157	13313	
806165208	Existing Customer	47	M	1	Doctorate	Divorced	$60K - $80K	Blue	42	5	2	0	20979	1800	19179	

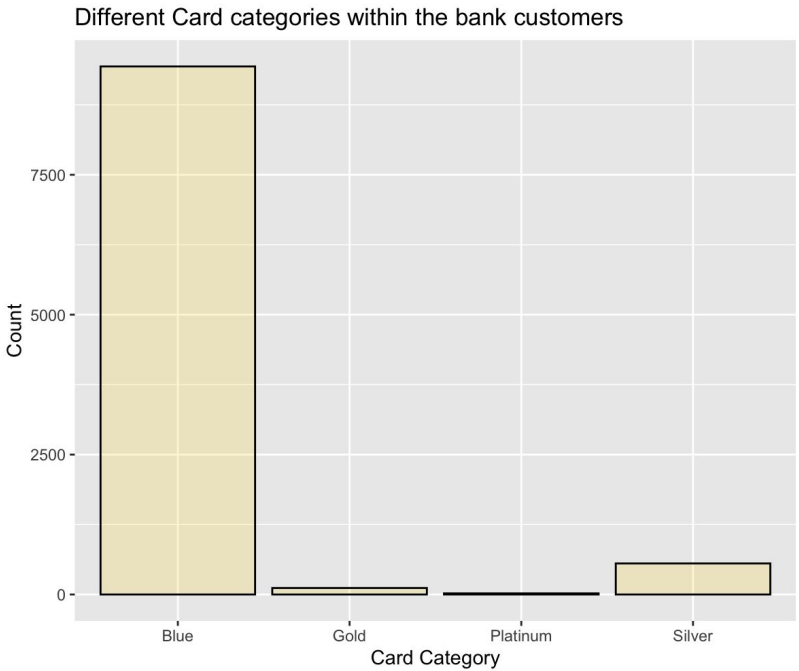
About the dataset - Key Variables

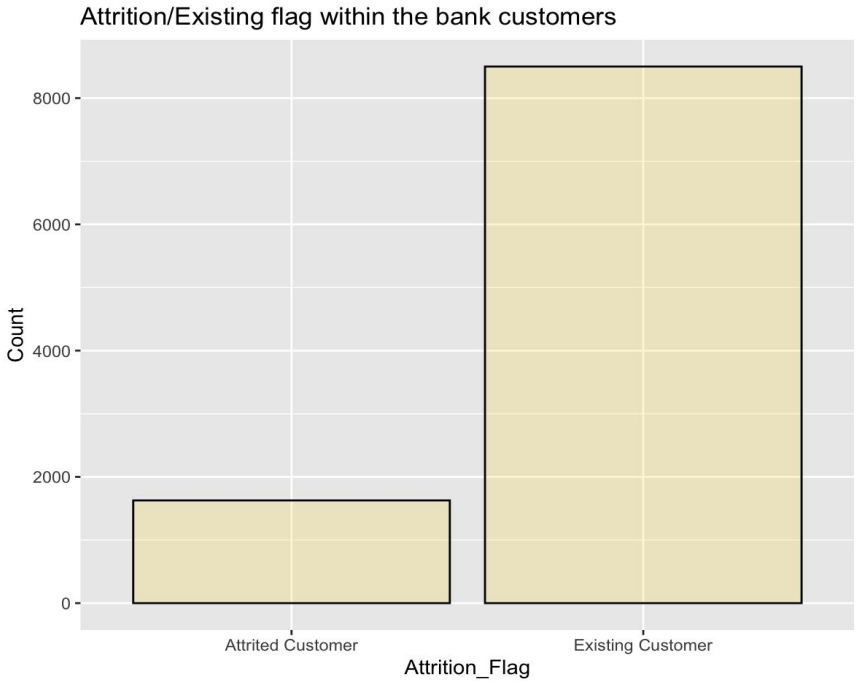
Data description is as below:

Variable	Type	Description
Clientnum	Num	Client number. Unique identifier for the customer holding the account
Attrition_Flag	char	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Num	Demographic variable - Customer's Age in Years
Gender	Char	Demographic variable - M=Male, F=Female
Dependent_count	Num	Demographic variable - Number of dependents
Education_Level	Char	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status	Char	Demographic variable - Married, Single, Unknown
Income_Category	Char	Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown)
Card_Category	Char	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Num	Months on book (Time of Relationship)
Total_Relationship_Count	Num	Total no. of products held by the customer
Months_Inactive_12_mon	Num	No. of months inactive in the last 12 months
Contacts_Count_12_mon	Num	No. of Contacts in the last 12 months
Credit_Limit	Num	Credit Limit on the Credit Card
Total_Revolving_Bal	Num	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Num	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Num	Charge in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Num	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Num	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Num	Charge in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Num	Average Card Utilization Ratio


Gender distirbution within the bank customers
Count
4000
2000
0
F
M
Gender


Education Level within the bank customers
Count
3000
2000
1000
0
College
Doctorate
Graduate
High School
Post-Graduate
Uneducated
Unknown
Education_Level


Different Card categories within the bank customers
Count
7500
5000
2500
0
Blue
Gold
Platinum
Silver
Card Category


Attrition/Existing flag within the bank customers
Count
8000
6000
4000
2000
0
Attrited Customer
Existing Customer
Attrition_Flag

Data Cleaning

Specific data changed into binary values

- Gender
- Marital Status

Remove irrelevant data

- Client Number
- Naive Bayes Classifier

Handle Missing Data

- Removed null values

Linear Regression

$$Avg\_Utilization\_Ratio = b_0 + b_1 * Credit\_Limit + b_2 * Months\_on\_book + b_3 * Total\_Revolving\_Bal + b_4 * Total\_Trans\_Amt$$

- Assumption: Average Utilization Ratio can be used to indicate a customer's likelihood of attrition

- Dependent Variable: Average Utilization Ratio
- Independent Variables: Credit Limit, Months on Book, Total Revolving Balance, Total Transaction Amount in the last 12 months

Linear Regression

R-Squared	Adjusted R-Squared
0.651	0.6508

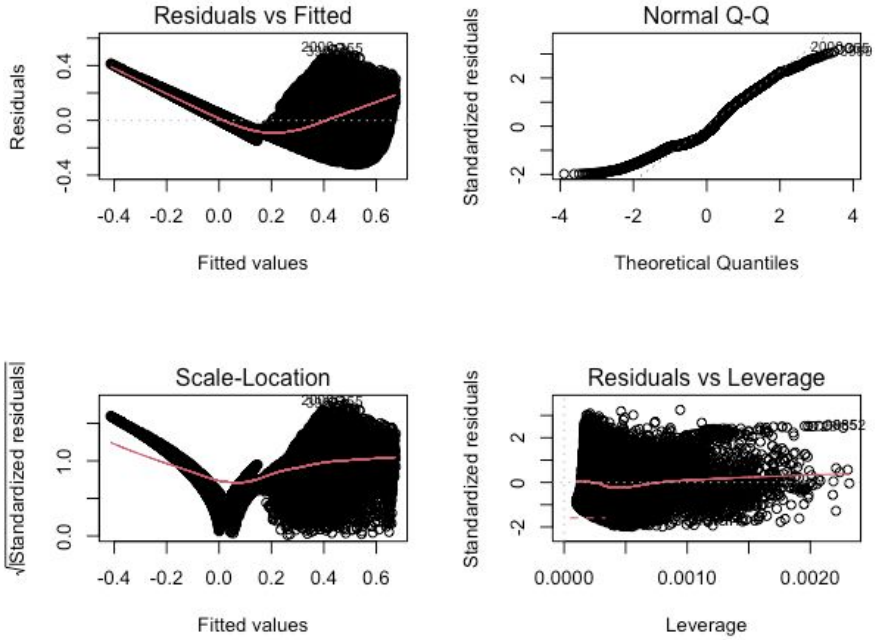
	Estimate	Std. Error	t value	Pr>|t|
(Intercept)	1.792e-01	8.172e-03	21.924	< 2e-16
Credit Limit	-1.528e-05	1.809e-07	-84.466	< 2e-16
Months on Book	-3.741e-04	2.029e-04	-1.844	0.0652
Total Revolving Balance	2.192e-04	1.992e-06	110.062	< 2e-16
Total Transaction Amount (Last 12 months)	-3.137e-06	4.850e-07	-6.468	1.04e-10

Linear Regression: Interpretation & Analysis

- As credit limit increases by $1.00, average utilization rate will decrease by 0.00001528, keeping all else constant
- As total revolving balance increases by $1.00, average utilization rate will increase by 0.0002192, keeping all else constant
- As total transaction amount over the last twelve months increases by $1.00, average utilization rate will decrease by 0.000003137, keeping all else constant

- All of the independent variables decreased the average utilization rate, except total revolving balance
  - Aligns with expected result

Linear Regression



- Residuals vs. fitted plot indicates heteroscedasticity
- Slightly tailed Q-Q plot suggests there could be non-linearity

- Nonlinear transformations need to be considered to achieve a more normal distribution and a more constant variance

Model 1: Linear-Linear Model

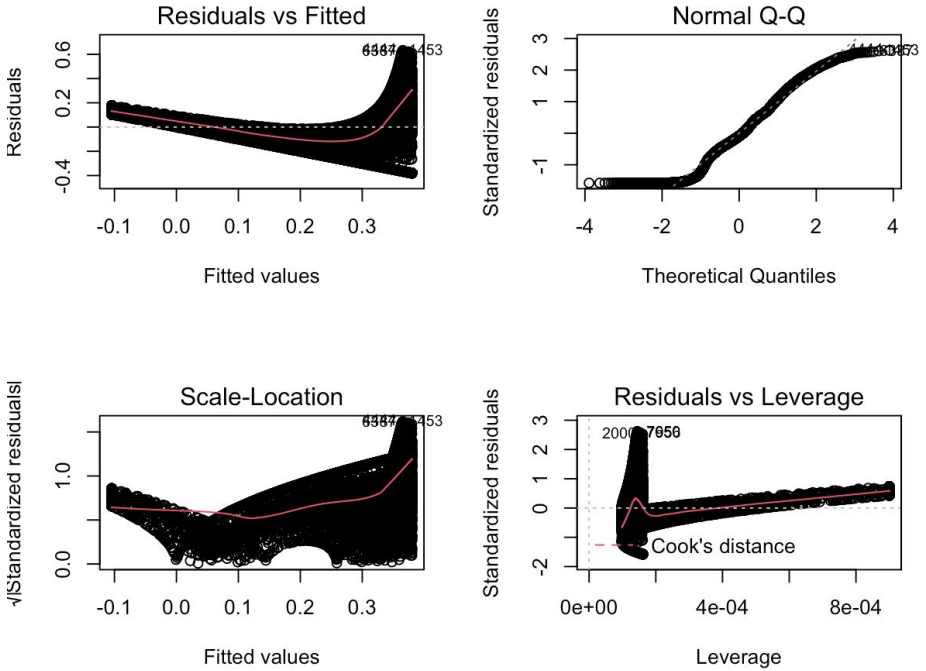
	Estimate	Std. Error	t value	Pr>|t|
Intercept	4.014e-01	3.309e-03	121.3	<2e-16
Credit Limit	-1.465e-05	2.640e-07	-55.5	<2e-16

R-Squared	Adjusted R-Squared
0.2333	0.2332

Model 1 : *Average_Utilization_Ratio = b0 + b1*Credit_Limit*

- As credit limit increases by $1.00, average utilization ratio decreases by 1.465e-05 units, holding all other factors constant

Model 1: Linear-Linear Model



Model 2: Linear-Log Model

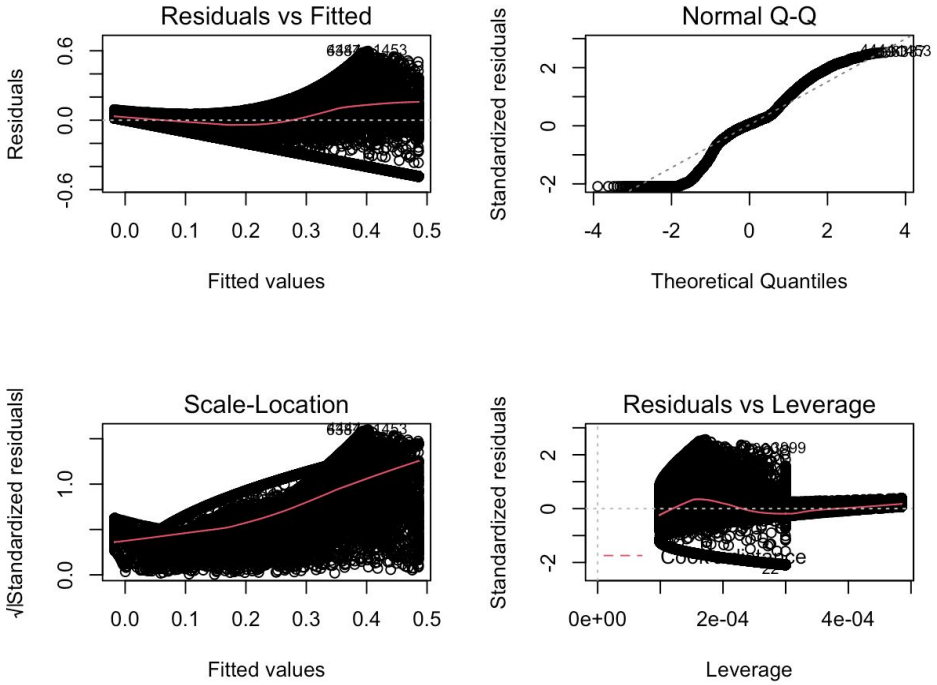
	Estimate	Std. Error	t value	Pr>|t|
Intercept	1.636603	0.021435	76.35	<2e-16
log(Credit Limit)	-0.158276	0.002477	-63.90	<2e-16

R-Squared	Adjusted R-Squared
0.2874	0.2873

Model 2 : *Average_Utilization_Ratio = b0 + b1*log(Credit_Limit)*

- As credit limit increases by 1%, average utilization ratio decreases by 0.00158276 units, holding all other factors constant

Model 2: Linear-Log Model



Model 3: Log-Linear Model

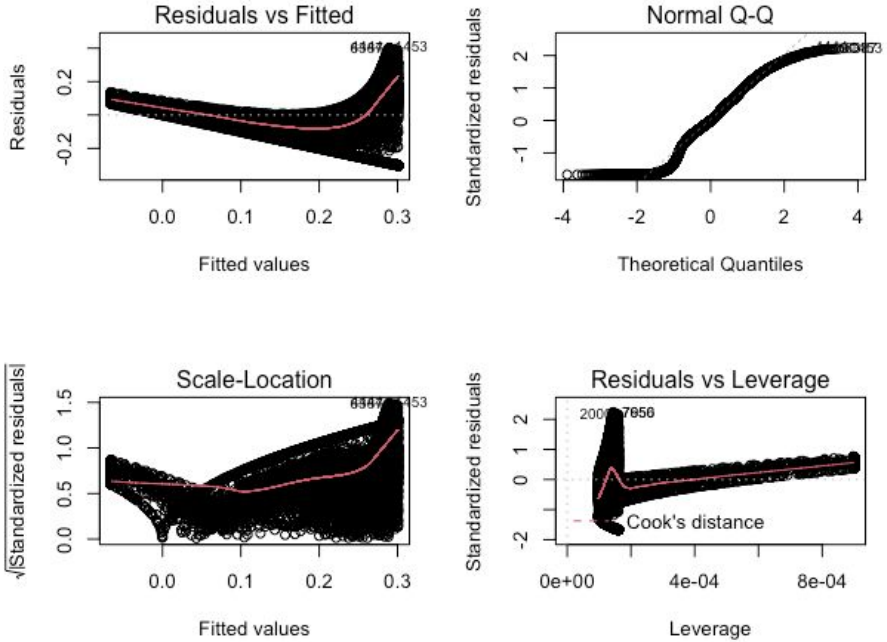
	Estimate	Std. Error	t value	Pr>|t|
Intercept	3.164e-01	2.469e-03	128.19	<2e-16
Credit Limit	-1.107e-05	1.969e-07	-56.19	<2e-16

R-Squared	Adjusted R-Squared
0.2377	0.2376

Model 3 : log(*Average_Utilization_Ratio) = b0 + b1*Credit_Limit*

- As credit limit increases by $1.00, average utilization ratio decreases by approximately 0.001107 units, holding all other factors constant

Model 3: Log-Linear Model



Model 4: Log-Log Model

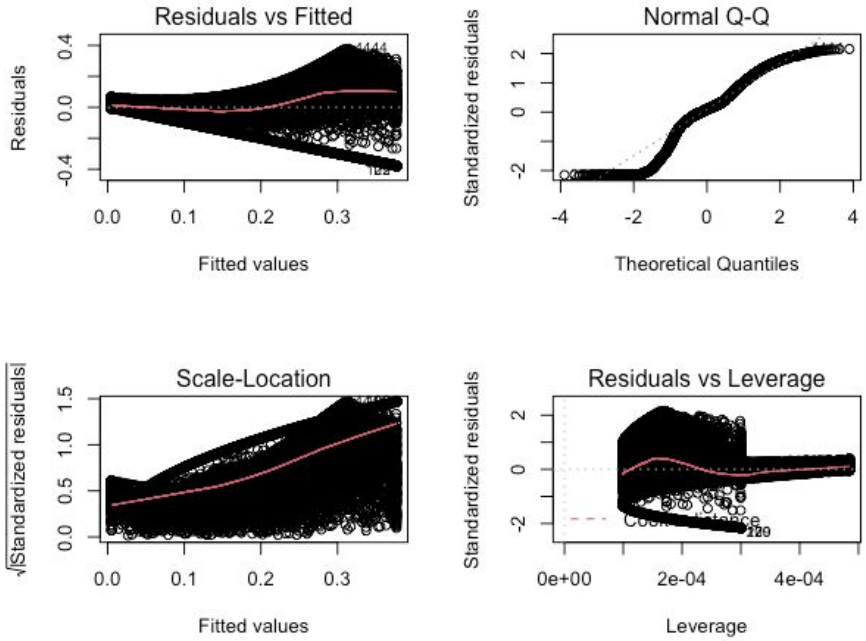
	Estimate	Std. Error	t value	Pr>|t|
Intercept	1.227279	0.016119	76.14	<2e-16
log(Credit Limit)	-0.116972	0.001863	-62.80	<2e-16

R-Squared	Adjusted R-Squared
0.2803	0.2803

Model 4 : log(*Average_Utilization_Ratio) = b0 + b1*log(Credit_Limit)*

- As credit limit increases by 1%, average utilization ratio decreases by 0.116972%, holding all other factors constant

Model 4: Log-Log Model



Comparing Nonlinear Models

	R-Squared	Adjusted R-Squared
Model 1: Linear-Linear	0.2333	0.2332
Model 2: Linear-Log	0.2874	0.2873
Model 3: Log-Linear	0.2377	0.2376
Model 4: Log-Log	0.2803	0.2803

Model 2 showed the largest relative R-squared value, but was **not strong enough** to provide a clear conclusion

Logistic Regression: Creating Dummy Variables

We would like to understand what factors affect the chances of a particular customer leaving the bank.

-Income category

Binary values (1/0) for different kinds of incomes with $120K+ as the base value.

-Card Category

Binary values (1/0) for different types of cards(Blue, Gold, Platinum) with silver as the base value.

-Attrition Flag

Binary values (1/0) - 1 for attrited customers and 0 for existing customers.

Logistic Regression: Results

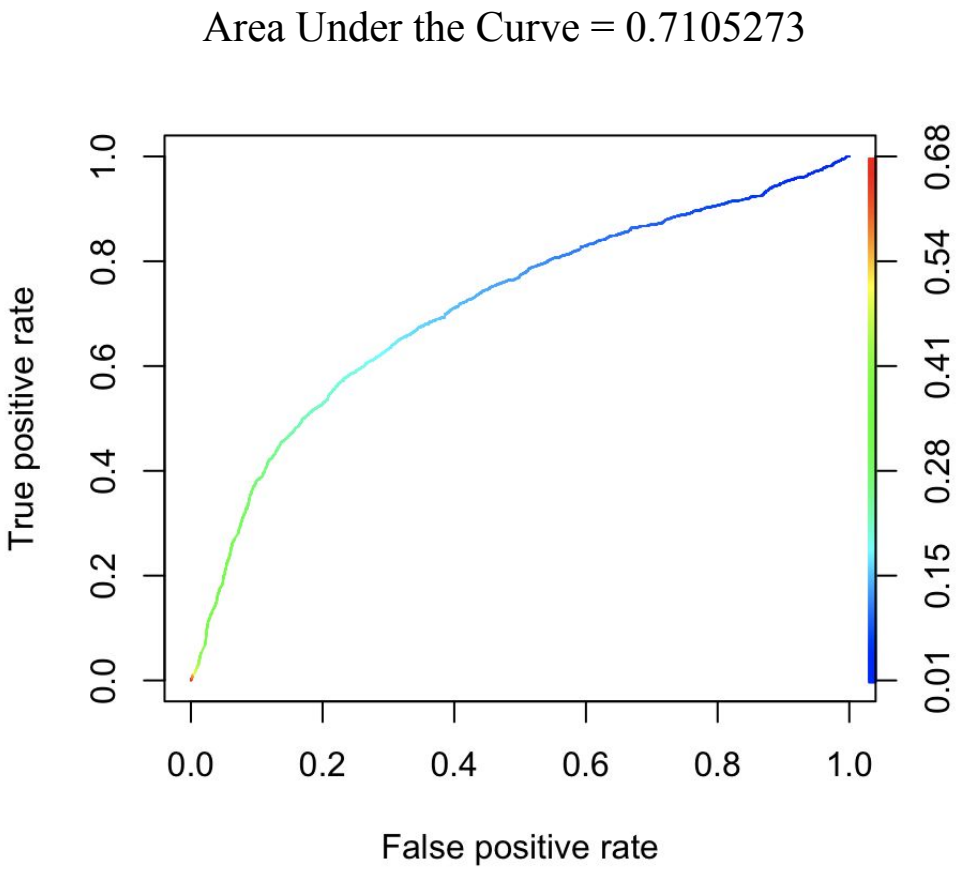
```
Call:
glm(formula = Attrited ~ Avg_Utilization_Ratio + Credit_Limit +
    Months_Inactive_12_mon + Less_than_40K + I40_60K + I60_80K +
    I80_120K, family = "binomial", data = bankchurnersdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5012  -0.6226  -0.4735  -0.3298   2.7756

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.657e+00  1.110e-01 -14.923  < 2e-16 ***
Avg_Utilization_Ratio  -2.759e+00  1.328e-01 -20.777  < 2e-16 ***
Credit_Limit           -3.648e-05  4.019e-06  -9.075  < 2e-16 ***
Months_Inactive_12_mon  3.950e-01  2.692e-02  14.670  < 2e-16 ***
Less_than_40K           7.998e-02  8.636e-02   0.926  0.35435
I40_60K                -1.184e-01  9.761e-02  -1.213  0.22530
I60_80K                -3.190e-01  1.030e-01  -3.098  0.00195 **
I80_120K               -4.128e-02  9.677e-02  -0.427  0.66968
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8927.2  on 10126  degrees of freedom
Residual deviance: 8182.1  on 10119  degrees of freedom
AIC: 8198.1

Number of Fisher Scoring iterations: 5
```

Logistic Regression: Interpretation & Analysis

- If the utilization ratio increased by one unit, we can expect to see an approximate decrease of 2.759 in the log odds and odds of a customer leaving the bank.
- If the credit limit increased by one unit ($1), we can expect to see an approximate decrease of 0.00003648 in the log odds and odds of a customer leaving the bank.
- If the number of months inactive increased by one unit (1 month), we can expect to see an approximate increase of 0.395 in the log odds and odds of a customer leaving the bank.
- A customer being within the salary range of $60K-$80K decreases the log odds and odds of a customer leaving the bank by approximately 38%. --- ((exp(0.3190)-1)*100)

Note: other coefficients not shown here due to being statistically insignificant with a p-value greater than 0.05.

ROC Curve


Area Under the Curve = 0.7105273
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.15
0.28
0.41
0.54
0.68

Confusion Matrix

	0	1	Total
0	6579	1921	8500
1	703	924	1627
Total	7282	2845	10127

Calculated with cutoff = 0.20

Sensitivity and Specificity

Calculated with cut-off = 0.20

Sensitivity: True Positive Rate

56.79%

The test did well in correctly identifying customers who have left the bank.

Specificity: True Negative Rate

77.4%

The test was strong at correctly identifying customers who have not left the bank.

Precision and Accuracy

Calculated with cut-off = 0.20

Precision:

32%

The test has many false positives.

Accuracy:

74%

The test has good accuracy and can predict customer attrition relatively well.

Conclusion - Summary

Variables with greatest impact on customer attrition:

- average utilization ratio
- number of months inactive

Notable trends of bank's current customer base:

- Customers within an income range of $60k-$80k most likely to be long term loyal customers
  - Possibly because their needs match best with the bank's current offerings

Conclusion - Recommendations

Target advertising towards an audience within 60k-80k income bracket

Specific actions:

- Run small promotions to attract customers of that income range
- Introduce a referral rewards system
  - People are friends with people with socioeconomic similarities

We believe these would be a good efforts in trying to decrease the current customer attrition(turnover) rate.