A Regression and Visual analysis

Credit Card Customers: Bank Churners

Codex Lambda

Executive Summary

The bank has seen a sizeable amount of churning within its customer base (16.40%). To uncover the underlying factors behind this phenomenon, logistic regression analysis was applied, and a supervised learning model was trained. The regression results indicate churning to be associated with different demographic segments, such as female customers. Key metrics include reduced usage and transaction activity by the customers, as well as an increase in contact towards the customers from the bank. With the assistance of the proposed predictive model, the bank could be given the opportunity to reduce churning by a substantial amount, as the model was able to predict almost 52% of the actual churning cases in the test dataset.

Data Preparation

The dataset was imported into RStudio, where the columns concerning Naive Bayes classification were removed. The Dependent Variable (DV) was converted into a logical variable and correlation between the model's variables was calculated. The variables "Avg_Open_To_Buy", "Avg_Utilization_Ratio", "Total_Ct_Chng_Q4_Q1" and "Total_Trans_Ct" were found to correlate strongly with one or more variables of the model and were subsequently emitted. The variable "Customer_Age" was investigated separately, through visual means, as it was found to be correlated with the variable "Months_on_book" and was omitted. The resulting dataset retains 16 out of the 23 initial columns.

Analysis Method

Logistic regression was utilized to uncover the factors affecting churning. Through backwards selection, an optimal model was built, which is shown in Table 1. The model was created using the library MASS and its Fitting Generalized Linear Models (glm), as shown in Figure 1.

```
lr = glm(Attrition_Flag ~ Gender + Dependent_count +
           Education_Level + Income_Category +
           Card_Category + Total_Relationship_Count +
           Months_Inactive_12_mon + Contacts_Count_12_mon +
           Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 +
           Total_Trans_Amt, family = binomial)

summary(lr)
```

Figure 1. The code associated with the regression model corresponding to the total bank churning dataset.

Following the initial model, investigation of geographic data was conducted. In detail, subsets of the original dataset were obtained for the female customers, customers with families, customers with an education level equivalent to a doctorate, customers with an income of $120 thousand or more and owners of a gold or platinum credit card.

For the **female customer subset**, the model was reduced to nine variables from the eleven found in the main model. The variable "Gender" is omitted by default, due to the subset's properties. The variable "Income Category" was found non-significant and was removed from the model. For

the **family owner subset** (customers with "Dependent_count" > 1) no exclusions were made, other than the subset's representative variable "Dependent_count". For the **doctorate customers subset**, the variable "Dependent_count" was not used, together with "Education_level". The included variables for the **$120K+ subset** and the **gold & platinum card subset** are shown in Figure 2. Additional analysis of the dataset was conducted through visual means, using Microsoft's PowerBI.

Table 1. Logistic Regression model for the bank churning dataset.

Coefficients	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*	
(Intercept)	4.39E-01	2.87E-01	1.532	0.12553	
GenderM	-3.99E-01	1.24E-01	-3.212	0.00132	**
Dependent_count	6.23E-02	2.48E-02	2.51	0.01208	*
Education_LevelDoctorate	4.46E-01	1.71E-01	2.606	0.00917	**
Education_LevelPost-Graduate	3.71E-01	1.70E-01	2.186	0.02881	*
Income_Category$40K - $60K	-4.36E-01	1.59E-01	-2.748	0.00599	**
Income_Category$60K - $80K	-4.02E-01	1.47E-01	-2.733	0.00628	**
Income_Category$80K - $120K	-1.19E-01	1.42E-01	-0.841	0.40021	
Income_CategoryLess than $40K	-3.67E-01	1.72E-01	-2.134	0.03283	*
Income_CategoryUnknown	-5.35E-01	1.92E-01	-2.791	0.00525	**
Card_CategoryGold	8.53E-01	3.12E-01	2.735	0.00624	**
Card_CategoryPlatinum	1.25E+00	6.46E-01	1.939	0.05255	.
Card_CategorySilver	3.50E-01	1.46E-01	2.392	0.01677	*
Total_Relationship_Count	-4.44E-01	2.29E-02	-19.397	< 2e-16	***
Months_Inactive_12_mon	4.91E-01	3.48E-02	14.081	< 2e-16	***
Contacts_Count_12_mon	4.45E-01	3.18E-02	13.988	< 2e-16	***
Total_Revolving_Bal	-8.66E-04	4.00E-05	-21.654	< 2e-16	***
Total_Amt_Chng_Q4_Q1	-9.98E-01	1.74E-01	-5.734	9.80E-09	***
Total_Trans_Amt	-2.91E-04	1.85E-05	-15.688	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' ' 0.1 ' ' 1

```
glm(Attrition_Flag ~ Total_Relationship_Count +
      Months_Inactive_12_mon + Contacts_Count_12_mon +
      Total_Revolving_Bal + Total_Trans_Amt, family = binomial)


lr = glm(Attrition_Flag ~ Months_Inactive_12_mon + Contacts_Count_12_mon +
  Total_Revolving_Bal + Total_Trans_Amt, family = binomial)
```

Figure 2. Variables included in the $120K (top) and gold & platinum card (bottom) subsets.

Regression Analysis Results

The regression analysis provided insights on the factors contributing to churning. For the main/general model, it was found that the important factors were: **(1) gender**, where **female** customers are more prone to churn, **(2) family owners**, **(3) education level**, where **Doctorate and Post Graduate** levels are of importance, **(4)** the **Income Category of $120K+**, **(5) owners of Gold, Platinum or Silver cards**, wherein Gold & Platinum users are found to attribute to churning more prominently, **(6)** customers with a **low number of products** held, **(7)** customers who have been **inactive** for longer than the average and **(8)** customers who were **contacted more frequently** than the average.

Subset analysis results differ slightly from the main dataset. Firstly, for the **female customers subset**, the factors 2, 3 (only Doctorate), 5, 6, 7 and 8 were found to drive churning. The most significant churning factors in this category were owning a Gold or Platinum card and the reduction in yearly spending. Secondly, for the **family owner subset**, factors 1 (female), 3 (Doctorate), 4, 5, 6, 7, 8 and having a lower transaction amount (Q4 over Q1) were significant. The most significant factors were owning a gold or platinum card, and the customer's transaction amount. Thirdly, for the **doctorate subset**, factors 1 (female), 4, 6, 7, 8 were significant, with the income category of $120K+ and the female gender attributing the most. Finally, the **$120K+ subset** churn is affected by factors 6, 7 and 8, while the **gold and platinum card subset** is affected by factors 7 and 8 alone. The collection of these models can be seen in Table 2.

Table 2. Main and subset Regression results

	Main dataset	Female Customers	Family Owners	Doctorate	$120K+	Gold and Platinum Card
Coefficients:	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
(Intercept)	0.4393	0.9343**	0.8633**	0.9511	-0.9370	-2.2220.
GenderM	-0.3991**		-0.2966.	-1.0600.		
Dependent_count	0.0623*	0.0950**				
Doctorate	0.4462**	0.6952**	0.3613.			
Graduate	0.0977	0.2028	0.1190			
High School	0.1115	0.2216	0.1159			
Post-Graduate	0.3707*	0.0720	0.1381			
Uneducated	0.1124	0.0562	0.1069			
Unknown	0.2192.	0.2938	0.2490			
Income $40K - $60K	-0.4357**		-0.2977	-1.0630		
Income $60K - $80K	-0.4022**		-0.4117*	-0.9459		
Income $80K - $120K	-0.1190		-0.0733	-0.5486		
Income Less than $40K	-0.3668*		-0.2178	-0.7321		
Income Unknown	-0.5346**		-0.3208	-1.3740.		
Card Gold	0.8527**	1.0410.	0.9524**			
Card Platinum	1.2530.	2.5020**	1.4430.			
Card Silver	0.3502*	0.4063.	0.4458**			
Relationship_Count	-0.4441***	-0.3877***	-0.4630***	-0.3673***	-0.4132***	
Months_Inactive_12_mon	0.4905***	0.4811***	0.5230***	0.6615***	0.6543***	0.5955*
Contacts_Count_12_mon	0.4446***	0.4170***	0.4238***	0.4940***	0.3394**	0.5076.
Revolving_Bal	-0.0009***	-0.0008***	-0.0008***	-0.0007***	-0.0009***	-0.0010**
Amt_Chng_Q4_Q1	-0.9976***	-1.2640***	-1.1860***	-0.7999		
Trans_Amt	-0.0003***	-0.0006***	-0.0004***	-0.0004***	-0.0002***	-0.0001

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '. ' 0.1 ' ' 1

Predictive Model

Based on the main dataset, two subsets were created to train a model would be able to predict the churning probability of the bank's existing customers. The "training" subset was attributed 8400 records and the "test" subset 1000 records, out of the 9400 records in the main dataset. The number of records in each subset aims to adhere to an 9-1 ratio.

The methods tested to arrive to the final predictive model were logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and K-nearest neighbors (KNN).

The results of the methods are shown in Table 3. The KNN method did not yield presentable results, therefore it is not be included.

Table 3. Comparison of the different methods applied to create predictive models.

	Logistic Regression (threshold = 0.35)	LDA	QDA
Actual Churning Predicted	51.72%	36.55%	30.34%
False Positives	44.62%	28.38%	45.67%
Error Rate	13.10%	11.30%	13.80%

Out of the presented models, most promising is found to be the logistic regression model. This model showcases a much higher churning prediction rate than the other two models, while keeping false positives on par with the QDA model. The validity of the model can be tested more thoroughly on an extensive dataset. Error Rates should not be expected to be reduced, there is rather the possibility that they would increase for different configurations of the model.

Conclusions & Recommendations

To retain the length of the current report short, visual analysis and relevant findings are transferred to the presentation slides. In brief, the bank should cater towards churning demographics, providing products and services that fit their needs and nuances. Female Customers should be offered such services, with increased attention to the $60K+ annual salary demographics, which should be incentivized to become a customer.

The bank could also focus on providing family programs for customers with a higher number of dependents. In addition, benefits and bonuses from each credit card type should be assessed, as there is evidence of churning in the male demographic (analysis on the male demographic is included in the visual presentation) for blue card owners. Moreover, the gold and platinum card owners are also churning to an increased number, therefore re-thinking of the perceived customer value each credit card type provides could address such issues.

The most importance metrics across most segments are found to be the reduction of yearly spending, together with the average months of credit card inactivity. To a lesser extent, the number of products owned by the customer, along with the number of times the customer was contacted by the bank, affect most segments and the main model. It should be noted that differences between metric significance should not be taken for granted, as future data could affect the importance of some key metrics.

The company should also push for an expansion of the data analytics efforts, in order to discover additional metrics and expand data analysis to other sections of the business. High importance metrics about churning are bound to extend beyond credit card data, while an assessment of the current data gathering model could provide additional clarity to the current situation.

Apart from increasing the effort on diagnostic analytics, predictive models could assist in reducing the effect of churning. Timely prevention with the support of machine learning would be able to provide access to untapped insights and increase contact efficiency with the customer.

Limitations

The analysis is limited to the data provided. Given a different context, the analysis may not hold and the results may lead to negative outcomes for the bank. An increased amount of data, both in customer quantity and number of metrics would assist in validating the effectiveness of the model. Moreover, the changing condition of the market, economy and culture could negatively affect the performance of the analysis.

Finally, data diversity and inclusivity of different metrics could prove detrimental for the efficiency of the presented models. The current dataset does not offer high diversity towards the ways metrics are set to address churning. The analysis is limited to credit card churning, for the current timeframe.