Identifying Pilgrim Bank's Loyal Customer

Team 7:

Michael Isble
Karl Olson
Johan Rong
Ampun Janpengpen
Hyun Kim

Executive Summary

Our objective with this project was to see if we could tease out information from a dataset of customers at Pilgrim Bank, a small retail bank in Texas. The project was strictly pedagogical. We received the dataset from P.K. Kannan and had no access to Pilgrim's management.

The dataset contained nearly 32,000 data points and had eight variables – *District*: referring to one of three zip codes around the banks' retail location, *Profit*: indicating how much the bank made from each customer in 1999, *Age*: broken down into seven categories, *Income*: broken down into nine categories, *Tenure*: a numerical variable indicating years at the bank as of 1999, *Online*: a binary variable indicating whether or not a Pilgrim customers banked online, *Bill Pay*: a binary variable indicating whether or not a customer used Pilgrim's online bill pay service, and *Retained*: our binary response variable indicating whether or not a customer was retained by Pilgrim in 2000.

Approximately 20% of the data points did not list *Age* and *Income*. While there is a strong correlation between not listing this information and a customer not being retained, little else is can be deduced about this group as a whole. None of the predictive variables is well correlated with the "not listers".

Our analysis indicates that the district in which a customer banks has little bearing on whether or not they will be retained. This indicates that all three branches likely provide relatively similar service. Additionally, the use of Pilgrim's online bill payment service did not seem to affect customer retention levels.

On the other hand, customers who banked online were more likely to stay with Pilgrim. Additionally, as might be expected, tenure with the bank had an impact on retention. The longer a customer has been with Pilgrim, the more likely they are to stay with the bank. Older patrons are also less likely to move than are younger patrons. Considering Pilgrim's relatively rural location, this may be caused by younger customers' higher tendency to move out of the area. Interestingly, customers who generated higher levels of profit for Pilgrim tended to be retained. This could be because management puts an emphasis on retaining these higher margin clients by offering them concentrated customer service or by offering excellent rates on CDs and other products that those who pay more for banking might use more frequently. Conversely, this tendency could also be because these customers have higher switching costs (e.g., lost days of interest).

A conversation with decision makers at Pilgrim would be necessary to make solid recommendations as to how the bank can increase its retention rates. Some of the possible suggestions we would bring up in that conversation are that they should advertise and promote their online services. If it is not already doing so, Pilgrim might also want to create special programs targeted at higher income and age groups to further increase their tendency to stay with company. Finally, Pilgrim might want to consider creating programs that target the lower income populations who seem less likely to stay with the bank. This last recommendation, however, could be tempered by management's impression on the value of retaining lower-income customers. Lower income customers do equate with lower profit levels per customers. These lower rates might make appealing to this group unattractive.

Technical Summary

As is typical of most data analysis endeavors, our process was highly iterative. In particular, we investigated a number of methods of accounting for the significant number of data points without records for the Income and Age variables. What follows is a description of the best analysis process we have identified and the resulting model used to help Pilgrim determine how best to retain customers.

- Missing Value Processing
  - Of the 31,633 data points in our dataset nearly 20% did not have values recorded for Age and Income. We interpolated the most frequently cited value, which was similar for retained and nonretained customers in both categories. Missing income values were recorded as "6" and missing Age values were recorded as "3" (Table A). Additionally, we deleted one data point that was missing information for several variables in addition to Age and Income.
- Data Transformation
  - The distribution of two variables, Profit and Tenure, is skewed. To account for this log forms were taken (Graph A).
- Graphs
  - A graphical analysis showed that the categorical variables – Online, Bill Pay and District - did not seem any influence on whether or not customers were retained (Graph B).
  - The numerical variables, particularly Log(Tenure) and Log(Profit) did illustrate some differences between retained customers and non-retained customers (Graph C).
- Chi-square test & Correlation Test
  - Chi-square test result: All variables except 'Bill Pay' showed dependence with a response variable
  - Correlation Test: Log(Profit), Age, Log(Tenure) showed relatively high correlation coefficient.
- Logistic Regression
  - We chose to perform a logistic regression as our dependent variable is binary and the purpose of our analysis is categorical, not predictive.
  - The predictor variables used in the logistic regression are log(Profit), Age, Income, Log(Tenure), District, Online and Bill Pay. While our Chi-square test indicated Bill Pay would not prove significant, we included it to test whether there would be a discrepancy between the regression and the Chi-square analysis.
  - We ran a logistic regression including all predictors, best subset selection.
  - In order to maximize the accuracy of our model, we dropped insignificant predictors one at a time to get the best output. As expected Bill Pay proved insignificant and was the first dropped. District also proved insignificant. We found that the following is the best model:

The Logistic Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.64380628	0.25709355	0.01227386	*
Log(Profit)	0.29747277	0.10401043	0.00423603	1.34645176
Online	0.42755309	0.09118895	0.00000275	1.53350055
AGE	0.25735554	0.02353481	0	1.29350495
INCOME	-0.05267424	0.0145958	0.00030754	0.94868898
Log (Tenure)	1.06416845	0.08408901	0	2.89842772

Residual df	9994
Residual Dev.	8347.935547
% Success in training data	83.92
# Iterations used	10
Multiple R-squared	0.05350931

- The result is basically consistent with the graph, chi-square analysis, and correlation analysis.

- Implication
  - With increasing Profit, Tenure, and age, the possibility of being retained is increasing.
  - Online users are more likely to stay than customers who do not bank online.
- Possible Recommendations
  - We call the following recommendations “possible” because, without real knowledge of Pilgrim’s business, it is hard to determine what side effects they might have. A conversation with the bank’s management would allow us to provide more concrete recommendations.
    - Create products that target the lower-income groups the bank has trouble retaining.
    - Provide customized services for the most profitable age groups.
    - Campaign and Encourage customers to use online banking.

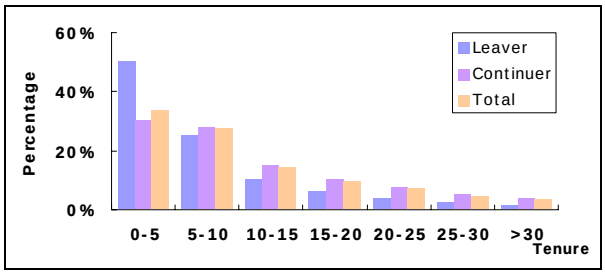
Appendices

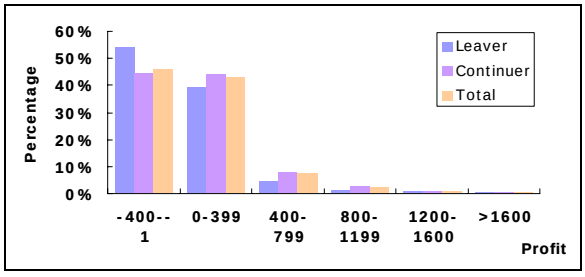
Table A

Count of INCOME	RETAIN CUST.		Grand Total
INCOME	0	1	
1	213	1831	2044
2	76	733	809
3	245	2326	2571
4	191	2121	2312
5	188	2181	2369
6	364	5049	5413
7	231	2921	3152
8	136	1606	1742
9	163	2797	2960
10	3431	4830	8261
Grand Total	5238	26395	31633

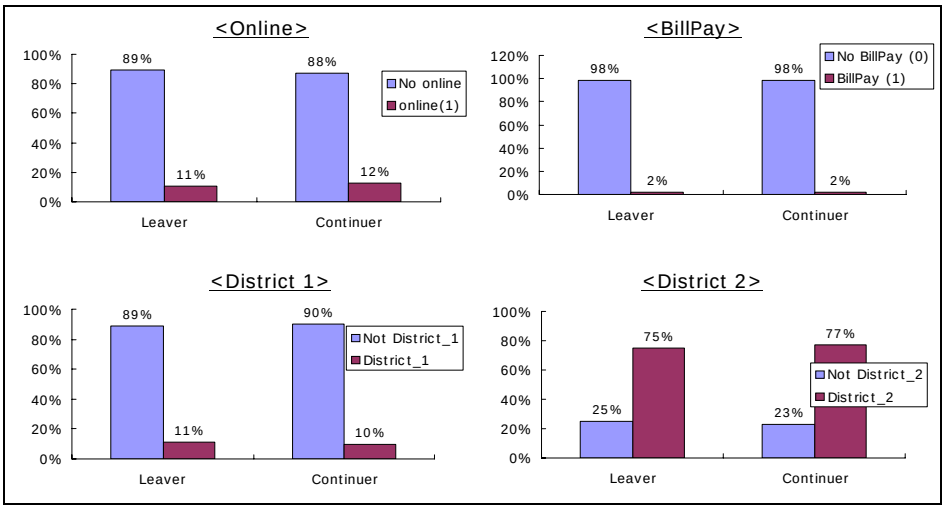
Count of INCOME	RETAIN CUST.		Grand Total
Age	0	1	
1	126	583	709
2	405	3245	3650
3	395	4995	5390
4	336	5040	5376
5	218	3018	3236
6	152	2138	2290
7	165	2528	2693
8	3441	4848	8289
Grand Total	5238	26395	31633

Graph A





Graph B



Graph C