Team: Yecheng Li, Doris Long, Vera Wang, Jikun Zhou
Class: BUS 111A
Project: Pilgrim case study

1. What is Pilgrim Bank's data problem? What is the final managerial objective?

Pilgrim Bank's senior management is currently reconsidering bank's internet strategy -- whether to charge service fee for those customers using online channel or offer with lower service charge to engage customers. To make the decision, the key point is to answer if online customers could bring higher profit or secure a higher retention rate. In our report, we described the dataset received from P.K. Kannan, and further conducted regression and correlation test to see whether online customers could bring higher profit or have associated with higher retention rate. If the analysis shows online customers are indeed better customers, the senior management would decide to offer rebates or lower the service charges for customers using online banking.

Major Issues with Database

The current dataset mainly have two problems:
(1) Lack of specific information about the calculation of profit.

As online banking might reduce cost of serving a customer and increase fee revenue by engaging customers' transaction with convenience, it is crucial to analyze related factors in the equation of profit calculation. However, the dataset only includes the final number of profit rather than specific components of it.

(2) Contains missing values.

At least 20% of the consumer information (/Dataset points) are incomplete and missed one or more information in "Age", "Income", or "Billpay".

2. Description of Variables

"ID" simply means the customer ID, which is an identity, and it is a nominal variable.

"District" is also a nominal variable because it represents geographic regions that are assigned into different numbers (1100,1200, and 1300), but there is no implied order among these categories.

"Profit" indicates how much the bank makes from customer and is calculated using the formula

(Balance in Deposit Accounts)*(Net Interest Spread) + (Fees) + (Interest from Loans) - (Cost to serve)

Since profit is obtained through mathematical calculation, it is a ratio variable.

“Age” is an ordinal variable. The age of customer are divided into 7 categories, starting from 1 to 7. “1” represents customers younger than 15 years old, following by “2” represents 15-24 years old. “3” represents 25-34 years old, “4” is for a range between 35 and 44 years old, “5” is for a range between 45 and 54 years old. “6” represents people age from 55 to 64 years old, and “7” represents 65 years and older. It is an ordered category that can be ranked but has no exact value.

The ordered variable “Income” utilizes number 1 to 9 to represent individual customer's income levels. “1” represents a range of income less than $15,000. “2” means an income range of $15,000 - $19,999. “3” means an income range of $20,000-$ 29,999. “4” means an income range of $30,000-$39,999. “5” means an income range of $40,000-$49,999. “6” means an income range of $50,000-$74,999. “7” means an income range of $75,000-$99,999. “8” means an income range of $100,000-$124,999, and “9” represents income level of $125,000 and more. Since the intervals of this variable are not equal, “Income” is an ordinal variable.

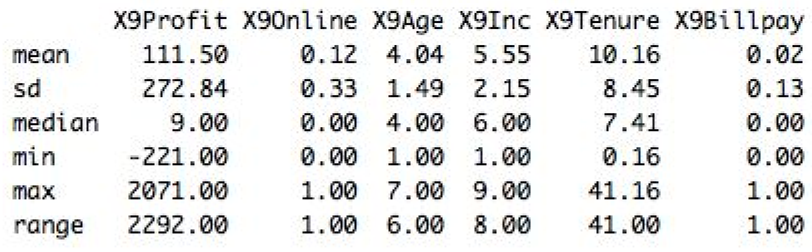
“Tenure” indicates the length of years that consumers stay with the bank as of 1999. It is a ratio value because it can be calculated with mathematical calculation.

“Online” is a binary variable indicating whether a Pilgrim customer uses online banking or not. 0 represents the customer does not use online banking and 1 represents he or she does. The variable “Online” is also a nominal variable because they just represent two individual categories that cannot be ranked or compared.

“Bill Pay” is a binary variable indicating whether or not a customer uses Pilgrim’s online bill pay service. It is also a nominal variable. 0 represents there has been transactions in the customer’s account, while 1 represents there is no transaction at all.

3. Data Summary: A table similar to Exhibit 4 from Pilgrim Bank Case A

This summary gives the mean, median, standard deviation, min, max and range for 1999 Profit, Age, Income, Online, Bill Pay, and Tenure.



```
       X9Profit X9Online X9Age X9Inc X9Tenure X9Billpay
mean     111.50     0.12  4.04  5.55    10.16      0.02
sd       272.84     0.33  1.49  2.15     8.45      0.13
median     9.00     0.00  4.00  6.00     7.41      0.00
min     -221.00     0.00  1.00  1.00     0.16      0.00
max     2071.00     1.00  7.00  9.00    41.16      1.00
range   2292.00     1.00  6.00  8.00    41.00      1.00
```

4. Solution for missing data

Among 31,634 data points in the dataset, nearly 20% missed of values of “Age” and “Income”. Simply deleting this portion of would significantly decrease our sample size. Instead, we replaced missing value with the median value of 1999 “Age” and “Income”, which is 4 and 6 respectively. Furthermore, we deleted those who missed values of 1999 “Age” and “Income” and left the bank in 2000(those who have no “Billpay” and no “Online Banking” data in 2000). Other than that, there are still 19 observations that stay in the bank but have no “Profit” data. However, since the data in 1999 would be the most important information for the regression and correlation analysis, we currently would keep those 19 observations of future reference.

5. Provide histograms/density plots for key variables, such as customer profitability

Age & Profit

From the boxplot between age and profit, we can tell the median profit in category “7” is much higher, followed by “6”, “5”, “3”, “4”, “2”, and “1”. The range of category “7” from 1st quartile and the 3rd quartile is also the largest, followed by “6”, “5”, “3”, “4”, “2”, and “1”.

Income & Profit

From the boxplot between income and profit, the median profit in category “9” is the highest, followed by “8”, “7”, “5”, “6”, “4”, “3”, “2”, and “1”If we look at the median of profit level of all income categories, there is a slight curvilinear relationship between income and profit. The higher income is, the higher profit the bank can generate from the customer, and slope is getting larger.

Histogram of Profits and Zoomed In

In 1999, Pilgrim Bank earned total $3,527,276 from from 31,634 customers. The profit ranged from $-221 to $2071, averagely $111.5 per customer with a standard deviation of 272.8 and median of $9, which indicates this variable is far stretched out. As the X-axis represented the profit range from -200 to 2000 in dollar, and Y-axis represented the frequency of each profit amount. According to the Histogram of Profit, we can see the fluctuation among each customers; it might due to individual differences on consuming habit, or the complexity formula to calculate profit. Generally, Pilgrim Bank earn positive profit from about 60% of customers.

Histogram of Online

In 1999, among 31,634 customers, 3854 customers were using Pilgrim Bank's service, which was 12.18% of the total dataset point. It is estimated that partly because of the popularization of personal computer back in 1999. Although in 1997, the percentage of household owning computers increase to 35%, the majority households still has no personal computer.

Histogram of District

All 31,634 dataset points were allocated into three different Districts: 1100, 1200, and 1300. Among 31,634 customers, 3142 customers were from district 1100; 24342 customers were from district 1200; and 4150 customers were from district 1300. It shows that most customers were from district 1200, which indicates Pilgrim Bank might own subsidiary banks than other two districts.

Histogram of Billpay

Billpay means the electronic service under the online banking service. In 1999, among 12% of 31,634 customers were using online banking service, only 528 customers used Billpay service. So in total, only 1.6% of customers used this service. Additional to the low popularization of personal computer, the distrust of electrical bill pay in 1999 is also a major reason

6. Create bivariate frequency distributions (tables or plots) for key variables

In order to provide more detailed analysis for the relationship of tenure and profits with other factors, we decide to create two new variables, named as Tenure.Level and profits.Level.

Profit.Level buckets are as follows: 1 = profitability less than 0; 2 = 0-100; 3 = 100-200; 4 = 200-400; 5 = 400-600; 6 = 600-800, 7 = 800-1000, 8 = 1000-1200, 9=1200-1400, 10=1400-1600, 11=1600-1800, 12 = 1800-2000, 13= profitability larger than 2000.

Tenure.Level buckets are as follows: 1 = less than 3 years; 2 = 3-6 years; 3 = 6-9 years; 4 = 9-12 years; 5 = 12-15 years; 6 = 15-18 years, 7 = 18 - 21 years, 8 = 21-24 years, 9=24-27 years, 10=27-30 years, 11=30-33 years, 12=33-36 years, 13= 36-39 years, and 14=39-42 years.

7. Discuss what the data patterns indicate, and what this could mean for the managers at Pilgrim Bank?

From the table “1999 Income with Online and Billpay”, we notice that there is a larger percentage of customers use online banking in customer category with higher income level than those who are in the lower income level. 26% of customers from category 9 (over $125,000 annual income) use online banking, while only 13.5% of customers from category 1 (less than $15,000 annual income) use the service. In the meantime, younger people tend to use online banking more frequently than older people. Compared to 27.8% people younger than 15 years old using online banking, only 6.7% people older than 65 years old use online banking.

From the table “1999 Income with Online and Billpay”, we observe that customer with Income level 6 ($50,000 - $74,999) had most online uses and electronic bill pay uses. However, if we look back to plot “Box-Plot of Profit Distribution by Income Cont.(Zoomed In)”, customers with income level 6 generated a medium profit near zero, which is very low compared to the level 5. Level 6 has a slightly decrease after an increasing trend of median values from income level 1-level 5. Therefore, the group of customers that used online banking and electronic billpay generate relatively low profit for the bank. A similar observation can be found in the “1999 Age with Online and Billpay”. The group of customers in age level 4 (35 - 44years) had most online uses and electronic uses. The plot “Box-Plot of Profit Distribution by Age Cont.(Zoomed In)” shows that the same group of people generated a relatively low profit for the bank.

A more direct view can be concluded in the table of “profit level and online & billay”. In this table, we separated the profit into 13 different levels. The observation is clear that the group of people who had most online uses and electronic bill pay uses generated a profit in level 1 (profit is less than 0). In conclusion, the customers who had the most online uses and electronic bill pay uses did not generate much profit for the bank and should be charged with a higher fee.