Pilgrim case study

Yecheng Li, Doris Long, Vera Wang, Jikun Zhou

November 27, 2017

The URL for our Team GitHub repository is

link: https://github.com/Brandeis-BUS111-FinalProject/Pilgrim-Final.git

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 on online service to discourage the using of online channel, or to offer lower service to engage customers using online channel. To make the decision, the key point is to answer if online customers are better. In our report, we described the dataset received from Erica Dorstamp, and further conducted correlation and regression test to see whether online customers could bring higher profit or secure higher retention rate for Pilgrim. 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. Other way, Pilgrim could raise the service charge to buffer the cost of offering online banking service.

However, the current dataset mainly has 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. There exists omitted variable bias because some detailed information of profit calculation is missing.

(2) Contains missing values: At least 20% of the consumer information 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.

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.

‘BillPay’ 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. Handling of the 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.

```
# Read the given dataset
consumerDB = read.csv("dataset.csv") ### read the given dataset
# Check who stay with the bank in 2000: 1 means that consumers stay with the bank,
# while 0 means that consumers leave the bank
consumerDB$retention =1
consumerDB[is.na(consumerDB$X0Online) & is.na(consumerDB$X0Billpay),]$retention = 0
# Find the median for 1999 Age and Income
AgeMedian_1999 = median(consumerDB$X9Age,na.rm = TRUE)
IncomeMedian_1999 = median(consumerDB$X9Inc,na.rm = TRUE)
# Present the data for 1999 Age/Income median
AgeMedian_1999
```

```
## [1] 4
```

```
IncomeMedian_1999
```

```
## [1] 6
```

```
consumerDB[is.na(consumerDB$X9Age),]$X9Age = 4
consumerDB[is.na(consumerDB$X9Inc),]$X9Inc = 6
```

4. Statistic Summary

(1) Statistics Summary for 1999 Data

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

```
# Sort the concumerDB and get a Table that sepcifically contains data for 1999
statsTable1999= consumerDB[,2:6]
X9Billpay = consumerDB[,10]
statsTable1999= cbind(statsTable1999,X9Billpay)
Summary_Table=t(describe(statsTable1999))
Summary_Table = round(Summary_Table,2)
Summary_Table_New = Summary_Table[c(3:5,8:10),c(1:6)]
Summary_Table_New
```

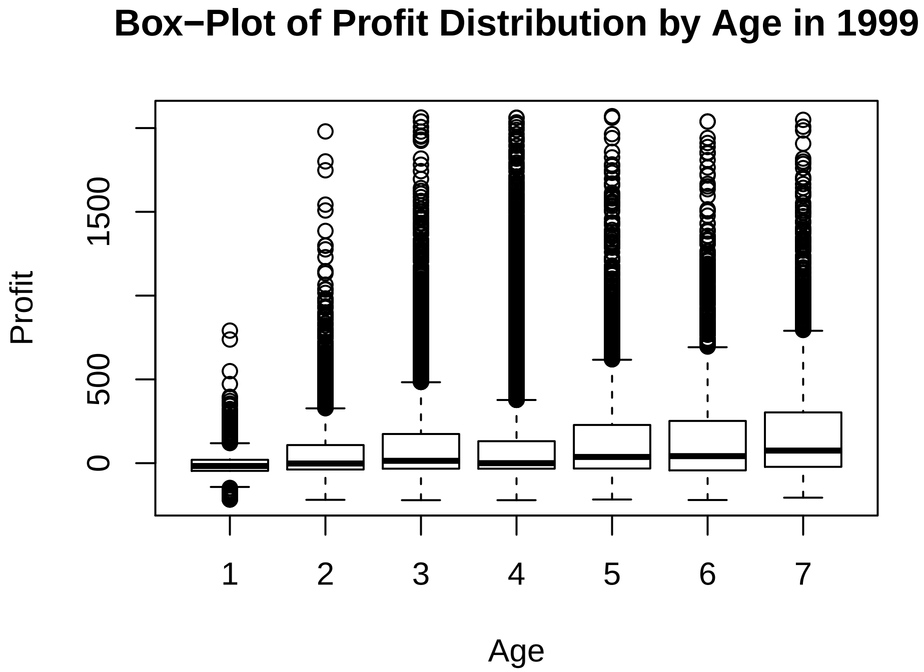
```
##         X9Profit X9Online X9Age X9Inc X9Tenure X9Billpay
## mean      111.50     0.12  4.03  5.60    10.16      0.02
## sd        272.84     0.33  1.41  2.03     8.45      0.13
## median      9.00      0.00  4.00  6.00      7.41      0.00
## min      -221.00      0.00  1.00  1.00      0.16      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
```

(2) Visual Description of Statistic Summary

a.Histogram of Age

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’.

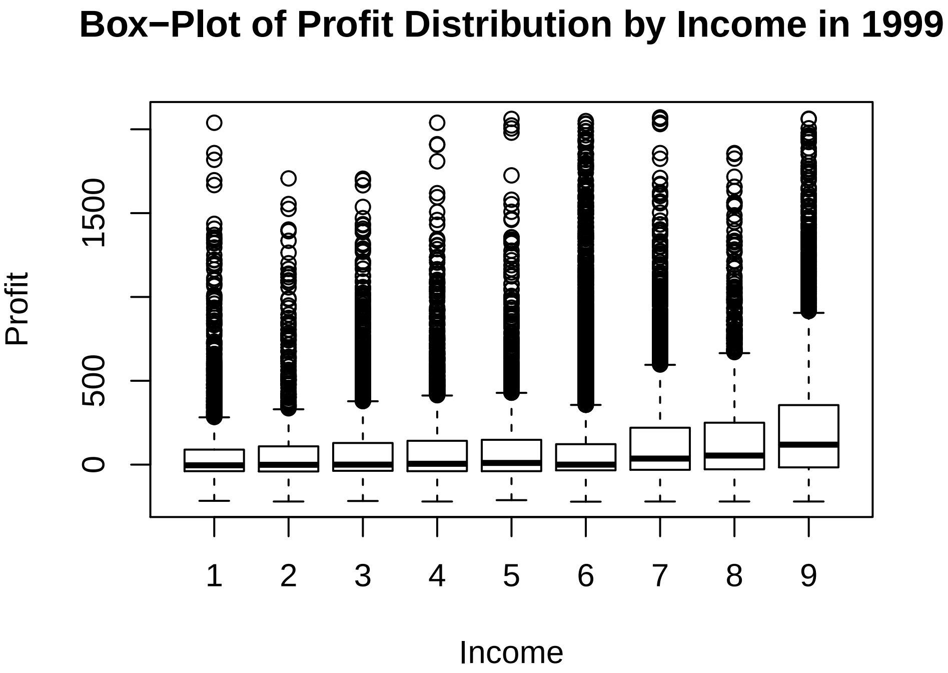
```
# This is a boxplot graph for Profit& Age
boxplot(X9Profit~X9Age, data = consumerDB,
        main = "Box-Plot of Profit Distribution by Age in 1999",
        xlab = "Age", ylab = "Profit") # Sets X and Y Axes
```



b.Histogram of Income

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.

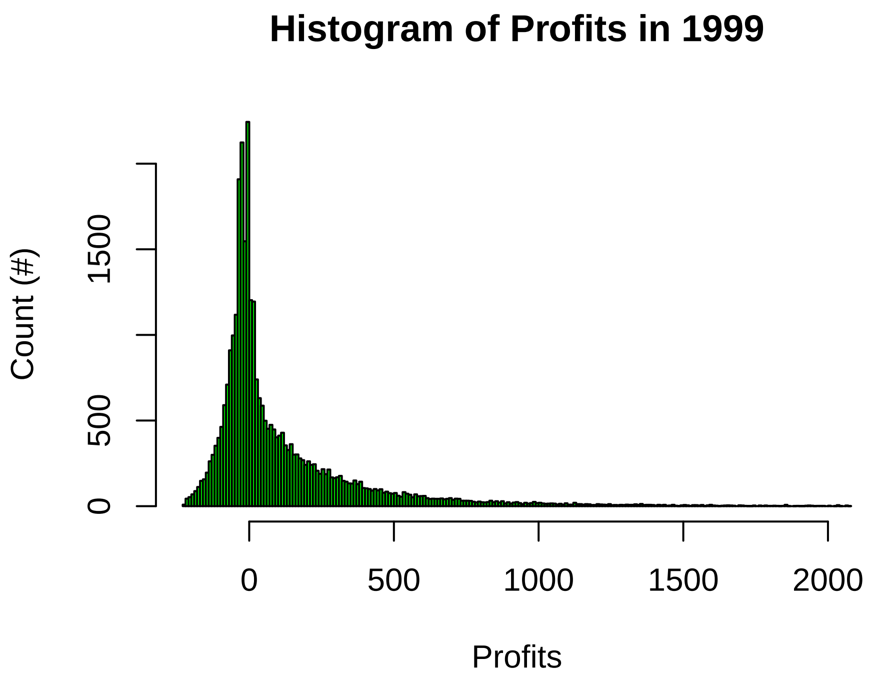
```
# This is a boxplot graph for Profit& Income
boxplot(X9Profit~X9Inc, data = consumerDB,
        main = "Box-Plot of Profit Distribution by Income in 1999",  ## Sets Title to Plot
        xlab = "Income", ylab = "Profit") # Sets X and Y Axes
```



c.Histogram of Profits

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.

```r
# Histogram of Profits
hist(consumerDB$X9Profit, main = "Histogram of Profits in 1999",
     xlab = "Profits", ylab = "Count (#)", col = "green", n = 200)
```



(3) Data Patterns Summary

From the above table, we can see a positive relation between the income levels and uses of online banking and electronic billpay. The more income one customer earns, the more likely that customer will use online banking and electronic billpay. We can interpret this as customers may find a easier way to manage their money from online banking and electronic billpay.

Age groups and the uses of online banking and electronic billpay, on the contrary, appears to be a negative relation. The younger one customer is, the more likely that customers will use online banking and electronic billpay.

The similar relationships can be applied to the retention rate as well. The higher the income level and age group, the more likely customers will stay with the bank.

In conclusion, customers in income group 9 are more likely to use online banking and electronic billpay and to stay with the bank. Customers in age group 7 are more likely to use online banking and electronic billpay and to stay with the bank.

```
# 1999 Income with online and billpay customer
summary_Income_1999_Online = table(consumerDB$X9Inc,consumerDB$X9Online)
# summary_Income_1999
summary_Income_1999_Online = round(summary_Income_1999_Online/rowSums(summary_Income_1999_Online),2)
summary_Income_1999_Online
```

```
##
##          0    1
##   1 0.92 0.08
##   2 0.92 0.08
##   3 0.89 0.11
##   4 0.89 0.11
##   5 0.87 0.13
##   6 0.89 0.11
##   7 0.85 0.15
##   8 0.84 0.16
##   9 0.82 0.18
```

```
summary_Income_1999_Billpay = table(consumerDB$X9Inc,consumerDB$X9Billpay)
summary_Income_1999_Billpay = round(summary_Income_1999_Billpay/rowSums(summary_Income_1999_Billpay),2)
summary_Income_1999_Billpay
```

```
##
##        0    1
##   1 0.99 0.01
##   2 1.00 0.00
##   3 0.99 0.01
##   4 0.99 0.01
##   5 0.98 0.02
##   6 0.98 0.02
##   7 0.98 0.02
##   8 0.98 0.02
##   9 0.97 0.03
```

```
# 1999 Age with online and billpay customer
summary_Age_1999_Online = table(consumerDB$X9Age,consumerDB$X9Online)
# summary_Age_1999
summary_Age_1999_Online = round(summary_Age_1999_Online/rowSums(summary_Age_1999_Online),2)
summary_Age_1999_Online
```

```
##
##        0    1
##   1 0.81 0.19
##   2 0.78 0.22
##   3 0.85 0.15
##   4 0.88 0.12
##   5 0.91 0.09
##   6 0.95 0.05
##   7 0.96 0.04
```

```
summary_Age_1999_Billpay = table(consumerDB$X9Age,consumerDB$X9Billpay)
# summary_Age_1999
summary_Age_1999_Billpay = round(summary_Age_1999_Billpay/rowSums(summary_Age_1999_Billpay),2)
summary_Age_1999_Billpay
```

```
##
##        0    1
##   1 0.99 0.01
##   2 0.97 0.03
##   3 0.97 0.03
##   4 0.99 0.01
##   5 0.99 0.01
##   6 0.99 0.01
##   7 0.99 0.01
```

```
retention_1999 = table(consumerDB$X9Inc,consumerDB$retention)
retention_1999 = round(retention_1999/rowSums(retention_1999),2)
retention_1999
```

```
##
##        0    1
##   1 0.10 0.90
##   2 0.09 0.91
##   3 0.09 0.91
##   4 0.08 0.92
##   5 0.08 0.92
##   6 0.28 0.72
##   7 0.07 0.93
##   8 0.08 0.92
##   9 0.05 0.95
```

```
retention_Age = table(consumerDB$X9Age,consumerDB$retention)
retention_Age = round(retention_Age/rowSums(retention_Age),2)
retention_Age
```

```
##
##        0    1
##   1 0.18 0.82
##   2 0.11 0.89
##   3 0.07 0.93
##   4 0.28 0.72
##   5 0.07 0.93
##   6 0.07 0.93
##   7 0.06 0.94
```

5. Mean profitability of years 1999 and 2000 customers using online banking or electronic billpay

To compare the mean profitability of customers for the years 1999 and 2000 by their enrollment status in online banking or electronic billpay, we conducted four independent t-test.

(1) Compare the profitability of 1999's customer using online banking or not. Null hypothesis: Mean profit for year 1999's customers using online banking = Mean profit for year 1999's customers not using online banking
Alternative Hypothesis: Mean profit for year 1999's customers using online banking != Mean profit for year 1999's customers not using online banking

```
t.test(consumerDB[consumerDB$X9Online == 0,]$X9Profit, consumerDB[consumerDB$X9Online ==1,]$X9Profit)
```

```
##
##  Welch Two Sample t-test
##
## data:  consumerDB[consumerDB$X9Online == 0, ]$X9Profit and consumerDB[consumerDB$X9Online == 1, ]$X9
## t = -1.2124, df = 4882.1, p-value = 0.2254
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.389887   3.628706
## sample estimates:
## mean of x mean of y
##  110.7862  116.6668
```

According to the independent t-test, we failed reject the null hypothesis, p-value = 0.2254 > 0.05 at the 95% confidence interval. Then we can conclude that there is no significant difference between the mean profit for year 1999's customers using online banking and mean profit for year 1999's customers not using online banking. That is to say, using online banking in 1999 did not increase or decrease the profit significantly.

(2) Compare the profitability of 1999's customer using electronic billpay or not. Null hypothesis: Mean profit for year 1999's customers using electronic billpay = Mean profit for year 1999's customers not

using electronic billpay Alternative Hypothesis: Mean profit for year 1999's customers using electronic billpay != Mean profit for year 1999's customers not using electronic billpay

```
t.test(consumerDB[consumerDB$X9Billpay == 0,]$X9Profit, consumerDB[consumerDB$X9Billpay ==1,]$X9Profit)
```

```
##
##  Welch Two Sample t-test
##
## data:  consumerDB[consumerDB$X9Billpay == 0, ]$X9Profit and consumerDB[consumerDB$X9Billpay == 1, ]$X
## t = -5.9092, df = 539.19, p-value = 6.097e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -113.69415  -56.96329
## sample estimates:
## mean of x mean of y
##  110.0785  195.4072
```

According to the independent t-test, we rejected the null hypothesis, p-value = 6.097e-09 < 0.05 at the 95% confidence interval. Then we can conclude that there is significant difference between the mean profit for year 1999's customers using electronic billpay and mean profit for year 1999's customers not using electronic billpay. That is to say, using electronic billpay in 1999 did have significant effect on customers' profit.

```
t.test(consumerDB[consumerDB$X9Billpay == 0,]$X9Profit, consumerDB[consumerDB$X9Billpay ==1,]$X9Profit,
```

```
##
##  Two Sample t-test
##
## data:  consumerDB[consumerDB$X9Billpay == 0, ]$X9Profit and consumerDB[consumerDB$X9Billpay == 1, ]$X
## t = -7.1317, df = 31632, p-value = 5.064e-13
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -65.64791
## sample estimates:
## mean of x mean of y
##  110.0785  195.4072
```

According to the two sample t-test, we reject the null hypothesis, p-value = 5.064e-13 < 0.05 at the 95% confidence interval. Then we can conclude that the mean profit for year 1999's customer using electronic bill is more than customers not using electronic billpay.

(3) Compare the profitability of 2000's customer using online banking or not.

Null hypothesis: Mean profit for year 2000's customers using online banking = Mean profit for year 2000's customers not using online banking

Alternative Hypothesis: Mean profit for year 2000's customers using online banking != Mean profit for year 2000's customers not using online banking

```
t.test(consumerDB[consumerDB$X0Online == 0,]$X0Profit, consumerDB[consumerDB$X0Online == 1,]$X0Profit)
```

```
##
##  Welch Two Sample t-test
##
## data:  consumerDB[consumerDB$X0Online == 0, ]$X0Profit and consumerDB[consumerDB$X0Online == 1, ]$X0
## t = -3.7637, df = 8995.7, p-value = 0.0001685
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -31.65474  -9.97370
## sample estimates:
```

```
## mean of x mean of y
##  140.6967  161.5109
```

According to the indepedent t-test, we rejected the null hypothesis, p-value = 0.0001685 < 0.05 at the 95% confidence interval. Then we can conclude that there is significant difference between the mean profit for year 2000's custumers using online banking and mean profit for year 2000's custumers not using online banking. That is to say, using online banking in 2000 did have significant effect on customers' profit.

```
t.test(consumerDB[consumerDB$X0Online == 0,]$X0Profit, consumerDB[consumerDB$X0Online == 1,]$X0Profit,
```

```
##
##  Two Sample t-test
##
## data:  consumerDB[consumerDB$X0Online == 0, ]$X0Profit and consumerDB[consumerDB$X0Online == 1, ]$X0
## t = -3.4589, df = 26394, p-value = 0.0002716
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -10.91593
## sample estimates:
## mean of x mean of y
##  140.6967  161.5109
```

According to the two sample t-test, we reject the null hypothesis, p-value = 0.0002716 < 0.05 at the 95% confidence interval. Then we can conclude that the mean profit for year 2000's customer using online banking is more than customers not using online banking.

(4) Compare the profitability of 2000's customer using electronic billpay or not.

Null hypothesis: Mean profit for year 2000's customers using electronic billpay = Mean profit for year 2000's customers not using electronic billpay Alternative Hypothesis: Mean profit for year 2000's customers using electronic billpay != Mean profit for year 2000's customers not using electronic billpay

```
t.test(consumerDB[consumerDB$X0Billpay == 0,]$X0Profit, consumerDB[consumerDB$X9Billpay == 1,]$X0Profit
```

```
##
##  Welch Two Sample t-test
##
## data:  consumerDB[consumerDB$X0Billpay == 0, ]$X0Profit and consumerDB[consumerDB$X9Billpay == 1, ]$X
## t = -5.7965, df = 442.47, p-value = 1.289e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -146.76352  -72.44092
## sample estimates:
## mean of x mean of y
##  141.7334  251.3357
```

According to the independent t-test, we rejected the null hypothesis, p-value = 1.289e-08 < 0.05 at the 95% confidence interval. Then we can conclude that there is significant difference between the mean profit for year 2000's customers using electronic billpay and mean profit for year 2000's customers not using electronic billpay. That is to say, using electronic billpay in 2000 did have significant effect on customers' profit.

```
t.test(consumerDB[consumerDB$X0Billpay == 0,]$X0Profit, consumerDB[consumerDB$X9Billpay == 1,]$X0Profit
```

```
##
##  Two Sample t-test
##
## data:  consumerDB[consumerDB$X0Billpay == 0, ]$X0Profit and consumerDB[consumerDB$X9Billpay == 1, ]$X
## t = -5.7943, df = 26035, p-value = 3.47e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -78.48773
## sample estimates:
## mean of x mean of y
##  141.7334  251.3357
```

According to the two sample t-test, we reject the null hypothesis, p-value = 0.0002716 < 0.05 at the 95% confidence interval. Then we can conclude that the mean profit for year 2000's customer using billpay is more than customers not billpay.

6. Transition matrix

```
# Note that there is no such a type in both 1999 and 2000
# Online = 0 & Billpay = 1
# Therefore, we didn't show this type here
consumerDB$status_1999 = 1
consumerDB[consumerDB$X9Online==1&consumerDB$X9Billpay==0,]$status_1999 = 2
#consumerDB[consumerDB$X9Online==0&consumerDB$X9Billpay==1,]$status_1999 = NONE
consumerDB[consumerDB$X9Online==1&consumerDB$X9Billpay==1,]$status_1999 = 3

consumerDB$status_2000 = 1
consumerDB[consumerDB$X9Online==1&consumerDB$X9Billpay==0,]$status_2000 = 2
consumerDB[consumerDB$X9Online==1&consumerDB$X9Billpay==1,]$status_2000 = 3
consumerDB[is.na(consumerDB$X0Online)&is.na(consumerDB$X0Billpay),]$status_2000 = 4
rowNames = c("1999 Type 1","1999 Type 2","1999 Type 3")
TransitionMatrix = table(consumerDB$status_1999,consumerDB$status_2000)
TransitionMatrix
```

```
##
##         1     2     3     4
##   1 23107     0     0  4673
##   2     0  2860     0   466
##   3     0     0   448    80
```

```
round(prop.table(TransitionMatrix,1),3)
```

```
##
##         1     2     3     4
##   1 0.832 0.000 0.000 0.168
##   2 0.000 0.860 0.000 0.140
##   3 0.000 0.000 0.848 0.152
```

Type 1 represents customers with no online account and no billpay; Type 2 represents customers with online account but no billpay; Type 3 represents customers with both online account and billpay; Type 4 represents those customers leave the bank.

So 16.8% of those customers with no online account and no billpay in 1999 leave the bank in 2000; 14% of those customers with online account but no billpay in 1999 leave the bank in 2000; and 15.2% of those customers with both online account and billpay in 1999 leave the bank in 2000. Although there is no significant differences among these percentages, the small discrepancy shows that those who have no online account and no billpay in 1999 have the higher possibility to leave the bank. On the other hand, customers who use online accounts are more likely to stay with the bank. Therefore, promoting usage of online accounts will increase the retention rate.

7. Regression analysis

```
lm_profit_Online_Billpay = lm(X9Profit ~ X9Online+X9Billpay, data = consumerDB)
summary(lm_profit_Online_Billpay)
```

```
##
## Call:
## lm(formula = X9Profit ~ X9Online + X9Billpay, data = consumerDB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415.41 -144.79 -101.79   52.21 1960.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  110.786      1.636  67.732  < 2e-16 ***
## X9Online      -6.619      5.002  -1.323    0.186
## X9Billpay     91.240     12.771   7.144 9.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.6 on 31631 degrees of freedom
## Multiple R-squared:  0.001661,   Adjusted R-squared:  0.001597
## F-statistic: 26.31 on 2 and 31631 DF,  p-value: 3.843e-12
```

A regression was calculated to see the the use of online banking and billpay has any effect on profit in year 1999. Using the online banking will decrease profit by 6.619 unit. The use of billpay will increase profit by 91.240 unit. When customers do not use online banking or billpay service, the estimated profit mean is 110.786.

The coefficients for X9Online and X9Billpay are significant. That is saying, there is significant difference on customer profitability between the use of the online banking&billpay and not using the online banking& billpay. R square is 0.001661, meaning that 0.16% of the variance in profit can be explain by using Online banking and billpay. Adjusted R-squared is 0.001597, which is used for testing a goodness-of-fit. Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in this model. The adjusted R-squared increases only if the new predictors improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

```
lm_retention_Online_Billpay = lm(retention ~ X9Online+X9Billpay, data = consumerDB)
summary(lm_retention_Online_Billpay)
```

```
##
## Call:
## lm(formula = retention ~ X9Online + X9Billpay, data = consumerDB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8599  0.1401  0.1682  0.1682  0.1682
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.831785   0.002226 373.606  < 2e-16 ***
## X9Online     0.028106   0.006809   4.128 3.67e-05 ***
## X9Billpay   -0.011407   0.017384  -0.656    0.512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3711 on 31631 degrees of freedom
## Multiple R-squared:  0.0005608,  Adjusted R-squared:  0.0004976
## F-statistic: 8.874 on 2 and 31631 DF,  p-value: 0.0001403
```

A regression was calculated to see the the use of online channel and billpay has any effect on tenure in year 1999. Using the online service could increase the average tenure length by 0.028106 year. Using the billpay will decrease tenure length by 0.011407 year. When customers do not use online banking service, the estimated tenure mean is 0.831785 year.

The coefficients for X9Online and X9Billpay are significant.That is saying, there is significant difference of retention between the use of online banking& billpay and not using online banking& billpay. R square is 0.0005608, meaning that 0.05% of the retention can be explain by using Online banking. Adjusted R-squared is 0.0004976, which is used for testing a goodness-of-fit. Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in this model. The adjusted R-squared increases only if the new predictors improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

8. Profit Models and Retention Model

```
# Profit Models
# Note: We omitted the summary for Online, Age, Inc and District and only kept the best one
# In question 7, lm_profit_Online_Billpay is our base model
lm_profit_Age = lm(X9Profit ~ factor(X9Age) + X9Online * factor(X9Age)
                   + X9Billpay * factor(X9Age), data = consumerDB)

lm_profit_Inc = lm(X9Profit ~ factor(X9Age) + factor(X9Inc)
                   + X9Online * factor(X9Age) + X9Online * factor(X9Inc)
                   + X9Billpay  *factor(X9Age) + X9Billpay * factor(X9Inc), data = consumerDB)

lm_profit_Tenure = lm(X9Profit ~ factor(X9Age) + factor(X9Inc)
                      + X9Online * factor(X9Age) + X9Online * factor(X9Inc)
                      + X9Billpay * factor(X9Age) + X9Billpay * factor(X9Inc)
                      + X9Tenure * factor(X9Age) + X9Tenure * factor(X9Inc), data = consumerDB)
lm_profit_District = lm(X9Profit ~ factor(X9Age) + factor(X9Inc) + factor(X9District)
                     + X9Online * factor(X9Age) + X9Online * factor(X9Inc)
                     + X9Tenure * factor(X9Age) + X9Tenure * factor(X9Inc)
                     + X9Billpay * factor(X9Age) + X9Billpay * factor(X9Inc),  data = consumerDB)
# District is the best model that we found, so the summary of District is presented here
summary(lm_profit_District)
```

```
##
## Call:
## lm(formula = X9Profit ~ factor(X9Age) + factor(X9Inc) + factor(X9District) +
##     X9Online * factor(X9Age) + X9Online * factor(X9Inc) + X9Tenure *
##     factor(X9Age) + X9Tenure * factor(X9Inc) + X9Billpay * factor(X9Age) +
##     X9Billpay * factor(X9Inc), data = consumerDB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -647.95 -138.45  -64.79   50.43 1995.38
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -40.8481    17.6102  -2.320  0.02037 *
## factor(X9Age)2            12.3771    17.3372   0.714  0.47529
## factor(X9Age)3            47.0050    16.8367   2.792  0.00524 **
## factor(X9Age)4            21.9674    16.1256   1.362  0.17312
## factor(X9Age)5            54.6974    17.6124   3.106  0.00190 **
## factor(X9Age)6            55.5712    18.5954   2.988  0.00281 **
## factor(X9Age)7           104.5463    18.2847   5.718 1.09e-08 ***
## factor(X9Inc)2            37.6414    17.3938   2.164  0.03047 *
## factor(X9Inc)3            25.2579    12.2989   2.054  0.04001 *
## factor(X9Inc)4            24.2390    12.8603   1.885  0.05947 .
## factor(X9Inc)5            34.3050    12.8593   2.668  0.00764 **
## factor(X9Inc)6            30.4342    10.0514   3.028  0.00246 **
## factor(X9Inc)7            79.1756    12.2245   6.477 9.51e-11 ***
## factor(X9Inc)8            78.0857    14.3440   5.444 5.25e-08 ***
## factor(X9Inc)9           164.4515    12.7580  12.890  < 2e-16 ***
## factor(X9District)1200    21.3721     5.1027   4.188 2.82e-05 ***
## factor(X9District)1300     9.7805     6.2480   1.565  0.11750
## X9Online                 -13.8521    32.9151  -0.421  0.67387
## X9Tenure                   0.5058     2.3843   0.212  0.83200
## X9Billpay                 17.7281   126.4302   0.140  0.88849
## factor(X9Age)2:X9Online   21.5903    28.1402   0.767  0.44295
## factor(X9Age)3:X9Online   29.8100    28.1253   1.060  0.28920
## factor(X9Age)4:X9Online   22.0115    27.3391   0.805  0.42075
## factor(X9Age)5:X9Online   51.2329    31.0653   1.649  0.09912 .
## factor(X9Age)6:X9Online   19.0992    37.7698   0.506  0.61309
## factor(X9Age)7:X9Online   -9.4176    39.5705  -0.238  0.81189
## factor(X9Inc)2:X9Online  -36.0803    41.8709  -0.862  0.38886
## factor(X9Inc)3:X9Online  -10.7654    29.0258  -0.371  0.71072
## factor(X9Inc)4:X9Online  -20.0702    29.6372  -0.677  0.49829
## factor(X9Inc)5:X9Online  -28.6879    28.8233  -0.995  0.31960
## factor(X9Inc)6:X9Online   -8.6970    24.8699  -0.350  0.72657
## factor(X9Inc)7:X9Online   11.6428    27.1460   0.429  0.66800
## factor(X9Inc)8:X9Online   12.4175    29.7785   0.417  0.67669
## factor(X9Inc)9:X9Online   -3.0760    27.1254  -0.113  0.90971
## factor(X9Age)2:X9Tenure    3.5996     2.5492   1.412  0.15795
## factor(X9Age)3:X9Tenure    4.1527     2.4177   1.718  0.08587 .
## factor(X9Age)4:X9Tenure    6.0760     2.3585   2.576  0.00999 **
## factor(X9Age)5:X9Tenure    4.3673     2.3854   1.831  0.06713 .
## factor(X9Age)6:X9Tenure    5.8101     2.3944   2.427  0.01525 *
## factor(X9Age)7:X9Tenure    4.9452     2.3830   2.075  0.03798 *
## factor(X9Inc)2:X9Tenure   -3.1475     1.1812  -2.665  0.00771 **
## factor(X9Inc)3:X9Tenure   -1.4141     0.8894  -1.590  0.11186
## factor(X9Inc)4:X9Tenure   -1.2377     0.8982  -1.378  0.16825
## factor(X9Inc)5:X9Tenure   -1.4041     0.9181  -1.529  0.12619
## factor(X9Inc)6:X9Tenure   -0.3268     0.7501  -0.436  0.66312
## factor(X9Inc)7:X9Tenure   -1.9234     0.8883  -2.165  0.03038 *
## factor(X9Inc)8:X9Tenure   -0.3869     1.0336  -0.374  0.70818
## factor(X9Inc)9:X9Tenure   -2.0723     0.8908  -2.326  0.02001 *
## factor(X9Age)2:X9Billpay  -6.3746   109.4344  -0.058  0.95355
## factor(X9Age)3:X9Billpay   0.9050   108.9342   0.008  0.99337
## factor(X9Age)4:X9Billpay  22.9973   108.3206   0.212  0.83187
## factor(X9Age)5:X9Billpay  78.0622   114.9151   0.679  0.49695
## factor(X9Age)6:X9Billpay  78.7168   125.3504   0.628  0.53003
## factor(X9Age)7:X9Billpay  79.8174   127.2477   0.627  0.53049  
## factor(X9Inc)2:X9Billpay  49.9158   183.8667   0.271  0.78602  
## factor(X9Inc)3:X9Billpay   3.7280    82.8194   0.045  0.96410  
## factor(X9Inc)4:X9Billpay  53.1964    90.9915   0.585  0.55880  
## factor(X9Inc)5:X9Billpay  12.4437    83.3989   0.149  0.88139  
## factor(X9Inc)6:X9Billpay  15.6217    71.7517   0.218  0.82765  
## factor(X9Inc)7:X9Billpay -19.9543    77.4343  -0.258  0.79665  
## factor(X9Inc)8:X9Billpay  43.9907    82.9948   0.530  0.59609  
## factor(X9Inc)9:X9Billpay 156.2135    74.6981   2.091  0.03651 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 263.3 on 31572 degrees of freedom
## Multiple R-squared:  0.07066,    Adjusted R-squared:  0.06886
## F-statistic: 39.35 on 61 and 31572 DF,  p-value: < 2.2e-16
```

Our best model for profitability prediction is the District model. The summary for this regression is presented.The intercept, 1999Age, Income level, District 1200, and tenure (from Age level 3-7) are significant. Analysis for coefficient,

For example, when a customer starts to use online banking, the expected 1999Profit will be -54.6 (intercept), when others are all zero and when hold other variable constant. Using the coefficient of factor(X9Age)4:X9Tenure as an example, 6.0760 means that the profit difference between age group 1 and age group 4 is 6.076, if we hold other constant. Among those statistically insignificant coefficient, we will interpret the significant one. For example, factor(X9Inc)2:X9Online =-36.0803. It means the profit difference between a customer in Income Level 2 with online banking account and a customer in income level 1 with online banking is -36.08, eliminating other effects on age, distrct, billpay, or tenure.
The R-square is 0.07066, which means that 7.07% of 1999Profit can be explained by the variations in x variables.

(4) Major Statistics Summary

```
set.seed(12345678)
# Creating Trainning, Validation and Test Sets
randOrder = order(runif(nrow(consumerDB)))
training.data = subset(consumerDB,randOrder < .9 * nrow(consumerDB))
validation.data =  subset(consumerDB,randOrder >= .9*nrow(consumerDB) &
                            randOrder < .95*nrow(consumerDB))
test.data = subset(consumerDB,randOrder >= .95 * nrow(consumerDB))
```

We first created a subset by randomly choosing 5% from the original 1999 year's data points. This is used as the validation for the following model we created. 5% of the data will be the test dataset, while the 90% of the data will be the training dataset.

```
# Prediction errors among different profit models
# Model lm_profit_Online and Billpay
predicted.profit1 = predict(lm_profit_Online_Billpay, validation.data)
prediction.error1 = sqrt(mean((predicted.profit1-validation.data$X9Profit)^2))
# Model lm_profit_Age
predicted.profit2 = predict(lm_profit_Age, validation.data)
prediction.error2 = sqrt(mean((predicted.profit2-validation.data$X9Profit)^2))
# Model lm_profit_Inc
predicted.profit3 = predict(lm_profit_Inc, validation.data)
prediction.error3 = sqrt(mean((predicted.profit3-validation.data$X9Profit)^2))
# Model lm_profit_Tenure
```

```
predicted.profit4 = predict(lm_profit_Tenure, validation.data)
prediction.error4 = sqrt(mean((predicted.profit4-validation.data$X9Profit)^2))
# Model lm_profit_District
predicted.profit5 = predict(lm_profit_District, validation.data)
prediction.error5 = sqrt(mean((predicted.profit5-validation.data$X9Profit)^2))
```

Comparison Table for Profit Models

```
# Creating a comparison table for profit models
comparison.table.profit = matrix(c(summary(lm_profit_Online_Billpay)$adj.r.square,
            AIC(lm_profit_Online_Billpay), BIC(lm_profit_Online_Billpay), prediction.error1,
            summary(lm_profit_Age)$adj.r.square, AIC(lm_profit_Age), BIC(lm_profit_Age),
            prediction.error2,
            summary(lm_profit_Inc)$adj.r.square, AIC(lm_profit_Inc), BIC(lm_profit_Inc),
            prediction.error3,
            summary(lm_profit_Tenure)$adj.r.square, AIC(lm_profit_Tenure), BIC(lm_profit_Tenure),
            prediction.error4,
            summary(lm_profit_District)$adj.r.square, AIC(lm_profit_District),
            BIC(lm_profit_District), prediction.error5), nrow = 5, ncol = 4, byrow = TRUE)
comparison.table.profit = round(comparison.table.profit,6)
colnames(comparison.table.profit) = c("Adj.r.square", "AIC", "BIC", "Prediction Error")
rownames(comparison.table.profit) = c("lm_profit_Online_Billpay", "lm_profit_Age",
            "lm_profit_Inc", "lm_profit_Tenure", "lm_profit_District")
comparison.table.profit
```

```
##                          Adj.r.square      AIC      BIC Prediction Error
## lm_profit_Online_Billpay     0.001597 444590.7 444624.1         284.2532
## lm_profit_Age                0.022323 443945.1 444129.0         280.9747
## lm_profit_Inc                0.047833 443132.7 443517.3         277.4723
## lm_profit_Tenure             0.068291 442460.5 442970.6         275.7149
## lm_profit_District           0.068864 442443.1 442969.9         275.8065
```

We compared the adjusted R-square, AIC, BIC, and Prediction Error. It can tell that the District Model has the highest adjusted R-square (0.068864), lowest AIC (442443.1) and lowest BIC (442969.9), although the prediction error of District is slightly higher that of Tenure model. The lower the AIC, BIC, and Prediction Error, the better the model. The District model has the highest adjusted R square, implying that we chose the numbers of predictors properly. We concluded that the District Profit Model fit the validation subset the best, and so it is the most appropriate Profit Model.

```
# Retention Models
# Note: We omitted the summary for Billpay, Age, and Inc and we only kept the best one.
# In question 7, lm_retention_Online_Billpay is our base model

lm_retention_Age = lm(retention ~ factor(X9Age) + X9Online * factor(X9Age) +
                        X9Billpay * factor(X9Age), data = consumerDB)

lm_retention_Inc = lm(retention ~ factor(X9Age) + factor(X9Inc)
                    + X9Online * factor(X9Age) + X9Online * factor(X9Inc)
                    + X9Billpay  *factor(X9Age) + X9Billpay * factor(X9Inc), data = consumerDB)

lm_retention_Tenure = lm(retention ~ factor(X9Age) + factor(X9Inc)
                       + X9Online * factor(X9Age) + X9Online * factor(X9Inc)
                       + X9Billpay * factor(X9Age) + X9Billpay * factor(X9Inc)
                       + X9Tenure * factor(X9Age) + X9Tenure * factor(X9Inc), data = consumerDB)
```

```
lm_retention_District = lm(retention ~ factor(X9Age) + factor(X9Inc) + factor(X9District)
                      + X9Online * factor(X9Age) + X9Online * factor(X9Inc)
                      + X9Billpay * factor(X9Age) + X9Billpay * factor(X9Inc)
                      + X9Tenure * factor(X9Age) + X9Tenure * factor(X9Inc), data = consumerDB)

summary(lm_retention_District)
```

```
##
## Call:
## lm(formula = retention ~ factor(X9Age) + factor(X9Inc) + factor(X9District) +
##     X9Online * factor(X9Age) + X9Online * factor(X9Inc) + X9Billpay *
##     factor(X9Age) + X9Billpay * factor(X9Inc) + X9Tenure * factor(X9Age) +
##     X9Tenure * factor(X9Inc), data = consumerDB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05965  0.02240  0.07293  0.19917  0.43876
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.8201421  0.0232908  35.213  < 2e-16 ***
## factor(X9Age)2             0.0764454  0.0229296   3.334 0.000857 ***
## factor(X9Age)3             0.0932870  0.0222677   4.189 2.81e-05 ***
## factor(X9Age)4            -0.1094644  0.0213273  -5.133 2.87e-07 ***
## factor(X9Age)5             0.0824916  0.0232937   3.541 0.000399 ***
## factor(X9Age)6             0.0864089  0.0245938   3.513 0.000443 ***
## factor(X9Age)7             0.0999539  0.0241828   4.133 3.59e-05 ***
## factor(X9Inc)2             0.0036814  0.0230045   0.160 0.872858
## factor(X9Inc)3             0.0007333  0.0162661   0.045 0.964042
## factor(X9Inc)4             0.0040464  0.0170087   0.238 0.811957
## factor(X9Inc)5             0.0271762  0.0170073   1.598 0.110073
## factor(X9Inc)6            -0.1511221  0.0132937 -11.368  < 2e-16 ***
## factor(X9Inc)7             0.0310692  0.0161677   1.922 0.054655 .
## factor(X9Inc)8             0.0333122  0.0189710   1.756 0.079105 .
## factor(X9Inc)9             0.0538892  0.0168733   3.194 0.001406 **
## factor(X9District)1200     0.0204469  0.0067487   3.030 0.002449 **
## factor(X9District)1300     0.0099356  0.0082634   1.202 0.229235
## X9Online                  -0.0425715  0.0435325  -0.978 0.328120
## X9Billpay                 -0.2829182  0.1672127  -1.692 0.090662 .
## X9Tenure                   0.0034119  0.0031534   1.082 0.279266
## factor(X9Age)2:X9Online   -0.0244849  0.0372174  -0.658 0.510616
## factor(X9Age)3:X9Online   -0.0364836  0.0371977  -0.981 0.326697
## factor(X9Age)4:X9Online    0.0347862  0.0361578   0.962 0.336024
## factor(X9Age)5:X9Online   -0.0033114  0.0410861  -0.081 0.935763
## factor(X9Age)6:X9Online   -0.0440221  0.0499532  -0.881 0.378180
## factor(X9Age)7:X9Online   -0.0222598  0.0523348  -0.425 0.670596
## factor(X9Inc)2:X9Online    0.0327681  0.0553773   0.592 0.554040
## factor(X9Inc)3:X9Online    0.0271961  0.0383887   0.708 0.478678
## factor(X9Inc)4:X9Online    0.0590690  0.0391972   1.507 0.131828
## factor(X9Inc)5:X9Online    0.0529981  0.0381208   1.390 0.164458
## factor(X9Inc)6:X9Online    0.0903105  0.0328921   2.746 0.006042 **
## factor(X9Inc)7:X9Online    0.0561011  0.0359025   1.563 0.118157
## factor(X9Inc)8:X9Online    0.0908486  0.0393842   2.307 0.021076 *
## factor(X9Inc)9:X9Online   0.0595998  0.0358752   1.661 0.096662 .
## factor(X9Age)2:X9Billpay  0.1676330  0.1447346   1.158 0.246787
## factor(X9Age)3:X9Billpay  0.0999068  0.1440731   0.693 0.488035
## factor(X9Age)4:X9Billpay  0.0750757  0.1432615   0.524 0.600250
## factor(X9Age)5:X9Billpay  0.1124512  0.1519833   0.740 0.459371
## factor(X9Age)6:X9Billpay  0.1509187  0.1657846   0.910 0.362656
## factor(X9Age)7:X9Billpay  0.2197129  0.1682940   1.306 0.191722
## factor(X9Inc)2:X9Billpay  0.1000287  0.2431766   0.411 0.680825
## factor(X9Inc)3:X9Billpay  0.1992932  0.1095345   1.819 0.068851 .
## factor(X9Inc)4:X9Billpay  0.2004510  0.1203426   1.666 0.095789 .
## factor(X9Inc)5:X9Billpay  0.1131664  0.1103009   1.026 0.304909
## factor(X9Inc)6:X9Billpay  0.1642711  0.0948966   1.731 0.083452 .
## factor(X9Inc)7:X9Billpay  0.1966571  0.1024123   1.920 0.054836 .
## factor(X9Inc)8:X9Billpay  0.0414830  0.1097664   0.378 0.705492
## factor(X9Inc)9:X9Billpay  0.1678097  0.0987935   1.699 0.089406 .
## factor(X9Age)2:X9Tenure  -0.0037732  0.0033715  -1.119 0.263079
## factor(X9Age)3:X9Tenure  -0.0010763  0.0031975  -0.337 0.736412
## factor(X9Age)4:X9Tenure   0.0038403  0.0031193   1.231 0.218279
## factor(X9Age)5:X9Tenure  -0.0013578  0.0031548  -0.430 0.666922
## factor(X9Age)6:X9Tenure  -0.0016894  0.0031668  -0.533 0.593714
## factor(X9Age)7:X9Tenure  -0.0023745  0.0031517  -0.753 0.451223
## factor(X9Inc)2:X9Tenure  -0.0006993  0.0015622  -0.448 0.654424
## factor(X9Inc)3:X9Tenure  -0.0004027  0.0011763  -0.342 0.732066
## factor(X9Inc)4:X9Tenure   0.0002006  0.0011880   0.169 0.865917
## factor(X9Inc)5:X9Tenure  -0.0008362  0.0012143  -0.689 0.491028
## factor(X9Inc)6:X9Tenure   0.0032766  0.0009921   3.303 0.000958 ***
## factor(X9Inc)7:X9Tenure  -0.0007066  0.0011749  -0.601 0.547564
## factor(X9Inc)8:X9Tenure  -0.0015622  0.0013670  -1.143 0.253115
## factor(X9Inc)9:X9Tenure  -0.0014272  0.0011782  -1.211 0.225772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3482 on 31572 degrees of freedom
## Multiple R-squared:  0.1216, Adjusted R-squared:  0.1199
## F-statistic: 71.66 on 61 and 31572 DF,  p-value: < 2.2e-16
```

First we create a new binary variable called retention. If retention is 0, it means the customers leave the bank. If retention is 1, customers stay in the bank.

Finally, we add Aga, Inc, Tenure, District along with their interactive terms into the regression. The term intercept is significant. It represent the probability of retention in the base case, where customer is in age group 1, income group 1, and district 1100 and variable Tenure equals to zero. And when a customer starts to use billpay service, holding other base case variable unchanged, the intercept will decrease by 0.28. It means that the probability of retention will decrease by 28%. The coefficient of factor(X9Inc)6:X9Tenure (0.0032) is the difference of the Tenure effect on probability of retention between groupA (income level1& Age level1& do not use online or billpay service, the base case) the groupB (income level6 & Age level1 & do not use online or billpay service). The R-square is 0.1216, which means that 12.16% of Profit can be explained by the variations in Online banking, Age bucket, Income level bucket, online service, billpay service and Tenure.

```r
# Prediction errors among different retention models
# Model lm_retention_Online_Billpay
predicted.retention1 = predict(lm_retention_Online_Billpay, validation.data)
prediction.error.retention1 = sqrt(mean((predicted.retention1-validation.data$retention)^2))
# Model lm_retention_Age
predicted.retention2 = predict(lm_retention_Age, validation.data)
```

```
prediction.error.retention2 = sqrt(mean((predicted.retention2-validation.data$retention)^2))
# Model lm_retention_Inc
predicted.retention3 = predict(lm_retention_Inc, validation.data)
prediction.error.retention3 = sqrt(mean((predicted.retention3-validation.data$retention)^2))
# Model lm_retention_Tenure
predicted.retention4 = predict(lm_retention_Tenure, validation.data)
prediction.error.retention4 = sqrt(mean((predicted.retention4-validation.data$retention)^2))
# Model lm_retention_District
predicted.retention5 = predict(lm_retention_District, validation.data)
prediction.error.retention5 = sqrt(mean((predicted.retention5-validation.data$retention)^2))
```

We calculated the predicted errors for all retention models, and the prediction error for our final retention Model is 0.347791 To compare the errors from more perspectives, we made the following table.

Comparison Table for Retention Models

```
# Creating a comparison table for retention models
comparison.table.retention = matrix(c(summary(lm_retention_Online_Billpay)$adj.r.square,
              AIC(lm_retention_Online_Billpay), BIC(lm_retention_Online_Billpay),
              prediction.error.retention1,
              summary(lm_retention_Age)$adj.r.square, AIC(lm_retention_Age),
              BIC(lm_retention_Age), prediction.error.retention2,
              summary(lm_retention_Inc)$adj.r.square, AIC(lm_retention_Inc),
              BIC(lm_retention_Inc), prediction.error.retention3,
              summary(lm_retention_Tenure)$adj.r.square, AIC(lm_retention_Tenure),
              BIC(lm_retention_Tenure), prediction.error.retention4,
              summary(lm_retention_District)$adj.r.square, AIC(lm_retention_District),
              BIC(lm_retention_District), prediction.error.retention5), nrow = 5,
              ncol = 4, byrow = TRUE)

comparison.table.retention = round(comparison.table.retention,6)
colnames(comparison.table.retention) = c("Adj.r.square", "AIC", "BIC", "Prediction Error")
rownames(comparison.table.retention) = c("lm_retention_Online", "lm_retention_Age",
         "lm_retention_Inc", "lm_retention_Tenure", "lm_retention_District")
comparison.table.retention
```

```
##                       Adj.r.square      AIC      BIC Prediction Error
## lm_retention_Online       0.000498 27057.89 27091.33         0.367540
## lm_retention_Age          0.071509 24744.56 24928.53         0.355896
## lm_retention_Inc          0.100339 23770.70 24155.35         0.350401
## lm_retention_Tenure       0.119673 23098.44 23608.52         0.347958
## lm_retention_District     0.119923 23091.47 23618.28         0.347791
```

We compared the adjusted R-square, AIC, BIC, and Prediction Error. It shows that the our District Online Retention Model has the lowest AIC of 23091.48, greatest Adj.r.square 0.119923, and lowest Prediction Error of 0.347791. Although BIC is slightly higher than the previous model, we would like to choose lm_retention_district as our final prediction model.

9. Predicted likelihood of retention

The following code predicts the likelihood of retention in year 2000.

```r
Prediction_Retention_1999 = predict(lm_retention_District, consumerDB, se.fit = TRUE)
hit_rate = data.frame (Prediction_Retention_1999$fit, consumerDB$retention)
head(hit_rate)
```

```
##   Prediction_Retention_1999.fit consumerDB.retention
## 1                     0.6466501                    0
## 2                     0.9666658                    1
## 3                     0.9690903                    1
## 4                     0.6036924                    0
## 5                     0.9531999                    1
## 6                     0.7373059                    1
```

```r
hit_rate$predicted_retention_sorted = 0
hit_rate[hit_rate$Prediction_Retention_1999.fit >= 0.5,]$predicted_retention_sorted = 1
hit_rate$hit_times = 0
hit_rate[hit_rate$consumerDB.retention == hit_rate$predicted_retention_sorted,]$hit_times = 1
# create a table with percentage
round(table(hit_rate$hit_times)/sum(table(hit_rate$hit_times)),3)
```

```
##
##     0     1
## 0.165 0.835
```

Since the our retention prediction model offers the possibility of a customer's status. If the possibility is greater than .5, we identify this customers as they will stay with the bank; while possibly smaller than .5 as this customer will leave the bank. Then we compare this to the actual data in 2000 year, and we found out that we made the correct prediction for 84% of the dataset.

10. Customer profitability

Database is a subset that only contains 26396 observations and 14 variables. It's a subset from the original dataset. According to the request, Database2000 only contains those observations that don't have any missing data in 2000 Online, 2000 profit and 2000 billpay.

```r
Database2000 = consumerDB[,c(4,5,6,7,8,9,11,12)]
Database2000 = Database2000[!is.na(Database2000$X0Online)&!is.na(Database2000$X0Profit)
                           &!is.na(Database2000$X0Billpay),]
Database2000$X9Tenure=Database2000$X9Tenure+1
colnames(Database2000)=c("X9Age","X9Inc","X9Tenure", "X9District","X0Profit",
                         "X9Online", "X9Billpay","retention")

Prediction_Profit_2000 = predict(lm_profit_Tenure, Database2000)
Database2000$predicted_profit = Prediction_Profit_2000
prediction.error.profit = sqrt(mean((Prediction_Profit_2000-Database2000$X0Profit)^2))
prediction.error.profit
```

```
## [1] 382.795
```

To test the fitness of our Profit Model and Retention Model, we extracted all the datapoints in year 2000 and the income, age, and district datapoint from year 1999 as the base dataset. Then we used the Profit Model and Retention Model we designed to predict the Profit and Retention Status in year 2000. Comparing the predicted Profit and Retention Status to the actual Profit and Retention Status in 2000.

According to the result, the prediction error of our Profit Model is 382.795. Meanwhile, 84% of the prediction is correct for our Retention Model, which is pretty accurate. In general, our Profit Model and Retention Model could fairly predict the profit and retention status of Pilgrim Banks' customers.

11. Standard error of prediction

Using the standard error of prediction from the above prediction analysis, construct the 95% confidence interval for each customer's predicted profitability. Assuming Pilgrim Bank's managers know customers' enrollment status at the start of 2000.

```
predicted.profit = predict(lm_profit_Tenure, test.data, se.fit = TRUE)
test.data = test.data[,c(4,5,6,7,8,9,11,12)]
colnames(test.data)=c("X9Age","X9Inc","X9Tenure", "X9District","X0Profit",
                         "X9Online", "X9Billpay","retention")
test.data$predicted.profit = predicted.profit$fit
test.data$se.fit = predicted.profit$se.fit
# Upper limit of confidence interval for each predicted y
test.data$upper.limit = test.data$predicted.profit + qnorm(0.975)*test.data$se.fit
head(test.data$upper.limit)
```

```
## [1] 130.26831 226.23922 279.17999  60.73635 133.70830 265.51770
```

```
# Lower limit of confidence interval for each predicted y
test.data$lower.limit = test.data$predicted.profit - qnorm(0.975)*test.data$se.fit
head(test.data$lower.limit)
```

```
## [1] 100.35696 199.41186 245.31252  48.76199 106.75223 219.03291
```

```
# Upper limit on total profits is the sum of upper limits for individual profits
TotalProfitUpperLimit = sum(test.data$upper.limit)
TotalProfitUpperLimit
```

```
## [1] 205038.8
```

```
# Lower limit on total profits is the sum of lower limits for individual profits
TotalProfitlowerLimit = sum(test.data$lower.limit)
TotalProfitlowerLimit
```

```
## [1] 147825.5
```

In order to estimate the total profit range, we run our Profit Model and extracted 95% of the dataset's profit of the lower and upper limit. After we added all the predicted upper limited, we got the highest total profit of 205038.8; and in the same way, we got the lowest total profit of 147825.5 So the range of 2000's profit is from 147825.5 to 205038.8

12. Summarization and Recommendation

Our regression of Profit Model and Retention Model shows that when people using online banking, the profit and the retention tend to increase, except for the base cases in different models. It shows that online customers are indeed better customers, and the senior management should offer rebates or lower the service charges for customers using online banking, as these customers will have relative higher possibility to stay with the bank and bring more profit to Pilgrim Bank.

To promote a more efficient recommendation to increase the profitability, we tend to focus more on a very specific groups - customers with high profitability but less likely to stay in the bank in 2001. Here is what we did:

```
profitSorted = consumerDB[!is.na(consumerDB$X0Profit),]
profitMedian = median(profitSorted$X0Profit)
highProfit = profitSorted[profitSorted$X0Profit >= profitMedian,]
highProfit = highProfit[,c(4:9,11,12)]
colnames(highProfit)=c("X9Age","X9Inc","X9Tenure", "X9District","X0Profit",
                       "X9Online", "X9Billpay","retention")
predict_retention_2001 = predict(lm_retention_District, highProfit, se.fit = TRUE)
highProfit$fit = predict_retention_2001$fit
targetCustomer = highProfit[highProfit$fit <= quantile(highProfit$fit, .25),]
length(targetCustomer$fit)
```

```
## [1] 3315
```

```
targetProfitPercent = sum(targetCustomer$X0Profit)/sum(profitSorted$X0Profit)
targetProfitPercent
```

```
## [1] 0.2564424
```

```
targetCustomerPercent = length(targetCustomer$fit)/length(profitSorted$X0Profit)
targetCustomerPercent
```

```
## [1] 0.1255872
```

After the segmentation, customers in our target segments represent 13% of the customers who stay with the bank in 2000 and 26% of the total profit in 2000. To make them more profitable for the bank, we would like to use our tired services. For example, we can offer discounts on mortgage rates or higher interest rates on certificates of deposit. Tired service like are aiming to increase them retention in the bank, which, consequently, make them more profitable for the bank.