Pilgrim case study

Yecheng Li, Doris Long, Vera Wang, Jikun Zhou

November 17, 2017

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.

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

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 are incomplete and missed one or more information in 'Age', 'Income', or 'Billpay.

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.

```
# Read the given dataset
consumerDB = read.csv("dataset.csv") ### read the given dataset
# Check who stay with the bank in 2000: 1 mean stay with the bank, 0 mean
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
```

```
# Check who didn't left the bank in 2000, and the income or age in 1999 were missing
consumerDB$fixAge = (consumerDB$retention==1) & is.na(consumerDB$X9Age)
consumerDB$fixIncome = (consumerDB$retention==1) & is.na(consumerDB$X9Inc)
# For "fixAge"== TRUE, we substitute "NA" to be "4", which is the median
# For "fixIncome"== TRUE, we substitute "NA" to be "6", which is the median
consumerDB[consumerDB$fixAge,]$X9Age = 4
consumerDB[consumerDB$fixIncome,]$X9Inc = 6
```

```
# Sort the concumerDB and get a Table that sepcifically contains data for 1999
statsTable1999= consumerDB[,2:6]
X9Billpay = consumerDB[,10]
statsTable1999= cbind(statsTable1999,X9Billpay)
```

4. Major Takeaways from Interim Deliverable-I

(1) Statistics Summary for 1999 Data

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.

```
Summary_Table=t(describe(statsTable1999))
Summary_Table = round(Summary_Table,2)
# This summary gives the mean, median, standard deviation, min, max and range
# for 1999 Profit, Age, Income, Online, Billpay, and Tenure
Summary_Table_New = Summary_Table[c(3:5,8:10),c(1:6)]
Summary_Table_New
```

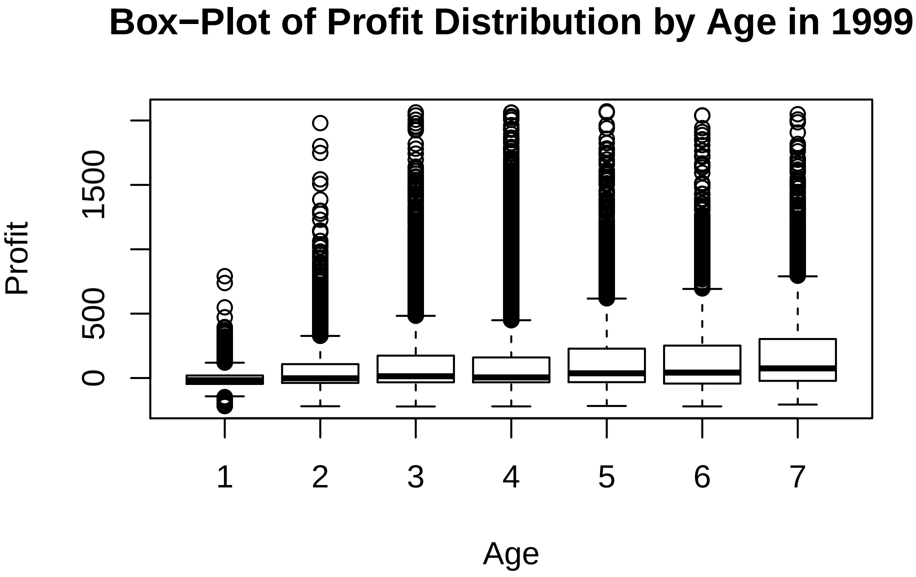
```
##        X9Profit X9Online X9Age X9Inc X9Tenure X9Billpay
## mean     111.50     0.12  4.04  5.55    10.16      0.02
## sd       272.84     0.33  1.49  2.15     8.45      0.13
## median     9.00     0.00  4.00  6.00     7.41      0.00
## min     -221.00     0.00  1.00  1.00     0.16      0.00
## 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
```

(2) Graphic 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'.

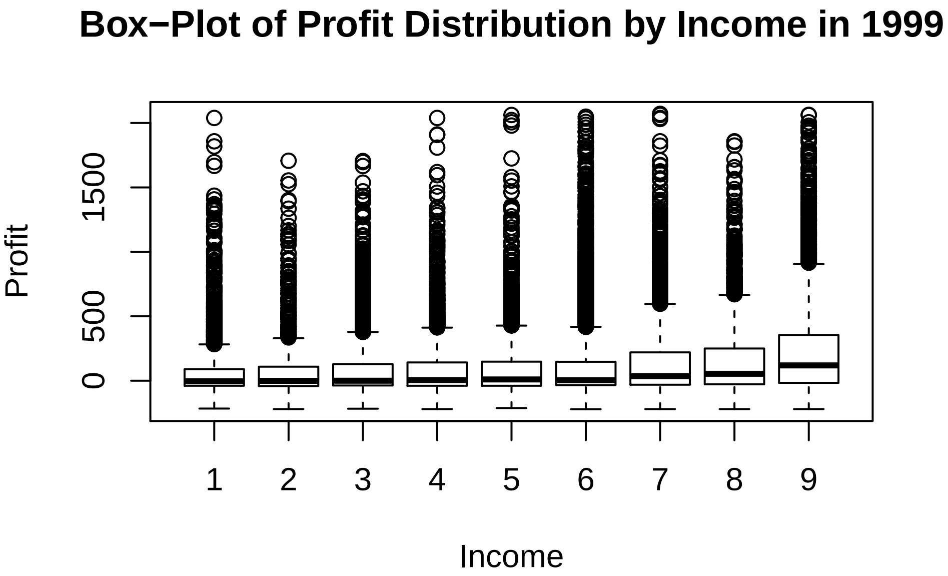
```
# This is a boxplot graph for Profit& Age
boxplot(X9Profit~X9Age, data = consumerDB,
        main = "Box-Plot of Profit Distribution by Age in 1999",
        xlab = "Age", ylab = "Profit") ### Sets X and Y Axes
```



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.

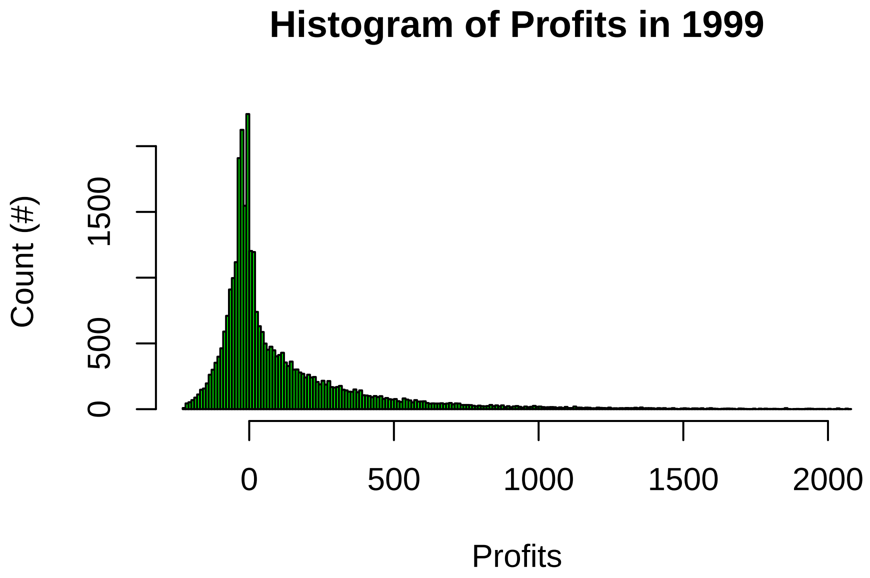
```
# This is a boxplot graph for Profit& Income
boxplot(X9Profit~X9Inc, data = consumerDB,
        main = "Box-Plot of Profit Distribution by Income in 1999",  ## Sets Title to Plot
        xlab = "Income", ylab = "Profit") ### Sets X and Y Axes
```



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) Summary data of income and age

```
# 1999 Income with online and billpay customer
summary_Income_1999 = table(consumerDB$X9Inc,consumerDB$X9Online)
# summary_Income_1999   Note: We only show the percentage of different income level here.
summary_Income_1999_2 = round(summary_Income_1999/rowSums(summary_Income_1999),2)
summary_Income_1999_2
```

```
##
##        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 0.88 0.12
##   7 0.85 0.15
##   8 0.84 0.16
##   9 0.82 0.18
```

```
# 1999 Age with online and billpay customer
summary_Age_1999 = table(consumerDB$X9Age,consumerDB$X9Online)
# summary_Age_1999      Note: We only show the percentage of online and billpay customers here
summary_Age_1999_2 = round(summary_Age_1999/rowSums(summary_Age_1999),2)
summary_Age_1999_2
```

```
##
##        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
```

(4) Major Statistics Summary

```
# Sort the table and ignore those data points that miss 1999 income or age value
# Name a new consumer Database "consumerDB2"
consumerDB2_temp1=consumerDB[!is.na(consumerDB$X9Inc),]
consumerDB2_temp2=consumerDB2_temp1[!is.na(consumerDB2_temp1$X9Age),]
consumerDB2 = consumerDB2_temp2[!is.na(consumerDB2_temp2$X0Online)|!is.na(consumerDB2_temp2$X0Profit)
                                |!is.na(consumerDB2_temp2$X0Billpay),]
```

(5) Data Patterns Summary

From the table '1999 Income with Online and Billpay', larger percentage of customers with higher income level use online banking than those with 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 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’, customer with Income level 6 ($50,000 - $74,999) had most online uses and electronic bill pay uses. However, from ‘Box-Plot of Profit Distribution by Income Cont’, customers with income level 6 generated a medium profit near zero, which is low compared to the level 5. Therefore, the group of customers who 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 Income in 1999’ shows that the same group of people generated a relatively low profit for the bank.

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.

5. Mean profitability of years 1999 and 2000 customers using online banking or electronic billpay or not

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 proforbility of 1999’s customer using online banking or not. Null hypothesis: Mean profit for year 1999’s custumers using online banking = Mean profit for year 1999’s custumers not using online banking
Alternative Hypothesis: Mean profit for year 1999’s custumers using online banking != Mean profit for year 1999’s custumers 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$X9Online == 1, ]$X9P
## 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 indepedent 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 custumers using online banking and mean profit for year 1999’s custumers not using online banking. That is to say, using online banking in 1999 did not have significant effect on customers’ profit.

(2) Compare the proforbility of 1999’s customer using electronic billpay or not. Null hypothesis: Mean profit for year 1999’s custumers using electronic billpay = Mean profit for year 1999’s custumers not using electronic billpay Alternative Hypothesis: Mean profit for year 1999’s custumers using electronic billpay != Mean profit for year 1999’s custumers 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$X9Billpay == 1, ]$
## 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 indepedent 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 custumers using electronic billpay and mean profit for year 1999's custumers not using electronic billpay. That is to say, using electronic billpay in 1999 did have significant effect on customers' profit.

(3) Compare the proforbility of 2000's customer using online banking or not. Null hypothesis: Mean profit for year 2000's custumers using online banking = Mean profit for year 2000's custumers not using online banking
Alternative Hypothesis: Mean profit for year 2000's custumers using online banking != Mean profit for year 2000's custumers 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.

(4) Compare the proforbility of 2000's customer using electronic billpay or not. Null hypothesis: Mean profit for year 2000's custumers using electronic billpay = Mean profit for year 2000's custumers not using electronic billpay Alternative Hypothesis: Mean profit for year 2000's custumers using electronic billpay != Mean profit for year 2000's custumers 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$X9Billpay == 1, ]$
## 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 indepedent 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 custumers using electronic billpay and mean profit for year 2000's custumers not using electronic billpay. That is to say, using electronic billpay in 2000 did have significant effect on customers' profit.

Use of the online channel effect on the regression analysis.

```r
lm_profit = lm(X9Profit ~ X9Billpay, data = consumerDB)
summary(lm_profit)
```

```
##
## Call:
## lm(formula = X9Profit ~ X9Billpay, data = consumerDB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415.41 -144.08 -101.08   51.92 1960.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  110.078      1.546  71.213  < 2e-16 ***
## X9Billpay     85.329     11.965   7.132 1.01e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.6 on 31632 degrees of freedom
## Multiple R-squared:  0.001605,   Adjusted R-squared:  0.001574
## F-statistic: 50.86 on 1 and 31632 DF,  p-value: 1.013e-12
```

A simple linear regression was calculated to see the the use of online channel has any effect on profit in year 2009.The difference between not using the online billpay and using the online billpay increase by 85.329 on profit. When customers do not use online billpay service, the estimated profit mean is 110.078. There is a significant regression equation was found (p= 1.013e-12 < 0.05).That is saying, there is significant difference on customer profitability bewtween the use of the online channel and not using the online billpay. That is to say, customer using online billpay service generate 85.329 more profit compared to customers not using billpay.

```r
lm_retention = lm(X9Tenure ~ X9Billpay, data = consumerDB)
summary(lm_retention)
```

```
##
## Call:
## lm(formula = X9Tenure ~ X9Billpay, data = consumerDB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.021  -6.431  -2.771   4.569  30.979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.18056    0.04793 212.417  < 2e-16 ***
## X9Billpay   -1.07044    0.37097  -2.885  0.00391 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.453 on 31632 degrees of freedom
## Multiple R-squared:  0.0002631,  Adjusted R-squared:  0.0002315
## F-statistic: 8.326 on 1 and 31632 DF,  p-value: 0.003911
```

A simple linear regression was calculated to see the the use of online channel has any effect on tenure in year 2009.The difference between not using the online billpay and using the online billpay could decrease the average tenure length by 1.07044. When customers do not use online billpay service, the estimated tenure mean is 10.18 years. There is a significant regression equation was found (p= 0.0039 < 0.05).That is saying, there is significant difference on tenure years bewtween the use of the online channel and not using the online billpay. That is to say, the average tenure lenght for customer using online billpay service is 1.07 shorter than compared to customers not using billpay.

7. Profit Models and Retention Model

```
### Creating Trainning, Validation and Test Sets
randOrder = order(runif(nrow(consumerDB2)))
training.data = subset(consumerDB2,randOrder < .9 * nrow(consumerDB2))
validation.data =  subset(consumerDB2,randOrder>=.85*nrow(consumerDB2)&randOrder<.95*nrow(consumerDB2))
```

We first created a subset by randomly choosing 10% from the orginal 1999 year's datapoints. This is used as the validation for the following model we created.

```
# Profit Models
# Note: We omitted the summary for Billpay, Age, Inc, Tenure and only kept the best one.
lm_profit_Billpay = lm(X9Profit ~ X9Billpay, data = training.data)
lm_profit_Age = lm(X9Profit ~ factor(X9Age) + X9Billpay * factor(X9Age), data = training.data)
lm_profit_Inc = lm(X9Profit ~ factor(X9Age) + factor(X9Inc) + X9Billpay * factor(X9Age)
                   + X9Billpay * factor(X9Inc), data = training.data)
lm_profit_Tenure = lm(X9Profit ~ factor(X9Age) + factor(X9Inc) + X9Billpay * factor(X9Age)
                      + X9Billpay * factor(X9Inc) + X9Tenure * factor(X9Age) +
                        X9Tenure * factor(X9Inc), data = training.data)
lm_profit_final = lm(X9Profit ~ factor(X9Age) + factor(X9Inc) + factor(X9District) +
                       X9Billpay * factor(X9Age) + X9Billpay * factor(X9Inc) +
                       X9Billpay * factor(X9District) + X9Tenure * factor(X9Age) +
                       X9Tenure * factor(X9Inc) + X9Tenure * factor(X9District),
                     data = training.data)
summary(lm_profit_final)
```

```
##
## Call:
## lm(formula = X9Profit ~ factor(X9Age) + factor(X9Inc) + factor(X9District) +
##     X9Billpay * factor(X9Age) + X9Billpay * factor(X9Inc) + X9Billpay *
##     factor(X9District) + X9Tenure * factor(X9Age) + X9Tenure *
##     factor(X9Inc) + X9Tenure * factor(X9District), data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -666.68 -145.84  -68.41   58.97 1992.44
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -33.3432    20.5392  -1.623  0.10452
## factor(X9Age)2                    11.5853    19.2661   0.601  0.54762
## factor(X9Age)3                    44.7976    18.6602   2.401  0.01637 *
## factor(X9Age)4                    24.8261    17.9808   1.381  0.16738
## factor(X9Age)5                    55.3383    19.5335   2.833  0.00462 **
## factor(X9Age)6                    48.0392    20.6764   2.323  0.02017 *
## factor(X9Age)7                   110.2162    20.2488   5.443 5.29e-08 ***
## factor(X9Inc)2                    39.9019    18.9245   2.108  0.03500 *
## factor(X9Inc)3                   25.5658    13.5346   1.889  0.05892 .
## factor(X9Inc)4                   22.1007    14.1343   1.564  0.11792
## factor(X9Inc)5                   28.0062    14.0424   1.994  0.04612 *
## factor(X9Inc)6                   33.8159    11.2173   3.015  0.00258 **
## factor(X9Inc)7                   83.1994    13.4319   6.194 5.96e-10 ***
## factor(X9Inc)8                   94.8442    15.7808   6.010 1.88e-09 ***
## factor(X9Inc)9                  168.5826    13.8909  12.136  < 2e-16 ***
## factor(X9District)1200           10.0329     9.4879   1.057  0.29032
## factor(X9District)1300           13.5929    11.5786   1.174  0.24042
## X9Billpay                        20.3135   159.7645   0.127  0.89883
## X9Tenure                         -0.6941     2.7427  -0.253  0.80023
## factor(X9Age)2:X9Billpay          4.3738   132.3368   0.033  0.97363
## factor(X9Age)3:X9Billpay         24.2930   131.4781   0.185  0.85341
## factor(X9Age)4:X9Billpay         37.3352   131.0845   0.285  0.77579
## factor(X9Age)5:X9Billpay         83.6991   136.4704   0.613  0.53967
## factor(X9Age)6:X9Billpay        138.9714   147.1883   0.944  0.34509
## factor(X9Age)7:X9Billpay         71.7827   145.7834   0.492  0.62245
## factor(X9Inc)2:X9Billpay         10.7021   217.4662   0.049  0.96075
## factor(X9Inc)3:X9Billpay         42.3448    94.8603   0.446  0.65532
## factor(X9Inc)4:X9Billpay         57.0997   102.7413   0.556  0.57838
## factor(X9Inc)5:X9Billpay         41.8418    96.2827   0.435  0.66388
## factor(X9Inc)6:X9Billpay         50.4240    83.3665   0.605  0.54529
## factor(X9Inc)7:X9Billpay         20.3257    89.2478   0.228  0.81985
## factor(X9Inc)8:X9Billpay        138.9710    96.6942   1.437  0.15067
## factor(X9Inc)9:X9Billpay        205.0662    85.8867   2.388  0.01696 *
## factor(X9District)1200:X9Billpay -45.3379    70.5308  -0.643  0.52035
## factor(X9District)1300:X9Billpay -35.7461    81.5820  -0.438  0.66127
## factor(X9Age)2:X9Tenure           3.8112     2.8986   1.315  0.18857
## factor(X9Age)3:X9Tenure           4.8110     2.7412   1.755  0.07926 .
## factor(X9Age)4:X9Tenure           6.5284     2.6781   2.438  0.01479 *
## factor(X9Age)5:X9Tenure           4.6870     2.7051   1.733  0.08318 .
## factor(X9Age)6:X9Tenure           6.5565     2.7160   2.414  0.01578 *
## factor(X9Age)7:X9Tenure           4.9458     2.7017   1.831  0.06717 .
## factor(X9Inc)2:X9Tenure          -3.2776     1.3149  -2.493  0.01268 *
## factor(X9Inc)3:X9Tenure          -1.6317     0.9911  -1.646  0.09970 .
## factor(X9Inc)4:X9Tenure          -0.7609     0.9905  -0.768  0.44237
## factor(X9Inc)5:X9Tenure          -1.2893     1.0144  -1.271  0.20374
## factor(X9Inc)6:X9Tenure          -0.1267     0.8411  -0.151  0.88024
## factor(X9Inc)7:X9Tenure          -1.8697     0.9934  -1.882  0.05983 .
## factor(X9Inc)8:X9Tenure          -1.3257     1.1572  -1.146  0.25198
## factor(X9Inc)9:X9Tenure          -2.4488     0.9983  -2.453  0.01417 *
## factor(X9District)1200:X9Tenure   1.1405     0.6459   1.766  0.07742 .
## factor(X9District)1300:X9Tenure  -0.5663     0.7868  -0.720  0.47165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 268.4 on 23722 degrees of freedom
## Multiple R-squared:  0.07033,    Adjusted R-squared:  0.06837
## F-statistic: 35.89 on 50 and 23722 DF,  p-value: < 2.2e-16
```

We first create a base model to compare the relationship bill pay and profit. Expected 1999Profit = 118.808 + 108.619*Billpay. Both intercept of the regression (beta0) and coeffcient on billpay (beta 1) are statistically significant. The adjusted R-square is 0.002488, which means that 0.25% of 1999Profit can be explained by 2009Billpay

Then holding other variables constant, we add age factor into the regression. All the age factors are significant, but the intercept and billpay are not statistically significant.The adjusted R-square is 0.02318, which means that 2.32% of 2009Profit can be explained by 1999Billpay and 1999Age Once again, holding previous variables constant, we add income level into the regression. The intercept of the regression, age factors, and the majority of income level are statistically significant, but billpay is not significant. The adjusted R-square is 0.04991, which means that 5.0% of 1999Profit can be explained by 1999Billpay ,1999Age and 1999Income level.

Moreover, we add tenure into the regression and hold other vairable constant. The majority of Age factor, Income level are significant. The adjusted R-square is 0.06928, which means that 6.93% of 1999Profit can be explained by 1999Billpay, 1999Age, 1999Income level, and Tenure.

Finally, we add 1999District into the regression holding other variables constant. The intercept, 1999Age, Income level are significant. When it is in age category 1, income level 1, District 1100, do not use billpay and leaves the bank, the expected 1999Profit decrease by -37.8276. Then if we hold other constant but increase the age category to the second category, the expected 1999Profit will increase 16.5702. If we hold other constant but increase the income level to the second stage, the expected 1999Profit will increase 38.3894. If we hold other constant but change the district to 1200, the expected 1999Profit will increase 10.5582. If we hold other constant but look at people use billpay , the expected 1999Profit will decrease 0.5655. If we hold other constant but look at people who stay at the bank, the expected 1999Profit will increase 0.6818. Among thoese statistically insignificant coeffcient, we will interpret the significant. For example, factor(X9Inc)9:X9Billpay =227.5110. It means it is the profit difference of a customer in Income Level 9 and uses Billpay, eliminating other effects on age, distrct, and tenure.
The adjusted R-square is 0.07021, which means that 7.02% of 1999Profit can be explained by 1999Billpay, 1999Age, 1999Income level, Tenure, and 1999District.

```
# Prediction errors among different profit models
# Model lm_profit_Billpay
predicted.profit1 = predict(lm_profit_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_final
predicted.profit5 = predict(lm_profit_final, validation.data)
prediction.error5 = sqrt(mean((predicted.profit5-validation.data$X9Profit)^2))
```

We calculated the predicted errors for all profit models, and the prediction error for our final Profit Model is 295.4667. To compare the errors from more perspectives, we made the following table.

Comparison Table for Profit Models

```
# Creating a comparison table for profit models
comparison.table.profit = matrix(c(summary(lm_profit_Billpay)$adj.r.square, AIC(lm_profit_Billpay), BIC
comparison.table.profit = round(comparison.table.profit,4)
colnames(comparison.table.profit) = c("Adj.r.square", "AIC", "BIC", "Prediction Error")
rownames(comparison.table.profit) = c("lm_profit_Billpay", "lm_profit_Age", "lm_profit_Inc", "lm_profit_
comparison.table.profit
```

```
##                     Adj.r.square      AIC      BIC Prediction Error
## lm_profit_Billpay       0.0022 334997.8 335022.0         268.1220
## lm_profit_Age           0.0219 334535.8 334656.9         264.1788
## lm_profit_Inc           0.0485 333896.1 334146.5         258.9908
## lm_profit_Tenure        0.0675 333431.8 333803.3         257.7076
## lm_profit_final         0.0684 333414.7 333834.7         257.6333
```

We compared the adjusted R-square, AIC, BIC, and Prediction Error. It can tell that the final Profit Model has the highest adjusted R-square of 0.0710, and lowest AIC of 333388.7, lowest BIC of 333808.6, and lowest Prediction Error of 286.4478. We condluded that the final Profit Model fit the validation subset the best, and so it is the most approporate Profit Model.

```
# Retention Models
# Note: We omitted the summary for Billpay, Age, and Inc and we only kept the best one.
lm_retention_Billpay = lm(retention ~ X9Billpay, data = training.data)
lm_retention_Age = lm(retention ~ factor(X9Age) + X9Billpay * factor(X9Age),
                      data = training.data)
lm_retention_Inc = lm(retention ~ factor(X9Age) + factor(X9Inc) + X9Billpay * factor(X9Age)
                      + X9Billpay * factor(X9Inc), data = training.data)
lm_retention_Tenure = lm(retention ~ factor(X9Age) + factor(X9Inc) + X9Billpay * factor(X9Age)
                         + X9Billpay * factor(X9Inc) + X9Tenure * factor(X9Age) + X9Tenure * factor(X9I
                         data = training.data)
summary(lm_retention_Tenure)
```

```
##
## Call:
## lm(formula = retention ~ factor(X9Age) + factor(X9Inc) + X9Billpay *
##     factor(X9Age) + X9Billpay * factor(X9Inc) + X9Tenure * factor(X9Age) +
##     X9Tenure * factor(X9Inc), data = training.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -8.27e-12 -5.00e-16  0.00e+00  2.00e-16  4.66e-14
##
## Coefficients:
##                           Estimate Std. Error    t value Pr(>|t|)
## (Intercept)              1.000e+00  3.817e-15  2.620e+14  < 2e-16 ***
## factor(X9Age)2          -4.122e-18  3.861e-15 -1.000e-03   0.9991
## factor(X9Age)3           5.305e-16  3.739e-15  1.420e-01   0.8872
## factor(X9Age)4           8.101e-16  3.602e-15  2.250e-01   0.8221
## factor(X9Age)5           5.754e-16  3.914e-15  1.470e-01   0.8831
## factor(X9Age)6           5.241e-15  4.142e-15  1.265e+00   0.2058
## factor(X9Age)7           2.581e-16  4.057e-15  6.400e-02   0.9493
## factor(X9Inc)2          -7.169e-17  3.787e-15 -1.900e-02   0.9849
## factor(X9Inc)3           6.104e-15  2.698e-15  2.262e+00   0.0237 *
## factor(X9Inc)4          -4.846e-16  2.832e-15 -1.710e-01   0.8641
## factor(X9Inc)5          -4.546e-16  2.813e-15 -1.620e-01   0.8716
## factor(X9Inc)6          -8.750e-17  2.230e-15 -3.900e-02   0.9687
## factor(X9Inc)7          -2.035e-16  2.672e-15 -7.600e-02   0.9393
## factor(X9Inc)8          -1.519e-16  3.137e-15 -4.800e-02   0.9614
## factor(X9Inc)9          -3.030e-16  2.743e-15 -1.100e-01   0.9121
## X9Billpay               -7.081e-16  3.045e-14 -2.300e-02   0.9814
## X9Tenure                 1.255e-16  5.415e-16  2.320e-01   0.8167
## factor(X9Age)2:X9Billpay -7.540e-17  2.647e-14 -3.000e-03   0.9977
## factor(X9Age)3:X9Billpay  3.396e-16  2.633e-14  1.300e-02   0.9897
## factor(X9Age)4:X9Billpay  4.520e-16  2.623e-14  1.700e-02   0.9862
## factor(X9Age)5:X9Billpay  7.899e-16  2.734e-14  2.900e-02   0.9770
## factor(X9Age)6:X9Billpay  1.347e-15  2.944e-14  4.600e-02   0.9635
## factor(X9Age)7:X9Billpay -6.558e-16  2.920e-14 -2.200e-02   0.9821
## factor(X9Inc)2:X9Billpay  9.042e-16  4.343e-14  2.100e-02   0.9834
## factor(X9Inc)3:X9Billpay  2.315e-15  1.877e-14  1.230e-01   0.9018
## factor(X9Inc)4:X9Billpay  3.769e-16  2.035e-14  1.900e-02   0.9852
## factor(X9Inc)5:X9Billpay  1.698e-16  1.918e-14  9.000e-03   0.9929
## factor(X9Inc)6:X9Billpay  2.425e-16  1.653e-14  1.500e-02   0.9883
## factor(X9Inc)7:X9Billpay  4.671e-16  1.765e-14  2.600e-02   0.9789
## factor(X9Inc)8:X9Billpay  2.580e-16  1.916e-14  1.300e-02   0.9893
## factor(X9Inc)9:X9Billpay  2.312e-16  1.694e-14  1.400e-02   0.9891
## factor(X9Age)2:X9Tenure   1.857e-17  5.809e-16  3.200e-02   0.9745
## factor(X9Age)3:X9Tenure  -7.679e-17  5.493e-16 -1.400e-01   0.8888
## factor(X9Age)4:X9Tenure  -1.110e-16  5.367e-16 -2.070e-01   0.8361
## factor(X9Age)5:X9Tenure  -7.535e-17  5.421e-16 -1.390e-01   0.8895
## factor(X9Age)6:X9Tenure  -6.144e-16  5.443e-16 -1.129e+00   0.2590
## factor(X9Age)7:X9Tenure  -7.075e-18  5.414e-16 -1.300e-02   0.9896
## factor(X9Inc)2:X9Tenure   1.601e-17  2.628e-16  6.100e-02   0.9514
## factor(X9Inc)3:X9Tenure  -9.437e-16  1.971e-16 -4.789e+00 1.69e-06 ***
## factor(X9Inc)4:X9Tenure   6.370e-17  1.981e-16  3.220e-01   0.7478
## factor(X9Inc)5:X9Tenure   5.485e-17  2.028e-16  2.700e-01   0.7869
## factor(X9Inc)6:X9Tenure   1.100e-17  1.667e-16  6.600e-02   0.9474
## factor(X9Inc)7:X9Tenure   2.578e-17  1.967e-16  1.310e-01   0.8957
## factor(X9Inc)8:X9Tenure   1.623e-17  2.291e-16  7.100e-02   0.9435
## factor(X9Inc)9:X9Tenure   3.159e-17  1.955e-16  1.620e-01   0.8716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.379e-14 on 23728 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4991
## F-statistic: 539.3 on 44 and 23728 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.

We first create a base model to compare the relationship bill pay and retention. Expected 1999Retention = 1.000e + (3.553eˆ-16)*Billpay. If the customer uses the electronic Billpay, the 1999's profit will increase by (3.553eˆ-16). The intercept of the regression (beta0) is statistically significant. The adjusted R-square is 0.5, which means that 50% of 1999Profit can be explained by 1999Billpay. Then holding other variables constant, we add age factor into the regression. All the age factors are significant, but the intercept and Billpay are not statistically significant.The adjusted R-square is 0.4997, which means that 49.97%% of 1999Profit can be explained by 1999Billpay and 1999Age Once again, holding previous variables constant, we add income level into the regression. The intercept of the regression, age factors, and the majority of income level are statistically significant, but Billpay is not significant. The adjusted R-square is 0.4994, which means that 50% of 1999Profit can be explained by 1999Billpay, 1999Age and 1999Income level.

Finally, we add 1999 District into the regression holding other variables constant. The intercept, 1999 Age, Income level are significant. When it is in age category 1, income level 1, do not use Billpay and leaves the bank, the expected 1999 Profit decrease by 1.000e. Then if we hold other constant but increase the age category to the second category, the expected 1999Profit will decrease by -1.989eˆ17. If we hold other constant but increase the income level to the second stage, the expected 1999Profit will increase 5.437eˆ-17. If we hold other constant but look at people use Billpay , the expected 1999Profit will decrease by 1.389eˆ-15. If we hold other constant but look at people who stay at the bank, the expected 1999Profit will increase 1.027eˆ-16. Among those statistically insignificant coefficient, we will interpret the significant. For example, factor(X9Inc)3:X9Tenure =4.8eˆ-06. It means it is the profit difference of a customer in Income Level 3 and

stays in the bank, eliminating other effects on age and income.
The adjusted R-square is 0.4991, which means that 49.91% of 1999Profit can be explained by 1999 Billpay, 1999 Age, 1999 Income level, and Tenure.

```
# Prediction errors among different retention models
# Model lm_retention_Billpay
predicted.retention1 = predict(lm_retention_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))
```

We calculated the predicted errors for all profit models. It can tell that that all the models have relatively low prediction error, so we changed the prediction error to log(prefiction error) to see the difference. And then 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_Billpay)$adj.r.square,
                                      AIC(lm_retention_Billpay), BIC(lm_retention_Billpay), log(predict

comparison.table.retention = round(comparison.table.retention,4)
colnames(comparison.table.retention) = c("Adj.r.square", "AIC", "BIC", "Prediction Error")
rownames(comparison.table.retention) = c("lm_retention_Billpay", "lm_retention_Age", "lm_retention_Inc"
comparison.table.retention
```

```
##                      Adj.r.square      AIC      BIC Prediction Error
## lm_retention_Billpay       0.5000 -1385187 -1385162         -30.5466
## lm_retention_Age           0.4997 -1385174 -1385053         -30.5451
## lm_retention_Inc           0.4994 -1385152 -1384901         -30.5428
## lm_retention_Tenure        0.4991 -1385191 -1384819         -30.5472
```

We compared the adjusted R-square, AIC, BIC, and Prediction Error. It can cell that the final Retention Model with Tenure has the lowest AIC of -1385191, lowest BIC of -1384819, and biggest log(Prediction Error) of -30.5471. (Since we changed prediction error to log(prefiction error), it now becomes the bigger the smaller error.) Then we condluded that the final Retention Model with tenure fit the validation subset the best, and so it is the most approporate Retention Model.

8. Test of the final Profit Model and Retention Model

```
Database2000 = consumerDB2[c(4,5,6,7,8,9,11,12)]
Database2000$X9Tenure=Database2000$X9Tenure+1
colnames(Database2000)=c("X9Age","X9Inc","X9Tenure", "X9District","X9Online", "X0Profit", "X9Billpay","

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

```
## [1] 145.9697
```

```r
Prediction_Retention_2000 = predict(lm_retention_Tenure, Database2000)
prediction.error.retention = sqrt(mean((Prediction_Retention_2000-Database2000$retention)^2))
prediction.error.retention
```

```
## [1] 5.41686e-14
```

To test the fitness of our Profit Model and Retention Model, we extracted all the datapoints in year 2000 and the income, age, and distrcit 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 precited 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 147.2005, which is about half of the prediction error we validated before. Meanwhile, the prediction error of our Retention Model is 5.416577e-14, which is very small. In general, our Profit Model and Retention Model could fairly predict the profit and retention status of Pilgrim Banks' customers.

9. Summarizatoin and Recommendation

Our regression of Profit Model and Retention Model shows that when people using electronic billpay, the profit and the retension tend to decrease. It shows thst online customers are not indeed better customers, and the senior management should not offer rebates or lower the service charges for customers using online banking, as these customers even have relative lower possibility to stay with the bank and bring less profit to Pilgrim Bank.