OCT 09, 2022

PILGRIM BANK

CASE STUDY

Howard Jiang (251119686)

PREPARATION – DEAL WITH MISSING VALUES

01

Find the mean values for Age, Income & 0Profit - Avg(AGE) = 4.046, Avg(INC) = 5.459 & Avg(0Profit) = 144.8 AND Find the mode values for 0Online & 0Billpay - mode(0Online) = 0 & mode(0Billpay) = 0

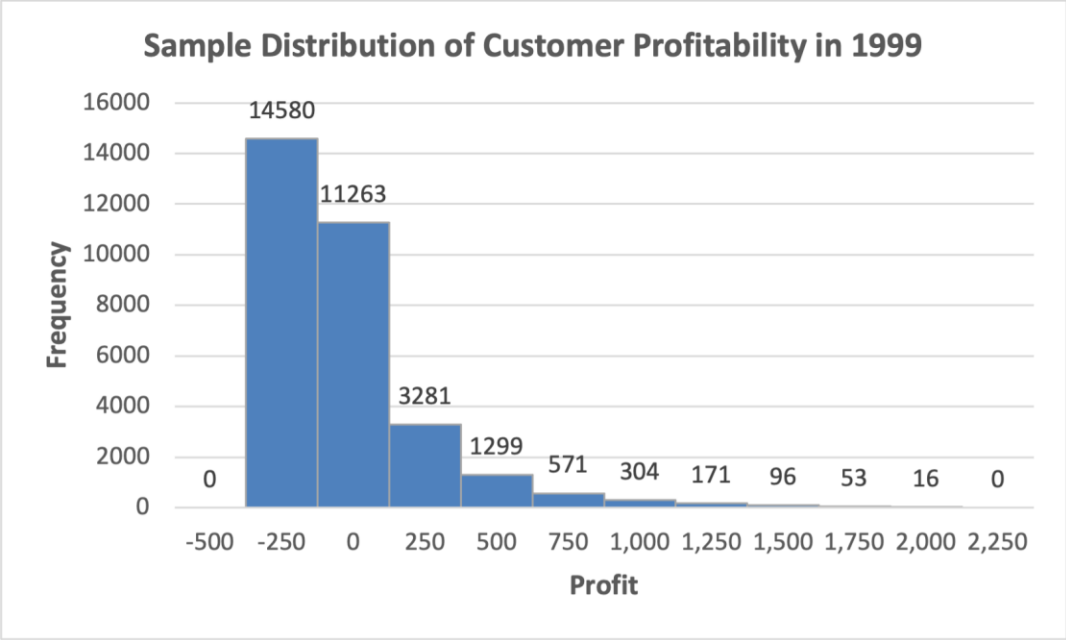
02

Add 5 dummy variables AgeExist, IncExist, 0ProfitExist, 0OnlineExist & 0BillpayExist to indicate whether a value for Age, Income, 0Profit, 0Online & 0Billpay exists or is missing for a given data point.

03

Replace the missing Age, Income & 0Profit with their average values AND Replace 0Online & 0Billpay with their mode values.

Q1. ANALYZING THE PROFITABILITY



According to 31634 sample customer data extracted from entire population data, the minimum and maximum customer profitability are $-221 and $2071, respectively.

Furthermore, it has an average of $111.50 and a standard deviation of $272.84.

Q1. 95% CONFIDENCE INTERVAL OF AVERAGE PROFITABILITY

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

According to the formula above, 95% Confidence Interval can be computed by the formula below:

$$111.5027 \pm 1.96\frac{272.8394}{\sqrt{31634}} = 111.5027 \pm 3.006732 = [108.496, 114.5094]$$

Therefore, 95% CI of average profitability for Pilgrim Bank's entire customer population is [108.496, 114.5094].

Q2. ONLINE USAGE IN 1999 VS. PROFIT IN 1999

Null Hypothesis: Status of customers online or offline does not affect the profit.

Alternative Hypothesis: Status of customers online or offline does affect the profit.

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              110.786      1.637  67.678   <2e-16 ***
as.factor(`90nline`)1      5.881      4.690   1.254     0.21
```

In order to conduct hypothesis Hesting, I performed a linear regression model between Online or Offline status in 1999 & profit in 1999.

Based on the ANOVA summary table above, we can find that the p-value is 0.21, which is larger than our level of significance 0.05. Therefore, we cannot reject H0, which results in the conclusion that whether customers shop online or offline do not affect the profit.

Q3. EXPLORE CUSTOMER DEMOGRAPHICS

We added more customer demographics variables, including Age, Income, Tenure, District, as well as two dummy variables we created, which are AgeExist & IncExist, to the new linear regression model.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -144.2333     7.9100 -18.234  < 2e-16 ***
as.factor(`9Online`)1        13.8233     4.6091   2.999  0.00271 **
`9Age`                       16.6701     1.1482  14.519  < 2e-16 ***
`9Inc`                       16.8530     0.7554  22.310  < 2e-16 ***
`9Tenure`                     4.7464     0.1918  24.742  < 2e-16 ***
as.factor(`9District`)1200   21.1941     5.0866   4.167 3.10e-05 ***
as.factor(`9District`)1300    7.9955     6.2582   1.278  0.20140
AgeExist1                     4.3905     8.2017   0.535  0.59243
IncExist1                    34.8812     8.2049   4.251 2.13e-05 ***
```

In accordance with the output above, we can find p-values for Age, Income, Tenure & District 1200 are all less than 0.05 while p-value for District 1300 is 0.2014, which is larger than 0.05. In conclusion, Age, Income, Tenure & District 1200 play significant roles in analyzing customer profitability for online and offline customers while District 1300 is not.

Moreover, if we control all other variables, profit for Online customers is $13.8233 greater than that for Offline customers.

Q3. EXPLORE CUSTOMER DEMOGRAPHICS

By analyzing Online & Offline customers separately, we can obtain results below:

Online

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -207.1351    26.0555  -7.950 2.44e-15 ***
`9Age`                      20.9471     3.8760   5.404 6.90e-08 ***
`9Inc`                      24.5726     2.2785  10.785  < 2e-16 ***
`9Tenure`                    4.7905     0.6765   7.081 1.69e-12 ***
as.factor(`9District`)1200  28.2414    18.1207   1.559    0.119
as.factor(`9District`)1300   2.2459    22.1374   0.101    0.919
AgeExist1                   34.4308    27.3720   1.258    0.209
IncExist1                   19.0860    27.2709   0.700    0.484
```

Offline

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -134.5029     8.2812 -16.242  < 2e-16 ***
`9Age`                      16.1954     1.2058  13.431  < 2e-16 ***
`9Inc`                      15.6983     0.8019  19.577  < 2e-16 ***
`9Tenure`                    4.7383     0.1998  23.719  < 2e-16 ***
as.factor(`9District`)1200  20.8974     5.2904   3.950 7.83e-05 ***
as.factor(`9District`)1300   9.0504     6.5135   1.389    0.165
AgeExist1                    1.4737     8.5894   0.172    0.864
IncExist1                   35.9334     8.5911   4.183 2.89e-05 ***
```

Similarly to the previous output, we can conclude that Age, Income, Tenure are significant in predicting the profit of Online customers while District 1200 & 1300 are both insignificant.

For predicting the profit of Offline customers, Age, Income, Tenure, and District 1200 are significant predictors of Offline customer profit, but District 1300 is not.

Q4A. ONLINE BANKING USAGE IN 1999 VS. PROFIT IN 2000

Null Hypothesis: There is no relationship between whether the customers are online or offline in 1999 and profit in 2000.

Alternative Hypothesis: There is a relationship whether the customers are online or offline in 1999 and profit in 2000.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           142.388      2.137  66.628   <2e-16 ***
as.factor(`9Online`)1  19.983      6.123   3.264   0.0011 **
```

The formula between whether the customers used online banking in 1999 and profit in 2000 is:

$$0Profit = 142.388 + 19.983 * 9Online$$

As the p-value for `9Online` = 0.0011, which is less than 0.05, so there is a statistical evidence to reject H0 and conclude that whether the customers used online banking in 1999 can significantly help to predict profit in 2000. Customers that used online banking in 1999 made 19.983 more profit in 2000 than those did not use online banking.

Q4A. ONLINE BANKING USAGE IN 1999 VS. PROFIT IN 2000

In order to make the output more reliable, we need to add more variables for control. The output is:

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  17.678287   8.868027   1.993  0.04622 *
as.factor(`9Online`)1        14.387583   5.141097   2.799  0.00514 **
`9Profit`                     0.721038   0.006271 114.972  < 2e-16 ***
`9Age`                        0.795347   1.284799   0.619  0.53589
`9Inc`                        7.721192   0.849089   9.094  < 2e-16 ***
`9Tenure`                     0.601360   0.216011   2.784  0.00537 **
as.factor(`9District`)1200    5.087946   5.674493   0.897  0.36992
as.factor(`9District`)1300    5.952774   6.979777   0.853  0.39374
AgeExist1                   -13.089900   9.147171  -1.431  0.15243
IncExist1                    -2.057870   9.153328  -0.225  0.82212
```

As the p-value for `9Online` = 0.00514, which is less than 0.05, so there is a statistical evidence to reject H0 and conclude that whether the customers used online banking in 1999 can significantly help to predict profit in 2000.

Q4B. ONLINE BANKING USAGE IN 1999 VS. WHETHER STAYS IN 2000

We define a new variable `stays` that takes 1 if profit in 2000 is not 'NA' and 0 otherwise and use it as the dependent variable in our model.

Linear Regression Model:

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.105e-01  1.015e-02  50.272  < 2e-16 ***
as.factor(`9Online`)1        1.050e-02  5.887e-03   1.784  0.07443 .
`9Profit`                    1.896e-05  7.181e-06   2.640  0.00829 **
`9Age`                       1.086e-03  1.471e-03   0.739  0.46020
`9Inc`                       3.197e-03  9.722e-04   3.288  0.00101 **
`9Tenure`                    3.832e-03  2.473e-04  15.495  < 2e-16 ***
as.factor(`9District`)1200 1.013e-02  6.497e-03   1.560  0.11883
as.factor(`9District`)1300 2.921e-03  7.992e-03   0.366  0.71473
AgeExist1                    1.720e-01  1.047e-02  16.425  < 2e-16 ***
IncExist1                    1.687e-01  1.048e-02  16.095  < 2e-16 ***
```

As the p-value for `9Online` = 0.07443, which is greater than 0.05, so there is no a statistical evidence showing whether the customers used online banking in 1999 can significantly affect whether customers stayed with the bank in 2000.

Q4B. ONLINE BANKING USAGE IN 1999 VS. WHETHER STAYS IN 2000

As in this model, we are supposed to predict a binary variable, so Logistic Regression model is better choice than Linear Regression model as it is a "Supervised Machine Learning" classification algorithm that can be used to predict the likelihood of a categorical variable based on a set of independent variable(s).

Logistic Regression Model:

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.908e-01  9.912e-02  -6.970 3.18e-12 ***
as.factor(`9Online`)1        1.531e-01  5.391e-02   2.840  0.00452 **
`9Profit`                    1.765e-04  7.185e-05   2.457  0.01401 *
`9Age`                       6.875e-02  1.600e-02   4.297 1.73e-05 ***
`9Inc`                       5.414e-02  1.061e-02   5.103 3.34e-07 ***
`9Tenure`                    3.757e-02  2.417e-03  15.543  < 2e-16 ***
as.factor(`9District`)1200   1.018e-01  5.499e-02   1.850  0.06427 .
as.factor(`9District`)1300   3.925e-02  6.806e-02   0.577  0.56419
AgeExist1                    1.122e+00  7.648e-02  14.675  < 2e-16 ***
IncExist1                    1.062e+00  7.629e-02  13.927  < 2e-16 ***
```

As the p-value for `9Online` = 0.00452, which is less than 0.05, so there is a statistical evidence showing whether the customers used online banking in 1999 can significantly affect whether customers stayed with the bank in 2000.

Q5A. ELECTRONIC BILLPAY USAGE IN 1999 VS. PROFIT IN 2000

Null Hypothesis: There is no relationship between whether the customers are online or offline in 1999 and whether customers stay with the bank in 2000.

Alternative Hypothesis: There is a relationship between whether the customers are online or offline in 1999 and whether customers stay with the bank in 2000.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                143.354      2.019  71.006  < 2e-16 ***
as.factor(`9Billpay`)1      88.007     15.627   5.632  1.8e-08 ***
```

The formula between whether the customers are online or offline in 1999 and the profit in 2000 is:

$$0Profit = 143.354 + 88.007 * 9Billpay$$

As the p-value = 1.8e-08, which is less than 0.05, so there is a statistical evidence to reject H0 and conclude that whether the customers used e-billpay in 1999 can significantly help to predict the profit in 2000. Customers that used e-billpay in 1999 made 88.007 more profit in 2000 than those did not use e-billpay.

Q5A. E-BILLPAY USAGE IN 1999 VS. PROFIT IN 2000

In order to make the output more reliable, we need to add more variables for control. The output is:

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     17.824410   8.869827   2.010  0.04449 *
as.factor(pilgrimbank$`9Billpay`)1  11.442741  13.819682   0.828  0.40767
as.factor(`9Online`)1              12.825990   5.476134   2.342  0.01918 *
`9Profit`                           0.720852   0.006275 114.868  < 2e-16 ***
`9Age`                              0.790009   1.284822   0.615  0.53864
`9Inc`                              7.710031   0.849200   9.079  < 2e-16 ***
`9Tenure`                           0.601906   0.216013   2.786  0.00533 **
as.factor(`9District`)1200          5.054393   5.674666   0.891  0.37310
as.factor(`9District`)1300          5.923297   6.979902   0.849  0.39610
AgeExist1                         -13.143938   9.147449  -1.437  0.15076
IncExist1                          -2.030215   9.153434  -0.222  0.82447
```

After adding more variables, our result is different. As the p-value for `9Billpay` = 0.40767, which is greater than 0.05, so there is no statistical evidence showing whether the customers used e-billpay in 1999 can significantly affected the profit in 2000.

Q5B. ELECTRONIC BILLPAY USAGE IN 1999 VS. WHETHER STAYS IN 2000

Use `stays` we defined in Q4B as the dependent variable in our model.

Linear Regression Model:

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          5.096e-01  1.014e-02  50.266  < 2e-16 ***
as.factor(pilgrimbank$`9Billpay`)1 -3.067e-02  1.580e-02  -1.942  0.05220 .
as.factor(`9Online`)1                1.966e-02  6.259e-03   3.141  0.00169 **
`9Profit`                            1.867e-05  7.173e-06   2.604  0.00923 **
`9Age`                               1.157e-03  1.469e-03   0.788  0.43066
`9Inc`                               3.245e-03  9.706e-04   3.343  0.00083 ***
`9Tenure`                            3.826e-03  2.469e-04  15.497  < 2e-16 ***
as.factor(`9District`)1200           1.059e-02  6.486e-03   1.633  0.10249
as.factor(`9District`)1300           3.900e-03  7.978e-03   0.489  0.62492
AgeExist1                            1.731e-01  1.046e-02  16.558  < 2e-16 ***
IncExist1                            1.675e-01  1.046e-02  16.010  < 2e-16 ***
```

As p-value for `9Billpay` = 0.0522, which is greater than 0.05, so there is no statistical evidence showing whether the customers used e-billpay in 1999 can significantly affect whether customers stayed with the bank in 2000.

Q5B. ELECTRONIC BILLPAY USAGE IN 1999 VS. WHETHER STAYS IN 2000

Like Q4B, we applied Logistic Regression Model to predict this binary variable:

Logistic Regression Model:

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -6.944e-01  9.915e-02  -7.004 2.49e-12 ***
as.factor(pilgrimbank$`9Billpay`)1 -2.835e-01  1.446e-01  -1.961 0.049849 *
as.factor(`9Online`)1              1.908e-01  5.766e-02   3.309 0.000937 ***
`9Profit`                          1.812e-04  7.193e-05   2.519 0.011754 *
`9Age`                             6.890e-02  1.600e-02   4.306 1.67e-05 ***
`9Inc`                             5.445e-02  1.061e-02   5.131 2.88e-07 ***
`9Tenure`                          3.753e-02  2.417e-03  15.526  < 2e-16 ***
as.factor(`9District`)1200         1.028e-01  5.500e-02   1.868 0.061714 .
as.factor(`9District`)1300         3.994e-02  6.807e-02   0.587 0.557338
AgeExist1                          1.124e+00  7.648e-02  14.694  < 2e-16 ***
IncExist1                          1.062e+00  7.628e-02  13.925  < 2e-16 ***
```

As the p-value for `9Billpay` = 0.049849, which is less than 0.05, so there is a statistical evidence showing whether the customers used e-billpay in 1999 can significantly affect whether customers stayed with the bank in 2000.

Q6. ONLINE CUSTOMERS W/O ELECTRONIC BILLPAY

CUSTOMERS WHO USED ONLINE BANKING IN 1999 WITHOUT ELECTRONIC BILLPAY		
	COUNT	AVERAGE PROFIT
STAYED ONLINE	2562	153
MOVED OFFLINE	764	145
W ELECTRONIC BILLPAY	225	198
W/O ELECTRONIC BILLPAY	3101	148

According to the table, for customers who used online banking in 1999 without e-billpay, in 2000, 2562 of them remain Online while 764 of them transitioning to Offline. Meanwhile, the average profit of those customers remain Online ($153) is also greater than those transitioning to Offline ($145).

And for the status of using e-billpay, although only 225 customers transitioning to use e-billpay, which is less than those who still not using e-billpay (3101), the average profit of them using e-billpay ($198) is greater than those not using it ($148).

Q6. ONLINE CUSTOMERS W ELECTRONIC BILLPAY

CUSTOMERS WHO USED ONLINE BANKING IN 1999 WITH ELECTRONIC BILLPAY		
	COUNT	AVERAGE PROFIT
STAYED ONLINE	421	252
MOVED OFFLINE	107	151
W ELECTRONIC BILLPAY	254	259
W/O ELECTRONIC BILLPAY	274	206

Based on the table, for those customers used online banking and e-billpay in 1999, 421 of them remain Online while only 107 of them transited to Offline. Additionally, the average profit of those customers remain Online ($252) is also greater than those transited to Offline ($151).

As for the status of using Electronic Billpay, there are 254 customers still decided to use e-billpay while 274 customers who still decided not to use e-billpay. However, customers that used e-billpay ($259) has greater average profit than those not using it ($206).

Q6. OFFLINE CUSTOMERS W/O ELECTRONIC BILLPAY

CUSTOMERS WHO USED OFFLINE BANKING IN 1999 WITHOUT ELECTRONIC BILLPAY		
	COUNT	AVERAGE PROFIT
STAYED OFFLINE	25507	141
MOVED ONLINE	2273	154
W ELECTRONIC BILLPAY	319	265
W/O ELECTRONIC BILLPAY	27461	141

According to the bar charts above, for customers who used offline banking in 1999 without e-billpay, although only 2273 of them transited to Online and 25507 of them remained Offline, the average profit of those Online customers ($154) is greater than those Offline customers($141).

And for the status of using e-billpay, although only 319 customers transitioning to use e-billpay, which is less than those who still not using e-billpay, the average profit of them using e-billpay ($266) is greater than those not using it ($141).

Note: In this case, there is no customers who used offline banking in 1999 with e-billpay so that we cannot analyze the performance of these offline customers in 2000.

Q6. STATUS IN 1999 VS. WHETHER STAYS IN 2000

CUSTOMERS' STATUS VS. WHETHER THEY STAYED WITH THE BANK IN 2000			
	NUMBER OF CUSTOMERS IN 1999	NUMBER OF STAYED CUSTOMERS IN 2000	RATIO
ONLINE w/o E-BILLPAY	3326	2860	86%
ONLINE w E-BILLPAY	528	448	85%
OFFLINE w/o E-BILLPAY	27780	23107	83%
OFFLINE w E-BILLPAY	0	/	/

Overall speaking, the customers' stickness of Pilgrim Bank is pretty high. No matter their online and e-billpay usage look like in 1999, more than 80% of them are willing to stay with the bank in 2000.

The retention rate of those Online customers is slightly higher than those Offline ones and Online customers also make more profit. However, the conversion rate of customers transferred from Offline to Online is low.

Similarly, customers without e-billpay in 1999 is more profitable if they adpted to e-billpay in 2000 but it seems more customers should be motivated to use e-billpay because only a small percentage of them converted.

Q6. FINAL THOUGHTS

As stated in the previous analysis, there is a trend of transitioning from Offline to Online & from non e-billpay to e-billpay. Also, generally, customers that use online banking and e-billpay are likely to make more profit than those do not use it. However, although the retention rate is high, the conversion rate is relatively low.

From our results, we can conclude that enlarging the promotional campaign to encourage consumers to make more use of online services and electronic purchase is an effective way to bring greater benefits to the bank. For example, using experiential marketing of online banking and electronic billpay allows customers to learn more about the convenience of these services and come up the idea to change.

However, since a large proportion of our data is missing and we replace these missing values by our own, there are likely to cause some biases. Therefore, filling in these blanks with exact valuesmay draw more appropriate conclusions. In addition, we have only 31634 valid applications to the extent that our analysis is also somewhat limited. If we can collect data for the next few years, we can conduct a more systematic trend analysis.