“Pilgrim Bank (A): Customer Profitability."

Alicja Wilk

BANA 271

Winter 2019

Abstract.

The paper analyzes consumer profitability and summarizes main findings as well as provides answers to the three core questions : What can we conclude about average customer, is there a difference in average profitability between online and offline customers and what role do demographics play in analyzing customer profitability. The case analysis is based on Harvard Business Review, Pilgrim Bank (A): Customer Profitability(1).

Data.

The data sample from 1999 was analyzed, seven main factors/variables were taken into consideration when looking at the data set. Visualizations to certain variables drawn through statistical software R and also statistics on the data were performed through R.

Among variables we have *Profit99* which indicates annual profit from 1999 per customer, dummy variable *Online99* which indicates whether a customer has a banking account online (1) or not(0), *Age99* (binned 1-7) which indicates the age of the customer: 1 = less than 15 years, 2=15-24 years; 3=25-34 years; 4=35-44 years; 5=45-54; 6=55-64, 7=65 and older.

Another variable *Inc99* (binned 1-9) indicates income per customer: 1= less than $15,000; 2 = $15,000-$19,999; 3 = $20,000-$29,999; 4 = $30,000-$39,999; 5 = $40,000- $49,999; 6 = $50,000-$74,999; 7 = $75,000-$99,999; 8 = $100,000-$124,999; 9 = $125,000 and more. *Tenure99* which indicates the length of one’s time with the bank. *District 99* which indicates geographic locations of the bank. The last variable taking into our analysis *Online99* which is a dummy variable whether a customer pays bills online (1) or not (0).

Performing descriptive analytics on the sample data (**Figure 1**) we have 31,634 customer IDs in our data that are the representation of the population for this study. An assumption is made that this sample is a good indicator due to good demographics variety and household characteristics. Also due to limited geographic we will assume that this is a good sample for this geographic area.

Figure 1.

	vars	n	mean	sd	median	trimmed	mad	min
ID	1	31634	15817.50	9132.09	15817.50	15817.50	11725.14	1.00
Profit99	2	31634	111.50	272.84	9.00	58.37	103.78	-221.00
Online99	3	31634	0.12	0.33	0.00	0.03	0.00	0.00
Age99	4	31634	3.99	1.48	4.00	3.91	1.48	1.00
Inc99	5	31634	5.39	2.09	5.76	5.44	1.84	1.00
Tenure99	6	31634	10.16	8.45	7.41	9.00	6.91	0.16
District99	7	31634	1203.19	47.91	1200.00	1203.90	0.00	1100.00
Billpay99	8	31634	0.02	0.13	0.00	0.00	0.00	0.00
Profit00	9	31634	136.19	366.66	23.00	75.31	110.36	-5643.00
Online00	10	31634	0.20	0.38	0.00	0.12	0.00	0.00
Billpay00	11	31634	0.03	0.16	0.00	0.00	0.00	0.00

	max	range	skew	kurtosis	se
ID	31634.00	31633	0.00	-1.20	51.34
Profit99	2071.00	2292	2.75	9.94	1.53
Online99	1.00	1	2.31	3.35	0.00
Age99	7.00	6	0.42	-0.39	0.01
Inc99	9.00	8	-0.23	-0.35	0.01
Tenure99	41.16	41	1.13	0.63	0.05
District99	1300.00	200	0.09	1.33	0.27
Billpay99	1.00	1	7.54	54.93	0.00
Profit00	27086.00	32729	17.88	1039.57	2.06
Online00	1.00	1	1.58	0.58	0.00
Billpay00	1.00	1	5.73	31.41	0.00

Missing variables.

In the dataset we have 8,289 missing variables for *Age99* and 8,261 for *Inc99*.

Age99	Inc99
8289	8261

There are no missing values for the reminding variables used in our study and analysis. Since 26% of our data is missing there is an assumption made that we will be replacing missing values with KNN method.

Imputation is chosen as our main method for both instances, no data points will be deleted or omitted. First, a simpler method is implemented where we substitute the missing values in *Age99* and *Inc99* with median values from both. Median value for *Age99* is 4 and median value for *Inc99* is 6.

Second, more effective method for this data set is KNN (k- nearest neighbor) machine learning algorithm method. KNN is widely used and has many advantages. K- neighbors are chosen based on a specific distance measure. The method is effective mostly due to the ability to predict discrete and continuous data. For the rest of this case analysis we will be referring to values produced using the KNN method dataset unless states otherwise.

Descriptive Statistics.

First glimpse into the data, through function *describe()* in R, gives us statistical insight into *mean, sd(standard deviation), median, trimmed, mad(median absolute deviation), min value, max value, range, kurtosis and se(standard error)*. (Figure 1)

We can infer that an average customer brings in $111.5 profit to Pilgrim Bank. Max profit per customer is $2,071 and min(negative) profit is -$221 which gives us range of $2,292 and standard error of 1.53 and customer profitability deviates about $272.84 from the mean value.

Average profitability of online vs. offline

Further we go into analyzing profitability based on the fact if the customer has an online account. The Table1. shown below is a summary statistics for Mean and STD for Online and Offline(Paper) customers. There are 3,854 online customers.

Table 1

	Online(Mean)	Online(Std)	Paper(Mean)	Paper(Std)
Profit99	116.6668396	283.6646371	110.786249	271.300975
Online99	1.0000000	0.0000000	0.000000	0.000000
Age99	3.3111331	1.2306981	4.088054	1.482561
Inc99	5.9043594	2.0430174	5.322556	2.083931
Tenure99	8.6537909	6.9410974	10.372024	8.622122
District99	1203.6325895	40.5936474	1203.124550	48.834941
Billpay99	0.1370005	0.3438925	0.000000	0.000000

On average online customers are more profitable, has been with the bank shorter than offline customers, probably due to the fact that younger customers are introduced to the online banking sooner than older customers who might have to switch(inconvenience). Online customers on average have shorter tenure and their income is slightly higher.

A statistical t-test was performed to analyze if there is a relationship between means of online customers versus offline customers. The results were as follows:

```
 Welch Two Sample t-test

data:  pilgrim_data.K$Profit99 by pilgrim_data.K$Online99
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 in group 0 mean in group 1
       110.7862        116.6668
```

Customers who did not have online account were on average $5.8806 less profitable than customers who did have an online bank account at Pilgrim Bank. On average customers with bank account online had ~ $117 profit and customers that were offline generated ~ $111 profit.

H0: The null hypothesis is that there is no significant difference in means between the two groups.

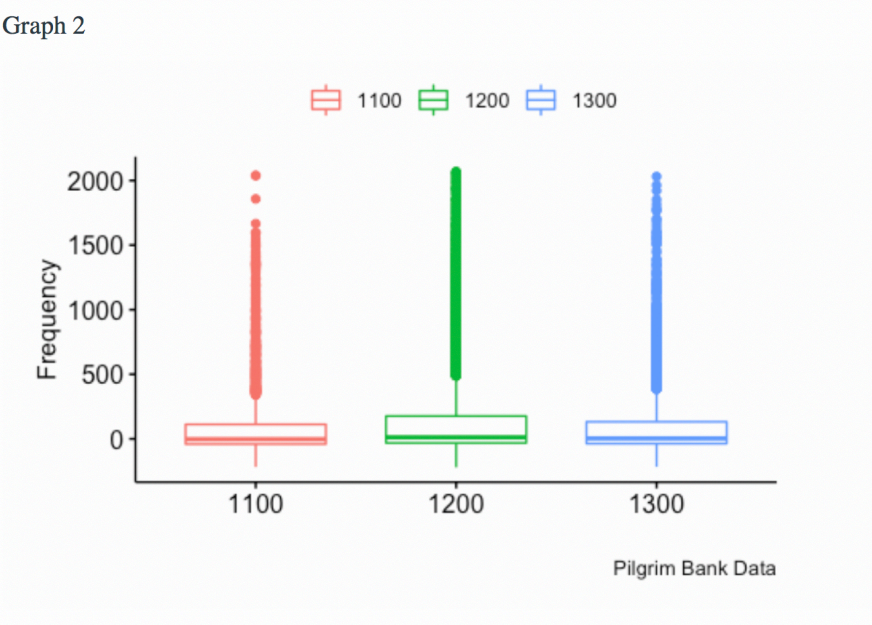
H1: The alternative hypothesis is that there is a significant (true) difference in means between two groups.

The t-value is t = -1.2124, df = 4882.1, p-value = 0.2254. With 95% Confidence Interval we fail to reject the null hypothesis and we conclude that at 5% level of significance, true difference in means between online and offline customers is not significant.

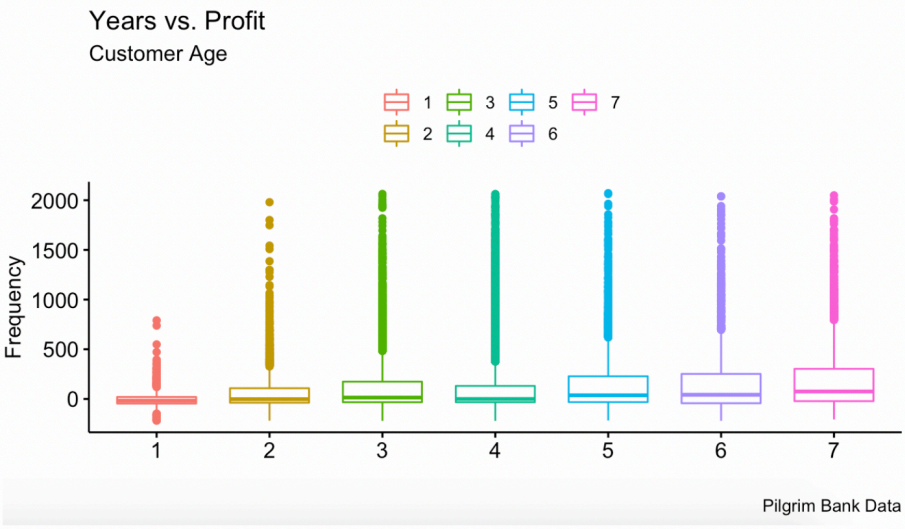
This statement is crucial to keep in mind when deciding on the Internet strategy going forward. The question that needs to be addressed is whether online customers are actually better customers. From this statistical test, all else equal, results from the data are not statistically significant for online customers. Visual relationship between *Profit* and *Online* outlined in Graph 1 and *Profit* and *District* in Graph 2 – both insignificant indicators of profitability for Pilgrim Bank and relationship between *Age* and *Profit* in Graph 3.

Graph 1

Graph 2



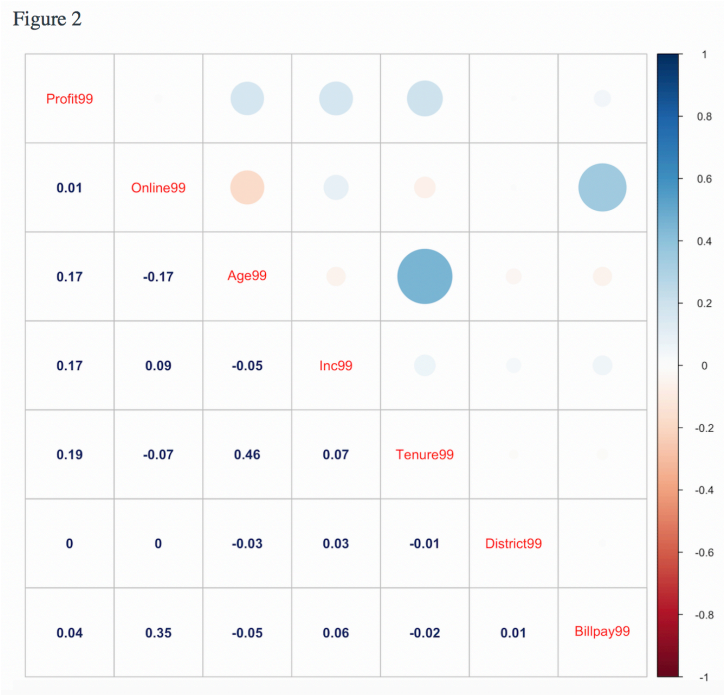
Graph 3



Correlation Analysis

Thereafter, correlation plot (Figure 2) is drawn to visually and numerically look at any, either positive or negative relationships between given variables in the dataset and the strength of this relationship.

Figure 2



Most of the variables are not correlated, sets of variables that stand out are *Age99 ~ Tenure99* with positive correlation of .46 and *Online99 ~ Billpay99* with positive correlation of .35. Both correlations are expected, customers do take out loans through their banks and maintain relationship and having an online account means that some customers will pay bills online. Surprisingly, correlation of .35 is smaller than expected in this instance.

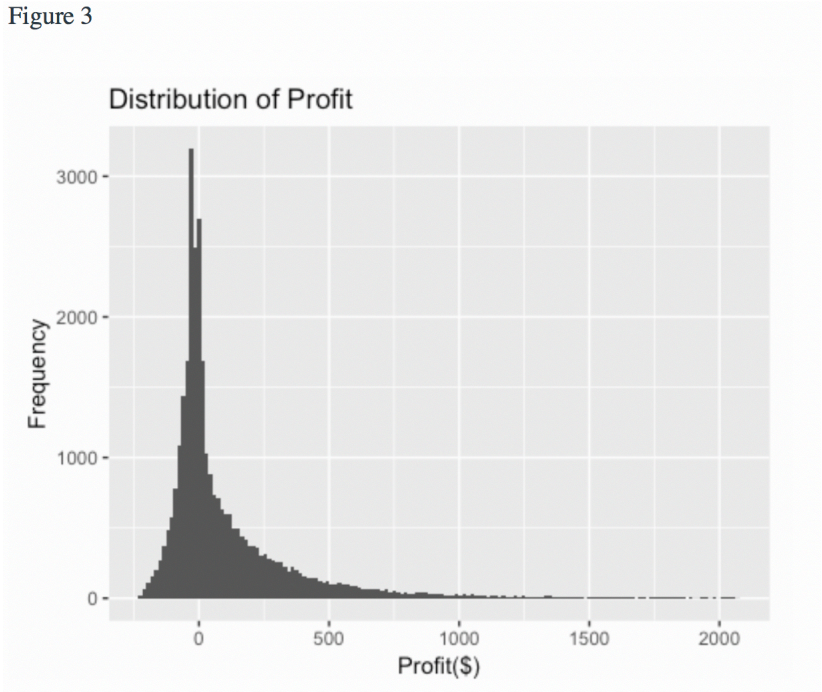
There is small positive correlation between *Profit99 and Age99, Inc99 and Tenure99*.

There is negative relationship between *Profit99 and Online99*.

Profit Distribution

Figure 3. shows that most of Pilgrim Bank customers are not profitable. Bank determines profitability based on a formula: "Profitability - (Balance in Deposit Accounts)*(Net Interest Spread) + (Fees) + (Interest from Loans) - (Cost to Serve) " (1). The probability chart shows that Pilgrim's most profit lies in -$100 to $200 range. 10% of all the customers generated about 70% of all the profit. The most profitable customers were those with higher Tenure and higher Age value. More profitable customers should be targeted at Pilgrims Bank.

Figure 3



Regression Analysis

Linear Regression is performed to further analyze customer profitability. From theory, all the variables included in the study(from 1999) do apply to be put into first predictive analysis. Four models will be built to be ensure best fitted regression model. *Profit99* as dependent variable.

In Model 1 both consumer segments were separately analyzed in linear regression.

Model 1

Online

```
Residuals:
    Min      1Q  Median      3Q     Max
-503.52 -158.59  -71.63   67.15 1950.76

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.019e+02  4.680e+01  -2.177   0.0295 *
Online99 == 1TRUE  6.369e+00  5.880e+00   1.083   0.2787
Age99              1.823e+01  1.245e+00  14.647  < 2e-16 ***
Inc99              1.774e+01  7.845e-01  22.610  < 2e-16 ***
Tenure99           4.020e+00  2.356e-01  17.057  < 2e-16 ***
District99         9.174e-03  3.846e-02   0.239   0.8114
Billpay99          8.289e+01  1.442e+01   5.747 9.18e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274.5 on 22805 degrees of freedom
  (8822 observations deleted due to missingness)
Multiple R-squared:  0.05879,    Adjusted R-squared:  0.05854
F-statistic: 237.4 on 6 and 22805 DF,  p-value: < 2.2e-16
```

Offline

```
Residuals:
    Min      1Q  Median      3Q     Max
-503.52 -158.59  -71.63   67.15 1950.76

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -95.511750  47.009213  -2.032   0.0422 *
Online99 == 0TRUE  -6.369098   5.879921  -1.083   0.2787
Age99              18.232303   1.244820  14.647  < 2e-16 ***
Inc99              17.736896   0.784469  22.610  < 2e-16 ***
Tenure99            4.019561   0.235648  17.057  < 2e-16 ***
District99          0.009174   0.038455   0.239   0.8114
Billpay99          82.885447  14.421623   5.747 9.18e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274.5 on 22805 degrees of freedom
  (8822 observations deleted due to missingness)
Multiple R-squared:  0.05879,    Adjusted R-squared:  0.05854
F-statistic: 237.4 on 6 and 22805 DF,  p-value: < 2.2e-16
```

Regression equations:

$$\text{Profit(online)} = -1.019e{+}02 + 6.369e{+}00\,Online99 + 1.823e{+}01\,Age99 + 1.774e{+}01\,Inc99 + 4.020e{+}00\,Tenure99 + 9.174e{-}03\,District99 + 8.289e{+}01\,Billpay99 + e$$

$$\text{Profit} = \text{Profit(offline)} = -95.511750 + -6.369098\,Online99 + 18.232303\,Age99 + 17.736896\,Inc99 + 4.019561\,Tenure99 + 0.009174\,District99 + 82.885447\,Billpay99 + e$$

Based on these two regression models we can say that online banking has smaller customer base. In both models the intercept is significant at 5% level of significance. The Adjusted R squared is very small. We

can conclude that although significant, the model does not explain the variation in the dependent variable(profit).

There is a big difference in the coefficients of particular variables and in the intercept, if we only take into account customers who have online banking but that could be due to sampling disposition. The evaluation is made that this is not the best model for this data. Model 2 (below) represents all variables for 1999 from the data set.

Model 2

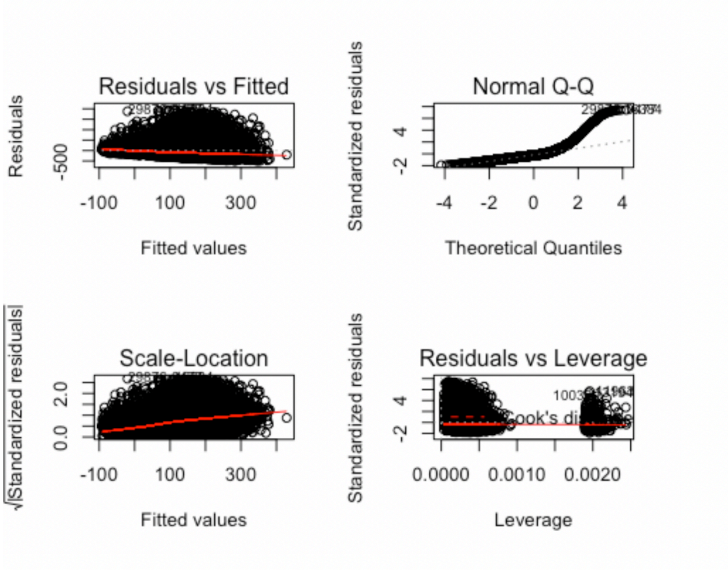
```
Residuals:
    Min      1Q  Median      3Q     Max
-502.93 -143.51  -64.59   53.55 1954.37

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -157.92222   37.63008  -4.197 2.72e-05 ***
Online99       7.00753    4.89834   1.431    0.153
Age99         22.12474    1.14524  19.319  < 2e-16 ***
Inc99         21.49119    0.71506  30.055  < 2e-16 ***
Tenure99       4.07879    0.19772  20.629  < 2e-16 ***
District99     0.01793    0.03086   0.581    0.561
Billpay99     77.06384   12.30956   6.260 3.89e-10 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262.7 on 31627 degrees of freedom
Multiple R-squared:  0.07338,    Adjusted R-squared:  0.07321
F-statistic: 417.5 on 6 and 31627 DF,  p-value: < 2.2e-16
```

This model indicates that all variables are significant, besides *Online99* and *District99*. Which indicates that there is no significant difference in locations and *having* online account on profit according to this model.

All values are significant and not strongly correlated(Figure 2) which is a good indicator of an unbiased model. The Adjusted R Square is improved slightly compared to the first model but is still low and again, we can infer variation in the dependent variable(Profit) can be only explained by 7% by the independent variables. F-statistics is significant at 5 % level of significance, which indicates overall good fit of the regression model.

$$\text{Profit} = -157.92 + 7.008\textit{Online99} + 22.125\textit{Age99} + 21.491\textit{Inc99} + 4.078\textit{Tenure99} + 0.01793\textit{District99} + 77.0639\textit{Billpay99} + e$$

Residuals and Errors:




```
Mean Absolute Error 137.7835

Mean Squared Error 68976.38

Root Mean Squared Error 262.6335
```

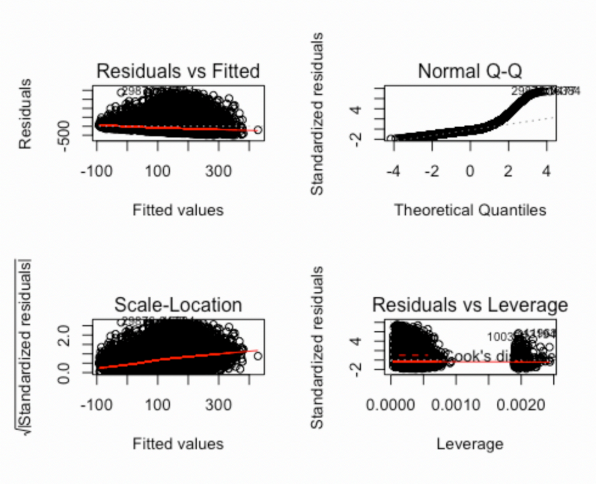
Taking out insignificant variables

To see if our model will improve we can take out the insignificant variables and compare models, residuals and errors.

Model 3

```
Residuals:
    Min      1Q  Median      3Q     Max
-503.05 -143.43  -64.68   53.60 1954.30

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -136.3383     5.9572 -22.886  < 2e-16 ***
Online99       6.9899     4.8982   1.427    0.154
Age99         22.1041     1.1447  19.310  < 2e-16 ***
Inc99         21.5033     0.7147  30.085  < 2e-16 ***
Tenure99       4.0792     0.1977  20.631  < 2e-16 ***
Billpay99     77.0955    12.3093   6.263 3.82e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262.7 on 31628 degrees of freedom
Multiple R-squared:  0.07337,    Adjusted R-squared:  0.07323
F-statistic: 500.9 on 5 and 31628 DF,  p-value: < 2.2e-16
```

Residuals and Errors:




```
Mean Absolute Error 137.7533

Mean Squared Error 68977.11

Root Mean Squared Error 262.6349
```

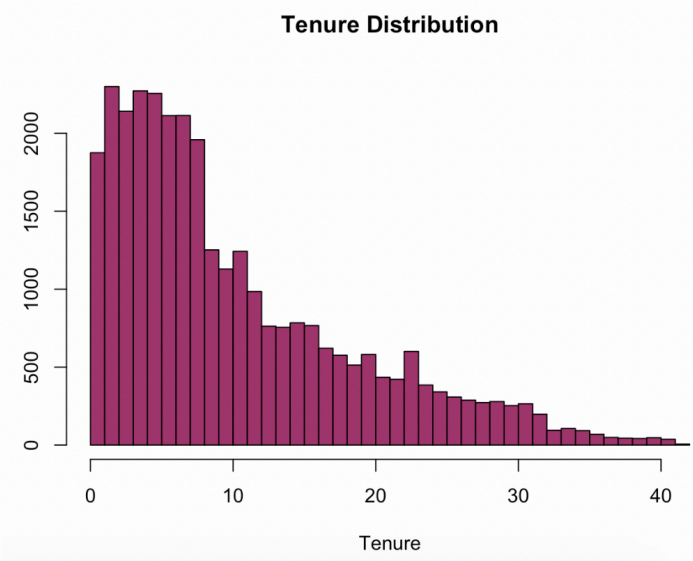
As we can see from the summary analysis, taking out insignificant variable(besides *Online*) did not change our model, besides the intercept that went down to ~ - 136.

Errors and Residuals remined similar. All the variables will be kept.

Taking a log

Another modification to our model is to take a log from the variable Tenure since the distribution is skewed to the right (Figure 4).

Figure 4



Model 4

```
Residuals:
    Min      1Q  Median      3Q     Max
-509.04 -146.66  -63.63   53.79 1968.92

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -208.50902   37.70873  -5.529 3.24e-08 ***
Online99             6.37268    4.90279   1.300    0.194
Age99               24.19525    1.11542  21.692  < 2e-16 ***
Inc99               20.92790    0.71992  29.070  < 2e-16 ***
log(Tenure99 + 1)   38.53027    1.99926  19.272  < 2e-16 ***
District99           0.02256    0.03089   0.730    0.465
Billpay99           77.06784   12.31997   6.256 4.01e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 262.9 on 31627 degrees of freedom
Multiple R-squared:  0.07182,    Adjusted R-squared:  0.07164
F-statistic: 407.8 on 6 and 31627 DF,  p-value: < 2.2e-16
```

```
Mean Absolute Error 141.5448

Mean Squared Error 69093.05

Root Mean Squared Error 262.8556
```

In this model our intercept has increased and so did the absolute errors. Adjusted R Square and significant variables remined significant.

Picking the model.

Based on Errors comparison as well as coefficient and logic, model 2 is our model for this case analysis. Since the R Squared is relatively small and there are not any significant changes between model, all variables from 1999 data should be included.

Business Implications, Recommendations and Conclusion

The original case had in mind one question: are online customers more profitable than offline? Through our data analysis we concluded that indeed they are more profitable, but that profitability is not significant. Other factors through demographics like how old the customers actually are (– younger customers are more likely to have an online account but older customers are more profitable), income also plays a significant role in customer profitability as well as tenure. These factors are something that Pilgrim might be overlooking when solely focusing on promoting online banking. Pilgrim's goal should be to generate more profit through existing customers and customer satisfaction. The management team should not be spending all their resources and money on trying to promote online banking unless customers will only pay their bills online. *Billpay* is one of the very significant variables that was not mentioned much in this case analysis as it was not the topic of it, nether less this aspect cannot be ignored because we can infer from previous models and statistics that having an online account is not significant but if the customer actually uses the online account to pay bills online might have an enormous implication for Pilgrim Bank and could save millions of dollars in profit each year by cutting on cost per customer.

Pilgrim Bank should evolve its analysis around demographic rather than focusing online and have better variables that could build a better predictive model with a stronger Adjusted R Square and better SE.

Recommendations

The marketing team should focus at the moment at collecting better data and more differentiated data that could help explain environment and therefore build better predictive models which would eventually transform into better profit. There should be more analysis done on how significant the *Billpay* is and how it can influence other variables. The management team should implement a policy that if you have an online banking you have to pay bills online. Older customers are more profitable but we can assume that plenty of the older customers do not know or trust computer systems and should be given free training and encouragement. Based on our analysis loyal customers should also be rewarded, if Tenure is plus ten years, customers should not be paying fees for online banking or assists. They are profitable somewhere else and customer loyalty should be prioritized is possible.

Banks should start behaving more like SASS companies where consumer satisfaction in above all. Amazon Services or PayPal have great customer satisfaction rates and that eventually transition into profit. Predictive model of a ‘perfect’ customer could be build if given a better variable data set and give more insight into the ’between the lines’ data that is so often disregarded by businesses.

References:

"Pilgrim Bank (A): Customer Profitability." Frei, Frances X.; Campbell, Dennis. Case No. 9–602–104. Published 10/19/2001, Revised 08/25/2005. **Harvard Business School Publishing,** (8 pages).