Pilgrim Bank Case Study

Mason Hansen

20207795

1/22/2019

Case Formulation

Based on the sample, this analysis will investigate what customer profitability will look like for the entire population of Pilgrim Bank customers.

- Descriptive Statistics will be analyzed for customer profitability in order to gain better insight on potential variables that might have influence customer profitability for Pilgrim Bank in the year 1999.

Then, the online banking channel data will be investigated to determine if online banking increases profitability or detracts from profitability. A Students' t-Test will be utilized to test the difference in mean profitability values between customers who use the online banking channel versus customers that do not use the online banking channel.

The hypothesis for this analysis is as follows:

> H_0: The average value of profitability for online banking and non-online banking customers is the same ($Profit_{online} = Profit_{offline}$)
>
> H_a: The average value of profitability for online banking and non-online banking customers is not equal ($Profit_{online} \neq Profit_{offline}$)

Finally, other statistical tests will be conducted to determine possible cofounding variables, such as customer demographics, that have potential influence in customer profitability.

- A linear regression model will be applied to investigate customer demographics and their impact on customer profitability for the year 1999.

Statistical Methods

For the purpose of this case analysis, the data will come from Pilgrim Bank and include information about their customers from the years 1999 and 2000. This analysis will not use the data from 2000 and focus entirely on the 1999 data. The 1999 variables included in the analysis are as follows:

- Profit from customers (in US Dollars)
- Online Banking Usage (Yes / Online or No / Offline)
- Customer Age (Binned by year: "< 15 years", "15-24 years", "25-34 years","35-44 years", "45-54 years", "55-64 years", "65+ years")
- Customer Income (Binned in dollars: "< 15k", "15-19.99k", "20-29.999k","30-39.999k", "40-49.999k", "50-74.999k", "75-99.999k", "100-124.999k", "125k+")
- Customer Loan Tenure (in years)
- Customer District
- Pay bills online (Yes / No)

The data set is not entirely complete, including missing values in ~25% of the sample variables age and income. In order to account for missing values for proper analysis, deletion of ~25% of a sample is too aggressive. For this analysis, a dynamic method of imputation was chosen as the accurate replacement of missing values. The method use is called K-nearest-neighbors imputation (KNN). This method was chosen because all of the missing values were either age or income, which were previously binned into ranges of values. Since both age and income are now represented as categorical variables, the precision of a central tendency imputation of either the median or mean would seem inadequate. KNN is an algorithm that classifies each data point based on every independent variable in the data set (profit, age, income, tenure, district, online) and measures the distance from each data point to the central value (mean) of each cluster. For example, if person A and person B have similar profit, income and tenure, but only person A is missing their age. The algorithm will calculate the distance between person A and person B and filling in the missing value with the value of the "nearest neighbor". In this scenario, person B is nearest to person A, so the missing age of person A will be imputed with person B's age value. This algorithm can handle continuous, categorical and binary data, which is the main reason why this method was chosen.

Descriptive Statistics

Descriptive statistics of the sample data set shows that the average value of profit across all customers are about $111.5. When accounting for customers that use online banking versus those who do not use online banking, the average value of profit for online customers is $116.67 and offline customers is $110.79. This might suggest that for customers that use the online banking channel generate sliglty more profit than those who do not use the online banking channel.

To better visualizle the data, a series of boxplots were plotted. Including each of variables Online, Age, Bill Pay and Income plotted against Profit. Outliers have been plotted in Gold. The variable District has been excluded from the visualization.

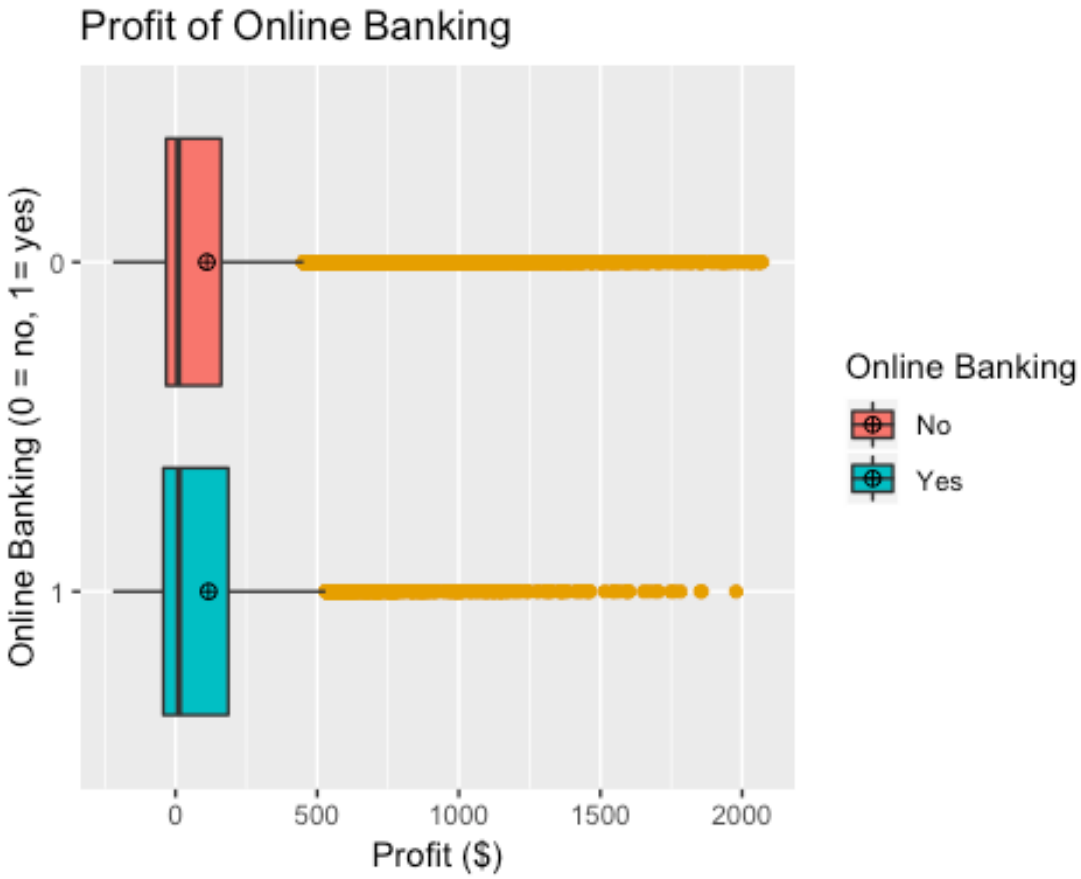

Figure 1. Boxplot (Profit x Online). Does not appear to have significant difference between the two groups.

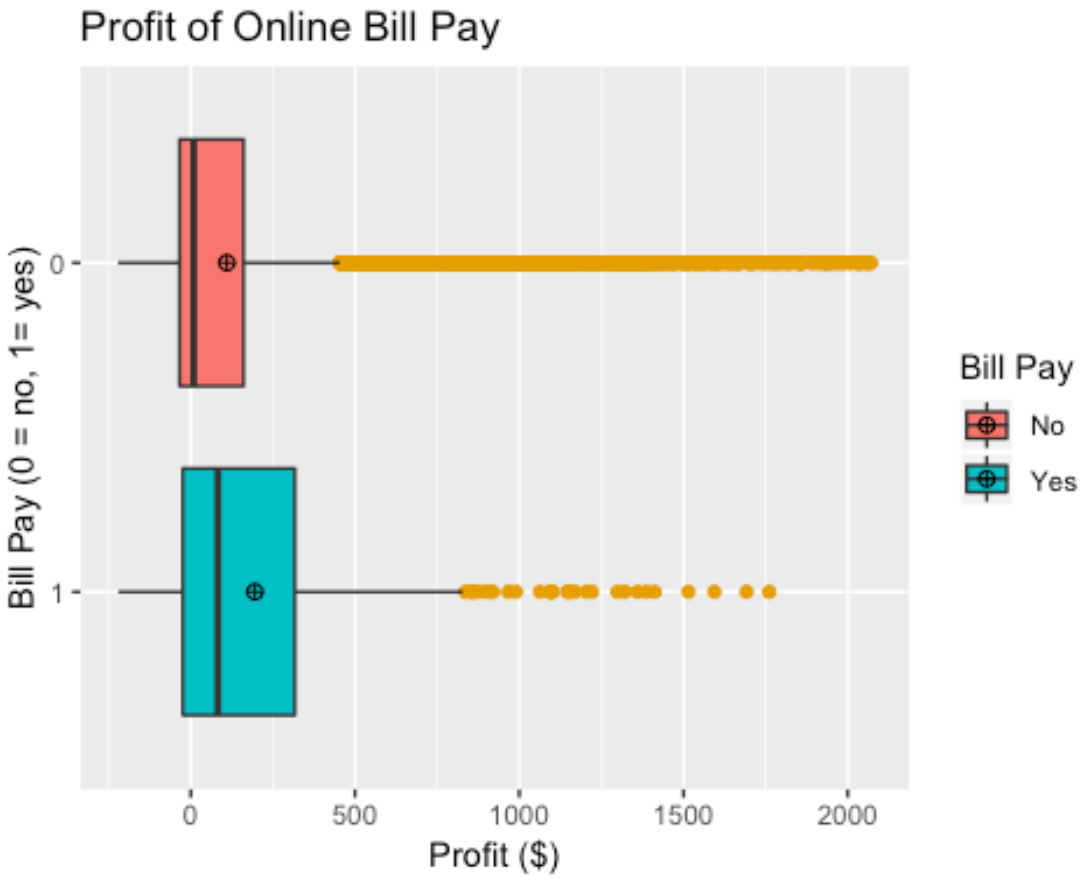

Figure 2. Boxplot (Profit x BillPay). Does not appear to have a significant difference between the two groups.

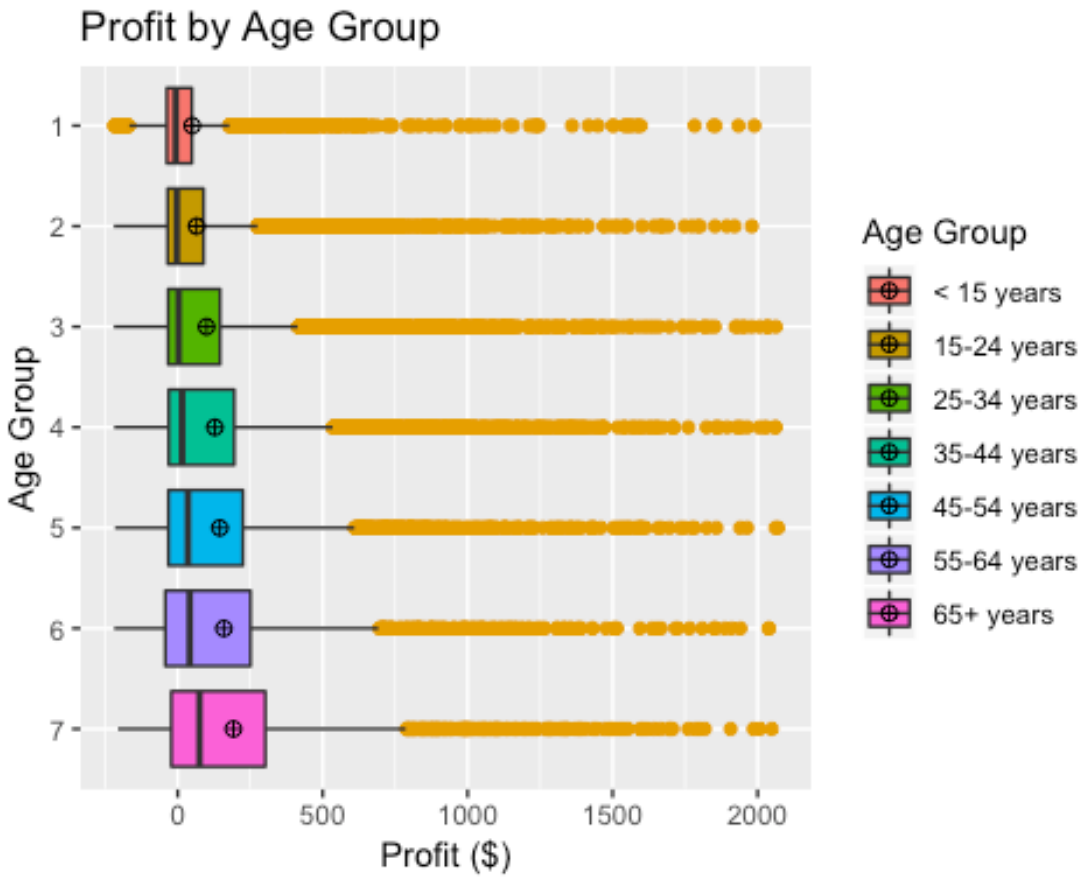

Figure 3. Profit x Age. Groups do not appear to be significantly difference from each other.

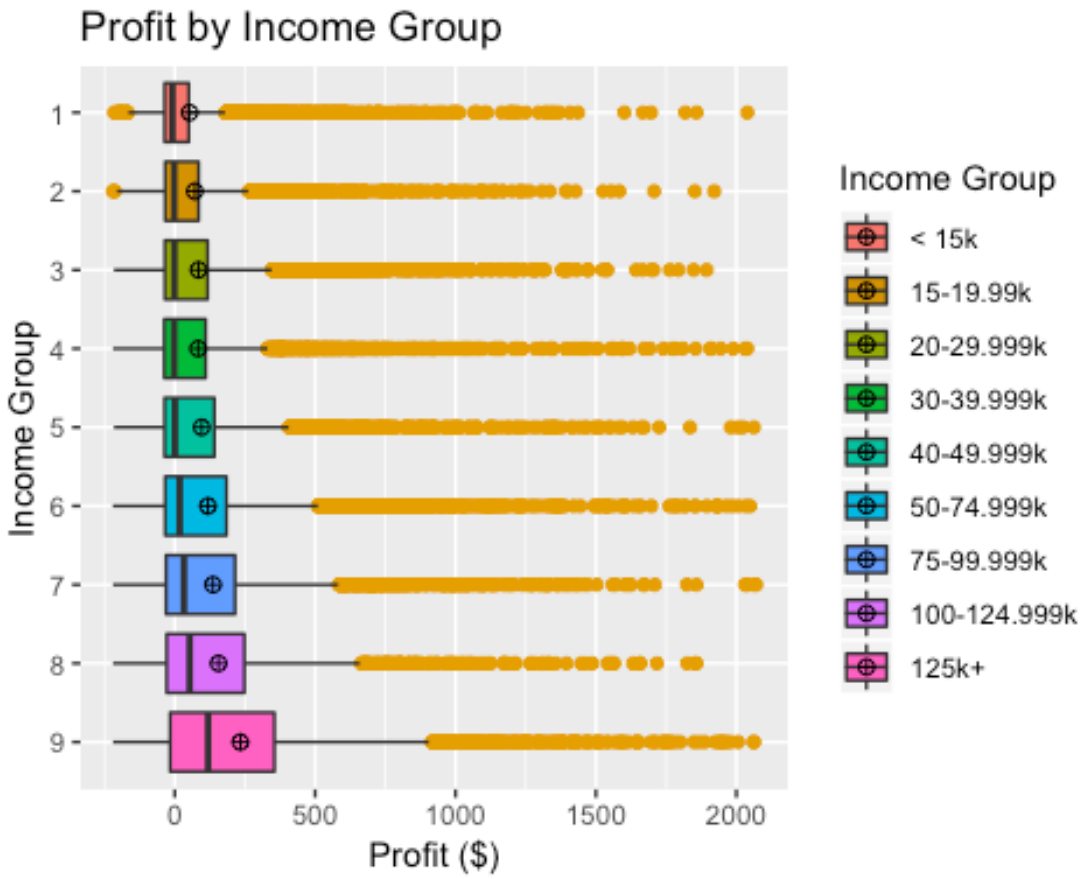

Figure 4. Profit x Income. Groups do not appear to be significantly different from each other.

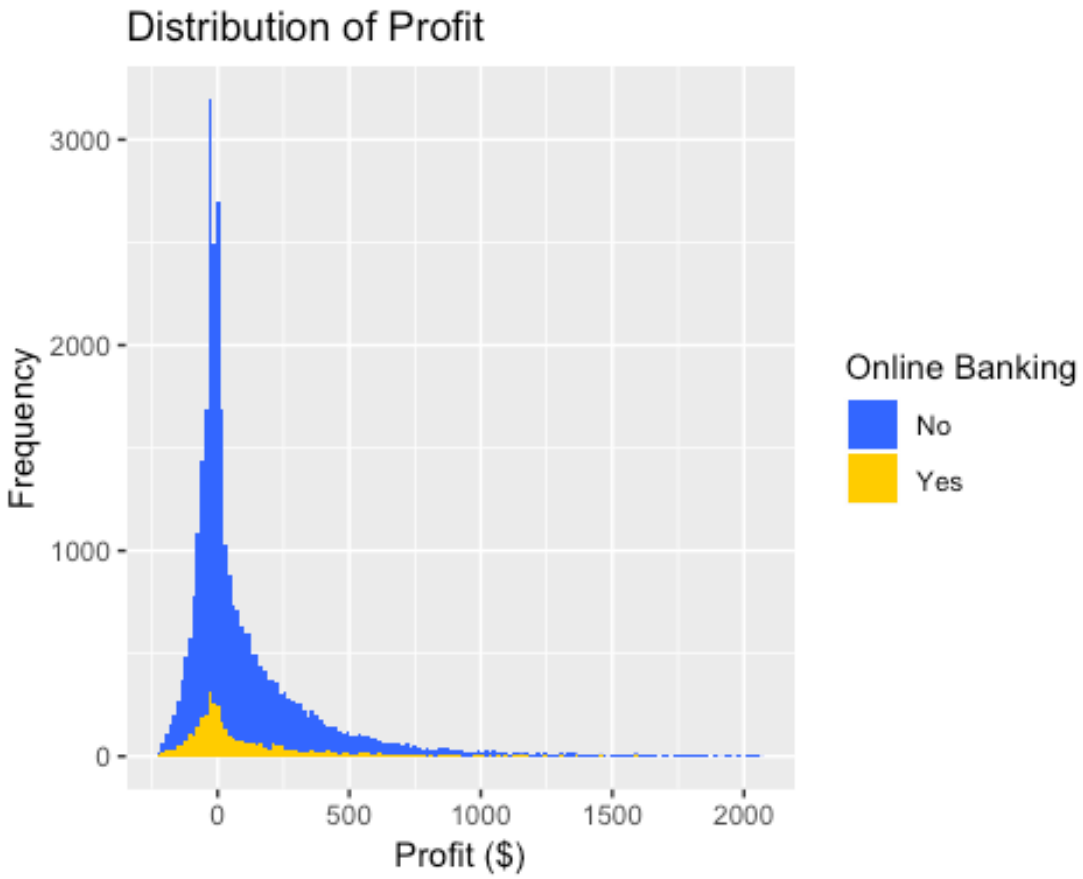

Figure 5. Density distribution of profit with online vs. offline users. Highly skewed-right. Also noted that there is a very small proportion of online banking customers.

Finally a clustering technique was applied to better visualize the classification of the Pilgrim Bank customers. This clustering technique is most useful for understanding what customer attributes tend to group together with other various customer attributes. Most Candidates will fall within these regions:

Summary Diagnostics

Number of Clusters:	5
Number of Points:	170
Between-group Sum of Squares:	25.595
Within-group Sum of Squares:	48.444
Total Sum of Squares:	74.039

		Centers		Most Common			
Clusters	**Number of Items**	**Avg. Profit**	**Avg. Tenure**	**Age**	**BillPay**	**District**	**Income**
Cluster 1	24	174.31	6.9212	55-64	No	1200	40-50k
Cluster 2	45	129.33	13.223	<15	Yes	1200	15-20k
Cluster 3	30	67.216	6.6452	45-54	Yes	1100	40-50k
Cluster 4	42	236.2	16.024	25-34	No	1100	<15k
Cluster 5	29	155.52	10.506	55-64	No	1100	125k+
Not Clustered	0						

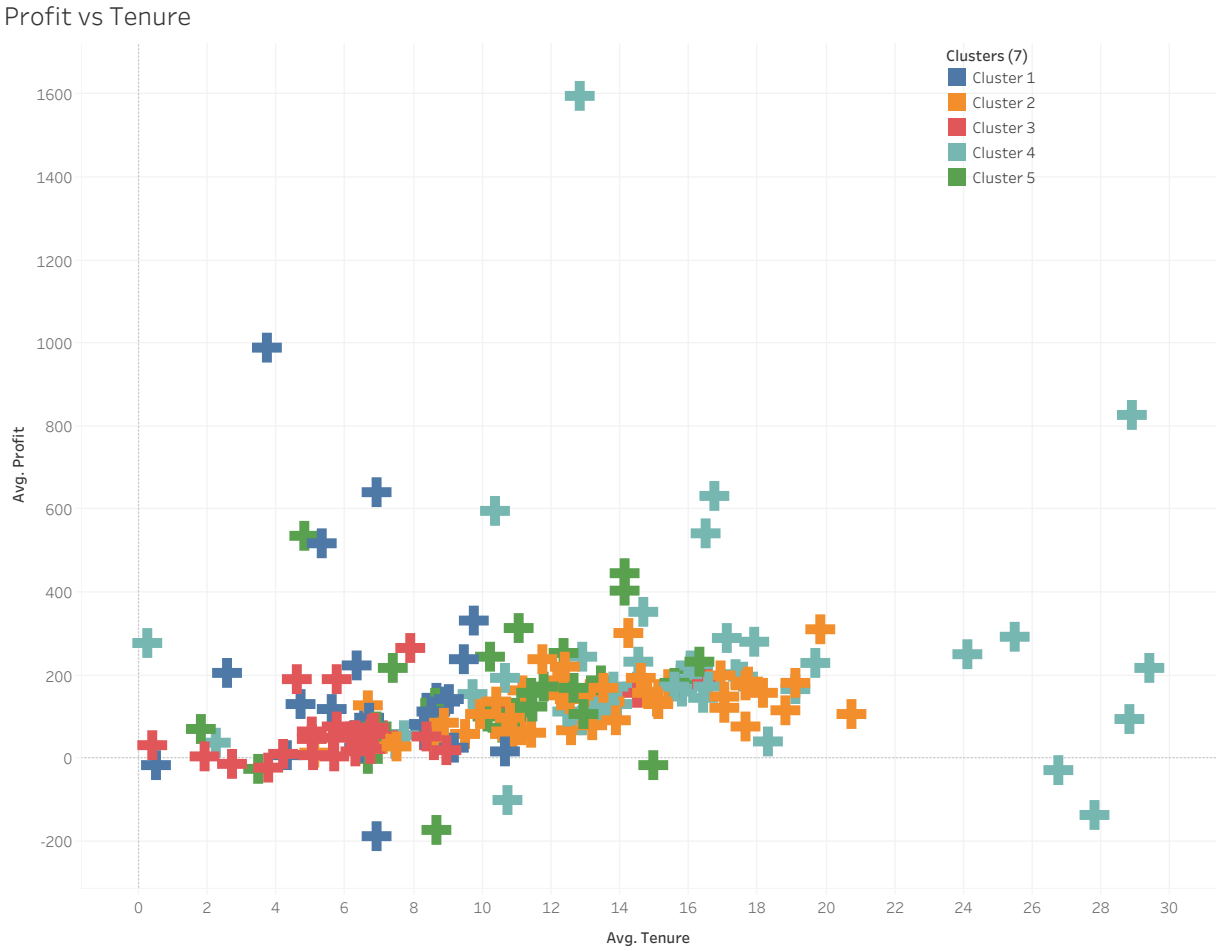

Figure 6. Clustering into 5 similar groups

t-Tests for Profitability

From the boxplot visualizations (Figures 1-4), it appears that none of the factors have a significant effect on customer profitability. The main analysis involves online versus offline customers and the effect on their profitability (Figure 1). A formal t-Test was conducted to check the significance of the effect of online banking.

> H_0: The average value of profitability for online banking and non-online banking customers is the same ($Profit_{online} = Profit_{offline}$)
>
> H_a: The average value of profitability for online banking and non-online banking customers is not equal ($Profit_{online} \neq Profit_{offline}$)

Results are as follows:

> t = -1.2124, df = 4882.1, p-value = 0.2254

The p-value is greater than the alpha level of 0.05, thus failing to reject the null hypothesis and unable to accept the alternative. With the sample data given, it cannot be concluded that there is any effect of online banking. There is no difference between the profitability of online versus offline banking customers.

Linear Regression

An initial linear model (Table 1, below) takes variables – Age, Billpay, District, Income, Online and Tenure – to help predict profitability outcome. Across all variables and all levels, most were statistically significant (p-val < 0.05). All except for customers that fall within Age bin #2, District 1300 or Online Banking which do not significantly add the linear model.

Each estimate from the model is the value added to the linear model given a one unit increase in continuous variables or the presence of an indicator variable. Every estimate that is significant to the model has a positive value.

The predictor variable online (1= yes, 0 =no), is non significant. This means that in predicting customer profitability, the fact that the customer uses online banking is irrelevant.

For example, if Pilgrim Bank wanted to predict the customer profitability of a customer that is 45 years old, uses Billpay, is from district 1200, has 70k income and has 4 years of tenure, the linear model would look like this

Y_{Profit} = Int + 35*Age5 + 78.4* BillPay1 + 17.7*D1200 + 49.9*Inc6 + 4.77*Tenure

= -27.2+35*1+ 78.4* 1 + 17.7*1 + 49.9*1+ 4.77*4

= $ 172.88

TABLE 1. Fitting linear model: X9Profit ~ X9Age + X9Billpay + X9District + X9Inc + X9Online + X9Tenure

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.2	8.09	-3.36	0.000786
X9Age2	4.16	6.3	0.66	0.509
X9Age3	19.8	6.26	3.16	0.00155
X9Age4	27.7	6.52	4.25	2.18e-05
X9Age5	35	7.19	4.87	1.13e-06
X9Age6	49.9	7.87	6.33	2.45e-10
X9Age7	84.5	7.68	11	3.89e-28
X9Billpay1	78.4	12.3	6.35	2.24e-10
X9District1200	17.7	5.12	3.45	0.00056
X9District1300	7.02	6.26	1.12	0.262
X9Inc2	18.4	7.46	2.46	0.0138
X9Inc3	26.3	6.18	4.26	2.09e-05
X9Inc4	24.2	6.19	3.92	8.98e-05
X9Inc5	36.1	6.42	5.62	1.96e-08
X9Inc6	49.9	5.9	8.46	2.89e-17
X9Inc7	68.1	6.68	10.2	2.34e-24
X9Inc8	85	7.96	10.7	1.33e-26
X9Inc9	155	6.99	22.2	9.14e-109
X9Online1	0.863	4.89	0.176	0.86
X9Tenure	4.77	0.191	24.9	1.08e-135

Business Implications

Pilgrim Bank's business relies on their ability to maximize customer profitability. The initial thought process was that there were inherent costs with certain banking platforms or channels that could impede the customer profitability. In this analysis, the online

banking channel was the center of focus. An initial look at the data suggested that customer profitability for online banking customers was higher than offline banking customers.

The difference in online versus offline customers turned out to be insignificant. After performing an official t-Test, the results were not significant enough to determine any difference between the two groups. From this sample, it is impossible to say that online banking customers generate more profit, yet since the average profit was higher for online users it is still something to consider in the future. If the fee structure for online banking changes, i.e. more transaction fees, account fees, late payments, etc…then the difference between online and offline could widen and potentially produce a significant effect. It should also be noted that online banking at this time is relatively novel and cannot compare to the quantity of offline customers in 1999. In this data sample, offline customers accounted for about 88% all customers.

If online banking cannot influence customer profitability, other variables must be able to explain some of the variability. From the results of the linear model, customer demographics (age, income, location) along with other customer parameters (billpay, tenure) significantly add to the model of customer profitability. While each variable consisted of several levels (i.e. age consists of 7 levels), almost all levels across each variable was significant. The only predictor variables that were not significant were age group #2, customers from district 1300 and online banking.

Pilgrim Bank could make much better business decisions with the information from the linear model. It can be assumed that Pilgrim Bank's largest business problems is customer retention and customer acquisition. From the data and resulting linear model, Pilgrim will be able to identify which customers they should be trying to retain and which customers they should be trying to attain. Since, there are many predictor variables that are both significant and have positive estimates that will add value to predicted profitability of any given customer, Pilgrim has many options and combinations of customer demographics or customer parameters to ensure that they are spending their money of the right quality of customer.

The ideal candidate to maximize profitability would have the following characteristics:

Age: 65+ years

BillPay: Yes

District: 1200

Income: $125k+

Tenure: Maximized (approx. max value in data set ~=41 years)

A candidate with similar attributes would result in a predicted profit of $ 503.9. This value should be considered as Pilgrim Bank's entitlement performance objective or best-case scenario given historical customer data. There was not enough external data given to determine a benchmark value, but Pilgrim Bank should aim to reach ~$504 profitability per customer from their current baseline value of ~$112 profitability per customer.