**BAMA 520 Customer Analytics**

**Group Assignment 3**

**Pilgrim Bank Case**

**Group 17**

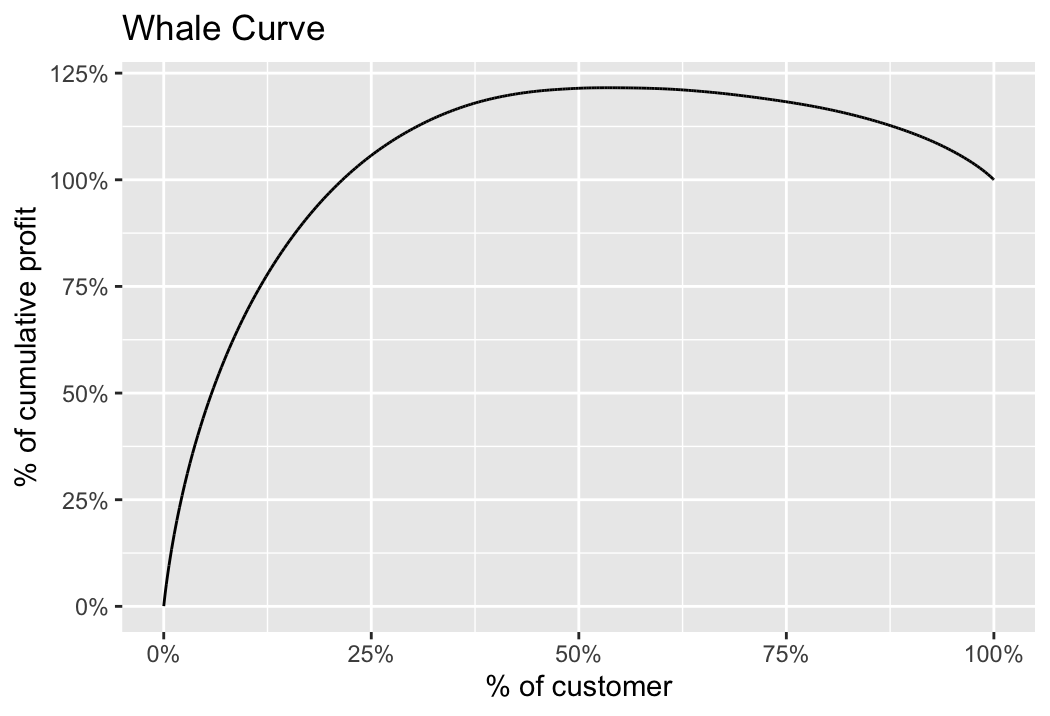
Xinyuan Wang

Iris Xing

Zhao Zhang

Robert Zhao

**1. Probability Skew (R-generated)**



**2. Heterogeneity in Customer Profitability**

**a) Determine the minimum percentage of customers that Pilgrim Bank would need to realize 100% of their current profit.**

21.54% of customers are required to realize 100% of their current profit.

**b) Determine the percentage of total profit Pilgrim Bank could realize by targeting profitable customers only. What is the percentage of customers they need to realize the maximum profit?**

53.21% of customers are required to realize this maximum profit.

**c) What might explain the variation in profitability across customers?**

As question a) and b) suggests, there is a considerable variation in profitability among customers, resulting in profitable and unprofitable customer groups. This variation in profitability might be explained by different customer demographics and customer behavior based on the bank industry’s profitability formula.

The first source of the revenue is the customers’ deposit balance. A businessman’s account balance may easily outperform a poor student’s with a limited balance. Revenue can also be generated through the fees charged. Depending on the different accounts and services, fees will vary across customers. For example, some customers are fined for late payments, comparing to those who always repay the debt on time. Additionally, loan interests, a bank’s most dominant assets also varied significantly across customers depending on the types, lengths, and interest rates of the loans.

In terms of the cost-to-serve, it involves transaction-related cost, which depends on customer banking activities and habits. For example, the more frequent a customer visits a bank, the more costly the customer is. The more frequently a customer utilizes online banking, the less expensive the customer is. Therefore, the variation in cost-to-serve explains the variation in profitability across customers as well.

**3. Estimate an intercept-only model. How do you interpret the parameter estimate for the intercept? Determine the 95% confidence interval for this parameter estimate. What precisely does this confidence interval mean?**

The intercept-only model has no explanatory variables. It does not explain anything of the response variable. It only provides the average value of our response variable, Profit99. The 95% confidence interval is from 108.496 to 114.5094. This confidence interval means that we’re 95% confident that the average response variable should fall between 108.496 to 114.5094.

**4. Estimate the six linear regression models below to answer 4a-4c, where “AgeMiss” is a dummy that indicates missing data on Age, “AgeMean” uses mean replacement with NA, and “AgeZero” uses zero replacement with NA. The same applies to “IncMiss”, “IncMean” and “IncZero**

**a) Provide a table that compares the coefficient of online99, income, and age. Between models 4 and 5, when you apply the third approach (discussed in class) to deal with missing data, does the mean or zero replacement affect the coefficient of age or income?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Online** | **Age** | **AgeZero** | **AgeMean** | **Income** | **IncZero** | **IncMean** |
| m1 | 17.4587 | 18.472 | N/A | N/A | 17.17 | N/A | N/A |
| m2 | 16.0814 | N/A | N/A | 16.2463 | N/A | N/A | 16.7374 |
| m3 | 8.7513 | N/A | 4.0054 | N/A | N/A | 9.1089 | N/A |
| **m4** | **13.8233** | N/A | N/A | **16.6701** | N/A | N/A | **16.853** |
| **m5** | **13.8233** | N/A | **16.6701** | N/A | N/A | **16.853** | N/A |
| m6 | 14.842 | N/A | 16.3998 | N/A | N/A | 74.8566 | N/A |

The mean or zero replacement does not affect the coefficient of age or income, as evident from the comparison between m4 and m5 models. Essentially no matter which imputation method we choose, no additional information is provided to the regression model.

**b) Are you treating “Age” and “Inc” as continuous or categorical in your analyses? Does it affect the coefficient of Online99?**

In the models we built, we treated the “Age” and “Inc” as continuous. Compared to the models whose variables are categorical, the coefficients of Online 99 didn’t change much. The main reason is that the categories in our case represented the natural order of age and income as they are of the continuous type.  Therefore, even if we treat them as factors in the models, they still bring in the continuous attributes.  Besides, one implicit assumption is that the customers within each of the same categories of ‘Age’ and ‘Inc’ tend to have similar profitabilities.

As the types of the variables ‘Age’ and ‘Inc’ don’t affect their importance/weighted in the model, the coefficient of online 99 won’t change much either.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Slope of online99** | **Model 1** | **Model 2** | **Model 3** | **Model 4** | **Model 5** | **Model 6** |
| Continuous (mi) | 17.4587 | 16.0814 | 8.7513 | 13.8233 | 13.8233 | 14.8420 |
| Categorical  (mi.c) | 17.0251 | 13.4983 | 13.4983 | 13.4983 | 13.4983 | 14.9452 |

*\*Please see the R file for the details of the models mi.c.*

**c)  What is your conclusion regarding the difference in profitability between customers who used online banking vs. not? Is the difference statistically significant?**

All models but mod3 show statistically significant (at 1%) and positive coefficient. Mod3 shows a statistically significant (at 10%) and positive coefficient. The difference in profitability between customers who used online banking vs. not is statistically significant.

**5.  Based on your analyses generate conclusions and recommendations for management. Does the online channel make customers more profitable? What is your recommendation to management regarding a pricing strategy for the online versus offline channel?**

Since using an online channel has a positive performance on a customer’s profitability, it is recommended that the management engage in promotional activities to acquire customers and help existing customers potentially switch channels. For the pricing strategy, the bank should encourage greater use of the online channel through cost-incentives rather than charging online users.

**6.  How well does customer channel use in 1999 predict how profitable customers will be in 2000?**

|  |  |  |
| --- | --- | --- |
|  | R-Square | P-Value |
| Model z2 (with online99) | 0.02475 | < 2.2e-16 |
| Model z2.c (without online99) | 0.02426 | < 2.2e-16 |

The p-values of both models are less than 2.2e-16, which means both models are significant.

All p-values of the explanatory variables in both models are less than 0.01, which means all variables are significant in this model. (Please refer to the R file for the details.) However, while comparing the two models, we realized that dropping the variable online99 doesn’t decrease the R-square much, 0.02475 - 0.02426 = 0.00049, which is only a 0.05% difference. We would like to say that online99 doesn’t help explain the dependent variable well. Therefore, customer channel use in 1999 is not a good predictor for customer profitability in 2000.

**7. How well does customer channel use in 1999 predict whether customers stay with the bank through 2000?**

|  |  |  |
| --- | --- | --- |
|  | AIC | Hit Rate |
| Model 1 (with online99) | 23601 | 79.35% |
| Model 2 (without online99) | 23602 | 79.51% |

We compared two models. The first model includes whether a customer used the Online channel in 1999 as an independent variable. This model results in a hit rate of 79.35%, which indicates that this model is good at predicting whether customers stay with the bank through 2000. The sensitivity rate is 82.02%, and that means 82.02% of times that the model detects a response when customers actually respond. The specificity rate is 65.71%, and that means 65.71% of times that the model predicts no response when customers actually have no response. And the False positive rate is 34.29%.

For the second model, with removed Online Channel 1999 from the independent variables with no change to the rest of the parameters. We achieved the following results:

Hit Rate -- 79.51%

Sensitivity Rate -- 82.29%

Specificity Rate -- 65.46%

False Positive Rate -- 34.54%

In conclusion, as we can see from the table, removing online99 does not worsen the model’s predictability, and the measures improve by a tiny amount.  Therefore, the variable online99 (customer channel use in 1999) does not predict how profitable customers will be in 2000.