MS -Business Intelligence & Analytics

Spring 2015

*April 20, 2016*

**Vivek-John Dominic Martins**

Home Work 4

**Ethics Statement**

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_

**Pilgrim Bank. Second part.**

**5.a. Evaluate the drivers of customer profitability for the year 2000 (Hint: you can evaluate the variables explored for profitability of 1999).**

On running a regression model of 1999 variable on 2000 customer profitability, I obtained the results that the 1999 Online Channel, Age ,Income, Existing Income data and Tenure are the drivers for cutomer profitabilty for the year 2000.

**5.b. Evaluate if the variable 9Profit should be included in the customer profitability analysis for 2000.**

9Profit should be included as the model predictability (R-squared value) improves, as well as it makes other factors such as Age and Income Exist redundant in the model.

**5.c. Evaluate the drivers of customer profitability for 2000 after adding electronic billpay using OLS and Random Forest regression. Compare these**

**two methods based on their Rsquared.**

The Random forest regression has an R-squared of 0.878, while the OLS has an R-squared of only 0.361. Therefore the Random Forest Regression model is a better predictor of customer profitabilty for 2000.

Also note the OLS showed that electrnoic billpay was not a significant driver of customer profitabilty for 2000.

**6. a Evaluate the drivers of customer retention for the year 2000. Compare OLS with logistic regression results and indicate which one of these two**

**Methods is more appropriate for this problem. Retain takes a value of 0 when 0Profit has a missing observation and 1 otherwise.**

The two models are very close, however the Logistic Regression has a higher R-squared of 0.1801 (versuse 0.177 for OLS). This also makes sense since logistic regression is used when the predicted variable is binary. In this case customer retention for the year 2000 is infact a binary variable.

**6.b. Calculate and rank the odds ratio (exp(coefficients of logistic regression)) of the logistic regression and explain the impact of the 9Online and**

**Billpay variables on customer's retention.**

The table below shows that the odds for retention are 1.210107 given that the customer uses the Online Channel. While the odds for retention are 0.549678 for customer that use the electronic bill pay option.

Odds Ratio rank

AgeExist 3.057265 1

IncExist 2.902960 2

9Online 1.210107 3

AgeAvg 1.071095 4

D1200 1.069993 5

IncAvg 1.055550 6

9Tenure 1.038118 7

9Profit 1.000195 8

D1100 0.967212 9

9Billpay 0.549678 10

const 0.519526 11

**7. Evaluate the effect of the online channel and billpay on customer's retention using a hidden Markov model (HMM) with the variables 9Online, 9Billpay, 0Online, 0Billpay and the new variable Retain.**

From the Transition matrix we can see that the probabilities for Online channel to Retain are slightly higher than for transitioning from Billpay to Retain.

Transition matrix

[[ 0.660746 0.09116221 0.16129807 0.0240987 0.06269502]

[ 0.66173943 0.09694053 0.15984874 0.02475077 0.05672052]

[ 0.64604009 0.09359798 0.16870987 0.02977233 0.06187974]

[ 0.6754386 0.0764411 0.16165414 0.02882206 0.05764411]

[ 0.65644172 0.09509202 0.16206544 0.02453988 0.06186094]]

**8. Build a transition matrix (online, billpay) from 1999 to 2000 from those different customers' states: those that were online, offline without electronic billpay, online with electronic billpay for 2000, customers who left the bank. Explain the billpay effect on customers' retention.**

From the transition matrix below we see that the probabilty of transitioning from electronic billpay in 1999 to retention is 0.05672052. While the transition probabilty from electronic billpay in 2000 to retention is 0.05764411

Transition matrix

[[ 0.660746 0.09116221 0.16129807 0.0240987 0.06269502]

[ 0.66173943 0.09694053 0.15984874 0.02475077 0.05672052]

[ 0.64604009 0.09359798 0.16870987 0.02977233 0.06187974]

[ 0.6754386 0.0764411 0.16165414 0.02882206 0.05764411]

[ 0.65644172 0.09509202 0.16206544 0.02453988 0.06186094]]