MS -Business Intelligence & Analytics

Fall 2015

*April 5, 2016*

**Vivek-John Dominic Martins**

Home Work 3

**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. First part.**

***1. Calculate average customer profitability with 95% confidence level***

Average customer profitability with 95% confidence level in 1999

Average: 111.502686982

CI: (108.49606998631779, 114.50930397840372)

Average customer profitability with 95% confidence level in 2000

Average: 144.827019245

CI: (140.1222561095272, 149.53178238115322)

***2.a. Evaluate if online channel has a significant impact on 1999 profitability (9Profit).***

On running an OLS regression a Probability (F-Statistic) of 2.1% was obtained. (Please note that all the hypothesis testing were carried out at significance level of 5%).This tells us that the model is not a good fit and hence the online channel does not have a significant impact on 1999 profitability.

***2.b. Does age help to explain if online channel has a significant impact on 1999 profitability?***

The overall model is a good fit and the coefficients are significantly not zero.

A regression equation of 9Profit= 17.1+27.2 (9Online) + 25.9 (9Age) was obtained.

From the correlation matrix 9Age is negatively correlated with 9Online. This together with the regression equation tells us that profitability increases with online channel usage and also increases with customer age. However as customer age increases more of the profit is explained by age than by the online channel and vice versa. Thus, Age does help explain the online channel’s impact on 1999 profitability.

***3. To test for bias of missing data, evaluate if missing data has an effect on profitability analysis:***

***3a. Evaluate the effect of online channel on 1999 profits when Age0 is included.***

The overall model is a good fit and the coefficients are significantly not zero.

A regression equation of 9Profit= 57+13.8 (9Online) + 17.7 (Age0) was obtained.

We can conclude that Age0 does help explain the online channel’s impact on 1999 profitability.

***3b. Evaluate if adjusting missing data using Age0 or AgeAvg is relevant. In both cases, it is still necessary to include the additional variable AgeExist to control for the missing data***

Age0:

The overall model is a good fit and the coefficients are significantly not zero. A higher R-squared of 0.024was obtained compared to the model from 3a. This indicates a better explanation.

A regression equation of 9Profit= 70.9+19.6 (9Online) + 25.6 (Age0) – 51.85(AgeExist) was obtained.

Thus adjusting the missing data using Age0 is relevant.

AgeAvg:

The overall model is a good fit and the coefficients are significantly not zero. A higher R-squared of 0.024was obtained compared to the model from 3a. This indicates a better explanation.

A regression equation of 9Profit= -32.7+19.6 (9Online) + 25.6 (Age0) + 51.74(AgeExist) was obtained.

Thus adjusting the missing data using AgeAvg is relevant.

However from the correlation matrix we can see that AgeExist is negatively correlated with 9Profit and positively correlated with 9Online. Thus the AgeAvg is a more realistic model. The AgeAvg model also has a higher R-square which tells us that the model explains for more variance in the data.

***3c. Repeat above steps with income. Evaluate if adjusting missing data using Inc0 or IncAvg is relevant. Include AgeExist and AgeAvg in the calculations.***

Inc0:

The overall model is a good fit and the coefficients are significantly not zero, except for the coefficient for AgeExist with a p-value of 0.08. The R-squared increases even more than the model in 3b to 0.044. This might be due to adding more variables in the model.

A regression equation of 9Profit= -38.2+11.9 (9Online) + 26.9 (AgeAvg) +18.8 (Inc0) – 63.6 (IncExist) was obtained.

Thus adjusting the missing data using Inc0 is relevant.

IncAvg:

The overall model is a good fit and the coefficients are significantly not zero, except for the coefficient for AgeExist with a p-value of 0.08. The R-squared increases even more than the model in 3b to 0.044. This might be due to adding more variables in the model.

A regression equation of 9Profit= -38.2+11.9 (9Online) + 26.9 (AgeAvg) +18.8 (IncAvg) – 63.6 (IncExist) was obtained.

Thus adjusting the missing data using IncAvg is relevant.

***4. Evaluate if online channel has a significant impact on 1999 profitability after controlling for demographic variables: age, income, tenure, and geographic district.***

***To evaluate the latter, create dummy variables D1100 and D1200 for districts 1100 and 1200 respectively from the variable 9District.***

The overall model is a good fit and the coefficients are significantly not zero, except for the coefficient for AgeExist and D1100 with a p-value of 0.592 and 0.201 respectively. The R-squared increases even more than the model in 3c to 0.062. This might be due to adding more variables in the model.

A regression equation of

9Profit= -136.2+13.8 (9Online) + 16.7 (AgeAvg) +16.85 (IncAvg) – 34.9 (IncExist) +4.7(9Tenure) +13.2(D1200)

was obtained.

This model shows that the online channel has an impact on 1999 profitability after controlling for age, income, tenure and geographic district. However the model finds customers living in the 1200 district the only significantly impactful geographical demographic.