Student ID: 250830704

Executive Summary for Chief Executive Officer

This report recommends that a machine learning model using a Logistic Regression algorithm be used in predicting customers expected to churn in the next month. As a measure of performance "Gross Benefit" will be used with reference to the amounts possibly generated had the model been applied in the final month of the data where churning of some customers occurred.

In calculating Gross Benefit, the sum of these values are determined:

1. Gross life-time profit after a customer correctly predicted as churn is retained using a discount
2. Money saved in future marketing campaigns to re-acquire the customers

In the first model a 20% discount is assumed to be sufficient to retain correctly predicted churn customers. The Gross Benefit of using this first model is $4,596,114 in which the optimal probability threshold to assign churn is at 0.05. As provided in the data, the average cost to acquire a new customer is $521, this is very large relative to the amount discounted at 20% ($6.82). As such, it makes most sense to be liberal in predicting that a customer will churn. This approach ensures that the model must be highly certain that the customer will not churn before not providing a discount. Exhibit 1 provides more detail on the first model's Gross Benefit calculation.

In the second model, the probability of retaining a churned customer increases as the discount provided increases. Similar to model 1, since the cost to acquire a new customer to offset a customer leaving is extremely high, it is recommended to provide a higher discount rate to increase probability of retention. The optimal discount rate is 30% and the Gross Benefit as a result of this is $2,496,190. Although a higher discount rate could be provided to customers predicted to churn, there is a diminishing return in probability of retention as the discount provided increases. The reason for this diminishing return is that a customer may have already made up their mind before a discount is given or that the reason for switching to a competitor could be for non-monetary purposes (better customer service or technology). Exhibit 2 provides more detail on the second model's Gross Benefit calculation.

As such, it is clear that leveraging a Logistic Regression machine learning model to predict churn is highly beneficial. As per the second model, which is more realistic, this report recommends providing a discount rate of 30% to customers predicted to churn. Further, a threshold to assign churn at 0.05 is recommended.

Student ID: 250830704

Executive Summary for Vice President of Data Analytics

The approach for deriving this model is to first complete a Logistic Regression, determine the significant and relevant variables then using these same variables to run a Naïve Bayes model. The performance of these models will be evaluated using the AUC of the ROC.

Logistic Regression

An initial Logistic Regression was run after cleaning the dataset and removing rows containing NA values. First, all variables which contained the total of other variables such as "total_ic_mou_6" and "total_vol_6" were removed since they were highly correlated with each other and its constituents that make up the totals already provides the same information in the model. A portion of the correlation matrix is provided in Exhibit 3.

Second, all variables which contained ratios were removed. These variables were generally insignificant since they explained the same information as the variables that make it up.

Third, only the most recent "happy" month was selected, it was found that the variables from the most recent "happy" month provided similar information to the second most recent "happy" month. After removing this, it was discovered that all "happy" month variables were generally insignificant compared to "sore" months. Even if a "happy" month was significant relative to the "sore" month, it likely provided the same information as the "sore" month. Further, if a customer is "happy" it is unlikely that their usage would provide any indication of becoming sore, since at that point they have no issues with the company.

Finally, revenue was found to be insignificant despite common knowledge dictating that price should indicate whether someone would want to switch companies. Further investigation on significance by creating dummy variables to indicate whether a particular customer was paying above or below the average price yielded poor results. As such, revenue was removed from the regression. A hypothesis is that since there are only three major tele-com companies in the market, it is likely that the same discounts are being provided at other competitors. The result of the final Logistic Regression is provided in Exhibit 4. The AUC of this Logistic Regression is 0.8721. The ROC curve and AUC calculated is provided in Exhibit 5.

Naïve Bayes

In creating a Naïve Bayes Model, all the variables needed to be converted into factors. This was achieved by determining the median value first of each variable. The median value was calculated since some variables in the dataset contained outliers and median is less sensitive to outliers compared to mean. The variables were then assigned factors based on whether they were above or below the median.

The Naïve Bayes Model yielded an AUC of the ROC of 0.8193, while successful, this was lower than the AUC from the Logistic Regression. Further, Naïve Bayes assumes that each variable is

independent which is not true for all variables. Therefore, Logistic Regression was selected as the most appropriate machine learning algorithm. The ROC and AUC calculation is provided in Exhibit 6.

Selecting a Threshold

In selecting a threshold to assign churn classification, the correct balance of sensitivity and specificity must be determined. As mentioned in the previous section, the cost to acquire a new customer is extremely high relative to the cost of retaining a customer. Therefore, it makes most sense to increase specificity at the cost of sensitivity. Using the ROC graph, the section that is optimal is in the arch of the curve, at a point where the true positive rate is higher. The optimal specificity is estimated to be approximately 0.87, while the optimal sensitivity is approximately 0.64. These optimal conditions are met when the threshold to assign churn classification is 0.05.

Student ID: 250830704

Exhibits

Exhibit 1: Gross Benefit Calculation for Model 1

Gross Profit ($/customer/month)	34.10
Gross Profit After Discount ($/customer/month)	27.28
Average Lifetime of Customer (months)	60
Number of Correctly Predicted Churn (customers)	2130
Gross Profit Over New Lifetime ($)	3,486,384
Money saved from retaining customer instead of acquiring new customer ($/customer)	521
Money saved from retaining customers instead of acquiring new customers ($)	1,109,730
Gross Benefits ($)	**4,596,114**

Exhibit 2: Gross Benefit Calculation for Model 2

Input for two-way data table:

Number of Correctly Predicted Churn (customers)	2130
Discount (%)*	5
Churning Retained (%)*	10
Customers Retained After Discount	213
Gross Profit ($/customer/month)	34.10
Gross Profit After Discount ($/customer/month)	32.40
Average Lifetime of Customer (months)	60
Gross Profit Over New Lifetime ($)	414,008
Money saved from retaining customer instead of acquiring new customer ($/customer)	521
Money saved from retaining customers instead of acquiring new customers ($)	110,973
Gross Benefits ($)	**524,981**

*Variable inputs in the two-way data table. The above table is simply an example calculation for the first row of the table below.

Student ID: 250830704

Output for two-way data table:

Discount Given (%)	Churning Retained (%)	Benefits ($)
5	10	524,981
10	25	1,257,978
15	40	1,925,605
20	50	2,298,057
25	56	2,451,800
30	60	2,496,190
35	63	2,483,723
40	65	2,420,937

Exhibit 3: Correlation Matrix of Variables Containing Totals

	total_ic_mou_6	total_ic_mou_7	total_ic_mou_8
total_ic_mou_6	1.00000000	0.80302924	0.709208695
total_ic_mou_7	0.80302924	1.00000000	0.820089904
total_ic_mou_8	0.70920869	0.82008990	1.000000000
total_vol_6	-0.02862071	-0.03496450	-0.027330443
total_vol_7	-0.04192542	-0.02702755	-0.016626504
total_vol_8	-0.02299030	-0.01055452	0.008520105

	total_vol_6	total_vol_7	total_vol_8
total_ic_mou_6	-0.02862071	-0.04192542	-0.022990298
total_ic_mou_7	-0.03496450	-0.02702755	-0.010554522
total_ic_mou_8	-0.02733044	-0.01662650	0.008520105
total_vol_6	1.00000000	0.68376968	0.642244032
total_vol_7	0.68376968	1.00000000	0.733564344
total_vol_8	0.64224403	0.73356434	1.000000000

Student ID: 250830704

Exhibit 4: Final Logistic Regression

```
    Min       1Q   Median       3Q      Max
-1.4522  -0.4020  -0.1857  -0.0487   6.2285

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.340e-01  4.249e-02  -7.862 3.77e-15 ***
onnet_mou_8   -9.023e-03  2.147e-03  -4.203 2.63e-05 ***
offnet_mou_8  -9.218e-03  2.143e-03  -4.302 1.69e-05 ***
roam_ic_mou_8  3.109e-04  2.443e-04   1.273 0.203070
roam_og_mou_8  9.695e-03  2.148e-03   4.514 6.37e-06 ***
loc_og_mou_8   6.909e-03  2.165e-03   3.192 0.001415 **
std_og_mou_8   8.030e-03  2.151e-03   3.734 0.000189 ***
isd_og_mou_8   6.593e-03  2.070e-03   3.185 0.001446 **
loc_ic_mou_8  -8.034e-03  4.209e-04 -19.087  < 2e-16 ***
std_ic_mou_8  -2.632e-03  5.505e-04  -4.781 1.75e-06 ***
isd_ic_mou_8  -1.324e-03  7.289e-04  -1.817 0.069185 .
vol_4g_8      -4.079e-03  2.657e-04 -15.352  < 2e-16 ***
vol_3g_8      -1.408e-03  8.289e-05 -16.984  < 2e-16 ***
aon           -2.192e-04  3.132e-05  -6.996 2.63e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16926  on 30000  degrees of freedom
Residual deviance: 12503  on 29987  degrees of freedom
AIC: 12531

Number of Fisher Scoring iterations: 8
```

Student ID: 250830704

Exhibit 5: ROC Curve and AUC Calculated for Logistic Regression

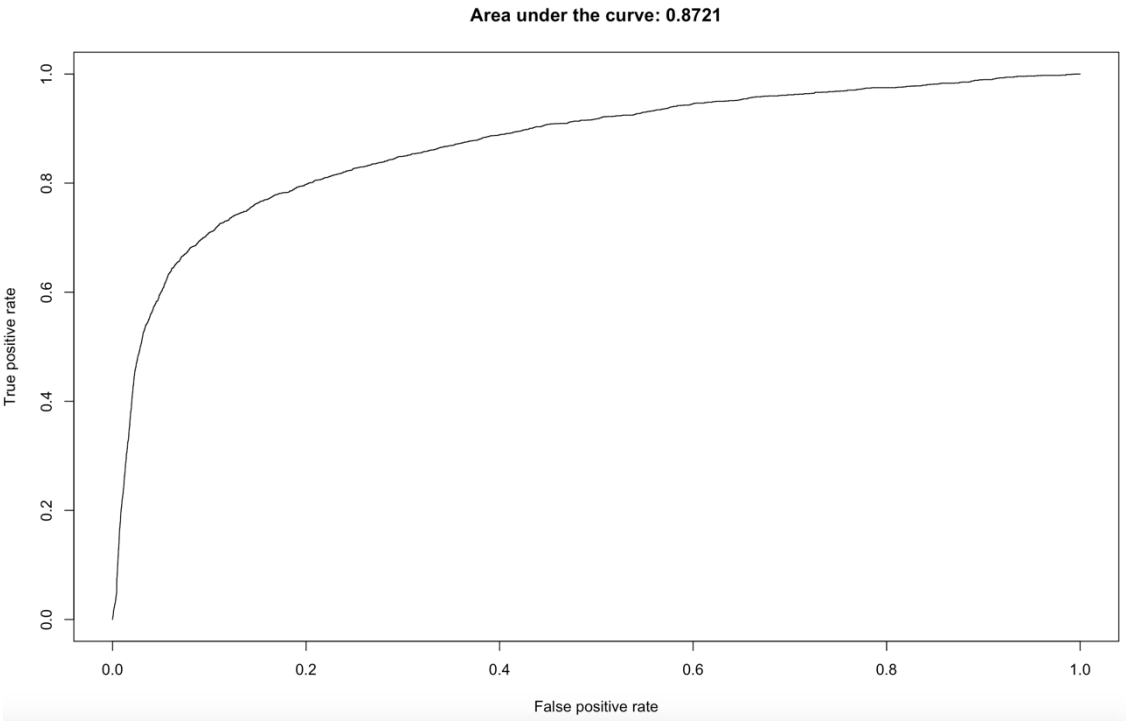

Exhibit 6: ROC and AUC calculation for Naïve Bayes