**NAME: Juan Villegas**

**Project 3A. Logistic Regression**

The file Chdata contains a subset of the churn data that was used in MAT 342. Focus on the variables VMail.Plan and Churn.

The target variable is Churn.

Churn = TRUE means the customer churned (Success)

Churn = FALSE means the customer stayed with the company (Failure)

The explanatory variable is Vmail.Plan

Vmail.Plan = yes means the customer has a voicemail plan

Vmail.Plan = no means the customer does not have a voicemail plan

1. a. Change Vmail into an indicator variable, Vmail.Plan\_Ind:

Vmail.Plan\_Ind = 1 if Vmail.Plan = yes

Vmail.Plan\_Ind = 0 if Vmail.Plan = no

b. Change Churn into an indicator variable, Churn\_Ind:

Churn\_Ind = 1 if Churn = TRUE

Churn\_Ind = 0 if Churn = FALSE

You will use this later. To report your answer to this question, show your code.

Chdata$Vmail.Plan\_Ind <- ifelse(Chdata$VMail.Plan == 'yes', 1, 0)

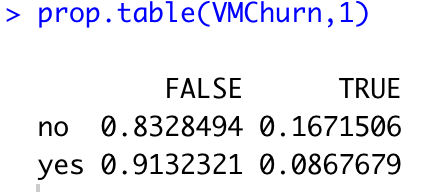
Chdata$Churn\_Ind <- ifelse(Chdata$Churn == 'TRUE', 1, 0)

Question 2 is straight out of MAT 315. (You can use the table command to make a two-way table and then prop.table command to find either row or column proportions.) You may also choose another method to answer this question.)

2. a. Among the customers with a voicemail plan, what proportion churned? (This is *p* or , your estimate for the population proportion of churn customers.) **Use at least 4 decimal places in your answers.**

The columns in my table represent with the customer churned

The Rows represent whether the customer has a voicemail plan or not.



b. Among the customers with no voicemail plan, what proportion churned?

0.1671506 of customers with no voicemail plan churned.

c. Estimate the odds that a customer who has a voicemail plan churns. Then calculate the log(odds). (Remember R’s log is the natural log, which is what we want.)’

The probability a customer has a voicemail plan and churns are 80/(80+842) = 0.0867679🡪

The Odds are .0867679/ 1 - 0.867679 = 0.09501188

The odds for a customer to churn with a voicemail plan is about 1/10.

The log(odds) a customer churns with a voicemail plan are log(0.09501188) = -2.353753

d. Estimate the odds that a customer who does not have a voicemail plan churns. Then calculate the log(odds).

The probability a customer does not have a voicemail plan and churns are 403/(2008+403)

🡪0.1671506

The Odds are 0.1671506/ (1- 0.1671506) 🡪 0.2006973

The odds for a customer to churn without a voicemail plan is about 2/10.

The log(odds) a customer churns without a voicemail plan are log(0.2006973) 🡪 -1.605957

**Logistic Regression Model:** 

The parameters in the model are and , and  is a binomial proportion that depends on the explanatory variable .

Logistic regression with an indicator variable (as we have here) is a very special case. It is important because many multiple logistic regression analyses focus on one or more such variables as the primary explanatory variables of interest. So, we can use this special case to understand a little more about the model.

What do we know about the model in this case, where *x* = the indicator variable for voice mail plan?

 = -2.353753

 B0 = -1.605957

Log odds for churners with plan = -2.353753

Log Odds for people without plan = -1.605957

In the second equation, we know that B0 is -1.605957

That means that the first equation will be -2.353753 +1.605957 = B1

B1 = -0.747796

3. Given the data, fit the logistic regression model *by hand*. In other words, find estimates  and for the population parameters  and .

=. -1.605957 - 0.747796x

Show your calculations to arrive at the solution.

See above

4. Let’s focus on the slope of your logistic regression model. Note: it is the difference between the log odds (for Churn) for those customers who have the voicemail and the log odds for customers who do not have voicemail plan.

-0.747796x

a. As mentioned in the lecture slides, most people are not comfortable thinking in the log odds scale. (Are you?) See class notes and provide the algebraic work needed to show:

 A piece of paper with writing

Description automatically generated with medium confidence

b. Find the value of (a) using your estimated slope. Interpret the results.

exp(-0.747796) = 0.4734088

The odds of churning for a customer who has a voicemail plan are almost half the odds compared to a customer who doesn’t have a voicemail plan.

OR

The odds of churning for a customer who has a voicemail plan is .474 times the odds compared to a customer who doesn’t have a voicemail plan.

5. Use R to fit the logistic regression model. You will need to use the indicator variables from question 1.

a. Write the equation of the model and compare it with the one that you calculated by hand. (Keep in mind, you rounded your answers along the way.) Report the code that you used.

Graphical user interface, text, application

Description automatically generated

The results I got are almost identical, R Just rounded to the nearest ten-thousandths.

b. Run the summary command and insert the output below.

Text

Description automatically generated

6. Next, we move on to a quantitative predictor variable, Day Minutes (Day.Mins).

a. Determine the estimated logit model.

Graphical user interface, text, application

Description automatically generated

b. Interpret the slope (coefficient of Day.Mins) in the context of log(odds) in the context of these data.

If the minutes you spend on your phone daily increase by one, we expect the log odds of churning to increase by 0.01127

c. It is difficult to interpret your answer to (b) because it involves changes to the log(odds). Recast your answer to (b) in terms of odds ratio .

> exp(0.01127) = 1.011334

The odds of churning are increased by a factor of 1.011334 for each minute increased in day minutes.

d. Restate (c) but using increases of 10-minutes in Day.Mins.

> exp(0.01127)^10

[1] 1.119296

The odds of churning are increased by a factor of 1.19296 for each 10 minutes increased in day minutes.

7. Next, let’s look at the output from the summary command.

a. Insert the output from the summary command.

Text

Description automatically generated

b. Note the coefficient for Day.Mins is quite small. Is the predictor variable Day.Mins significant in the model? Explain.

The predictor variable is considered significant because its P-value is a lot smaller than .05.

c. Give a 95% confidence interval for the slope of the model. Is 0 in this interval? Why should you not be surprised by the answer to this last question?

We are 95% confident that if a customer spends an extra minute on their phone daily, the log odds a customer churns is between 0.009375686 and 0.01319896.

We are not surprised that 0 is not in the interval because we rejected the null hypothesis so it can’t be 0.

Text

Description automatically generated with medium confidence

d. Apply an exponential transformation (exp) to the command that you used in (c). Interpret these results.

Text

Description automatically generatedWe are 95% confident that if a customer spends an extra minute on their phone daily the odds of churning are increased by a factor between 1.00941978 to 1.01328645.

8. a. Transform your equation in 6(a) to an equation for probability,  (or Text

Description automatically generated

(Intercept) Day.Mins

0.01927867 0.50281789

b. Overlay a plot of the sigmoidal curve (graph of answer to (a)) on a graph of the data. (See Powerpoint slide 13 for the code.)

Chart

Description automatically generated with medium confidence

9. Return to output from the summary command.

Text

Description automatically generated

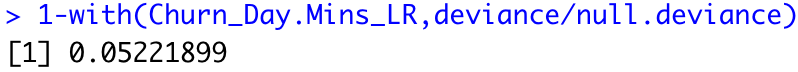
1. What is the value of AIC. (You’ll use this later in this Project.)

2618.3

b. Using the values for Null deviance and Residual deviance, calculate a value that could be thought of like R2. The Deviance / Null.Deviance .

> 1-with(Churn\_Day.Mins\_LR,deviance/null.deviance). =. [1] 0.05221899

Our regression model correctly classifies about 5% of the data in our dataset.



10. Next, you’ll form a model using both variables: Day.Mins and VMail.Plan\_Ind. Then you can compare this model to the models that you created earlier in this project.

a. Fit a logistic regression model. Write its equation.

Text

Description automatically generated with medium confidence (Intercept) Vmail.Plan\_Ind Day.Mins

-3.76672 -0.76771 0.01133

Y = -3.76672 - 0.76771(Voicemail Plan) + 0.01133(Day Mins)

1. Are both variables significant? Explain.Text

   Description automatically generated

Yes, both variables are significant because both of their p values are a lot less than .05.

1. Interpret the coefficients of the two variables in the context of these data. (Keep in mind that your interpretation of the slope for VMail.Plan\_Ind should include “given that the value of Day.Mins is held constant and similarly for slope for Day.Mins.)

Given that a customer does not have a voicemail plan, and they spend an extra minute on their phone daily, the log odds a customer churns is .0009802.

Given that a customer does not spend any extra daily minutes, and they have a voicemail plan, the log odds a customer churn is .2043130.

1. Compare the value of AIC for this model with the ones fit to single variables. Based on the AICs which model would you select as the better model, ones with one variable or the one with 2 variables?

Vmail Model AIC: 2618.3

Day Mins Model AIC: 2581.8

Both AIC: 2581.8

I noticed that DayMins Model and the model that included both variables had the same amount of AIC. Personally, because both the variables are significant, I would choose the model with 2 variables.

1. Compare our logistic version of R2 for all 3 models. What happened to R2 when you used 2 variables in the model instead of models based on only one variable?

Graphical user interface, text

Description automatically generated

The ‘R^2’ Of our model stayed the same as the voicemail churn model. It stayed the same as the one with the higher AIC. I found this strange.

11. Using the model that you fit in 10, predict the likelihood that the following customers will churn.

(Intercept). Vmail.Plan\_Ind Day.Mins

Y = 0.02260491 + 0.31697538(Vmail Plan) 0.50283349(Day.Mins)

1. A customer has no voicemail plan and uses 175 day-minutes.

> new.data1 <- data.frame(Vmail.Plan\_Ind = 0, Day.Mins = 175 )

> (predict1 <- predict(Churn\_VMail\_DayMins\_LR, newdata = new.data1, type = "response"))

0.1439009

If a customer has no voicemail plan and uses 175-day mins, the customer is 14.4% likely to churn

1. A customer has a voicemail plan and uses 175 day-minutes.

> new.data2 <- data.frame(Vmail.Plan\_Ind = 1, Day.Mins = 175 )

> (predict2 <- predict(Churn\_VMail\_DayMins\_LR, newdata = new.data2, type = "response"))

0.07236145

If a customer has a voicemail plan and uses 175-day mins, the customer is 7.24% likely to churn

1. A customer has no voicemail plan and uses 300 day-minutes.

> new.data3 <- data.frame(Vmail.Plan\_Ind = 0, Day.Mins = 300 )

> (predict3 <- predict(Churn\_VMail\_DayMins\_LR, newdata = new.data3, type = "response"))

0.4093863

If a customer has no voicemail plan and uses 300-day mins, the customer is 40.94% likely to churn

1. A customer has a voicemail plan and uses 300 day-minutes.

> new.data4 <- data.frame(Vmail.Plan\_Ind = 1, Day.Mins = 300 )

> (predict4 <- predict(Churn\_VMail\_DayMins\_LR, newdata = new.data4, type = "response"))

0.243385

If a customer has a voicemail plan and uses 300-day mins, the customer is 24.34% likely to churn