**Module 6 Group Assignment**

William Kelly

MBA 563

Group J073

In addition to being concerned about profit (as discussed in class), NANSE is concerned about predicting higher traffic in its stores. That is, NANSE wants to predict which weeks and stores will sell above the median number of units.

Use the code discussed in class to examine the effect of the same variables used in class to predict the target variable `high\_med\_units` (as discussed in the data description sheet listed above, this variable is an indicator variable that equals 1 when the number of units sold for that store for that week was above the median and 0 otherwise).

*Suggestion:* While you could just cut and paste the code used in the live session for this module, we encourage you to write all of the code from scratch. This is a technique we often use when looking up and using new code we are borrowing from someone else.

**Questions**

1.a**(0.5 points)** What feature/variable has the most negative statistically significant coefficient on the trained model summary?

* Intercept = -33.66
* “velocityNEW\_units\_per” = -34.00
* “promo\_units\_per” = -8.795
* The most negative statistically significant coefficient is velcoityNEW\_units\_per at -33.66 and the next negative coefficient is promo\_units\_per at -8.795.

1.b**(1 point)** Does selling a higher proportion of alternative beverages increase, decrease, or neither increase nor decrease the chance of having above median units sold? How do you know this?

* “altbev\_units\_per” = 10.39 w/ p = 2e-16 (statistically significant)
* -33.66 + 10.39 = -23.27
* The coefficient of alternative beverage is 10.36 so selling a higher proportion would increase chance of having above median units sold because it is a positive coefficient. But when adding to intercept it would make it -23.27 which would decrease due to the negative coefficient value.

1.c**(1 point)** Does selling a higher proportion of velocity B units increase, decrease, or neither increase nor decrease the chance of having above median units sold? How do you know this?

* “velocityB\_units\_per” = 5.603 w/ p = 0.1016 (Not significantly significant > 0.05)
* -33.66 + 5.603 = -28.06
* The coefficient for velocity B is 5.603 which is a positive coefficient meaning that selling a higher proportion of velocity B units would increase chance of having above median units sold but the coefficient is not significantly correlated due to the p value being greater than 0.05. Also, when adding to intercept the overall coefficient results in -28.096 which would overall decrease the chance due to negative coefficient.

1.d**(0.5 points)** Examine the accuracy of the predictions on the test data by answering whether there are more true positives or more true negatives?

* There are 1012 TP and 948 TN, so we can see that there are more True Positives as compared to True Negatives which shows an accuracy of 78% making the model fairly accurate.

1.e**(1 point)** If stores are sorted by the `store` feature in an ascending manner (lowest number first), which is the first store in the `full\_test` dataset that has a “WRONG!” prediction?

* The first store with the WRONG! prediction is store 186, size 966, region Ontario

2.**(1 point)** In the model training step, which data—training or testing—do we use and why (that is, explain why we split the data into training and testing subsets)?

* We use the training data in the model training step but also use the test data for checking the accuracy of the model. Thus, the point of splitting the data is to check whether the model that is created works not just on the data that we have, the training data, but also on the new data, the hold out data or the testing data as well. This avoids the situation in which our model fits the data perfectly but does not generalize to the other data. This is important in terms of machine learning because the goal of ML solutions is to speak not only to the data we have but also to predict future outcomes and future relationships b/t variables.

3.**(1 point)** The feature `region` has changed in the summary of the trained model. Further, only three regions show up in the summary of the model. The reasoning for this is that the `glm()` function automatically recognizes that `region` is a categorical variable (specifically a factor in R). This is discussed in our Coursera content. Thus, the `glm()` function has created “dummy variables” for the levels of `region`. Which level of the variable is not present here but rather accounted for in the intercept term?

* The variable that is shown as the intercept term is **Ontario** with an intercept value of -33.66 and p value of 2e-16

4.**(1 point)** Interpret the confusion matrix using the test / holdout data. Specifically, which of the four measures, Sensitivity, Specificity, Precision, or Negative Predictive Value has the highest value? Write a sentence that translates this value into words. That is, say something that starts like this: “this means this model is good at predicting...”.

* The measure of Sensitivity has the highest value at 79.4%
* This means that the model is good at predicting how many positives it got right, or in other words the Hit Rate/True Positive Rate.

5.**(1 point)** Interpret the confusion matrix. Specifically, which of the four measures, Sensitivity, Specificity, Precision, or Negative Predictive Value has the lowest value? Write a sentence that translates this value into words. That is say something that starts like this: “this means this model is not as good at predicting…”.

* The measure of Specificity has the lowest value at 76.3%
* This means that the model did not do as well at predicting how many negatives it got right as compared to the other values within the confusion matrix.

6.**(2 points)** Interpret the confusion matrix. In NANSE’s business setting, which of these measures does NANSE care about the most, sensitivity, specificity, precision, negative predictive value, or something else? Defend your answer in two or three sentences. There is no correct answer here, but you must successfully defend your answer to get credit.

* I think that NANSE cares the most about the Sensitivity due to the fact that the company would want to know how many times the model was able to predict the correct positive results. But also it would be useful to know the Accuracy through (TP+TN/(TP+TN+FP+FN)) which would show how accurate the model is overall.

**Submission\***

1. Answer these questions in an RMD file.
2. Keep all code, visuals and comments in this RMD file. Then knit it into an html.
3. Save your submission as groupname\_HW6.html.
4. Compress the .html file as a .zip or .rar file for submission.
5. Submit the .zip or .rar file on Compass2g before the deadline.