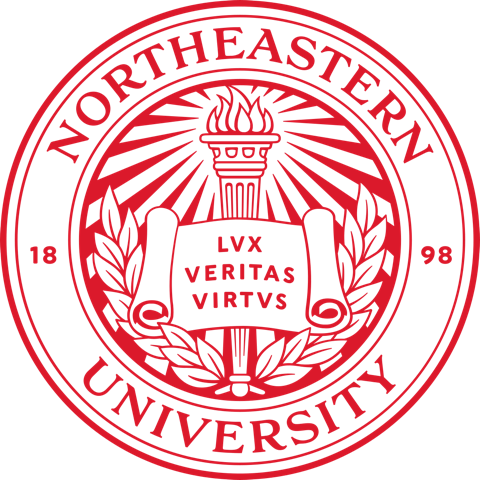
**Module 6**

**Final Exam**

Sunil Raj Thota



College of Professional Studies, Northeastern University ALY 6020 – Predictive Analytics

Prof. Justin Grosz

December 15, 2021

**Q1. Name 3 ways of how you can measure the accuracy of a classification model and the formula of how to do so.**

The 3 ways are Accuracy, Precision, and Recall. These three metrics are percentage metrics based on the absolute number you see in the confusion matrix and are calculated straight from the confusion matrix result to measure the accuracy of a classification model. Before going into the measures, we should know about the following ones:

**True Positive (TP):** Case when you correctly predict the positive result

**False Positive (FP):** Case when you predict the result to be positive but it is negative

**False Negative (FN):** Case then you predict the result to be negative but it is positive

**True Negative (TN):** Case when you correctly predict the negative result

1. **Overall Accuracy:** Accuracy is just the percentage of situations that you accurately anticipated overall. It is an excellent measure of how well a model performs overall.

**Overall Accuracy = (TP+TN)/(TP+TN+FP+FN)**

1. **Precision:** Precision measures the percentage of cases that are truly positive among those projected to be positive. It's utilized when you need to measure positive situations with extreme precision.

**Precision = TP/(TP+FP)**

1. **Recall:** The percentage of true positive instances correctly detected is measured by the recall. When the goal is to collect as many positive cases as possible, recall is used.

**Recall = TP/(TP+FN)**

**Q2. Which 3 algorithms did we cover in class that dummy variables would not be required for?**

In this class, we have covered the below 3 algorithms where dummy variables would not be required are:

1. **Decision Trees:** Because they provide a high-level overview of what is important when solving a predictive analytics problem, decision trees are popular. These rule-based models are especially popular because approaches for creating ensemble models, which are models made up of numerous algorithms, have been created. Entropy is used in some decision tree techniques, such as the C4.5 algorithm, while the Gini Index is used in the classification and regression tree algorithm (CART).
2. **Random Forests:** These models take a similar approach to the voting strategy that we discussed in nearest neighbors. It's a little more difficult to uncover significant variables in random forests, but there are a few features importance approaches that can help you figure out which variables are consistently present in each decision tree.
3. **Gradient Boosting Classifier:** Gradient boosting classifiers is a collection of machine learning techniques that combine numerous weaker models into a powerful, highly predictive output. Models of this type are popular because of their ability to accurately classify datasets. In most cases, decision trees are used to generate models for gradient boosting classifiers.

**Q3. True/False, KNN is only for classification problems?**

False. Because It's a supervised learning algorithm that may be used for both regression and classification. It's most commonly used in machine learning for classification problems. k-NN can also be used to solve regression problems. In this circumstance, the mean or median of the k-most comparable cases can be used to make a prediction. A common method of nearest neighbors is the K Method or KNN. There is no training time in KNN because it is a lazy learning algorithm. The K indicates how many nearby points you consider while deciding how to categorize the point.

**Q4. When interpreting the output of a multiple linear regression model, which of the following metrics should look at and why, (Coefficients, R-Squared, P-Values)**

P-Values, Coefficients, and R-Squared are the three most important statistics in any regression model. This will assist us in determining which business levers we can employ.

1. **P-Values:** P values are significant because they indicate if a variable is significant. If a variable's p-value is less than 0.05, we consider it significant. This is significant since the problem at hand only requires significant variables to be solved. Variables that aren't important can be eliminated.
2. **Coefficients:** Coefficients are significant because they reveal the magnitude of a variable's influence. This could be eye-opening for your firm in terms of figuring out what's driving or dragging performance. We only look at this after we've checked the p-value since even though the coefficient is huge if it's not significant, we won't want to utilize it in our analysis. The size of the coefficient for each independent variable in a simple or multiple linear regression. When the independent variable increases by one, this informs you how much the dependent variable is predicted to increase or decrease. The direction of the influence is determined by the sign of the coefficient (positive or negative).
3. **R-Squared:** The Coefficient of Determination, or R-squared, is the third measure, and it is the most widely used statistic for evaluating regression models. It tells us how accurate or inaccurate the model is. The higher the number, the surer we may be that the model correctly predicted any errors. We look at this statistic since it can be altered by adding or subtracting variables, and your model can still be useful even if R-squared isn't a high value (above 0.6).

**Q5. What is the key advantage/use case for using Support Vectors over Logistic Regression for Binary Classification? (Should be 2-3 sentence explanation)**

SVM works effectively with unstructured and semi-structured data, such as text and images, whereas logistic regression works with independent variables that have previously been defined.

SVM is based on data geometrical properties, whereas logistic regression is based on statistical methods. Overfitting is less of a concern in SVM. SVM seeks to locate the "best" margin (distance between the line and the support vectors) that separates the classes, reducing the possibility of data error. In high-dimensional spaces, SVM is more effective. SVM uses a small amount of memory. SVM is mostly used to optimize models. SVM can be used to solve the following problems: Image classification, Recognizing handwriting, Cancer detection.

**Q6. Clean the dataset and discuss how you cleaned the variables with missing values.**

To extract certain information from the dataset, the '?' sign is substituted by "NaN" during data filtering. Since the symbol has been changed with NaN, which stands for a null value, these can now be easily identified and computed to verify the sum of the missing values in the collection. There are 85 rows and 13 columns in this dataset. I have discovered 3 null values in the 'pricepercent' column and 24 null values in the ‘Multiple Pieces’ column during my analysis of the dataset. The describe function is used to check the dataset's statistics, such as mean, median, mode, and quartile values. Finally, a boxplot was utilized to look for outliers and suspicious data. We utilized the median value for 'pricepercent' and eliminated the ‘Multiple Pieces’ column to clean the dataset so that missing values do not affect our analysis and outcomes. I've opted to remove ‘Multiple Pieces’ from the data frame because it doesn't provide any valuable information.

**Q7. Find the significant variables using multiple logistic regression. If none are significant, feel free to expand it to .1. Of those that are significant, discuss their coefficients and which variable do you think would be most impactful from a business sense (can someone impact this or not if they are making a candy bar from scratch). This should be a minimum of one paragraph with a description of the package's uses, the technique used to find the significant variables, and a discussion on whether we can give advice/action.**

The significant variables found using Multiple Logistic Regression are fruity and winpercent. The p-value of fruity is 0.000 and the coefficient is -4.21. The p-value of winpercent is 0.072 and the coefficient is 0.03. Apart from the above ones, the next best p-value to consider is the p-value of the bar is 0.287 and the coefficient is 2.19. Fruit flavor appears to be negative. Negative coefficients indicate that fruity is less likely at that level of the chocolate than at the reference level. This can impact if a candy bar is made from scratch. The overall impact of winpercentage is slightly positive and the candy built from scratch will have a good impact by taking this into account. Chocolate is usually found in the form of candies in the shape of bars. Fruit-flavored candies, on the other hand, have the opposite effect. To summarize, a chocolate candy that is not fruity, bar-shaped, and has a positive winpercentage is likely to outperform the average. With the baseline configuration (i.e., including all variables as they came) we were able to attain an adjusted R squared of 57.1%.

I have used sklearn.linear\_model.LogisticRegression comes from the scikit-learn package. I have used Statsmodels, which is a Python module that includes several functions for estimating statistical models and running statistical tests. The Logit() method in Statsmodels is used to perform logistic regression. The Logit() function takes two inputs, y, and X, and returns the Logit object. After that, the model is fitted to the data. I have used the Logit function to fit the model with the necessary variables to see the p values and other statistics of each column. Used lbfgs method as the solver in the Logistic Regression method observed the model fit. And then I have calculated the Accuracy, Precision, Recall, and F1-Scores. The ratio of correctly classified subjects to the total number of subjects is known as accuracy. The ratio of accurately +ve labeled to all +ve labeled is known as precision. A recall is the proportion of those who are in reality to those who are appropriately +ve classified. Precision and recall are both taken into account while calculating the F1 Score. It's the precision and recall's average.

Finally, the accuracy obtained is 0.88 with a precision of 1 and recall as 0.75. Also, the ROC AUC score is 0.97. (See Appendix Figure 1)

**Q8. Run decision tree, random forest, and gradient boosting (1000 iterations) and address the accuracy of each of the three models This should be a minimum of two paragraphs with a description of the packages uses, a technique used to find the accuracy measure, what accuracy measure you used (and why) and a discussion on which model you think is the best representative of the model.**

Decision Tree – 76.5% (Appendix Figure 2)

Random Forest – 76.5 % (Appendix Figure 3)

Gradient Boosting – 82.3% (Appendix Figure 4)

I have used sklearn.tree.DecisionTreeClassifier comes from the scikit-learn package. From sklearn.metrics, I have imported the scores to measure accuracy\_score, precision\_score, recall\_score, mean\_squared\_error, confusion\_matrix, classification\_report, roc\_auc\_score, and plot\_confusion\_matrix. Plot\_tree to plot the Decision tree. I have used accuracy, recall, and precision scores as the accuracy measures.

The simplest intuitive performance metric is accuracy, which is just the ratio of properly predicted observations to all observations. The ratio of accurately predicted positive observations to total expected positive observations is known as precision. Yes, recall is defined as the proportion of accurately predicted positive observations to all observations in the class. I think Gradient Boosting Classifier is the best representative model because of its accuracy and AUC ROC Score which is 1. Gradient boosting is relatively resistant to overfitting; hence a big number usually yields better results. learning rate: learning rate reduces each tree's contribution by learning rate. Between learning rate and n estimators, there is a trade-off. The best model for this type of dataset.

The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. If-then-else sentences are commonly used as decision rules. The rules become more complex as the tree grows deeper, and the model becomes more accurate. Random Forest is an ensemble learning-based supervised machine learning technique. I built a Random Forest Classifier model in this project to predict the candy is chocolate or not. Gradient boosting is a type of ensemble machine learning method that can be used to solve classification and regression predictive modeling challenges. Decision tree models are used to create ensembles. To repair the prediction mistakes caused by past models, trees are introduced to the ensemble one at a time and fitted. The boosting model is a sort of ensemble machine learning model. The algorithm outputs hyperparameters that should, and possibly must be modified for each dataset. I have used 1000 iterations in this model.

**Q9. Build a matrix with all the models that include at least two measures of accuracy and speed for the models in 7 & 8. Compare the models and give us a reason for which model you think describes accuracy the best.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Speed** | **Accuracy** | **Precision** | **Recall** | **MSE** | **ROC AUC Score** |
| Logistic Regression | 0:00:00.012 | 0.882 | 1.00 | 0.75 | 0.12 | 0.97 |
| Decision Tree Model | 0:00:00.009 | 0.765 | 0.84 | 0.62 | 0.23 | 0.99 |
| Random Forest Model | 0:00:01.579 | 0.765 | 1.0 | 0.50 | 0.23 | 1.00 |
| Gradient Boosting Model | 0:00:00.575 | 0.823 | 1.0 | 0.62 | 0.17 | 1.00 |

I have used three benchmarks as accuracy, ROC AUC Score, and precision because accuracy will provide the exact efficiency of the model and Precision is one measure of a machine learning model's performance – the accuracy of a model's positive prediction. The number of true positives divided by the total number of positive predictions is known as precision (i.e., the number of true positives plus the number of false positives). We have to predict if the candy is chocolate or not. We have trained the dataset according to the chocolate variable. As per Table, we can see that the accuracy benchmark is highest in Logistic and Gradient boosting models with 88.2% and 82.3% respectively which is very high-level accuracy and very highly efficient. Gradient Boosting and Logistic also had good precision. When we move to the AUC ROC Score, we could see that Random Forest and Gradient Boosting have the highest scores which mean that model was the best to predict all the instances whether the candy was chocolate or not. R2 score provides the goodness of fit of a regression model. R2 score around the range of 0.5 to 0.75 is a good measure for fit. As per the conclusion based on the benchmark and metrics, I would suggest using the Gradient Boosting Classifier model to predict if the candy is chocolate or not.

**Q10. In two paragraphs, based on looking at the variable and model results, give feedback to a candymaker on what you’ve learned from the model. Do you feel one model does a better job than all the others? Does one model give us a better understanding of what to put in the candy bar? Do we know an ingredient or two that makes the best chocolate? In your two paragraphs, use these questions to frame a speech that you would give if you were presenting to the board of this chocolate company about your findings.**

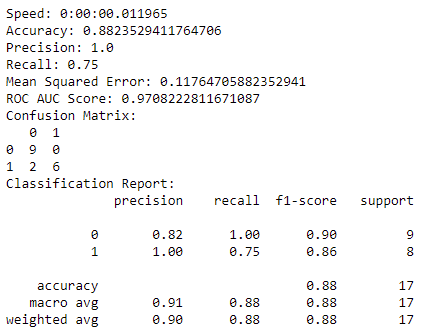
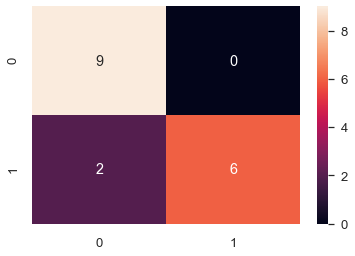
As a result, winpercent is the greatest and first option when choosing a candy. As for the second option, we've chosen fruity. Because candies in bars had a higher overall win rate than those in bits, we'll desire chocolate in bar form. We looked at a database to discover what characteristics characterize the most popular candies. The most common traits discovered are Winpercent, Fruity, and Bar. However, given we only have two options, making definite conclusions is difficult, therefore it's probably best to concentrate on the two traits that contribute the most to effectiveness. Winpercent, sugarpercent, and pricepercent were found to be unrelated. Based on the P-value and coefficient value of the multiple logistic regression utilizing logit, we determined that fruity, winpercent, and bar were a few of the most important ingredients. The coefficient values for winpercent and bar were the highest.

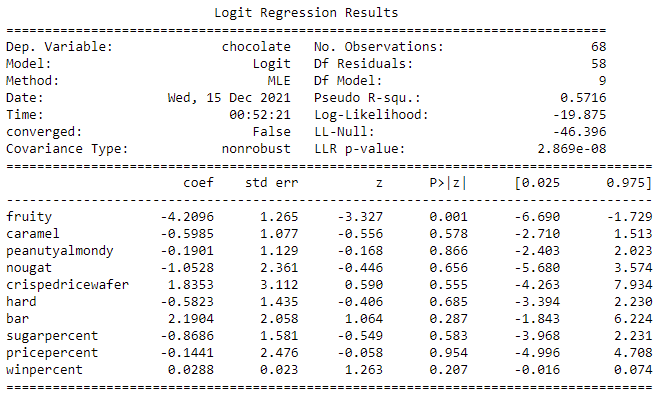
To develop the best candy, we looked for the three most crucial variables/ingredients that would result in the best candy combination and be appreciated by the most people. Fruity, bar and peanutyalmondy are the flavors. According to the models above, a candy with chocolate, bar, and peanutyalmondy is projected to do significantly better than normal. When we factor in chocolate, peanutyalmondy, and high sugar levels, the winning rate jumps to 72.12%. Because the holiday season is approaching, and candies sell the most during this time, now is the ideal moment to enter the candy industry and win significant market share using the ingredient analysis presented above. It would be good if run different models and compare the significant variables that are important in making the decision. It is suggested to consider the other model's results and form an inference in creating a new candy from scratch. Decision Trees and Gradient boosting Classifiers have similar features which are important and mainly add value in making a new candy from scratch.

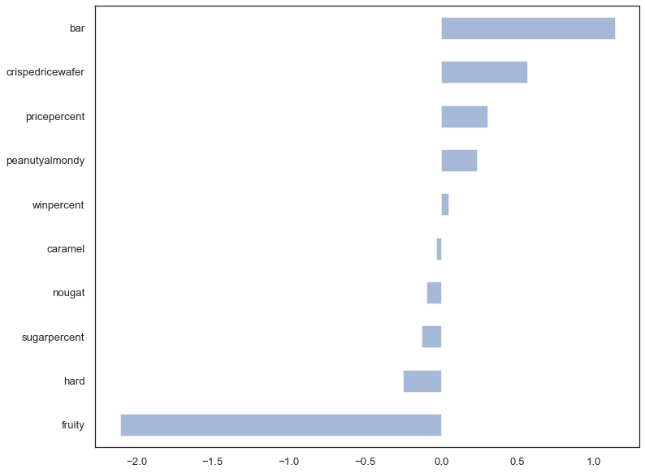
This dataset has given us a lot of critical thinking skills as well as the capacity to solve problems quickly. Also, apart from these, we'd like to dig deeper into the shopping patterns of our customers. Do they have any health concerns? Do people purchase what they desire or go for less-priced options? Combine the findings of this study with the outcomes of the preceding steps to make candy that people would like and buy. Bring in new candies to the market. In a real-world context, users may pay more attention to pricing and sugar content. Various groups of people shop in various ways. Some health-conscious customers will pay more for sugar-free candy, while others will choose the cheapest candy, regardless of sugar content or composition.

**Appendix:**

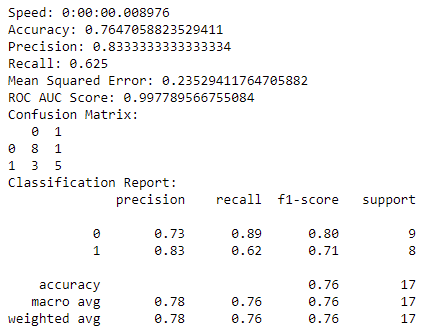
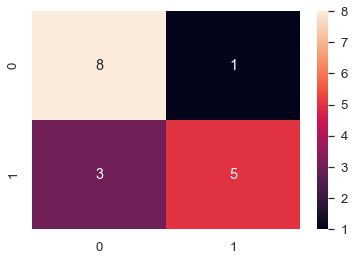
**Figure 1: Logistic Regression Results**

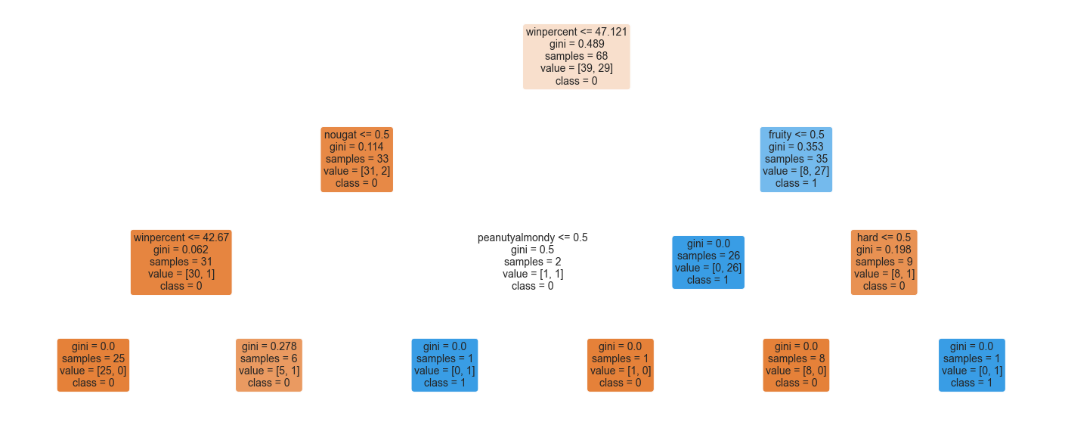
 

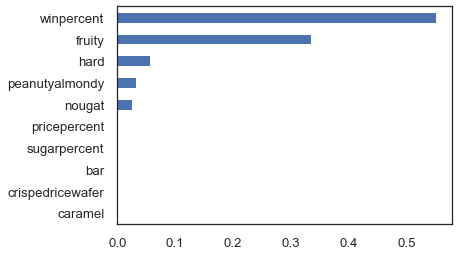




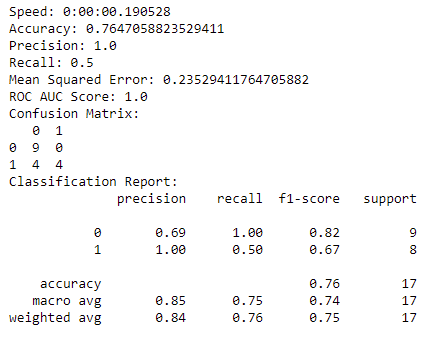
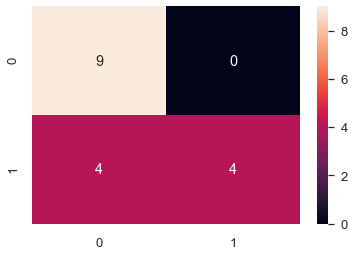
**Figure 2: Decision Tree Results**

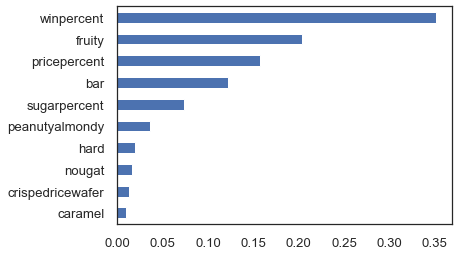
 



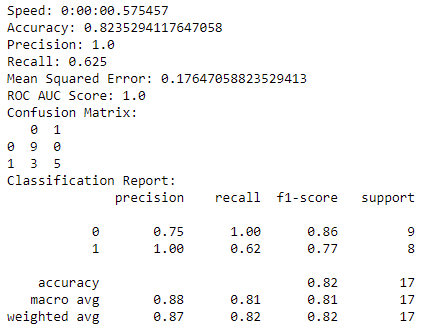
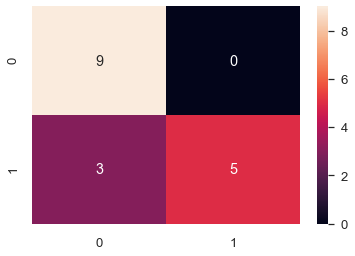


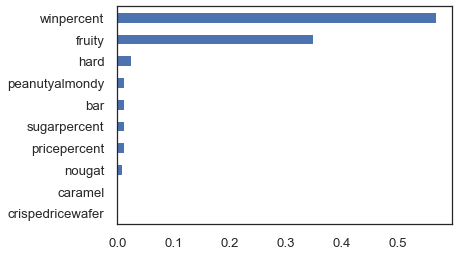
**Figure 3: Random Forest Results**

** **

****

**Figure 4: Gradient Boosting Results**

** **

****