Module 6 Final Assignment

Dhairyav Jatin Shah

College of Professional Studies, Northeastern University

Professor: Justin Grosz

February 17, 2023

1. The three ways to measure the accuracy of a classification model are :
   * 1. **Accuracy Score** : The accuracy score gives the ratio of the perfectly predicted results by all the predicted results. It is obtained by dividing the total number of predicted results in the dataset by the number of results that were correctly predicted. Accuracy score works most accurately when there are equal number of records in each sample. The formula to calculate accuracy is given by:
     2. **Classification Report** : A classification report involves the recall, precision, and F1-score.
        + 1. Precision is given by the number of correctly predicted positive results divided by the sum of total positive results predicted by the classifier. Precision is the ability of the model to not predict an instance positive which is actually negative.
          2. Recall is the number of positive results divided by all the relevant predicted values ( True Positives and False Negatives). Recall is the fraction of positives which were correctly predicted.
          3. F1 Score is given by the weighted harmonic mean of precision and recall. It calculates how many times the model made a correct prediction across the entire data set. The formula for each metric is given below:
     3. **Confusion Matrix** : Confusion matrix is a performance matrix used for machine learning classification models. The output of a Confusion Matrix can be a 2x2 matrix or more depending on the number of categories predicted. A 2x2 Confusion Matrix has 4 components True Positive( Actually True and Predicted Positive), True Negative (Actually False and Predicted Negative), False Positive (Actually False and Predicted Positive) and False negative (Actually True and Predicted Negative). A confusion matrix is also used to derive the above discussed performance matrix like accuracy, precision, recall, F1 score, etc,. A visual representation of a 2x2 Confusion Matrix is given below:

|  |  |  |
| --- | --- | --- |
|  | **Yes (Actual)** | **No (Actual)** |
| **Yes (Predicted)** | True Positive | False Positive |
| **No (Predicted)** | False Negative | True Negative |

1. There were multiple algorithms covered in this course out of which here are 3 algorithms that can be used for classification and regression problems.
   1. K- Nearest Neighbors: KNN is one of the most basic and commonly used classification and regression algorithm in machine learning. It is a non-parametric and lazy learning algorithm. A non-parametric algorithm is something which doesn’t require and assumption from the data. Since it doesn’t require any training data points from the data set it is called a lazy algorithm. The entire data set is used during testing which makes this algorithm slower and more memory consuming when it comes to larget data sets. In KNN, K is the number of nearest neighbors that need to be considered before clustering the new data point. Having a smaller value of K is more advisable, as the K increases greater number of training data points would be included in the data set which could make the classification uncertain due to the increase in input data points. KNN uses Euclidean Distance to measure the distance between the new data point and the training data points. The classification is done based on the lowest distance between the points.
   2. Random Forest Classifier : Random Forest classifier is an extension of the decision tree algorithm in supervised machine learning. It consists of large number of Decision Trees which are generated using multiple machine learning algorithms to obtain accurate results knows as ensemble methods. Each individual Decision Tree obtains a particular output class, the random forest then combines all the outputs generated by all the Decision Trees and obtains the result with the most voted class among the Decision Trees.
   3. Linear Regression : Linear Regression algorithm is most used machine learning technique for predicting data based on historical data provided. Linear regression finds correlation between the dependent variable and the independent variables also known as the predictors. This correlation is then used to find the significant features to predict the target variable closest to the regression line. In linear regression, the metric used to analyze the model’s performance is called the R-squared. It is also called as the coefficient of determination. R-squared values represents the closeness of the values predicted from the regression line. Greater R-squared value represents the predicted values are closer to the linear regression equation line which is formed. The standard formula for a linear regression is given by:
2. In a 2x2 confusion matrix the major areas we are trying to minimize are the Type I and Type II errors.
   1. Type I error are the false positives. A false positive is when the predicted value is Positive, but the actual is not Positive. For example, when someone takes a Covid test and tests positive but actually isn’t positive such a case is called a False Positive and a Type I error. We would want to minimize such an error as it may lead to a person taking a medication/treatment for something they don’t require and such steps can cause adverse effects of the health as well.
   2. Type II error are false negatives. A false negative is when the predicted value is Negative, but the actual is Positive. For example, when a person takes a Covid test and tests negative but, they are positive. In such a scenario one needs to take proper medication/treatment which the fail to take because of the false test results. Which might lead to adverse health effects in the future.

Both these cases of a confusion matrix can be dangerous while predicting any value. The choice of out of which of the two areas are more important and need to be reduced depends on the problem statement and the risks associated with the business problem.

1. Out the following metrics, the one metric we may not focus on is the R-squared, as it only determines the proportion of variance in the dependent variable that can be explained with the help of the independent variable. R-squared is calculated by the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared. R-squared displays how well the data fits near the regression line. Higher the value of R-squared the better the data is fitted. But there is one con of R-squared as well. We can get a false value of R-squared if we don’t check for multicollinearity and it exists. Multicollinearity may cause the model to give a higher R-squared.
2. In situations where the data is not linearly separable, there are outliers present or the data set has a high number of features/variables using Support Vector over Logistic Regression is preferable. Support Vector Machines (SVM) are less sensitive to outliers than Logistic Regression because SVM focuses on the margin that separates the two classes, whereas outliers in Logistic Regression can influence the decision boundary learned by the model. SVM also work better with data sets with higher features as there as changes of overfitting in the case of Logistic Regression.
3. For data cleaning, we start by dropping the columns **“winpercent”, “competitorname”** as they will not be contributing towards the prediction. In the next step we solve the null values present in the data frame. We have 24 null values in the column **“Multiple Pieces”** which we replace using the median of the column as it has just 2 values 0 and 1. Another column we have 3 null values for is the **“pricepercent”** for which we have used the mean parameter to impute the null values. After imputing the null values, we convert the **“Multiple Pieces”** into an integer and round 2 columns **“pricepercent” and “sugarpercent” up to 2 decimal places.**
4. To build a Logistic Regression Model we first plot a heat map to understand the correlation between the independent and dependent variables. From the heat map we can see that there is positive correlation between **“chocolate”** and other independent variables except for columns **“fruity”** and **“multiple pieces”** which have negative correlation with the dependent variable. We now check if the independent variables have multicollinearity. We use the Variance Inflation Factor (VIF) method to measure multicollinearity. We obtain that the following columns **“fruity”, “bar”, “sugarpercent” and “pricepercent”** have VIF value greater than 1 hence they may show characteristics of multicollinearity. Upon removing the columns, we fit the Logistic Regression model, where we split the data set into 80:20 train and test data set. The package we have used to build the model is from Statsmodels called Logit. From the results obtained from the model we can see that there are only 3 columns significantly contributing towards the prediction. In this case we have selected confidence interval as 90% which makes our p value as 0.1. Columns **“peanutyalmondy”** with p value **0.91** and a positive coefficient **0.5741**, **“crispedricewafer”** with p value **0.046** and a positive coefficient of **2.1094** and **“hard”** with p value **0.93** and a negative coefficient of -**0.7504** are the most significant columns in prediction. The positive coefficients show that if the candy has peanut/almonds or if it has crisp rice wafer then the chances of the candy to be a chocolate increases. The feature **“crispedricewafer”** having the lowest p value and highest coefficient depicts that it is the most significant feature to classify a candy into chocolate. One can use this feature to differentiate between chocolate and other candies. The results for the Logistic Regression Model are given below. From the classification report we obtain an accuracy of 88%. One recommendation that we can provide on the basis of this model is that when the manufacturer wants to classify chocolate among other candies they can use these variables out of which **“crispedricewafer”** the most significant variable would be the most helpful in classification.

Table

Description automatically generated with medium confidence

Table

Description automatically generated

1. Upon running Decision Tree, Random Forest and Gradient Boosting, the accuracy we obtained for each model is given as Decision Tree with 88%, Random Forest with 100% and Gradient Boosting with 94% accuracy. The accuracy measure that we have used to evaluate the model’s performance is accuracy\_score which is displayed in the classification report. The accuracy\_score metric represents the ratio of correctly predicted records to the total number of predicted records. We have used accuracy\_score to measure the performance of the model because our data set was almost balanced and in such cases this metric becomes the best choice as it is very easy to understand and interpret.

When we check the features used by each of these 3 models to predict if the candy is chocolate or not, we can see that **“fruity”** and **“pricepercent”** are the top 2 variables which are the most significant for all three columns. When we see the Decision Tree model, having 88% accuracy has used 5 variables to predict if the candy is chocolate or not. Where “fruity” is the most significant feature to classify. Whereas when we see the significant features for Gradient Boosting, having an accuracy of 94% has used all the 10 features to predict. At last, when we check the features used to predict in the Random Forest Classifier, having an accuracy of 100% we can see that it has used the feature “pricepercent” as the most significant whereas this column was the 2nd best feature for the other 2 models. We can say that when the model created multiple decision trees to evaluate the random forest model itself learned that “pricepercent” is an important variable, the classification of the candy can be done based on how much a chocolate would cost based on the most expensive candy. Another important feature that can be considered is the feature where it tells us whether the candy is fruity or not. This feature is also significant in all three models and can be very useful to predict the target variable.

1. The confusion matrix for all the models we have build is given below.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **LR** | | | **DT** | | | **RM** | | | **GB** | | |
|  | **Yes** | **No** |  | **Yes** | **No** |  | **Yes** | **No** |  | **Yes** | **No** |
| **Yes** | 9 | 0 | Yes | 9 | 0 | Yes | 9 | 0 | Yes | 9 | 0 |
| **No** | 2 | 6 | No | 2 | 6 | No | 0 | 8 | No | 1 | 7 |

The logistic regression model has an accuracy of 88%. The training time of this model was 0.0 and inference time is also 0.0. This states that the model was trained and tested very quickly but its performance accuracy is only 88%. Also from the confusion matrix we can see that the model predicted 2 records as False Negative and 0 False Positive, 9 True Positives and 6 True Negatives. We see a very similar situation in the decision tree model also, the accuracy of the model is again 88% and the confusion matrix that we obtained is also the same as the one obtained in Logistic Regression. When we observe the gradient boosting model, we can see that the accuracy increased to 94% and the model predicted only 1 record as False Negative and 0 False Positive, 9 True Positives and 7 True Negatives. The time recorded to train this model is .28 seconds. Finally, when we observe the Random forest classifier we can see that the model has obtained 100% accuracy. The training time recorded for training the model is 0.66 seconds and inference time is 0.4 which is the most taken by any of the 4 models. We can see that in the confusion matrix for random forest we don’t have values for False Positives and False Negatives. My recommendation for the manufacturer would be to use random forest classifier to classify if the candy is chocolate or not. One reason random forest classifier is able to predict with best accuracy is that it runs multiple decision trees and then selected the most voted class as an output.

1. As per our findings, to manufacture a candy bar the manufacturer can use multiple models to refer the features where in Logistic Regression we can see that Crisped Rice Wafer was the most significant among all other features. Hence while making chocolates the manufacturer can use crisped rice wafers. We can say that the model built by using the random forest classifier is the best fit to classify a candy into a chocolate or not. The most significant features we have obtained from the random forest model are “pricepercent”, “fruity”, “bar”, “sugarpercent”, “peanutyalmondy”.

While a manufacturer is understanding the process to make a candy the variables such as “hard” which indicates if the candy is hard or not, “bar” which indicates if the candy is a bar or not, “sugarpercent” which indicates the amount of sugar precent compared to the candy with the most, “fruity” which indicates if the candy is fruity or not. We can understand that if the candy is fruity then there are less chances of the candy to be a chocolate. Similarly, we can see that the hardness of the candy and if the candy is a bar or not can also be understood by the manufacture. Hence, these features shall help the manufacturer to produce from scratch. To find out what ingredient or two would make the best chocolate we can look at the Logistic Regression model along with the random forest classifier. We can see that Crisped Rice wafer can be one ingredient, along with peanuts and almonds can also be a good ingredient which we can say can be used to manufacture best chocolates.

**Appendix**

**Fig No 1 : Feature Importance of Decision Tree**

Chart

Description automatically generated

**Fig No 2 : Feature Importance of Random Forest**

**Chart, bar chart

Description automatically generated**

**Fig No 3 : Feature Importance of Gradient Boosting**

**Chart, funnel chart

Description automatically generated**