Watermelon Quality Prediction

COGS 109 Project

Introduction

Imagine that you were taking a trip to the countryside in a hot summer day, you stopped by a watermelon farm that tries to purchase the best watermelon in the world! What would you do to determine whether that is a good watermelon?

Many people make guesses based on the appearance of watermelon, such as how it looks, the patterns of the watermelon, the sound it makes when you tap it, or just simply following the instincts. If you're lucky, you might be getting a fresh and juicy watermelon. If you're out of luck, you might've bought a bland and tasteless watermelon — a waste of money. Therefore, to prevent such unfortunate events, we hypothesize a scheme that helps you characterize a good watermelon and saves your pocket! We will be introducing a machine-learning method that assists you to examine the watermelon qualities, to bring out the scientific relationships between these predictors, and to make your watermelon adventure more exciting in the future!

Data sets

We obtained the data set from Kaggle (Watermelon Quality Prediction). It contains 209 observations and 10 variables. In detail, this comprises of 7 categorical variables ("Color", "Root", "Sound", "Texture", "Belly_button", "Touch", "GB") and 3 numerical variables ("Density", "sugar_rate", "Num"). We will be using all variables except "Num" because it is not a factor affecting the watermelon quality. Additionally, the variable "GB" is an outcome label that determines the quality of the watermelons, "Yes" stands for good quality watermelons, and vice-versa. Out of the 209 observations, we observed 79 watermelons with good quality and 130 watermelons with bad quality. Intuitively, we can interpret that this is an imbalanced data set and might be biased to the bad watermelons.

Here's the denotation of the variables in the data set:

- "Num" : the count of the watermelon (numerical)
- "Color" : the external color (categorical: "Green", "Dark", "Light")
- "Root" : the degree of bending in the root (categorical: "rolled up", "curly", "straight")
- "Sound" : the sound heard when tapping it (categorical: "turbid", "low", "clear")
- "Texture" : the visibility of green stripes (categorical: "clear", "blurry", "very blurry")
- "Belly_button" : the depth of the central node (categorical: "sunken", "a little sunken", "flat")
- "Touch" : how the exterior feels (categorical: "slippery", "sticky")
- "Density" : the weight or water content (numerical)
- "sugar_rate" : the sweetness (numerical)
- "GB" : Good or Bad watermelon (categorical: "Good", "Bad")

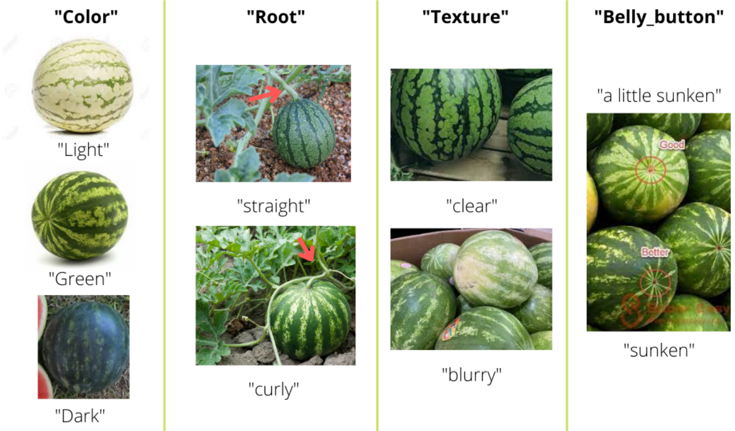


Fig. 1. Visualizations for some categorical watermelon predictors.

Hypothesis

Will belly button, sound and sugar rate be the best predictors to predict the quality of a watermelon?

Methods

Data Analysis Proposal

In this paper, we will be focusing on predicting watermelon qualities based on color, root, sound, texture, belly button, touch, density, and sugar rate. After which, we want to test our hypothesis to see which ones are the best predictors. We will be using logistic regression and K-Nearest Neighbors (KNN) as statistical modeling methods to conduct the prediction.

Label encoding: Before implementing both methods, we will first encode the categorical variables in our data set into representative numerical values using 'sklearn' packages and replace the original values with the dataset.

Cross-validation: Following encoding, We will be split our training and testing data into a $80-20$ split for validation process. The same training and testing sets will be used for both logistic regression and KNN to ensure consistency.

Logistic Regression: Considering that the outcome labels that we wanted to predict are composed of two categorical values ('Yes' and 'No'), applying logistic regression would be a good choice because it applies to binary classification. Additionally, We can statistically examine the association between the independent variables (predictors) in the data set and the dependent variable ('GB'). However, it would be biased if we conduct the modeling on the original data set because we observed an imbalanced data set. Therefore, we will include a random shuffling parameter(RandomState) in our modeling to increase the normality of the data set.

KNN: Moreover, choosing K-Nearest Neighbors (KNN) is also a good selection. It uses a non-parametric machine-learning algorithm that does not require a normally distributed data set. KNN predicts the label of the data by calculating the distance from x to all points in our data based on the k closest points. One advantage of using this method is that it makes the modeling more flexible and may result in a higher accuracy.

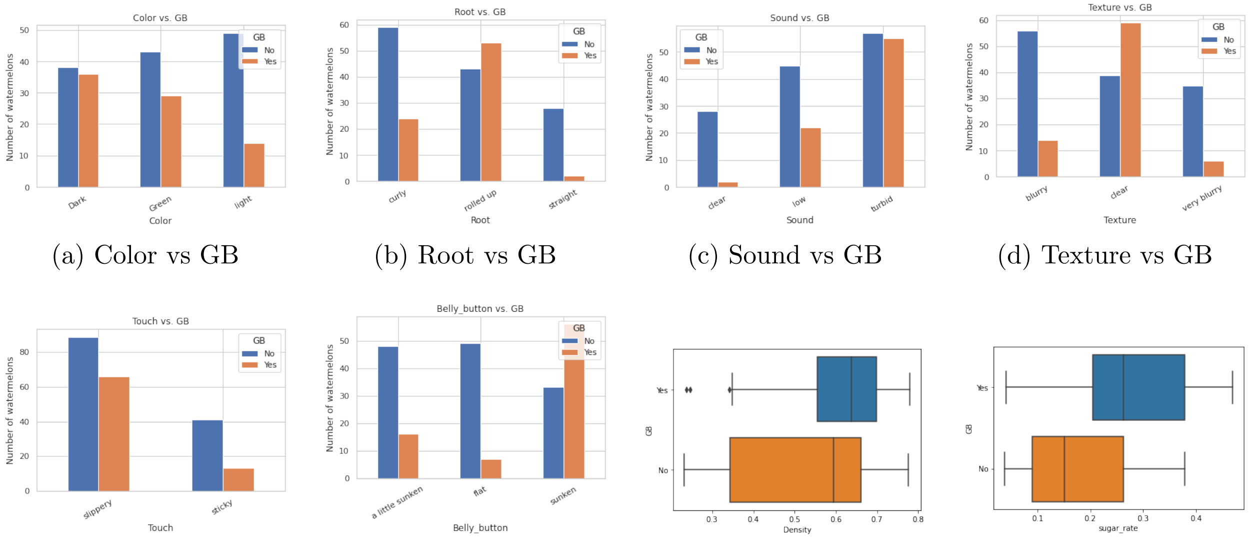
Selecting the best K for KNN: We will be tuning in a range of K from $1 - 40$ for the best K-estimates using the elbow method, we will get the best k value by finding the point with the minimum error rate in the plot. The plot will be included in the data visualization. By finding the best optimal value of K, we are able to maximize the prediction power of the KNN classifier.

Testing the models: After finishing the two modelings, we will compare the results using the test set to get the model estimation. We will use the confusion matrix and the model accuracy for assessing our modeling.

Exploratory Data Analysis

We will be using bar charts for the relationship between the categorical variable and the outcome, and box plot for the numerical variables. For example, *Fig. 2f* shows sunken belly buttons are probably related to the outcome variable since the "yes" (orange) bar is higher than the other categories which the "no" (blue) bar is dominant.

From the *Fig. 2g*, it is unclear to tell there is a relationship between density and GB since the boxes are overlapping most of the part. However, *Fig 2h* shows that there might be a relationship between sugar rate and GB since the boxes are overlapping little bit.



(a) Color vs GB (b) Root vs GB (c) Sound vs GB (d) Texture vs GB

(e) Touch vs GB (f) Belly_button vs GB (g) Density vs GB (h) sugar_rate vs GB

Fig. 2. Visualization for categorical predictors

Results

Model Selection

Logistic Regression: With the 80% training data set, we used the sklearn `LogisticRegression()` function to model the data. Since this model allows us to test the significance for individual regression coefficients, we created a coefficient table to validate which parameters are the most important in predicting watermelon quality. We additionally computed the log-odds ratio to show feature importance. After testing with the test set, we calculated the confusion matrix to analyze the accuracy, precision, and error rate of this model. We found these to be 66.66%, 54.54%, 33.33% respectively *(see Fig. 4)*.

KNN: Similar to logistic regression, using 80% of the data set for training, we plotted the RMSE values on this set so we can find the optimal K value. We looped through 40 possible K values using the we used the sklearn `KNeighborsClassifier`.

We chose $K = 3$ because it gave us the least error rate, or a local minimum, as shown in *Fig. 3.* Following the selection of the optimal K value, we calculated the confusion matrix to analyze the accuracy, precision, and error rate of this model. We found these to be 88.09%, 91.66%, 11.90% respectively *(see Fig. 4)*.

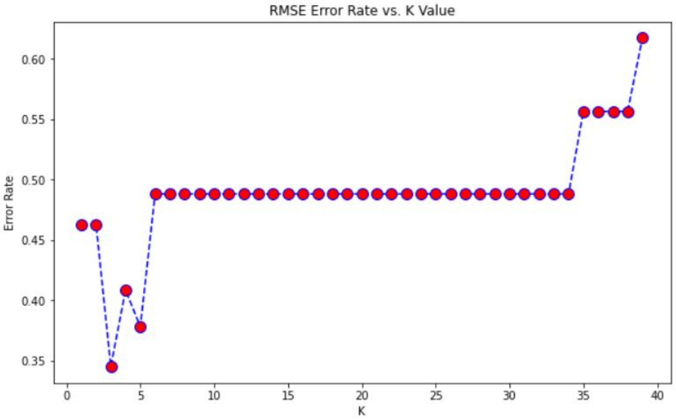
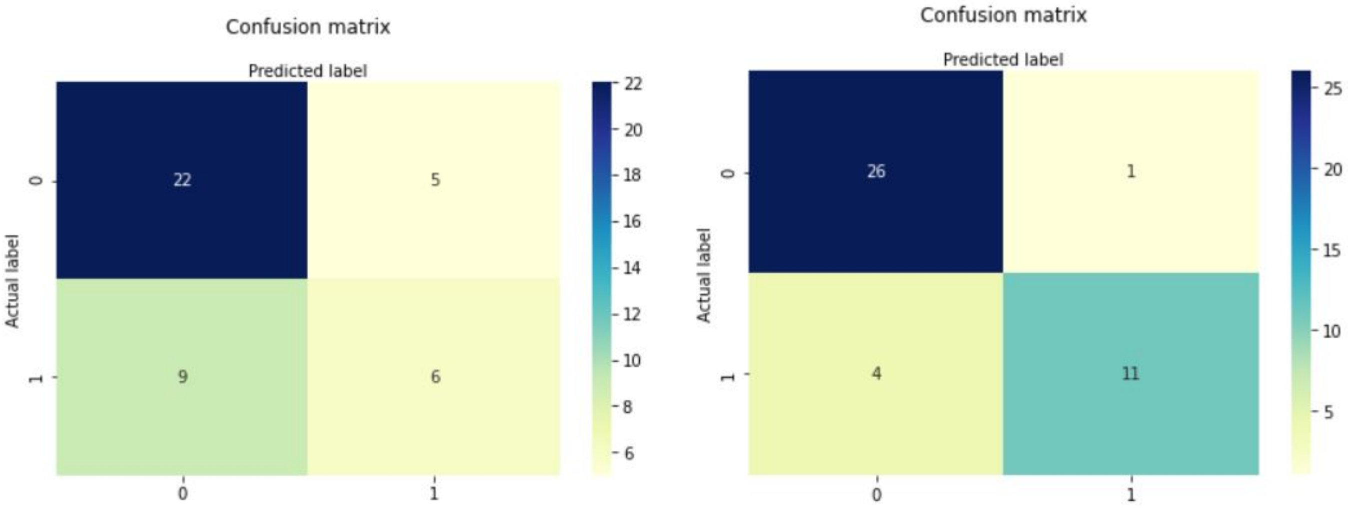


Fig. 3. RMSE Error Rate vs. K values

Model comparison: KNN gave us a lower error rate and higher precision and accuracy scores as opposed to logistic regression. However, it does not tell us about the predictor significance, which is important in evaluating our hypothesis. Even though logistic regression is not the better model at predicting good/bad watermelon, it gives us the coefficient table to determine predictor significance. Therefore, we chose logistic regression as the final model.



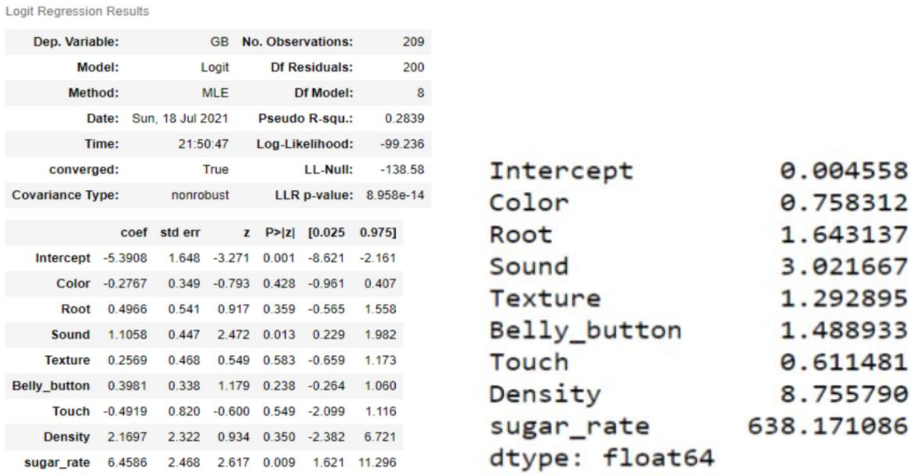
(a) Logistic Regression

(b) KNN

Fig. 4. Confusion matrix for Logistic Regression and KNN

Model Estimation

From the coefficient table in *Fig.5a*, we found that only sound and sugar rate have significant influence (p-values < 0.05) on predicting the qualities of the watermelons. The coefficient tells us that each one-unit change in sound will increase the log odds of having a good watermelon by 3.021667, and its p-value indicates that it is significant in determining good watermelons. Similarly, with each unit increase in sugar rate increases the log odds of having good watermelons by 638.171086 *(see Fig. 5b)*. We obtained a 66.67% accuracy after conducting the logistic regression modeling with the 80% training data and the 20% test data.



Logit Regression Results

Dep. Variable:	GB	No. Observations:	209
Model:	Logit	Df Residuals:	200
Method:	MLE	Df Model:	8
Date:	Sun, 18 Jul 2021	Pseudo R-squ.:	0.2839
Time:	21:50:47	Log-Likelihood:	-99.236
converged:	True	LL-Null:	-138.58
Covariance Type:	nonrobust	LLR p-value:	8.958e-14

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.3908	1.648	-3.271	0.001	-8.621	-2.161
Color	-0.2767	0.349	-0.793	0.428	-0.961	0.407
Root	0.4966	0.541	0.917	0.359	-0.565	1.558
Sound	1.1058	0.447	2.472	0.013	0.229	1.982
Texture	0.2569	0.468	0.549	0.583	-0.659	1.173
Belly_button	0.3981	0.338	1.179	0.238	-0.264	1.060
Touch	-0.4919	0.820	-0.600	0.549	-2.099	1.116
Density	2.1697	2.322	0.934	0.350	-2.382	6.721
sugar_rate	6.4586	2.468	2.617	0.009	1.621	11.296

(a) Regression results

```
Intercept           0.004558
Color               0.758312
Root                1.643137
Sound               3.021667
Texture             1.292895
Belly_button        1.488933
Touch               0.611481
Density             8.755790
sugar_rate        638.171086
dtype: float64
```

(b) Log odds

Fig. 5. Regression results and log odds for Logistic Regression

Conclusion

After conducting the logistic regression modeling, we discovered that sound and sugar rate are the most statistically significant predictors that best predict the quality of the watermelon as it has a p-value less than 0.05. The belly button predictor is proven not statistically significant predictor since its p-value is larger than 0.05. After we fitted the two models and compared, we learned that logistic regression acquired a 66.66% accuracy and KNN acquired a 88.09% accuracy. We also observed an error rate of 33.33% in the logistic model and 11.90% in KNN. Even though KNN gives a better prediction model, it does not let us know the parameter significance so we chose logistic regression.For researchers who are interested in this topic, it would be important to look into additional predictors that may influence the quality prediction e.g. weight. Also, it could also be more valuable to use scientific devices to measure the apparent watermelon qualities like color, sound, and texture. This may be more valuable in showing predictor significance as opposed to subjective terms like "very blurry" and "blurry" when describing texture for example.

Contribution

All the group mates contributed in researching, coding, and debugging; as well as putting the report together with the sources and graphs. Everyone equally contributed to the report composing and data analysis. In addition, Odemuno contributed to KNN classifying and the final presentation, Junkee contributed to logistic regression model, and Xiaoying contributed to model validations.

Codes: https://github.com/Janeiii/COGS109$_{P}roject/blob/main/Watermelon_{Q}uality_{P}rediction_{C}ode_{1(}1).ipynb$