**The Ultimate Halloween Candy Power Ranking Prediction using Decision Trees and Logistic Regression**

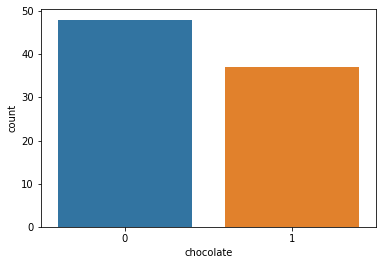
-Amey Vidvans

What’s the best (or at least the most popular) Halloween candy? That was the question this dataset was collected to answer. Data was collected by creating a website where participants were shown presenting two fun-sized candies and asked to click on the one they would prefer to receive. In total, more than 269 thousand votes were collected from 8,371 different IP addresses. The challenge here is using

The data set contains roughly equal number of chocolate and non-chocolate candy data, thus implying a non-biased dataset.

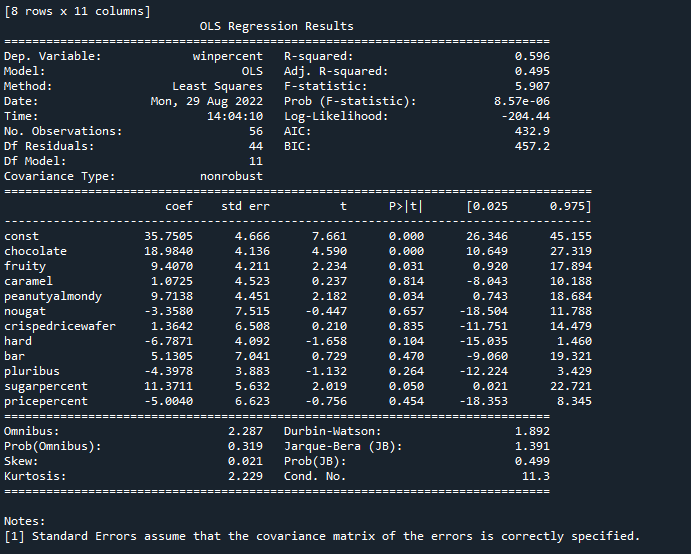
Q1) Is likeability related to chocolate presence?

Let us check the correlation between chocolate presence and likeability. Using pandas corr function, the correlation is found to be 0.63 indicating a significant relationship between

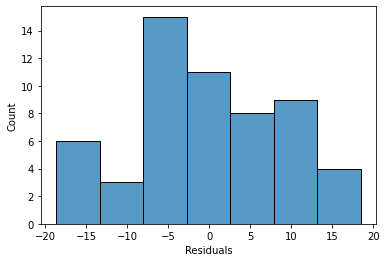


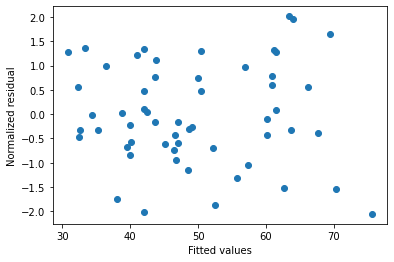
Q2) Can we predict likeability with chocolate presence and other variables?

Here we shall be using Linear Regression to predict the win percent. I have split the data into test and train datasets. Looking at the regression model below, we can see only a few variables are significant.



The errors resulting from the model are shown below and we can say they are roughly normally distributed indicating the linear model does well.





The normalized residuals indicate there is no heteroskedacity in the residuals. This also indicates the linear model is successful in predicting the win percent. Let’s use this model to predict the test data.

The error metrics are shown below.

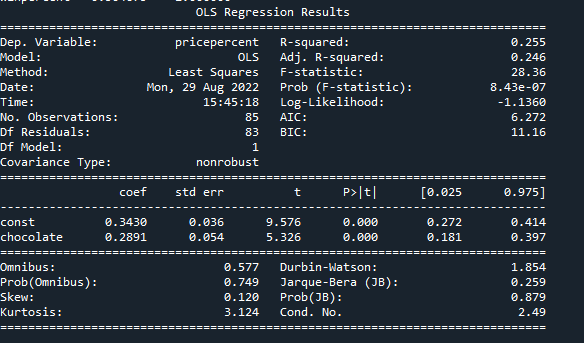
Mean Absolute Error: 9.475891721097527

Mean Squared Error: 143.26677128885373

Root Mean Squared Error: 11.969409813723221

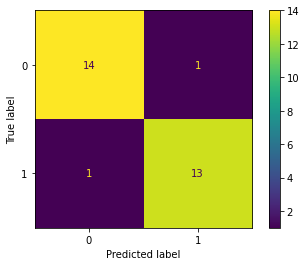
This will serve as the baseline for future comparison. I will not worry about overfitting due to inclusion of non-significant variables for now.

A few interesting observations from this model shows that adding chocolate increases the likeability of a candy. However, there is a tradeoff as the chocolate content is related to the price.

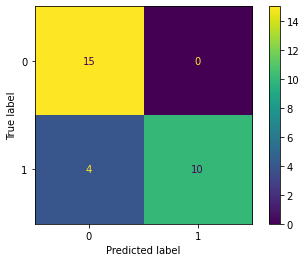


Can we predict chocolate presence from the rest of the data?

A simple logistic regression model is fitted on the train data, and the confusion matrix for the test data is shown below. It does seem the model is fairly well adapted to predicting the presence of chocolate. Lets see which observation throws off the model.



Now for decision trees,



It seems the decision trees are not great at predicting the presence of chocolate when it is present but succeed at accurately resolving the absence of chocolate when it is absent. Of course, we can try and optimize the decision trees but, I believe in simpler is better. Plus, it didn’t really affect the confusion matrices.

We can also compare decision trees to the linear regression model. Let’s do that and conclude this project.

Mean Absolute Error: 9.402064256581902

Mean Squared Error: 130.6323988621851

Root Mean Squared Error: 11.429453130495139

It seems the random forest regressors are better at predicting winpercent. Surprising result !