Sweet Victory: Predicting A Halloween Candy's Popularity based on its Attributes

AUTHOR
Group 8

PUBLISHED
May 2, 2023

Introduction and Data

Halloween is objectively the best time of the year. This glorious season is marked by spooky decor, scary movies on television and in theaters, crisp fall weather, and of course, candy. Whether gobbled up in earnest by children or secretly by adults, Halloween candy is a massive benefit to spooky season, but which Halloween candy is the greatest of them all?

While everyone has their own tastes, in October of 2017, FiveThirtyEight writer Walt Hickey tasked himself with answering this question objectively. Through an online knockout bracket between various popular Halloween candies, Hickey surveyed over 8000 IP addresses on 269,000 matchups. We cannot assume that this translated to 8000 people, as it's impossible to tell if more than one person shares an IP address) with over 269,000 choices - an average of about 33 per address. Sadly, the raw data sourced from the survey itself could not be sourced. Instead, we are provided with the results of Hickey's preliminary analysis, sorted by candy brand.

The details of the knockout bracket are as follows: One user (marked by their computer's IP address) is presented with two randomly selected fun-sized candy options and asked "Which would you prefer as a trick-or-treater?" The participant can either choose one of the two candies or choose to skip the question altogether, with no limit on the amount of responses one IP address can give. However, it is unlikely that any single participant significantly skewed the results, as it would take 3655 responses to give one's opinion about *every* candy combination. Quoth Hickey: *"We don't really need to care about the, say, hardcore Hershey fans attempting to rig the sample, because in order for someone to seriously dent their candy's outcome, they'd have to go through scores of irrelevant matchups."*

His results were outlined in a 2017 FiveThirtyEight article titled 'The Ultimate Candy Power Ranking', and posted to Github on Halloween of that same year. The data, organized by specific candy, details that candy's attributes (set as a series of 13 binary categorical variables), its price relative to its opponents, its sugar percentage, and its performance in the bracket. More specific information is given in the following codebook:

Codebook

Variable	Explanation
competitorname	The name of the candy
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Does it contain caramel?
peanutyalmondy	Does it contain nuts, or a nutty flavor?
nougat	Does it contain nougat?
crispedricewafer	Does it contain a 'crunch' factor, like crisped rice or a wafer?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Does it come in multiple pieces, like Skittles or M&Ms?
sugarpercent	The percentile of sugar as it falls under within the data set.
pricepercent	The unit price percentile compared to the rest of the set.
winpercent	The overall win percentage according to 269,000 matchups.
sour	Is it sour?
shaped	Is it an interesting shape, such as a bear or a worm?
complex	Does it have a complex flavor profile?
colorful	Is it brightly colored or multicolored?

Of course, our initial question is answered easily by simply looking at the data. Reese's has appeared to corner the candy market, occupying 4 out of the top 10 candies, with their original Peanut Butter Cup securing the top spot. But what is it *about* those cups that make them so popular?

In our research, we aim to answer the question: What attributes are most directly correlated with a candy's popularity (measured by win percentage), and how can we use them to create the most marketable candy possible? Based on a candy's attributes, can we predict how popular a candy might be?

Hickey's original research concludes that the presence of chocolate is the greatest predictor of a candy's success, but the answer might not be that simple. After all, Hershey's Milk Chocolate only placed 28th, below non-chocolates like Sour Patch Kids, Starburst, and Haribo Gold Bears. We posit that the greatest predictor of a candy's success is not the presence of one *specific* flavor, but rather a collection of *complex* flavors, colors, and/or shapes. The cardinal sin of Halloween is to be boring, and we believe that our research will corroborate that.

Hickey's research does not take these attributes into account, so some data wrangling was necessary in order to validate our hypothesis.

First, we removed the 'One quarter' and 'One dime' entries, as they aren't candy and Hickey admits in the article that he included them as a joke. (Though interestingly they performed decently well in the bracket - a dime won 32% of its matchups and a quarter won 46% of the time.) Then we added four new columns relating to four specific features that we think will be correlated to a candy's popularity - if a candy is sour, if it has an interesting shape, if it has a complex flavor profile, and if it's a bright color. Bright colors and sour flavors were added because the data set itself seems to be biased towards chocolate candies - chocolate bars were given 6 layers of analysis, while fruit candies were only given the attribute 'fruity'! A candy was labelled 'complex' if it contained two or more different flavor attributes.

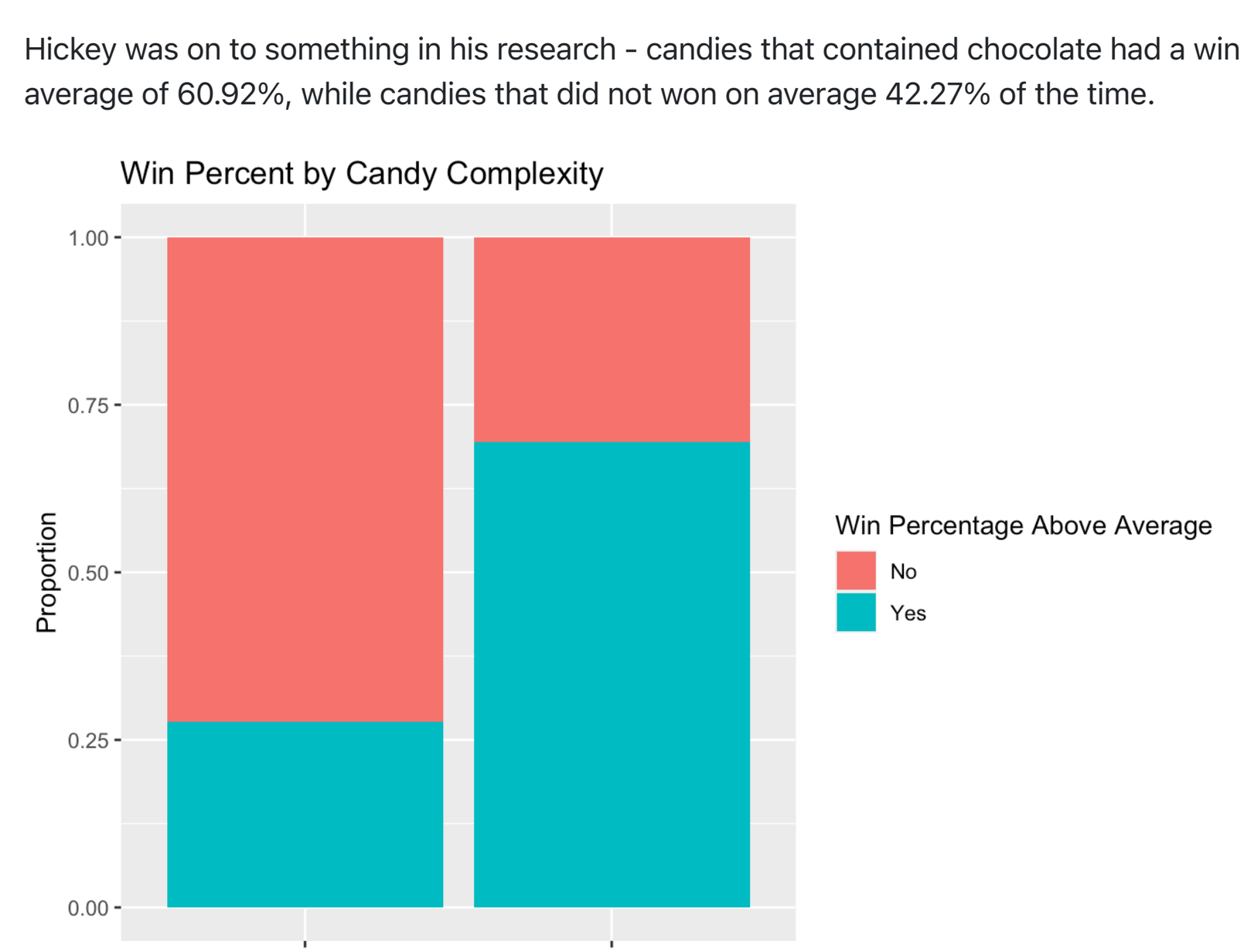
We plan on using all of these attributes when building our prediction model. These binary variables on their own do not provide too much information, but the unique combinations of them for each candy may help us uncover what unique flavor profile and physical attributes are most appealing to the public.

Methodology

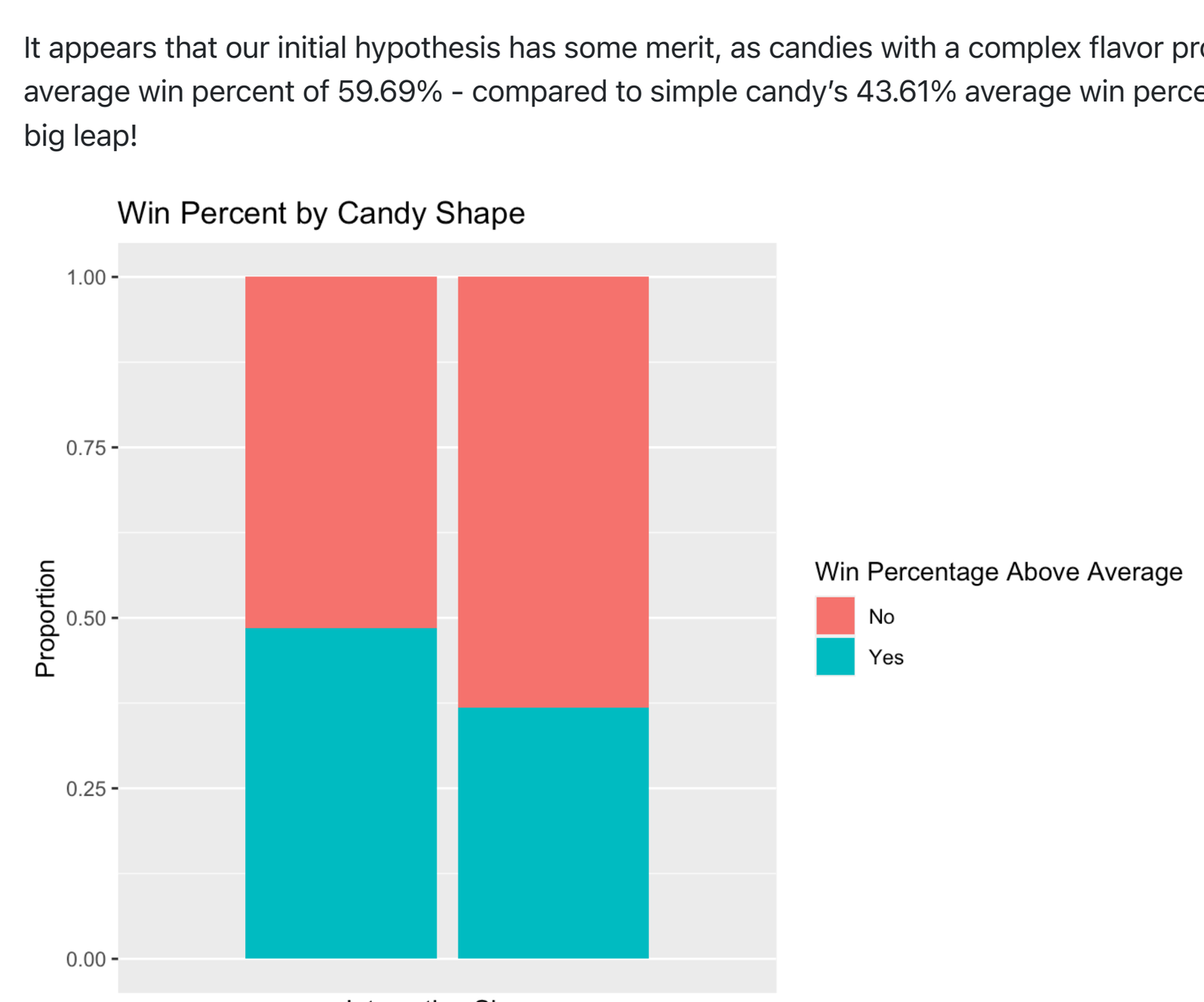
Our outcome variable is the candy's popularity, shown in the data set as `winpercent` - how often it was selected as a preferred option as opposed to its randomly selected competitor. It is a numerical variable, and is a representation of the probability that a candy is preferable to the others, and thus its 'popularity'. Because we are looking to understand how a candy's attributes might affect its popularity, our key exploratory variables include the candy's individual attributes - that is, all of the categorical columns of the dataset. Based on analysis of the correlation between attributes and popularity, we can then make predictions of a candy's popularity based on its attributes.

Lastly, for our preliminary analysis, we will create graphs to figure out the relationship between features of a candy (chocolate content, complexity, and shape) and its win percent. We hypothesize that win percent has a positive correlation with chocolate content, complexity, interesting shapes, and colorfulness.

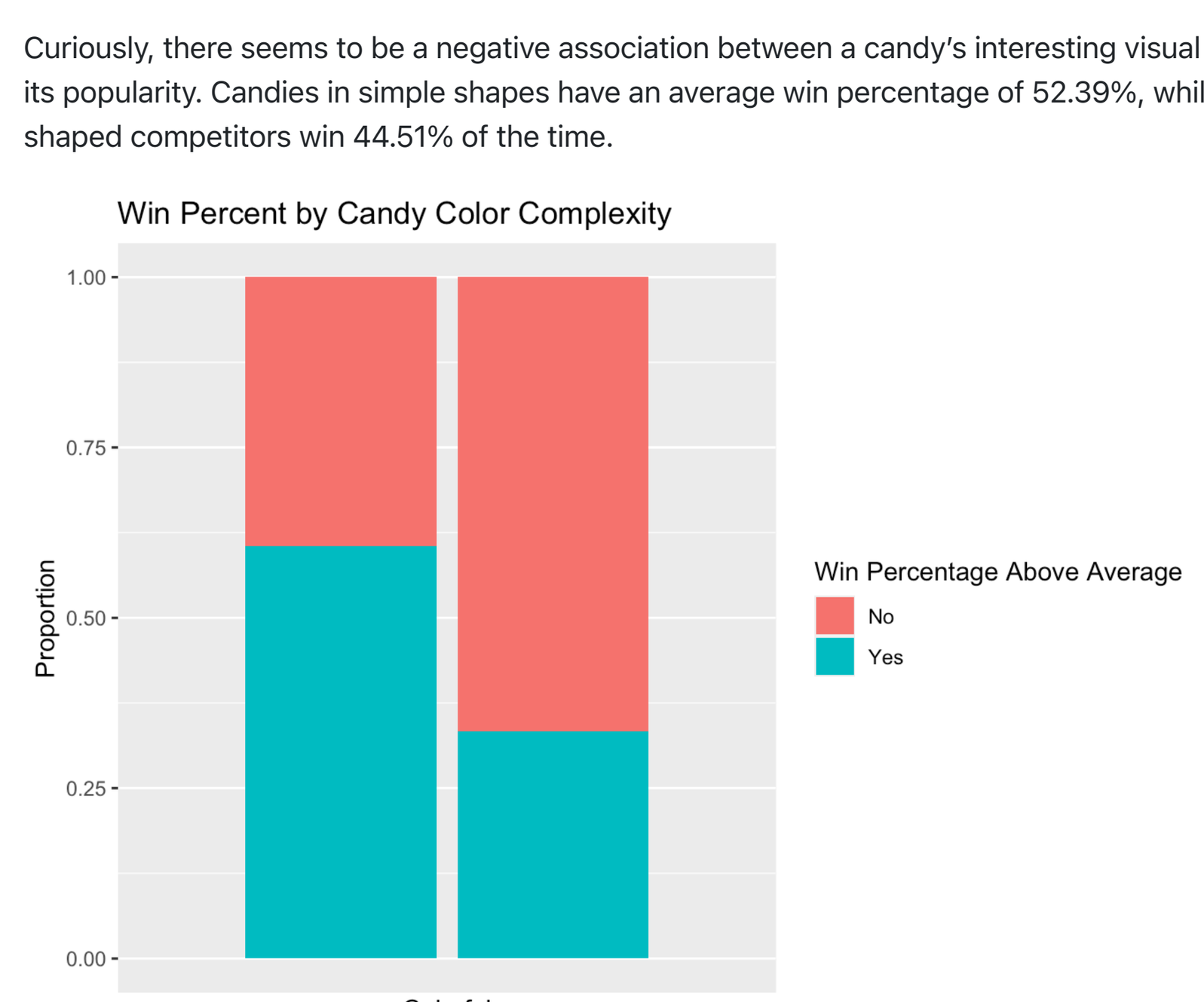
In order to present relevant summary statistics for the data, as well as run preliminary analysis of how different characteristics of a candy affects its win percent, we are using means and bar graphs to plot the relationship between a candy's characteristics and its win percent. Firstly, the overall average win percent is calculated to provide a threshold to determine whether a candy scores better than the average win percent. The bar graph, then, shows the proportion of the candy's win percent, according to the threshold, based on if it has the characteristic (1) or if it doesn't (0). The mean win percent for each candy's categories and whether or not they have the characteristics is also calculated.



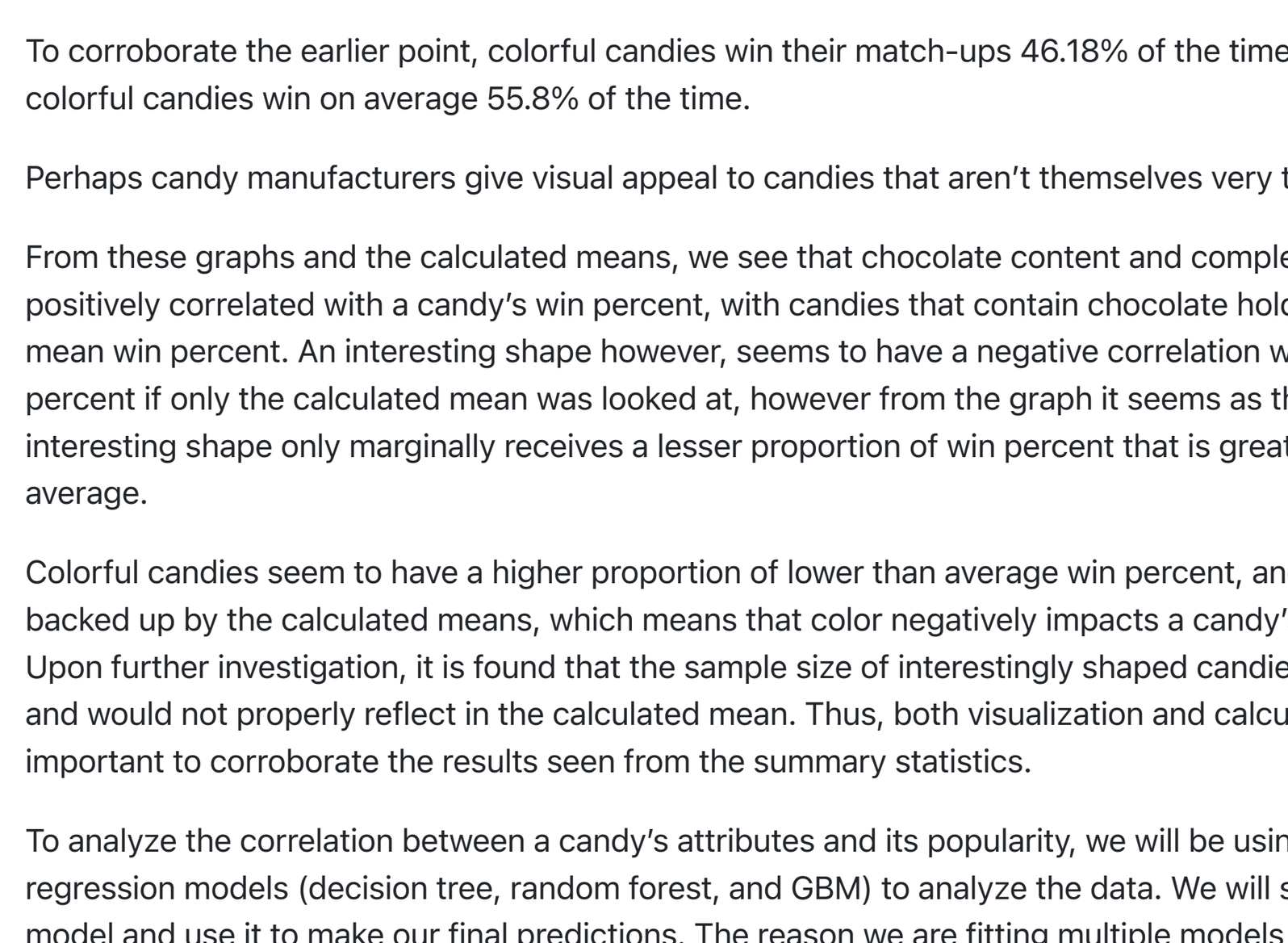
Hickey was on to something in his research - candies that contained chocolate had a win percent average of 60.92%, while candies that did not won on average 42.27% of the time.



It appears that our initial hypothesis has some merit, as candies with a complex flavor profile had an average win percent of 59.69% - compared to simple candy's 43.61% average win percentage, that's a big leap!



Curiously, there seems to be a negative association between a candy's interesting visual attributes and its popularity. Candies in simple shapes have an average win percentage of 52.39%, while their funky-shaped competitors win 44.51% of the time.



To corroborate the earlier point, colorful candies win their match-ups 46.18% of the time, and non-colorful candies win on average 55.8% of the time.

Perhaps candy manufacturers give visual appeal to candies that aren't themselves very tasty?

From these graphs and the calculated means, we see that chocolate content and complexity are positively correlated with a candy's win percent, with candies that contain chocolate holding the highest mean win percent. An interesting shape however, seems to have a negative correlation with a candy's win percent if only the calculated mean was looked at, however from the graph it seems as though an interesting shape only marginally receives a lesser proportion of win percent that is greater than the average.

Colorful candies seem to have a higher proportion of lower than average win percent, and this is also backed up by the calculated means, which means that color negatively impacts a candy's win percent. Upon further investigation, it is found that the sample size of interestingly shaped candies is very small and would not properly reflect in the calculated mean. Thus, both visualization and calculated means are important to corroborate the results seen from the summary statistics.

To analyze the correlation between a candy's attributes and its popularity, we will be using a variety of regression models (decision tree, random forest, and GBM) to analyze the data. We will select the best model and use it to make our final predictions. The reason we are fitting multiple models is because of our small data set. These models are all compatible with regression, and vary in complexity and accuracy, with interpretability often being the trade-off. Because almost all of our predictors are binary, we want to use more complex models that may be able to interpret these rather uninformative variables in a way that will lead to higher accuracy. We don't believe a linear regression model could perform this well. By using multiple models we hope to get the best performing machine possible. We hypothesize that win percent has a positive correlation with chocolate content, complexity, and interesting shapes.

Results

We choose to use Leave-One-Out Cross-Validation (LOOCV) instead of a traditional train-test split method for evaluating machine learning models due to the small size of our data set. With a small data set, a train-test split can result in a small sample size for training the model, leading to overfitting or underfitting depending on the number of observations and features in the data. LOOCV allows us to train the model on almost the entire data set, while still allowing for independent evaluation of the model's performance on each data point. This approach provides a more accurate estimate of the model's performance on new, unseen data, which is especially important when dealing with a small data set. Additionally, LOOCV eliminates the potential bias introduced by a random train-test split and provides a more reliable estimate of the model's generalization performance. Overall, LOOCV is a suitable approach for evaluating machine learning models when dealing with small data sets, as it provides a more accurate and reliable estimate of model performance.

Random Forest

After performing cross-validation to find the best Random Forest Regressor model, an RF with a max depth of 10, a minimum number of samples required for each internal node for a split of 2, and the number of estimators to be 200 is the best for predicting candy popularity using the candy data set. The model achieves an MSE of 133.92 and an R^2 score of 0.38. The Random Forest Regressor model uses an ensemble of decision trees to make predictions, which can capture nonlinear relationships and interactions between features that may not be captured by a linear model.

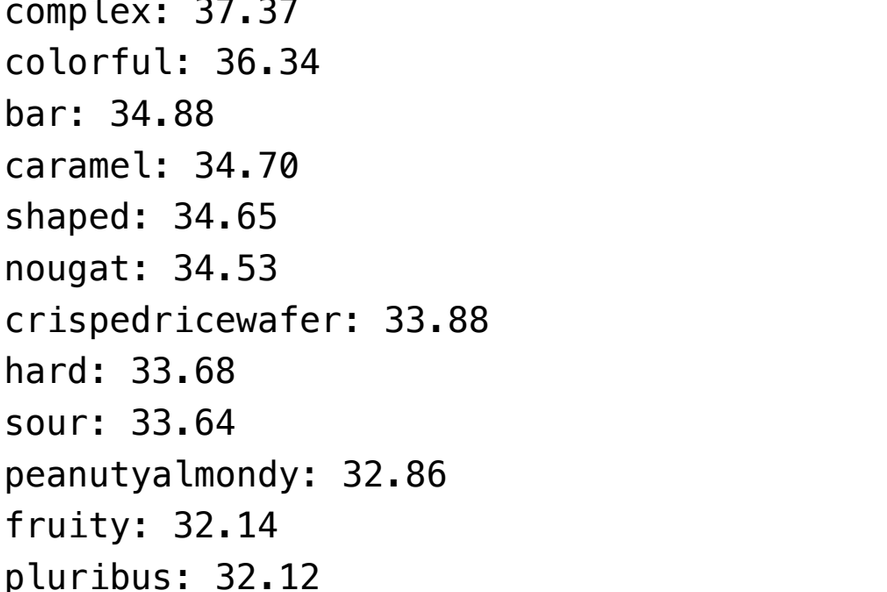
Decision Tree

For a Decision Tree Regressor model, it performs best when the max depth is 5, the minimum of samples needed to split an internal node is 10, and the minimum number of samples to be considered a leaf is 4. The model achieves an MSE of 142.21 and an R^2 score of 0.34, which is lower than the performance of the Random Forest Regressor model on the same data set. The Decision Tree Regressor model is a simpler model than the Random Forest Regressor model and can be useful for interpreting the relationships between features and the target variable. However, in this case, the Random Forest Regressor model performs better.

Gradient Boosting Regressor

The best hyperparameters for the GBM model are to have a learning rate of 0.1, a maximum depth of 7, the square root of the number of features used to determine the best split, 1 samples minimum for each leaf, 5 samples minimum samples for a split of an internal node, and 200 estimators. The MSE is relatively high at 138.73 and the R^2 score is 0.35.

In conclusion, the Random Forest Regressor model can best predict candy popularity using the candy data set. It accounts for about 37.68% of the variation in win percentage of a certain candy. The MSE is still relatively high, with an error of 133.92 on the entire model. This performs slightly better than the GBM, and much better than only using one decision tree.



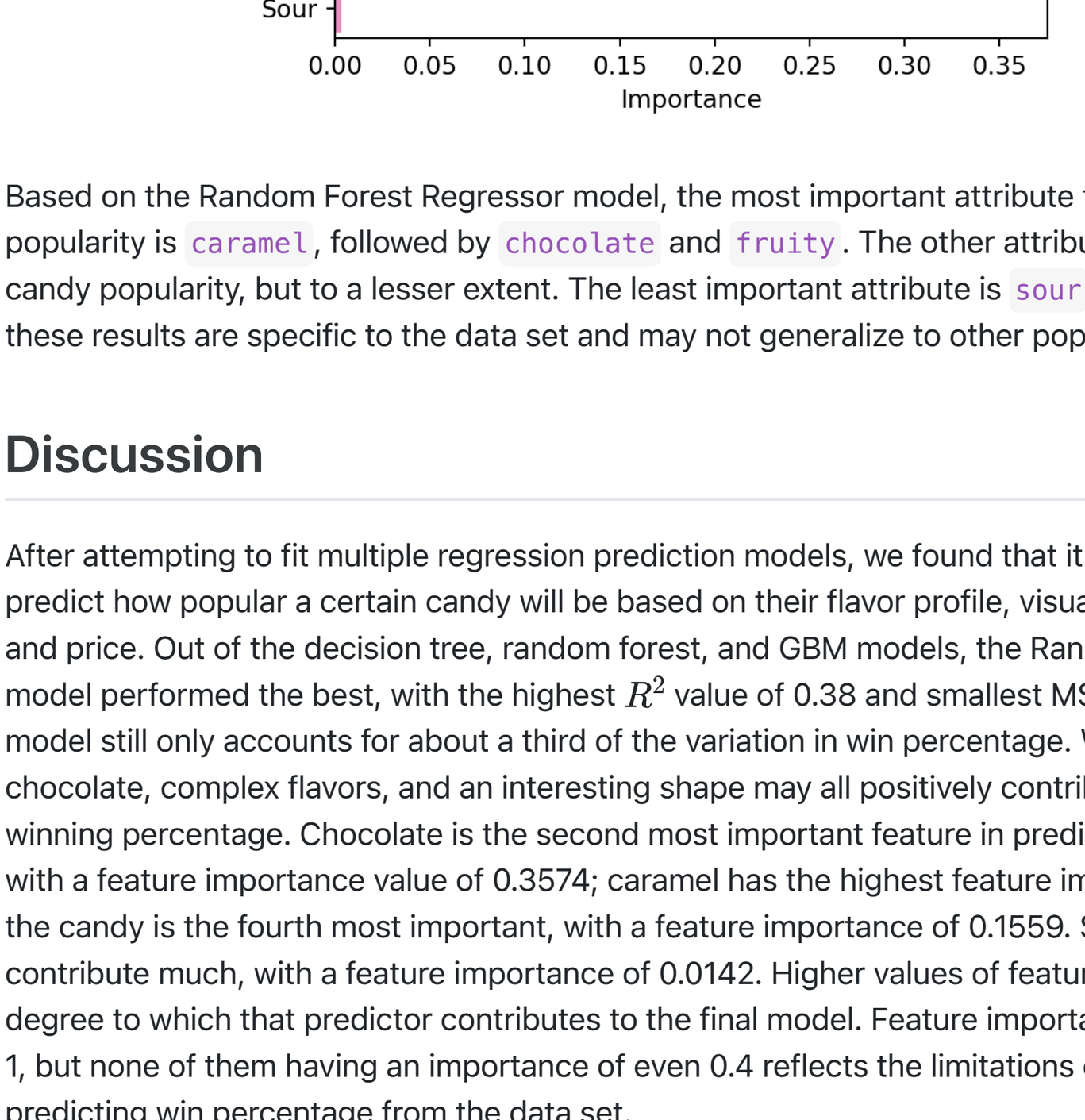
The Random Forest Regressor model's feature importance analysis shows that chocolate is the most important feature, with a weight of 0.376, followed by colorful and complex attributes with weights of 0.150 and 0.171, respectively. Other important features include fruity, caramel, peanutyalmondy, nougat, and sour. Pricepercent has the lowest feature importance weight, indicating that it has the least impact on candy popularity. These results suggest that a candy's taste and appearance are more important than its price in determining its popularity.

```
Winning Percentage with Only One Attribute:

chocolate: 47.16
sugarpercent: 40.87
complex: 37.37
colorful: 36.34
bar: 34.88
caramel: 34.70
shaped: 34.65
nougat: 34.53
crispedricewafer: 33.88
hard: 33.68
sour: 33.64
peanutyalmondy: 32.86
fruity: 32.14
pluribus: 32.12
pricepercent: 30.22
```

The Random Forest Regressor model was used to predict the popularity of different attributes of candy. The predicted popularity values for each attribute were then sorted in descending order. The attribute with the highest predicted popularity is chocolate with a score of 47.16, followed by a sugar percentile of 1 (the most sugary) with an expected winning percentage of 40.87 and colorful with a predicted winning percentage of 36.34 These results suggest that chocolate and high sugar content are strong predictors of candy popularity. A price percentile of 1 (most expensive) of a candy is predicted to be the least important attribute in predicting candy popularity. These results can be useful for candy manufacturers and marketers to understand which attributes of candy are most likely to be popular among consumers.

Feature Importances



Based on the Random Forest Regressor model, the most important attribute for predicting candy popularity is `caramel`, followed by `chocolate` and `fruity`. The other attributes may also play a role in candy popularity, but to a lesser extent. The least important attribute is `sour`. It's important to note that these results are specific to the data set and may not generalize to other populations or contexts.

Discussion

After attempting to fit multiple regression prediction models, we found that it is hard for machines to predict how popular a certain candy will be based on their flavor profile, visual and tactile presentation, and price. Out of the decision tree, random forest, and GBM models, the Random Forest Regressor model performed the best, with the highest R^2 value of 0.38 and smallest MSE of 133.92. However, this model still only accounts for about a third of the variation in win percentage. We hypothesized that chocolate, complex flavors, and an interesting shape may all positively contribute to predicting a candy's winning percentage. Chocolate is the second most important feature in predicting winning percentage, with a feature importance value of 0.3574; caramel has the highest feature importance of . Complexity of the candy is the fourth most important, with a feature importance of 0.1559. Shape does not seem to contribute much, with a feature importance of 0.0142. Higher values of feature importance indicates the degree to which that predictor contributes to the final model. Feature importance values range from 0 to 1, but none of them having an importance of even 0.4 reflects the limitations of our data and difficulty predicting win percentage from the data set.

To test our model, we defined candies with only one of 15 attributes being marked as present. Then, we predicted the winning percentage using our optimal Random Forest Regressor model. The results reveal that chocolate as the only attribute has the highest winning percentage of 47.16. It's hard to interpret these estimates in context (in regards to the predicted winning percentage with only that attribute present). For example, a candy cannot have an interesting shape with no other of the listed attributes (it would have no flavor). Under a similar interpretation, if a candy has a complex flavor profile but none of the individual flavors, then what does this complex flavor consist of? Our model predicts candies with a price percentile of 1 (most expensive) to have the lowest winning percentage, but this theoretical candy has no flavor or visual attributes. There is inherent interaction between these variables, which our predictions do not account for.

Because our data set is relatively small, we opted for a LOOCV (leave one out cross-validation) method to find the optimal model. This is a more computationally expensive algorithm but gives us more leverage with our small amount of data. The training data, with its inherent limitation of being rather small, limits the validity of our model. While we did use the LOOCV method to maximize each observation, 83 observations is still a very small amount to fit a model to. Something else that may limit the predictions for new data we can make is that the categories added to the original data set are mostly arbitrary. Analyst Sarah Kessler hand-picked candies that fit the criteria for each category out of the data set, imparting a not insignificant amount of human error. Perhaps a few different candies could have been categorized as colorful or interesting by a different human analyst, and that could have contributed to different scores in the `colorful` and `shaped` columns.

If we were able to start over with the project, we probably would try to find a new data set. There are not many observations in this data set, and while we tried sourcing the original, raw data, it proved unsuccessful. Having this raw data may have lead to a much more interesting analysis. If we were able to conduct the same survey as Hickey, we would ask about other demographic information like age, gender, location etc. Perhaps there's a stratified difference depending on who the candy consumer is (ex. children may like colorful, interestingly shaped candy, while adults may be more about the complexity of the flavor profile). Including factors other than the candy's attributes would also give us a richer data set with variables that aren't binary indicators. Like Hickey, we would probably deploy it online and try to gather a large, randomized sample size. The resulting data would also give us the results of each comparison, rather than the overall percentage of matchups each candy won. So, not only could we predict the absolute win percentage, but perhaps we could predict if a given candy would fall in the top 5 or bottom 5 (a question of classification). Only having 83 observations is an inherent limitation to building a model and performing predictions.

Citation

Hickey, W. (2017, October 27). *The ultimate halloween candy power ranking.* FiveThirtyEight. Retrieved April 16, 2023, from https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/