Expanding our candy brand

Philipp Spangenberg

October 2nd, 2020

1. Introduction
   1. Background

Lidl purchasing group wants to create a new store brand candy from scratch. In order to utilize a data driven approach, data was collected from a market research group. This data was made available to Lidl Analytics, in order to identify the main components and characteristics of the new candy.

* 1. Business Understanding/Problem Description

Based on consumer sentiment, a new candy needs to be created. A dataset was made available by FiveThirtyEight. It includes a variety of known candy products with a variety of characteristics. The goal is to maximize the dependent variable “winpercent” as an indicatior for the most popular product.

The data set will by analyzed in python (Jupyter notebooks).

1. Data Description

The dataset consist of 85 observations (individual candies) with 9 characteristics. They were provided in a csv file. Eight of which are considered independent variables (predictor variables), while “winpercent” is the dependent (target) variable. The variables are integers (binary 1 or 0) and floats (percentages):

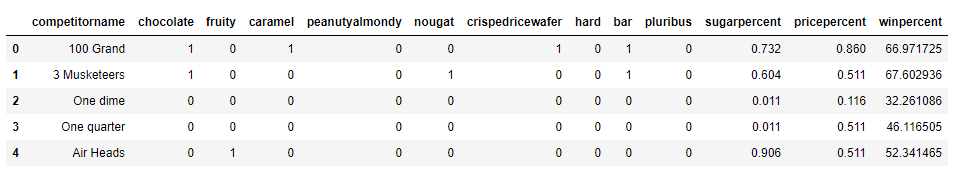


Figure 2.1 Loaded Data Set



Figure 2.2 Data Description

1. Methodology

After cleaning and preparing the data, the task of creating a model was broken down into several steps, in order to identify important characteristics.

Firstly, basic exploratory analysis was applied to the data. This was done by creating bar charts for the most popular candies. Secondly, a heatmap of the correlation of the different features was created. Finally, boxplots between the most impactful variables and the target variable were created.

The seaborn package was used to create statistical data visualization and the sklearn library was utilized to create the individual models.

* 1. Exploratory Data Analysis

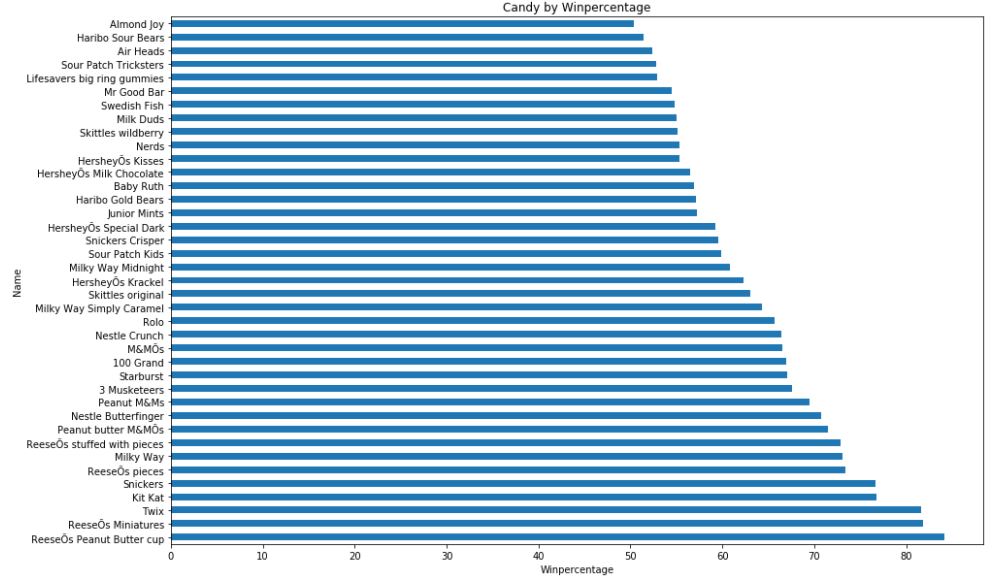


Figure 3.1.1

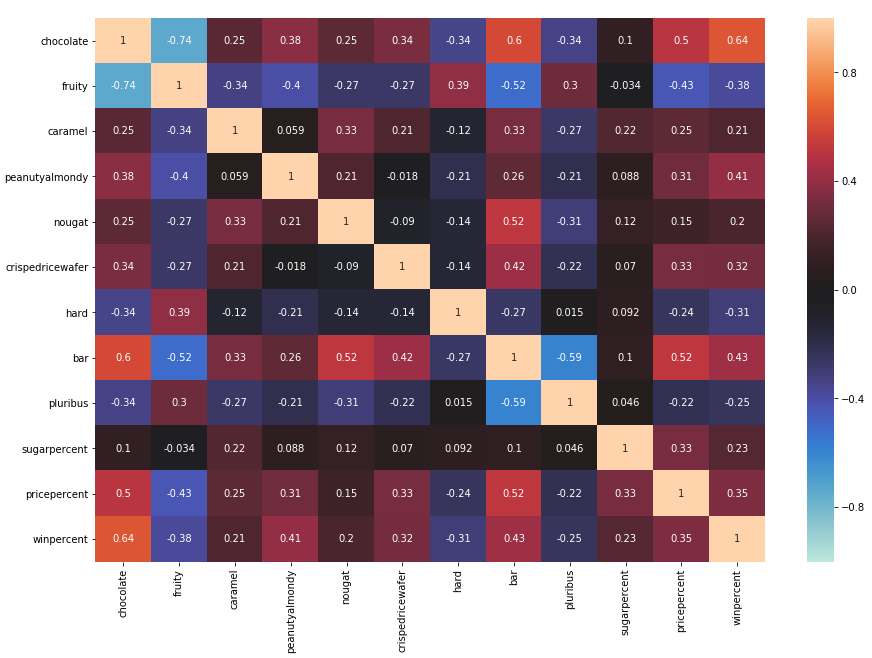


Figure 3.1.2

As can be observed in Figure 3.1.2, the strongest positive relations with winpercent are: chocolate (0.64), bar (0.43) and peanutyalmondy (0.41). The attributes fruity (-0.38) and hard bar (-0.31) were negatively correlated.

Based on the heat map, five features were investigated, using univariate analysis in the form of boxplots:

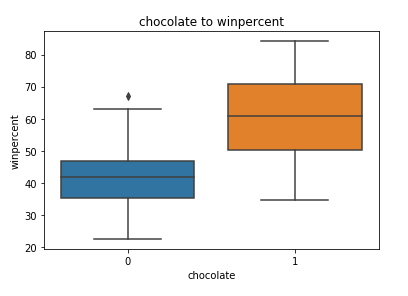


Figure 3.1.3

From the plot, we can deduce that a candy that contains chocolate has a median winpercent of over 60.

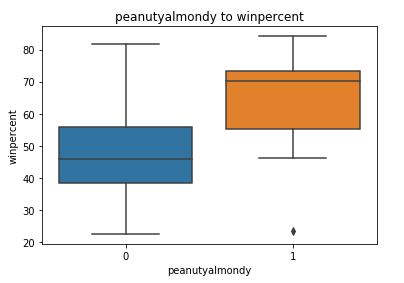


Figure 3.1.4

While a bar type candy has a median win rate of over 60, there are several outliers. This means that there are several very successful candies, which are not bars.

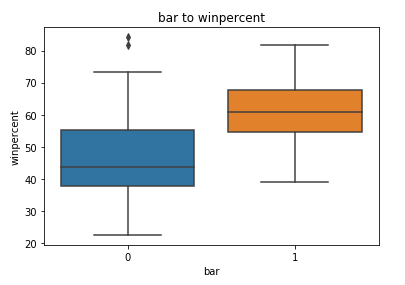


Figure 3.1.5

Candy containing peanuts, peanut butter or almonds have a very high median winrate of over 70. This feature might be of limited use, as it contains several distinct characteristics (e.g. nuts).

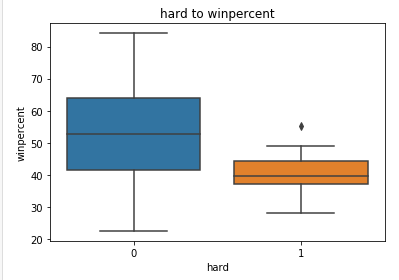


Figure 3.1.6

Hard candies have a lower than average win percentage and are in fact slightly negatively correlated with the target variable.

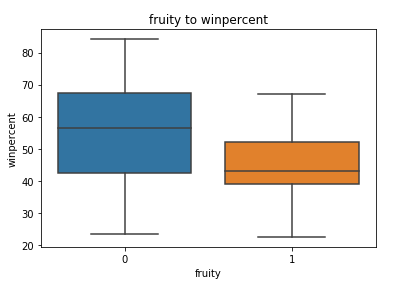
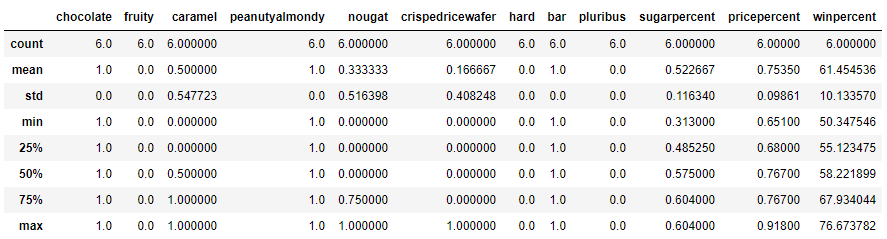


Figure 3.1.7

Fruity candy also seem to be negatively correlated to successful candy.

From the basic correlation matrix and the univariate analysis, we can deduce, that the "perfect candy" consists of a bar made with chocolate, peanuts, peanut butter or almonds. It is not fruity and not a hard bar. In order to test this hypothesis, we select all bars with the mentioned features and calculate the average win percentage:

v Figure 3.1.8

The average win percentage is 61.45 compared to 50.32 of the full data set.

* 1. Regression

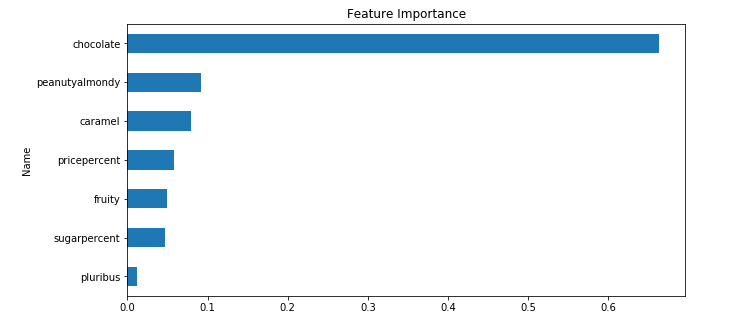
Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable:

Figure 3.2.1

As can be seen the features “chocolate” and “peanutyalmondy” are not only strongly positive related to the target variable, but are also useful in predicting the final result. However, since we only have 11 available features and a tiny sample size, no features were dropped in advance.

The model was split into a training and a test set, as is standard procedure for machine learning models. A test size of 30% was used to set up the split.

As we want to predict a single variable, while utilizing a set of independent variables a regression seems to be best suited to predict an optimal product. Since we have multiple predictor variables a "Multiple Linear Regression" or a "Polynomial Regression" seem to be the most intuitive fit.

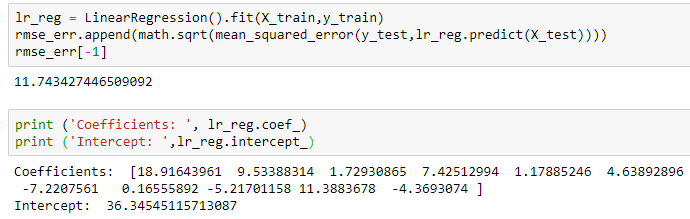


Figure 3.2.2 Multiple Regression

|  |  |
| --- | --- |
| **Candy Type** | **Coefficient** |
| chocolate | 18.92 |
| sugarpercent | 11.39 |
| fruity | 9.53 |
| peanutalmondy | 7.43 |
| crispedricewafer | 4.64 |
| caramel | 1.73 |
| nougat | 1.18 |
| bar | 0.17 |
| pricepercent | -4.37 |
| pluribus | -5.22 |
| hard | -7.22 |

Figure 3.2.3 Regression Coefficient

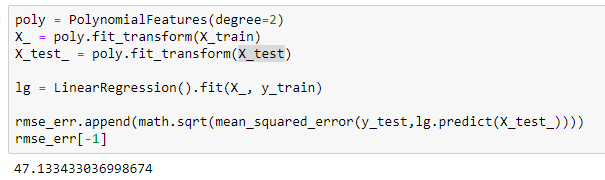


Figure 3.2.3 Polynomial Regression

1. Result

As we can see the multiple linear regression has a lower RMSE with 11.74 than the polynomial regression with 47.13 (2nd degree). Thus, it was deemed a better predictor of the model.

Interpreting the multiple linear regression, we can see that a candy without any of the given characteristics would have a win percentage of 36.35 (intercept). Adding chocolate would increase the win percentage by about 19 points. A hard bar, by contrast would subtract 7 points from the result.

1. Discussion

A point of improvement would be a larger data set with a richer variety of target variables. Furthermore, a more detailed breakdown of the target variable into customers would allow a segmentation of customer groups (e.g. health conscious likes low sugar). A more detailed analysis would also utilize a larger variety of machine learning algorithms.

1. Conclusion

In order to create the “perfect candy” from scratch it would seem logical to throw the most impactful characteristics from the regression together. However, several of the attributes also seem mutually exclusive (fruity and chocolate).

To draw from the EDA and the regression, it would make sense to create a candy containing chocolate, with an above average content of sugar and some type of nut (almond, or peanut). As fruit is negatively correlated with win percentage I would abstain from adding another flavor into the mix. Furthermore, creating a bar product seems to add to the appeal, while it seems to be clear that it needs to be a “soft bar”. Lastly, it would make sense to sell the product in a box (pluribus = 0).

Other factors might add to the appeal, however from personal experience a product which tries to be everything usually ends up being nothing.

1. Reference

[1] https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking