

# Predicting Ratings For a Variety of Dark Chocolates

Manvir Kohli, Julie Song, Kelvin Wong

2022-11-26

## Contents

<b>Summary</b>	<b>1</b>
<b>Introduction</b>	<b>1</b>
<b>Methods</b>	<b>1</b>
Data . . . . .	1
Analysis . . . . .	2
Results and Discussions . . . . .	2

## Summary

Four regression models were built using decision tree, support vector machine, ridge (linear regression) and k-nearest neighbors algorithms, and compared to determine the best model for predicting a chocolate's rating based on characteristics such as the number and type of ingredients in the chocolate, its amount of cocoa, location for the manufacturing company, memorable characteristics, and several others. The ratings are predicted based on a scale of 1 to 5. Our models performed \_\_\_\_\_ on a test set containing \_\_\_\_\_ observations...

## Introduction

How well-received a new chocolate may be, can depend on a variety of factors. The project present here is interested in predicting the rating for a type of dark chocolate on a scale of 1 to 5, given some of its characteristics. This information can be useful for product analysis, and may be able to aide in predicting the popularity of a new product. It may also be useful for determining some important factors in the development of the chocolate, and suggest characteristics to focus on in order to develop a successful product.

## Methods

### Data

The data set is provided by the Manhattan Chocolate Society, and was found and retrieved from the tidytuesday data project, specifically through this link. The data set contains observations for different types of dark chocolate, including the manufacturing company and its location, origin of the cocoa beans used to make the chocolate, the other ingredients in the chocolate, the amount of cocoa in the chocolate, and others. They have also provided a feature column that contains descriptive words relating to the characteristics of the chocolate flavor, and a final rating.

Before the models were built, the data also had to be processed. The information in the **ingredients** feature column was split apart into several columns, including one column for the number of ingredients in the chocolate, and one column for each of the recorded ingredients. The company location and cocoa bean

origin features were modified such that only around the 15 or 20 such categories in each were considered, while the remaining categories were considered “Other”. The amount of cocoa in the chocolate feature was also converted from a character data type to a numerical datatype.

## Analysis

Four regression models were built after splitting the original dataset into training and testing dataset in a 70%-30% split. Apart from the processing performed above, some features were also dropped due to being practically unique to each observations, such as the reference number and the company manufacturer. Table 1 shows the hyperparameters that were considered for each regression model. Hyperparameter optimization was performed for each model, using 20 iterations of 5-fold cross-validation each.

Table 1: Model Hyperparameters

Model	Hyperparameter
KNN	Leaf Size
	Number of Neighbors
	Weights
Ridge	Alpha
SVM RBF	C
	Gamma
Decision Tree	Max Depth

The initial exploratory data analysis for this dataset was performed in R using the following packages: dplyr, tidyverse, cowplot, knitr, and kableExtra. The modelling was performed in Python using the following packages, available in the provided environment file: ipykernel, ipython>=7.15, matplotlib>=3.2.2, scikit-learn>=1.1.3, requests>=2.24.0, graphviz, python-graphviz, eli5, shap, jinja2, altair\_saver, selenium<4.3.0, pandas<1.5, imbalanced-learn, pip, lightgbm, joblib==1.1.0, mglearn, psutil>=5.7.2, docopt-ng, dill, pycodestyle, and autopep8.

## Results and Discussions

The distributions for the numerical and categorical features were first analyzed to check for any unreasonable skewness. These distributions are present in Figures 1 and 2. The top 16 company locations and the top 23 origin locations for the cocoa beans were kept as individual categories, while the remaining categories were grouped as Other. It is revealed that the manufacturing company feature has an overwhelming number of categories, and even taking into considering only the top 50 companies, the Other category has far more points compared to the others. This feature was considered too unique to the observations, and was omitted from the modelling.

Table 2 shows the optimal hyperparameters, their associated values, and the resulting validation scores for each model. The hyperparameter optimization and cross-validation was performed for each of the models in this regression problem.

Table 2: Optimized Hyperparameters With Validation Scores

Model	Hyperparameter	Optimized Value
KNN	Leaf Size	1
	Number of Neighbors	2
	Weights	3
Ridge	Alpha	4
SVM RBF	C	5
	Gamma	6
Decision Tree	Max Depth	7

The models performed \_\_\_\_\_ on the test datasets (how they compare with the validation results, how they compare with each other, were they accurate) Table 3 shows a comparison of scores for the models. The scoring metric used was the negative mean absolute percentage error metric, hence the scores are negative. Looking only at their magnitudes, \_\_\_\_\_model had the smallest error (so was the best model, \_\_\_\_\_% difference compared to the actual rating)

There are several areas for improvement with these models. For instance, perhaps we could use a sentiment analysis package on the memorable characteristics text feature, and manually aggregate some of the - feature selection with ridge (based on coefficients) - sentiment analysis or something with the text feature