Chocolate Rating Prediction Model

Kolaković Hasib[a]

[a]*Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper, Slovenia*

Abstract

Predicting the binary classification of chocolate bar ratings using the ”Flavors of Cacao” dataset, based on 8 attributes: name of a company, geographical origin of the bar, a value linked to when the review was entered in the database, date of review, cocoa percentage, company location, bean type, and broad bean origin. The objective is to classify the ratings into two categories: high (> 3.25) and low (≤ 3.25). The data preprocessing steps involved handling missing values, encoding categorical variables, and converting the string formats into numeric formats. We used four machine learning models: XGBoost, Logistic Regression, Decision Tree, and Random Forest. The models were evaluated based on their accuracy and F1 scores, revealing that XGBoost and Random Forest outperform the other models in predictive performance.

1. Introduction

Predicting the binary classification of chocolate ratings using the ”Flavors of Cacao” dataset. The global chocolate industry is continually evolving, with consumer preferences significantly influencing market. Us understanding and predicting chocolate bar ratings can provide insights for manufacturers and marketers. We will use machine learning tools in python programming language for predictive analysis, enabling the classification and evaluation of complex datasets. This paper inspects the usefulness of three machine learning algorithms: Decision Tree, Logistic Regression and XGBoost, so we can classify chocolate ratings from the ”Flavors of Cacao” dataset.

The dataset contains attributes which can be used to predict the quality rating of chocolate bars. The objective is to classify the ratings into two categories: high (> 3.25) and low (≤ 3.25). Effective preprocessing techniques are crucial for preparing the data, including cleaning, encoding categorical variables, handling missing values, splitting the dataset into train and test set. This paper aims to transform the rating attribute into a binary classification problem.

We used the Decision Tree, Logistic Regression, and XGBoost classifiers to train and test our model. Based on their accuracy and F1 scores, we revealed that XGBoost outperforms both Decision Tree and Logistic Regression in terms of predictive performance. This research seeks to identify the most effective model for predicting chocolate ratings.

2. Repository to Project

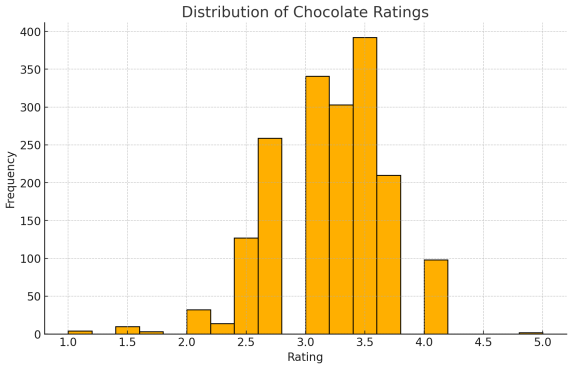
The project's repository can be found at: `https://github.com/Nuklearni-Burek/DataMiningProject`

3. Analysis & Data Preparation

First we analyzed the dataset. We want to ”meet” our dataset, that is to get some insight, to get the number of entries, attribute names, and finally to get description about count, mean, standard deviation, minimum, maximum, first and third quartile about numeric attributes in our dataset. This is achieved using the following code:

- print(data.head())
- print(data.info())
- print(data.describe())

We found out that there are 1795 instances of which: 1093 are equal or smaller than 3.25, and 702 are strictly greater than 3.25.



The next step is to prepare our data. We have to clean column names, that is to remove any backspaces. Next, we found that Cocoa Percentage column has type string.

It is easier to transform this type into float, for easier data handling. We need to handle missing values, and we found that there are two such instances(we can just drop them out of our dataset).

Encoding categorical variables is needed, because it is easier for computer to work with numeric, rather than string values. We introduce new attribute called Binary Rating.

The next step is to split the dataset into test and train sets, and start model preparation.

4. Model Chosen

For predicting data, we have chosen one linear model: Logistic Regression and four non-linear models: Decision Tree, XGBoost and Random Forest.

4.1. Decision Tree

The Decision Tree classifier is a simple model that has tree structure, where internal nodes represent features of a dataset and branches represent decisions.

Code:

```
decision_tree_model = DecisionTreeClassifier(
    random_state=42)
decision_tree_model.fit(X_train, y_train)
dectree_predictions = decision_tree_model.
    predict(X_test)
```

4.2. Logistic Regression

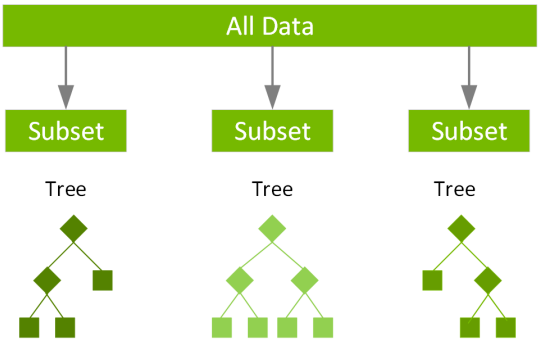
Logistic Regression is a linear model for binary classification. Number of maximum iterations of model is 200.

Code:

```
logistic_regression_model = LogisticRegression(
    max_iter=200,random_state=40)
logistic_regression_model.fit(X_train, y_train)
logreg_predictions = logistic_regression_model.
    predict(X_test)
```

4.3. XGBoost

XGBoost is a powerful gradient boosting algorithm that often outperforms other models in terms of predictive performance. In our example we used 150 estimators, which are trees of maximum depth 3.



Code:

```
xgb_model = xgb.XGBClassifier(n_estimators=150,
    max_depth=3, random_state=42)
xgb_model.fit(X_train, y_train)
xgb_predictions = xgb_model.predict(X_test)
```

4.4. Random Forest

Random Forest is a non-linear model that builds multiple decision trees using random subsets of data and features.

Code:

```
rf_model = RandomForestClassifier(random_state
    =42)
rf_model.fit(X_train, y_train)
rf_predictions = rf_model.predict(X_test)
```

5. Results Comparison

Based on their accuracy and F1 scores, we revealed that XGBoost outperforms Decision Tree, Logistic Regression, and Random Forest in terms of predictive performance. The following table summarizes the results:

Model	Accuracy	F1 Score
Decision Tree	0.57	0.66
Logistic Regression	0.62	0.74
XGBoost	0.67	0.75
Random Forest	0.67	0.76

Table 1: Comparison of model performance.

6. Conclusion

As we can see the XGBoost and Random Forest models outperformed Decision Tree and Logistic Regression models. Both had accuracy of 0.67, while Random Forest had a little bit higher F1 score. Logistic Regression had an accuracy of 0.62, while its F1 score was still high. Decision Tree model performed the worst, indicating that this is not the best choice for this task.

References

[1] IBM, "What is Logistic Regression?", https://www.ibm.com/topics/logistic-regression.
[2] Javatpoint, "Machine Learning Decision Tree Classification Algorithm", https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.
[3] NVIDIA, "What is XGBoost?", https://www.nvidia.com/en-us/glossary/xgboost/.
[4] IBM, "What is Random Forest?", https://www.ibm.com/topics/random-forest.