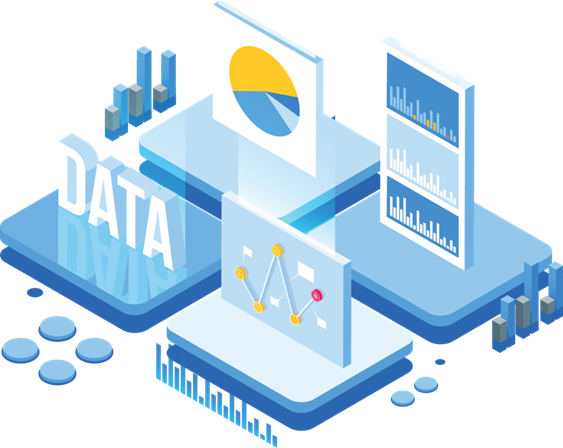
Recipe Recommendation by Consumers

Huan Nguyen

Data Mining

Dr. Ryan C. LaBrie, Ph.D.

March 11, 2025



# Objective

* Predicting customer opinion on the recipe based on their feedback.
* Customer feedback is significant for businesses that offer meal kits, or food bloggers offering recipes for viewers, etc.

# Process

* I use “Select Attributes” to only select “Stars” and “text” as they are customer’s score and feedback
* I use “Discretize” to create groups of customers that rate 4, 5 as good, 3 as neutral, and 0 and 1 as bad.
* A screenshot of a computer

  AI-generated content may be incorrect.A screenshot of a computer

  AI-generated content may be incorrect.Since I want to do supervised learning, I set “stars” as label
* Then using nominal to text so that I can run “process document from data”
* I use tokenize, transform case, filter stopwords, and stemming for preprocessing the data.
* Split data with 70% trained and 30% tested for the predicting model.
* Cross Validation to see which prediction model have higher prediction rate.
* I tried decision tree, naïve bayes, deep learning with split data and cross validation
* For the processing document from data, I choose TF-IDF as vector. I try TF, BTO and TO, but the prediction from TF-IDF is higher. TF-IDF works better for my data set.

# Algorithms

|  |  |  |
| --- | --- | --- |
|  | Split Data | Cross Validation |
| Decision Tree | 89.80% | 88.99% +- 0.59% |
| Deep Learning | Avg around 92.25% | 91.90% +- 2.79% |
| Naïve Bayes | 81.63% | 83.17% +- 3.16% |
| Gradient Boosted Tree | 89.29% | No Result |

* Deep Learning gives the best prediction rate whether doing split data or cross validation.

# Data Source

* The data is open for public use. The dataset originally includes 100 recipes with customer feedback for each one.

# Outcome

* I started the project by analyzing only 1 recipe. All customers’ feedback and key terms are different for each recipe.
* I am confident to use the same process to predict customers’ feedback on different recipes.

# 3 W’s

*What went well?*

Confident with the prediction rate, and the model can be applied for other recipes

TF-IDF improved the accuracy of the model

*What did not go well?*

A screenshot of a computer

AI-generated content may be incorrect. First, I want to do sentiment analysis to see whether customers have positive or negative feedback on the recipe. I saw that even though they rated “good” but the score was negative, so I looked into it. I noticed that the term “hate” might be the biggest issues because it gives such negative score. There are negative terms in the text, but some of them are not related to the recipe. I tried both sentiwordnet and vador, but same result. I understanded that it is due to lack of context awareness. For example, the text “I hate the texture, but I like the taste, so I use onion powder”. The model sees “hate” as a significant negative term, but as human, if we read the text, we will say that the customer like the recipe. I tried to find a way to filter this, but my computer crashed.

*What would I do differently?*

I should have spent more time on this project and tried out another tool. RapidMiner has its limits. I also want to try lemmatization, and I think it will give me better results. There was semantic analysis in RapidMiner, but I cannot used it.