

# Coffee, or tea?

A study on coffee and tea to  
chart the next consumer  
beverage strategy





The coffee market in Singapore is worth about

**\$2 billion in 2023\***

4th most consumed beverage across the world.



The tea market in Singapore is valued at

**~\$800mil in 2023\***

2nd most consumed beverage on Earth.

\*source: Statistica



You want to venture into a consumer beverage,  
**but which one to pick?**

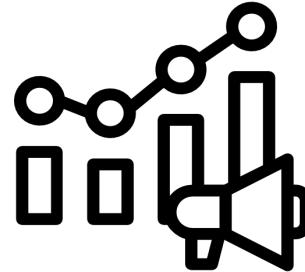
Once decided, what would be the right words for SEM (search engine marketing)?

# Project objectives



## Understand consumer behaviors

This includes their attitudes towards coffee and tea. What do they care about, and why?



## Chart marketing and sales strategies

Identify new business opportunities by capturing raw material and its supply chain



## Develop new products

Innovate new beverages and peripheral products according to customer preferences

# Methodology

1. Data Collection
2. Data Preprocessing
3. Exploratory Data Analysis
4. Model Selection
5. Model Evaluation
6. Summary and Recommendation



# Data Collection via Subreddits

The image shows two side-by-side screenshots of Reddit subreddits. On the left is the subreddit [r/Coffee](#), featuring a coffee cup icon and a blue header. The right is the subreddit [r/tea](#), featuring a teapot icon and a green header. Both subreddits have a 'Join' button at the top.

**r/Coffee (Left):**

- Pinned by Moderators:**
  - 26 [MOD] The Daily Question Thread (by u/menschmaschine5, 14 hours ago)
  - 10 [MOD] Show off your gear! - Battle-station Central (by u/menschmaschine5, 8 hours ago)
- Posts:**
  - Does RDT change anything beyond static reduction? (by u/CleariousHunt, 3 hours ago)
    - For context, I've recently added RDT into my process and obviously I'm noticing how much cleaner the grinding is, but I feel that the pourover cups. I've done since then have become.... Better? Like tasting notes and flavours from the coffee are becoming much more pronounced in all the right ways. I can't work out if it's a coincidence of me dialling in the grind at the same time or a placebo.
    - Has anyone experienced something similar or am I just going nuts?

**r/tea (Right):**

- About Community:** Tea! This subreddit is for discussion of beverages made from soaking camellia sinensis leaves (or twigs) in water, and, to a lesser extent, herbal infusions, yerba mate, and other tisanes.
- Posts:**
  - What's in your cup? Daily discussion, questions and stories - March 16, 2023 (Recurring)
  - Marketing Monday! - March 13, 2023 (Recurring)
  - This tea is going to get me to throw out all my shredded leaf English Breakfast (Photo)
- User Flair Preview:** Various-Fan7001

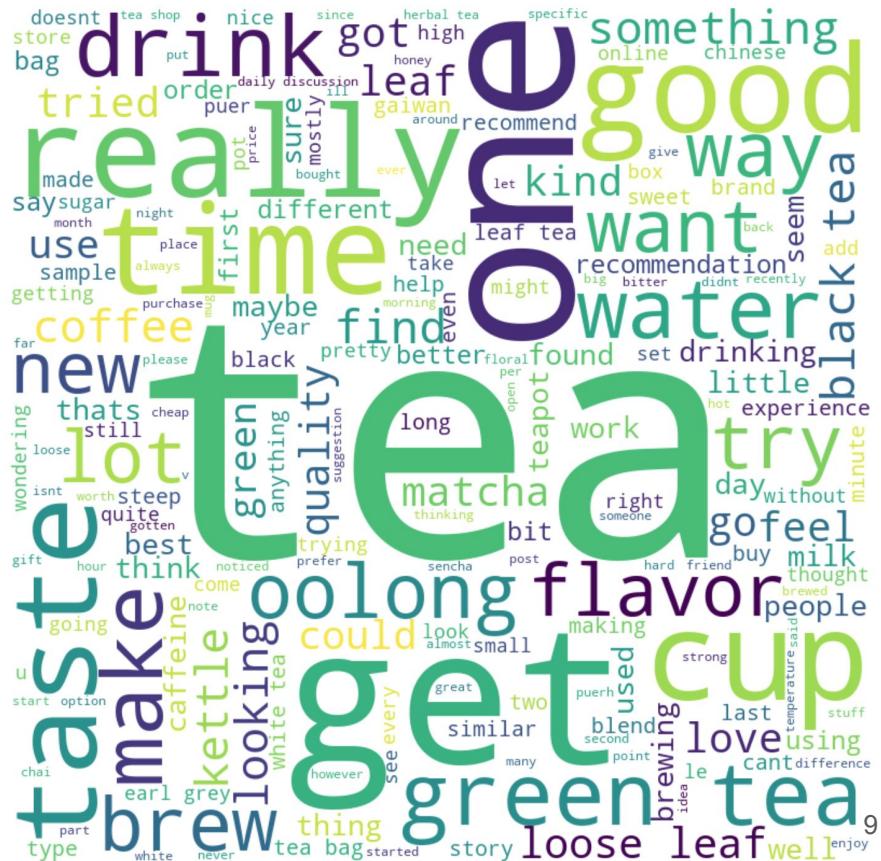
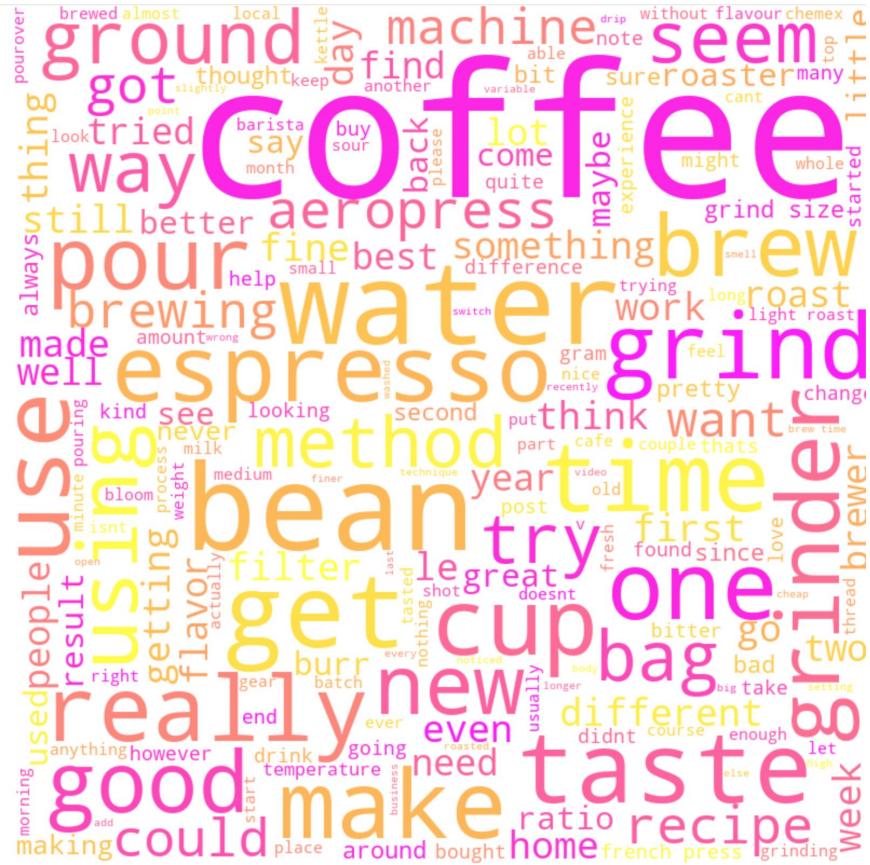
# Data Preprocessing

1. Remove punctuations, numbers, upper cases
2. Remove stop words and outliers (eg, non-English words)
3. Lemmatize
4. Tokenize
5. Vectorize with CountVectorizer and TF-IDF



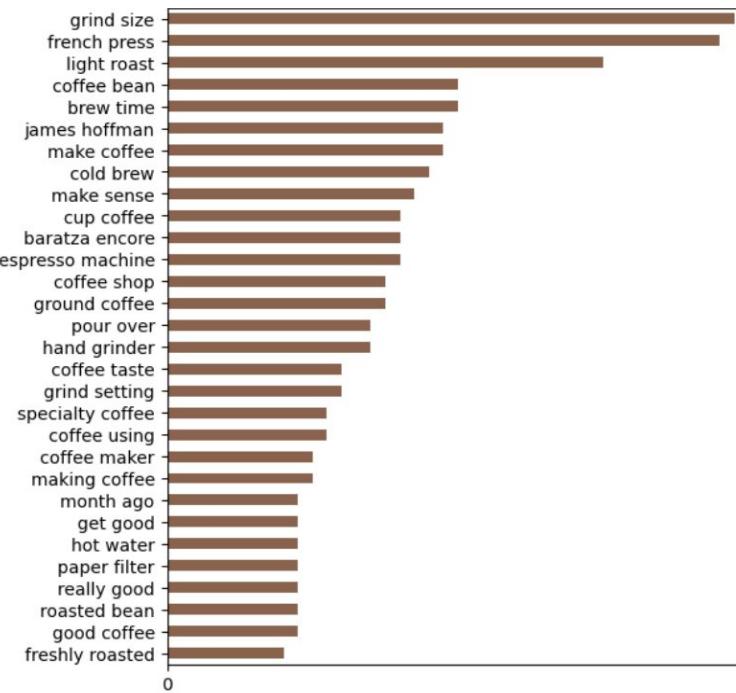
Since coffee has a higher market revenue, I will use Coffee = 1 (positive), Tea = 0 (negative)

# A quick glance of words among the documents



# For Coffee, consumers asks about :

Top 30 Frequently Bigram Words in Coffee Subreddit



- Techniques - brew time, grind size
- Styles of making coffee - cold brew, french press, pour over
- Tools - espresso machines, hand grinder
- James Hoffman - *who is he?*



# James Hoffmann

Barista



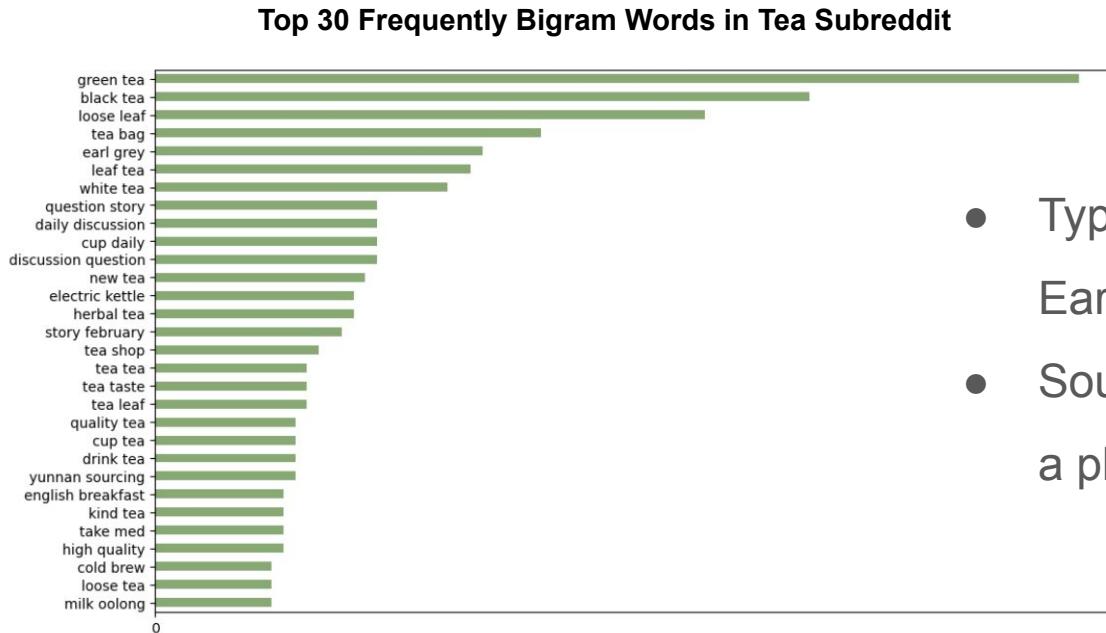
jameshoffmann.co.uk

James Alexander Hoffmann is an English barista, YouTuber, entrepreneur, coffee consultant, and author. Hoffmann first came to prominence after winning the World Barista Championship in 2007 and has since been credited as a pioneer of Britain's third-wave coffee movement. [Wikipedia](#)



Catch his Wired interview on [Youtube](#)

# Overall, Tea consumers asks about:



- Type of tea - green tea, black tea, Earl Grey, etc
- Source for raw material - either from a place (Yunnan), or a shop

# Model Selection

Start with seven classification techniques and perform cross-validations with tuned hyper-parameters using GridSearch

Stacking

Bagging

Boosting

Voting

Train multiple base model and use meta model to combine predictions

Train multiple instances of same base models on different subsets of the training data

Train multiple weak models sequentially, adjust weights to improve subsequent models

Combine predictions of multiple base models using majority votes

Naive Bayes, LogReg, SVM, Decision Tree

Random Forest

AdaBoost

K-Nearest Neighbor

# Model Evaluation Metrics

Precision

Recall

F1 Score

ROC-AUC score

Proportion of true coffee among **predicted** coffee posts

Proportion of true coffee among **actual** coffee posts

Harmonic mean between precision and recall

Trade-off between TPR/FPR, and its ability to distinguish both

Minimize false positive (classifying tea as coffee)

Minimize false negative (classifying coffee as tea)

Go for high F1 score: effective and correctly identify both posts

Go for high ROC-AUC score: evaluate models that can correctly classified more coffee posts

# Considerations on model selection and metrics

I would recommend a model that can  
**classify coffee better than tea.**

This is to avoid missed opportunities of learning insights about coffee consumers as it is more complex than tea consumers.

The model also has to be **flexible** enough to be fined tune for future



# Model Performance at a Glance

Category	Models	Test Score	Precision	Recall	F1 Score	ROC-AUC Score
Stacking	Naive Bayes	0.89	0.87	0.92	0.89	0.96
	LogReg	0.88	0.85	0.93	0.89	0.96
	Decision Tree	0.87	0.89	0.84	0.86	0.95
	SVM	0.89	0.88	0.90	0.89	0.94
Bagging	Random Forest	0.85	0.85	0.82	0.76	0.84
Boosting	AdaBoost	0.82	0.90	0.73	0.8	0.92
Lazy Learning	KNN	0.67	0.67	0.52	0.59	0.71

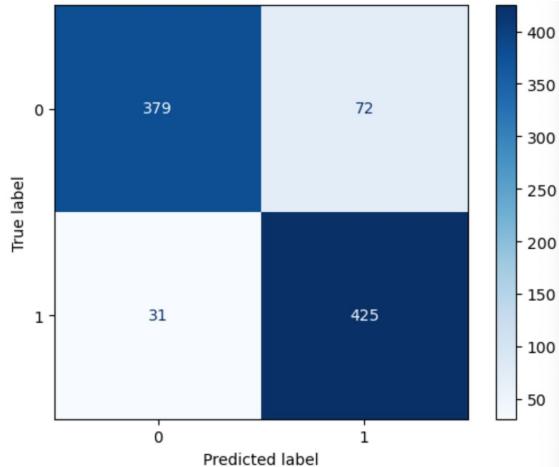
# Model Performance at a Glance

Category	Models	Test Score	Precision	Recall	F1 Score	ROC-AUC Score
Stacking	Naive Bayes	0.89	0.87	0.92	0.89	0.96
	LogReg	<b>0.88</b>	<b>0.85</b>	<b>0.93</b>	<b>0.89</b>	<b>0.96</b>
	Decision Tree	<b>0.87</b>	<b>0.89</b>	<b>0.84</b>	<b>0.86</b>	<b>0.95</b>
	SVM	<b>0.89</b>	<b>0.88</b>	<b>0.90</b>	<b>0.89</b>	<b>0.94</b>
Bagging	Random Forest	0.85	0.85	0.82	0.76	0.84
Boosting	AdaBoost	0.82	0.90	0.73	0.8	0.92
Lazy Learning	KNN	0.67	0.67	0.52	0.59	0.71

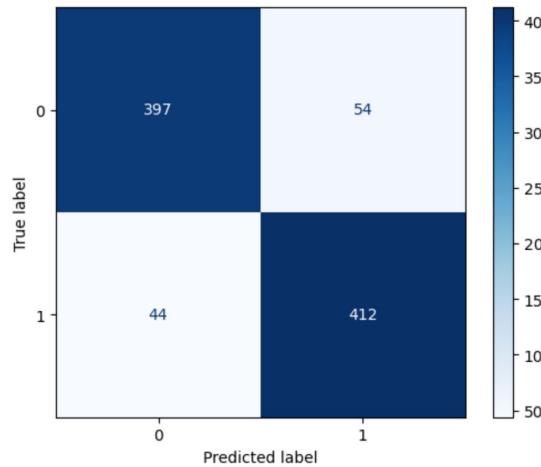
# Analyze chosen model with a Confusion Matrix

Since coffee has a higher market revenue, I assign **Coffee = 1, Tea = 0**

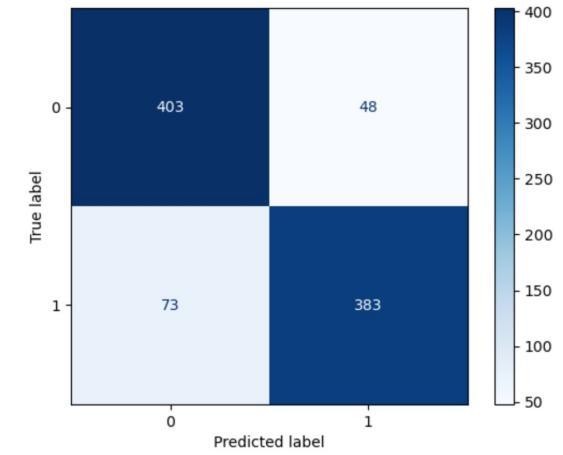
LogReg



SVM



Decision Tree



Efficient as a binary classification, may not perform with >2 categories

Can classify unseen data well, but computationally expensive

Capture complex non-linear relationships. Resilient against outliers

# Recommendation: Coffee

Coffee can be sold in a wide variety of manner. This adds more permutation in marketing and can reach more users with different preferences.

We can also create a unique brand out of the coffee styles.

Example: [Starbucks](#), indie coffee shops



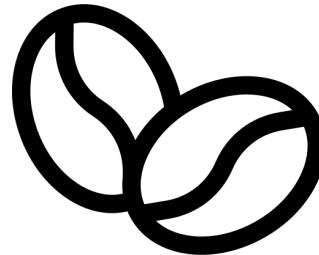
Also, there is a higher possibility of adding value to consumer experiences, by offering quality equipments and hiring good baristas

Example: [Chye Seng Huat](#) , [Nespresso](#)

# Further Enhancements



Include variety of data and capture the potential of new innovations - for example, Boba Tea



Embed words with with GloVe and Word2Vec besides tf-idf



Add **sentiment analysis** to understand consumers thoughts and feelings on both beverages

# Recommendation: Coffee



Fans queue for six hours to taste  
Mandopop star JJ Lin's Miracle Coffee



The pop-up in Singapore will run till Dec 29 and a flagship cafe is scheduled to open in the ArtScience Museum in 2023. ST PHOTOS: JAN LEE

# Coffee, or tea?

Answer: Coffee

