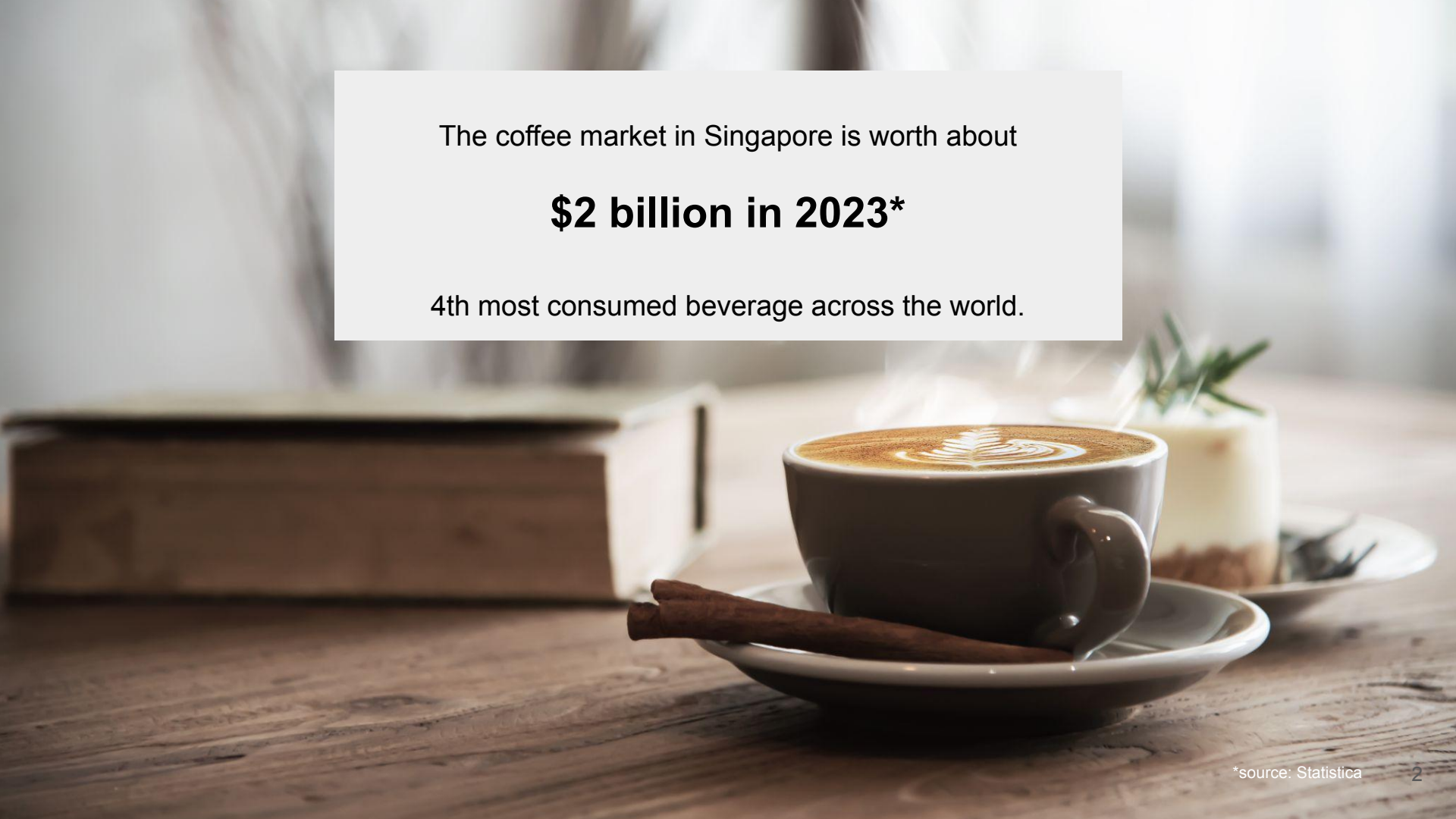
Coffee, or tea?

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.

*source: Statistica

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?

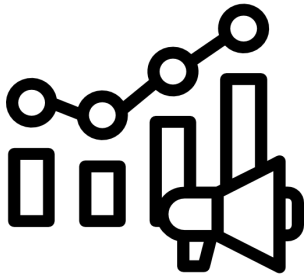


Chart marketing and sales strategies

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

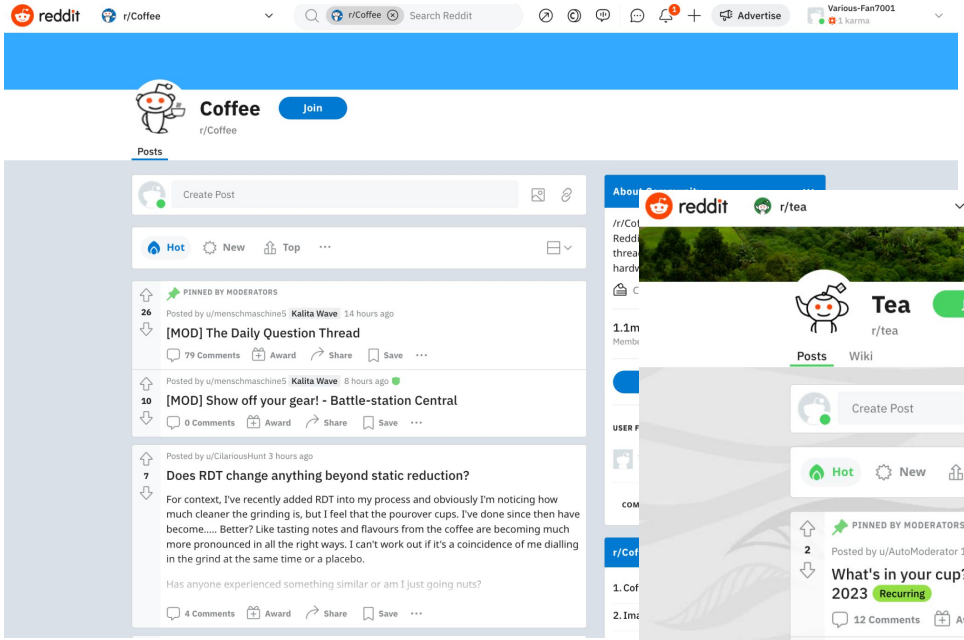
Develop new products

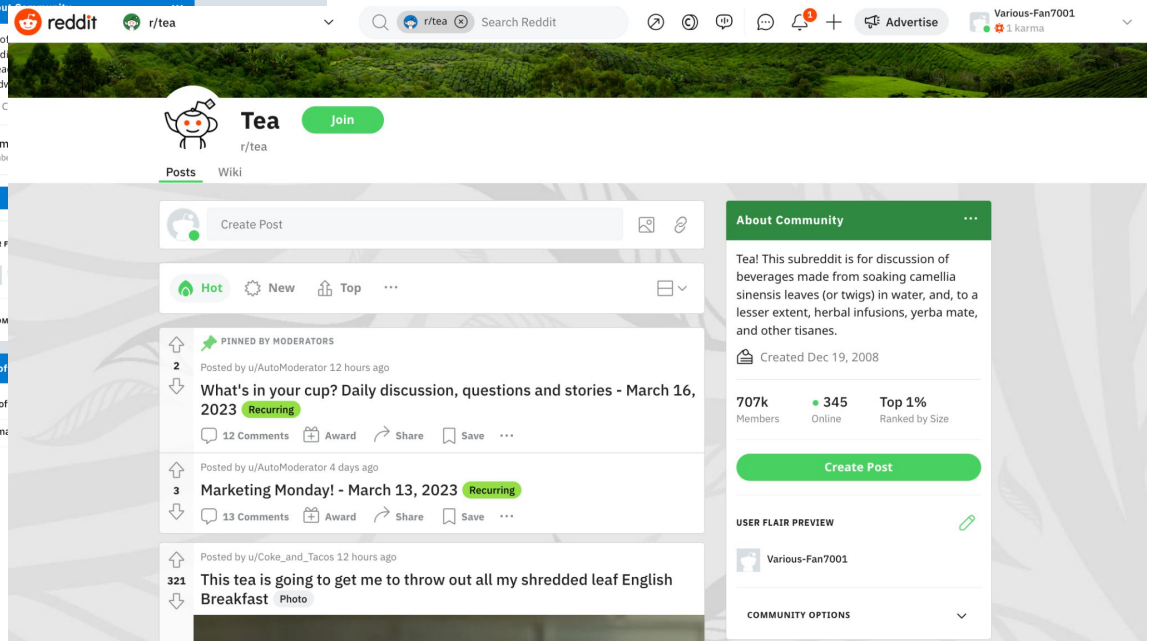
Innovate new beverages and peripheral products according to customer preferences

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



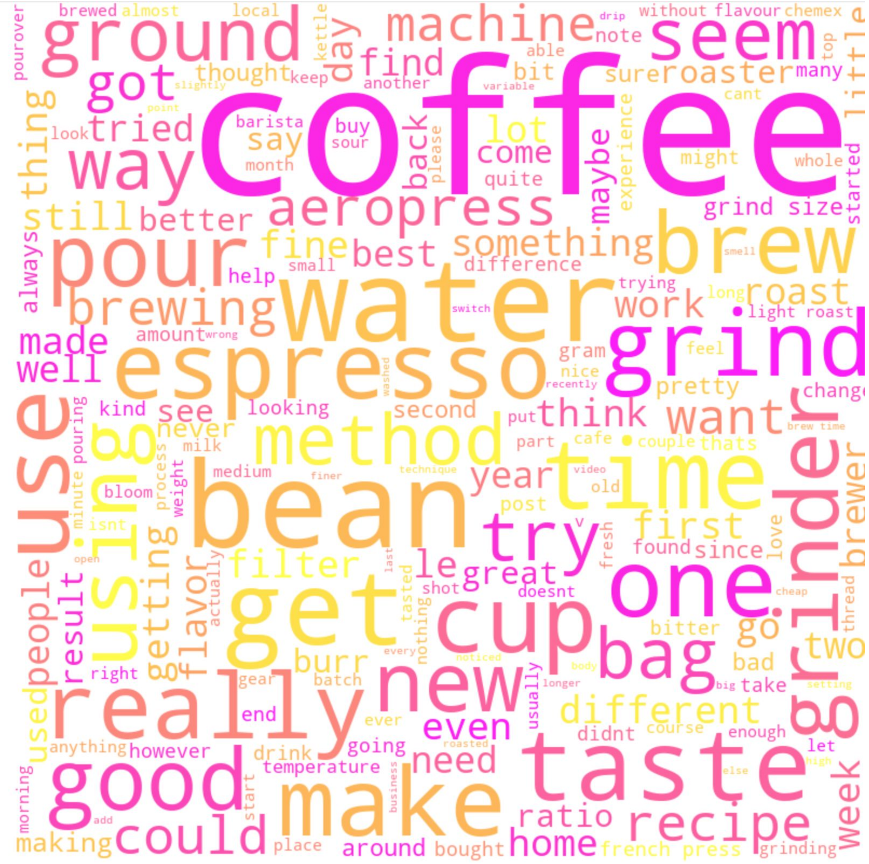


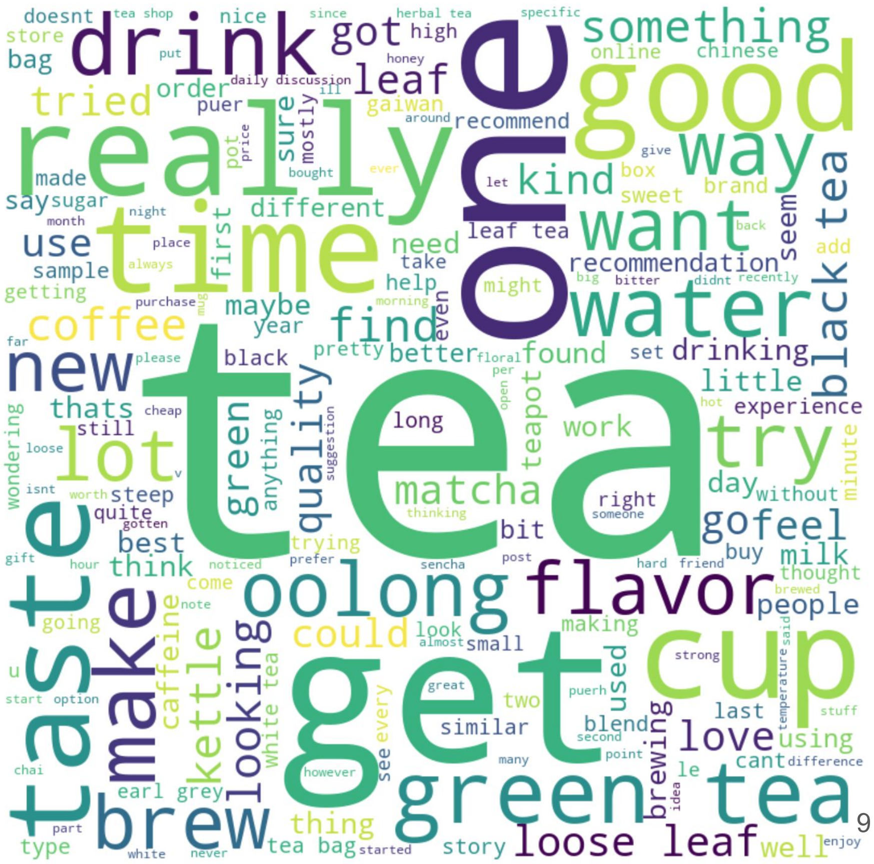
Data Preprocessing

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

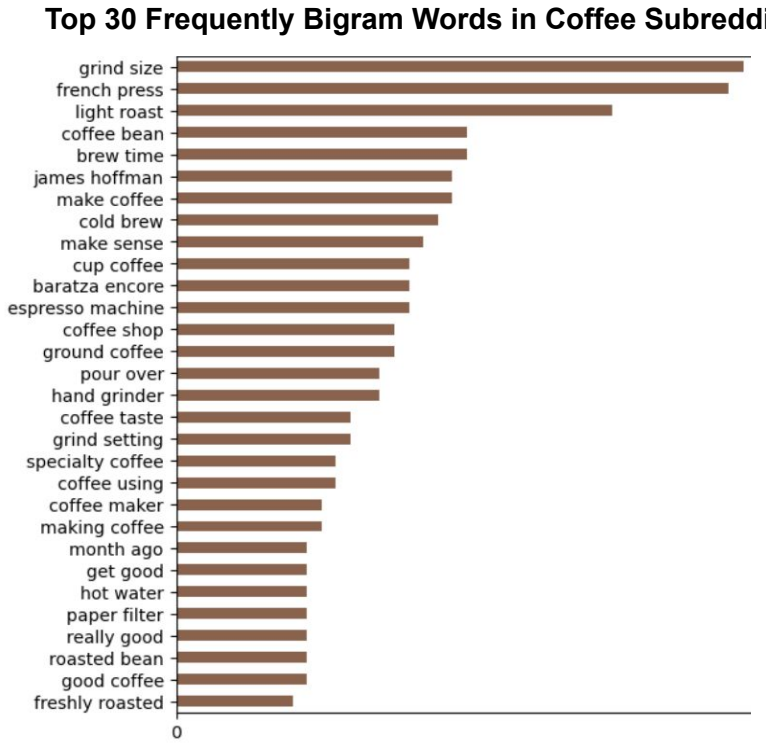
Since coffee has a higher market revenue, I will use Coffee = 1 (positive), Tea = 0 (negative)

A quick glance of words among the documents





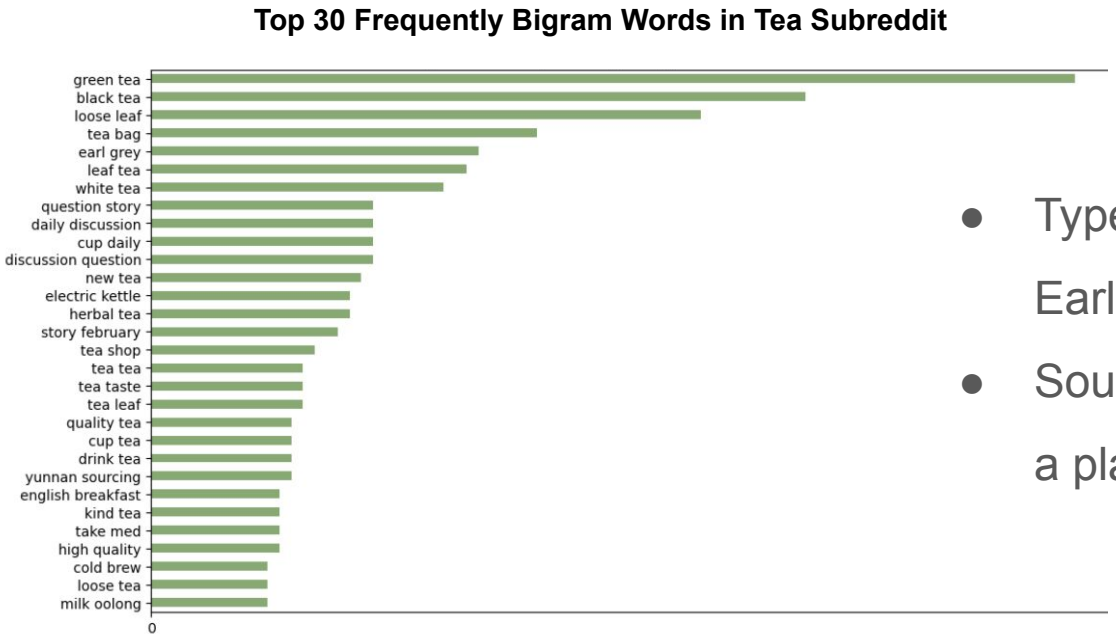
For Coffee, consumers asks about :



- 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?*

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

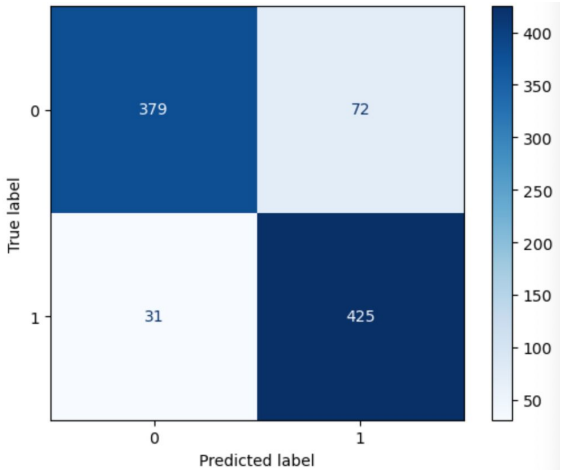
Model Performance at a Glance

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	**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

Analyze chosen model with a Confusion Matrix

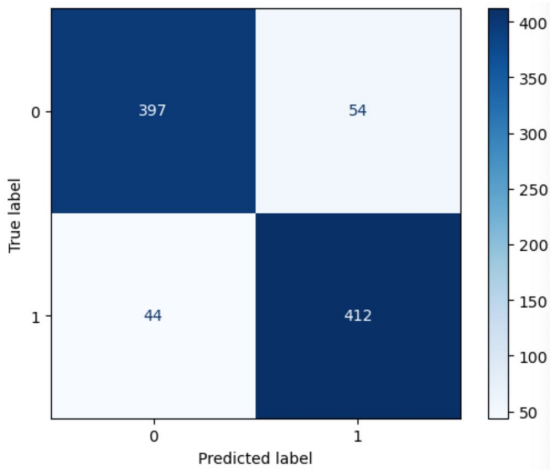
Since coffee has a higher market revenue, I assign **Coffee = 1**, **Tea = 0**

LogReg



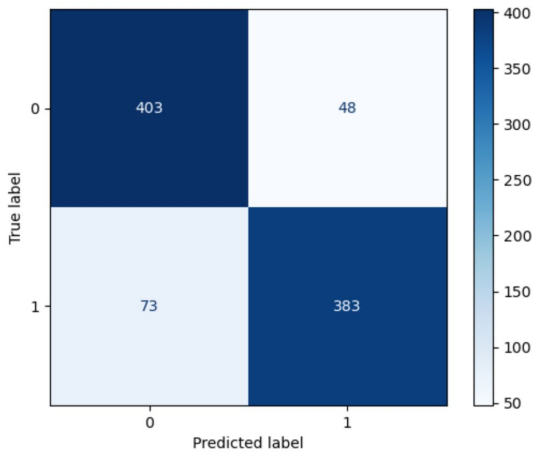
Efficient as a binary classification, may not perform with >2 categories

SVM



Can classify unseen data well, but computationally expensive

Decision Tree



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.

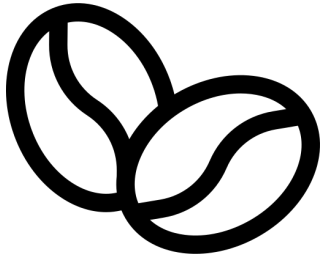
Example: Starbucks, indie coffee shops

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