Classifying Recipes into Cuisines

For Better Experiences and Authenticity

Michael Li

METIS®

Who benefits from knowing the cuisine of their foods?

Restaurants
- **Brand Image**
- Attracting Customers
  - Experiences
  - Authenticity

Food Delivery Companies and Meal Kit Companies
- Component of Recommendation System

Consumers
- Understand their cultural **Palates**
  - **Fridge**
  - **Lunch Spots**

GRUBHUB™

How to Classify Foods into Cuisines?

Using **Classification Models** to **predict cuisine** and identify **important ingredients**

Methodology

Gathering Data

Extracted Data from:
Yummly API
Kaggle Yummly

Data Format:
Cuisine, Ingredients

21 Cuisines
3000+ Ingredients
30,000 Rows

Kaggle, Pandas

Gathering Data

Extracted Data from:
Yummly API
Kaggle Yummly

Data Format:
Cuisine, Ingredients

21 Cuisines
3000+ Ingredients
30,000 Rows

Kaggle, Pandas

Data Cleaning Processing

Ingredients:
Removed **Numbers**
Removed **Punctuation**
Lowercase

NLP:
Vectorized words and weighted them against each other -- creates a sparse matrix of features

ReGex, NLTK

Gathering Data

Extracted Data from:
Yummly API
Kaggle Yummly

Data Format:
Cuisine, Ingredients

21 Cuisines
3000+ Ingredients
30,000 Rows

Kaggle, Pandas

Data Cleaning Processing

Ingredients:
Removed **Numbers**
Removed **Punctuation**
Lowercase

NLP:
Vectorized words and weighted them against each other -- creates a sparse matrix of features

ReGex, NLTK

Exploratory Data Analysis

Created plots and visualizations

Discovered **Class Imbalances**

Added weights to underrepresented Cuisine

Tableau and Matplotlib

Gathering Data	Data Cleaning Processing	Exploratory Data Analysis	Model Training and Validation
Extracted Data from: Yummly API Kaggle Yummly Data Format: Cuisine, Ingredients **21 Cuisines** **3000+ Ingredients** **30,000 Rows** Kaggle, Pandas	Ingredients: Removed **Numbers** Removed **Punctuation** **Lowercase** **NLP:** **Vectorized words** and weighted them against each other -- creates a sparse matrix of features ReGex, NLTK	Created plots and visualizations Discovered **Class Imbalances** **Added weights** to underrepresented Cuisine Tableau and Matplotlib	Split Data for Testing Created a basic KNN model as Base Model: **Accuracy 64%** Compared Base model With more complex models Adjusted Parameters and Class Weights Cross-Validated most promising models

Gathering Data	Data Cleaning Processing	Exploratory Data Analysis	Model Training and Validation	Testing
Extracted Data from: Yummly API Kaggle Yummly Data Format: Cuisine, Ingredients **21 Cuisines** **3000+ Ingredients** **30,000 Rows** Kaggle, Pandas	Ingredients: Removed **Numbers** Removed **Punctuation** **Lowercase** **NLP:** **Vectorized words** and weighted them against each other -- creates a sparse matrix of features ReGex, NLTK	Created plots and visualizations Discovered **Class Imbalances** **Added weights** to underrepresented Cuisine Tableau and Matplotlib	Split Data for Testing Created a basic KNN model as Base Model: **Accuracy 64%** Compared Base model With more complex models Adjusted Parameters and Class Weights Cross-Validated most promising models	Arrived at best model: **OVA Logistic Regression with class weights**


What did I Cook Up?

Best Model

OVA Logistic Regression with weighted features

Accuracy: 78%

Weighted Recall: 78%
Weighted Precision: 78%
F1: 78%

Weighted Features
Dramatically Increased Recall

Multiclass: One vs. All

Coefficients by Cuisine:

- Culturally Significant Ingredients

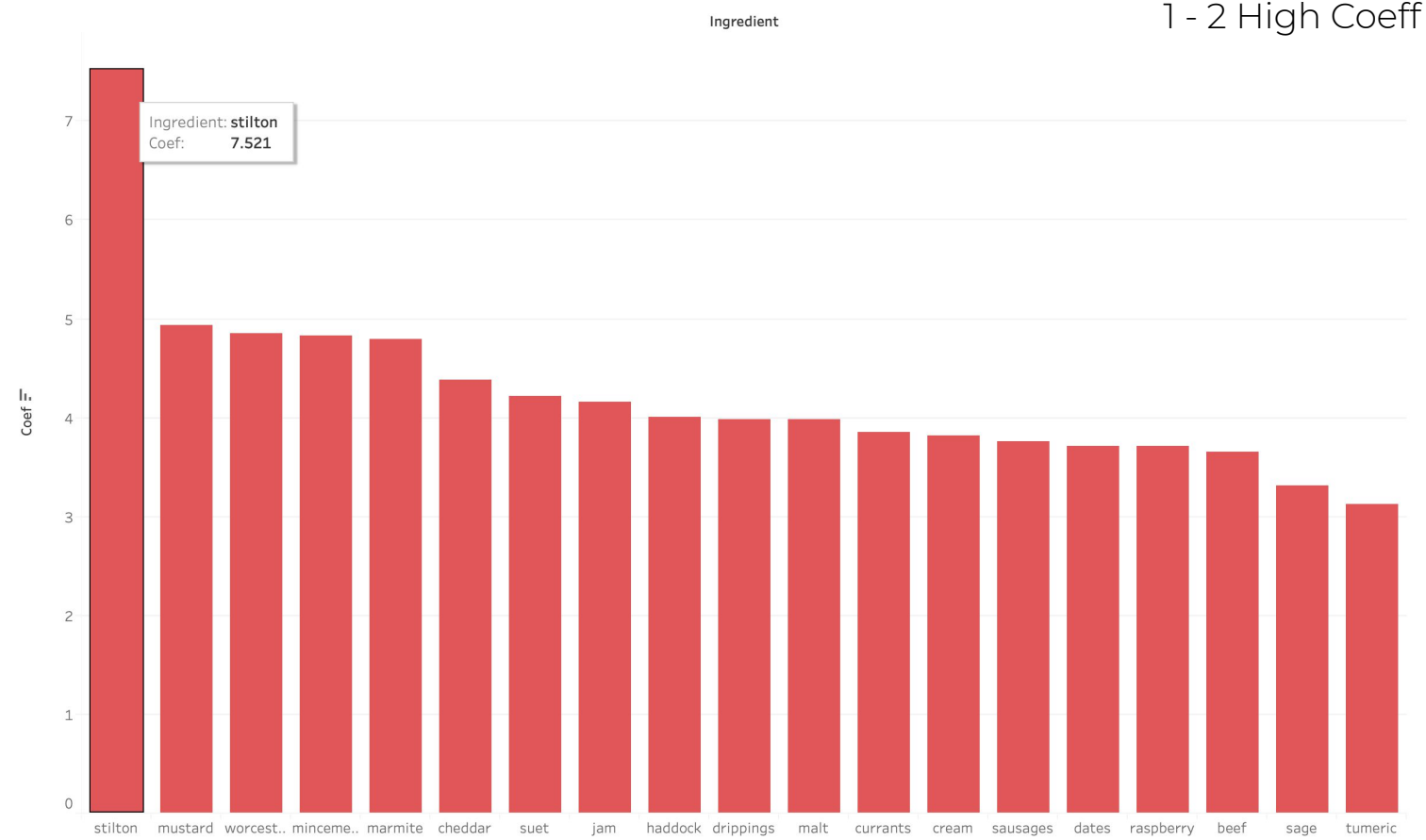
Indian Food 87% Precision

Italian Food 86% Precision

Mexican Food 92% Precision

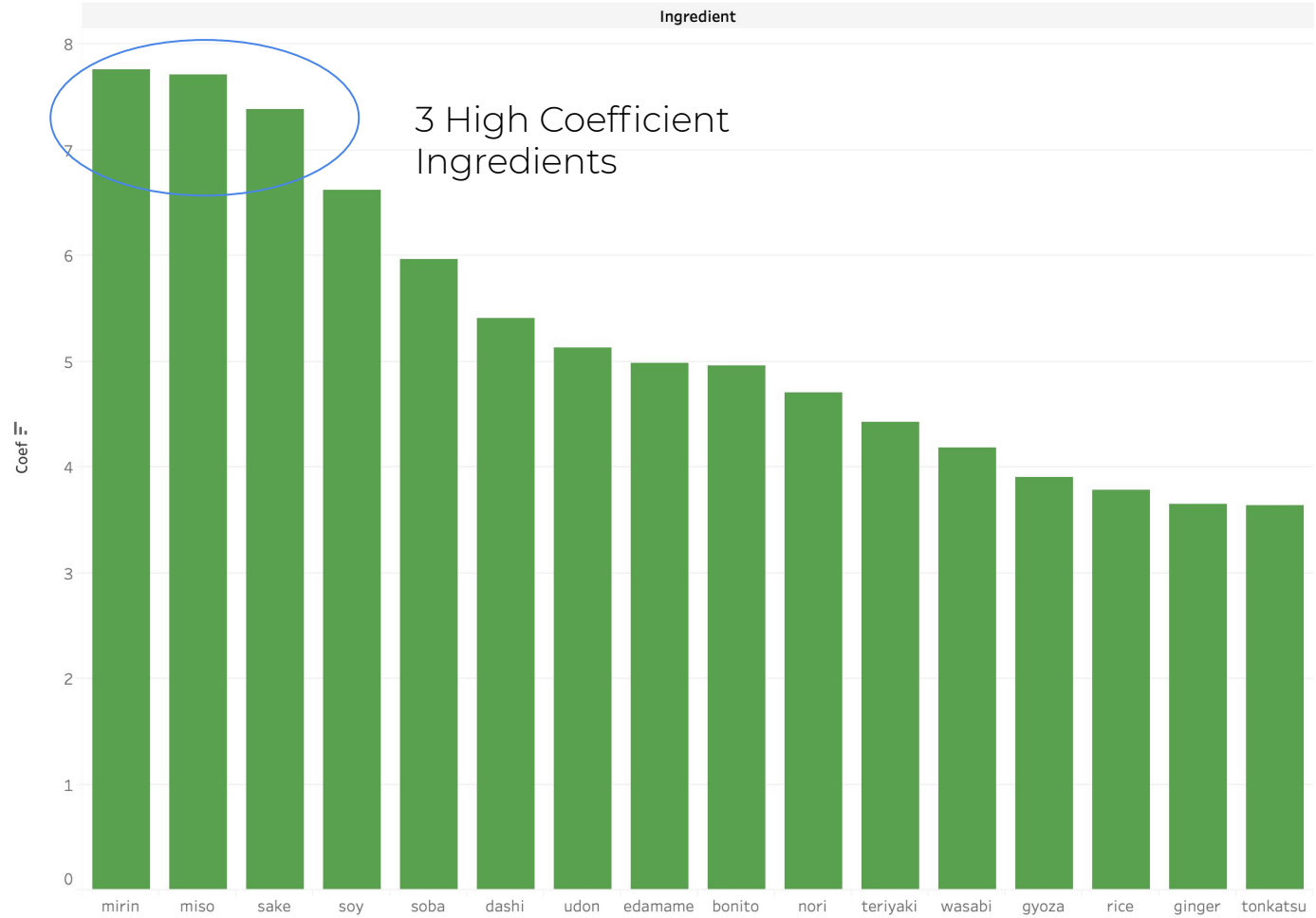
Chinese Food 74% Precision

Significant British Ingredients



1 - 2 High Coefficient Ingredients

Significant Japanese Ingredients

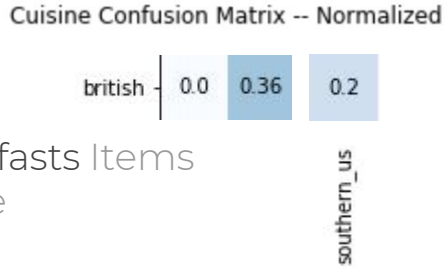


3 High Coefficient Ingredients

Original language names, more high coefficient ingredients

Insights Drawn from Misclassifications

- Many misclassifications with earlier models
- Southern United States
  - 20% Misclassified: British and Irish
  - Overlap of Ingredients: Potatoes, Fried Breakfasts Items
  - Southern Food Influenced by English Cuisine
    - Since Colonial Times
- Could be used for cross cultural recommendations
  - Hawaiian, Creole, Mediterranean Cuisines

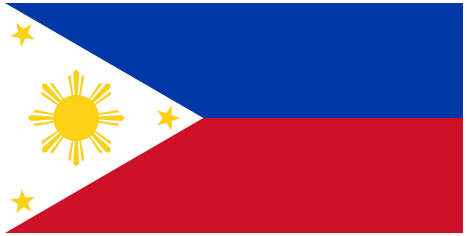


Testing my Model: Generating Top 3 Cuisines for a variety of Dishes

Chipotle: Chipotle Chicken Burrito Bowl

Chipotle: Chipotle Chicken Burrito Bowl

1% Spanish



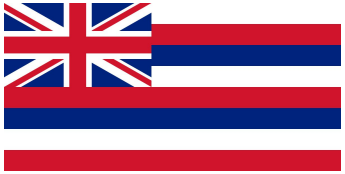
7% Filipino

90% Mexican

Korilla Street Cart: Korean Pork Burritos

Korilla Street Cart: Korean Pork Burritos

18% Mexican



3% Hawaiian

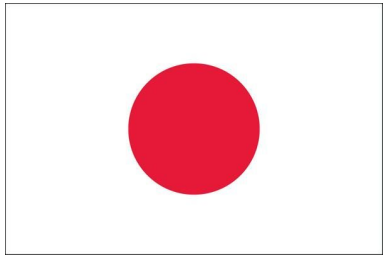
79% Korean

How does my Restaurant Fare?

Wiki Poke -- Traditional Bowl

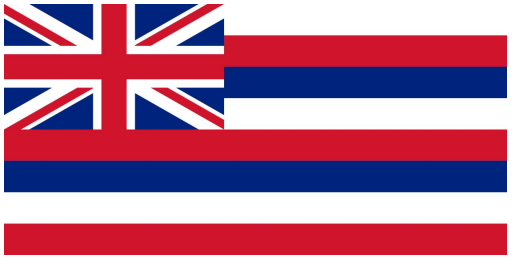
How does my Restaurant Fare?

Wiki Poke -- Traditional Bowl



17% Japan

2% Chinese



74% Hawaiian

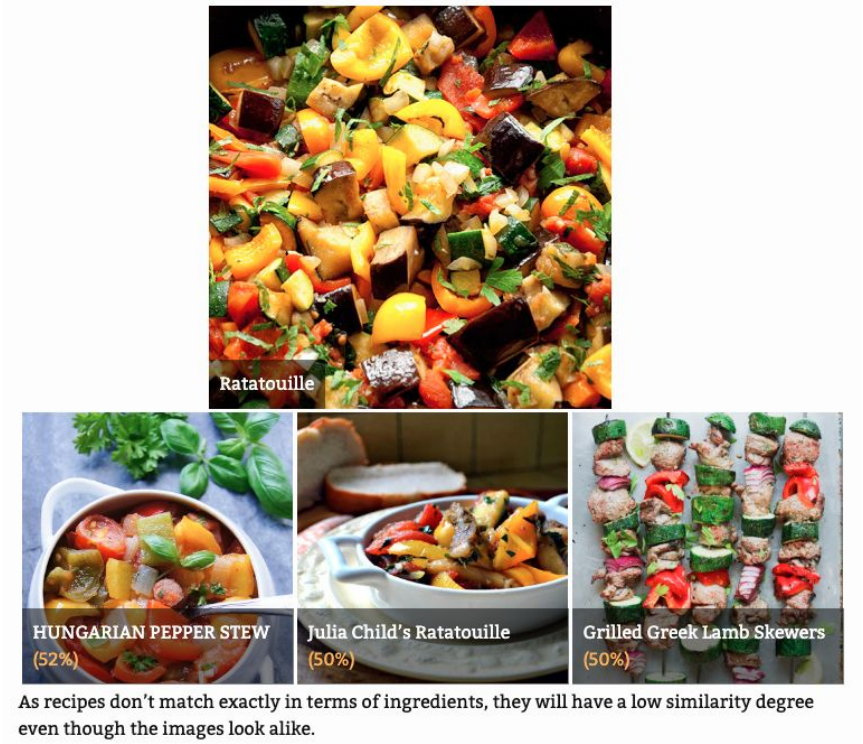
Takeaways

- **Ingredients** can be used to **classify cuisine** fairly **accurately**
- High **coefficients** of Log Model indicate **significance** of **ingredients** for a particular **cuisine**
- **Misclassifications** could be due to **cultural similarities** and **history**

Future Work:

Recommender System and App

- Recommend Dishes based on food choices
  - Cross Culturally
- Recommend Restaurants
  - Authenticity
  - Consumer Preference



As recipes don't match exactly in terms of ingredients, they will have a low similarity degree even though the images look alike.

Questions?