What's Cooking?

Cuisine Classification Model Based On Ingredients

Presenter: Haziel Andrade Ayala, Valentin Baltazar, An Nguyen

Data & preprocessing

1. Convert .json file (40k of rows) to dataframe
2. Data Preprocessing
   a. No missing data
   b. Imbalanced data: cuisine types
3. EDA & Visualization
4. Feature Engineering
   a. ‘Ingredients_count’ for EDA
   b. ‘Ingredients_strings’ for classification models
5. Train_Test_Split
6. Model Pipelines with CountVectorizer


```
{
"id": 24717,
"cuisine": "indian",
"ingredients": [
    "tumeric",
    "vegetable stock",
    "tomatoes",
    "garam masala",
    "naan",
    "red lentils",
    "red chili peppers",
    "onions",
    "spinach",
    "sweet potatoes"
]
},
```

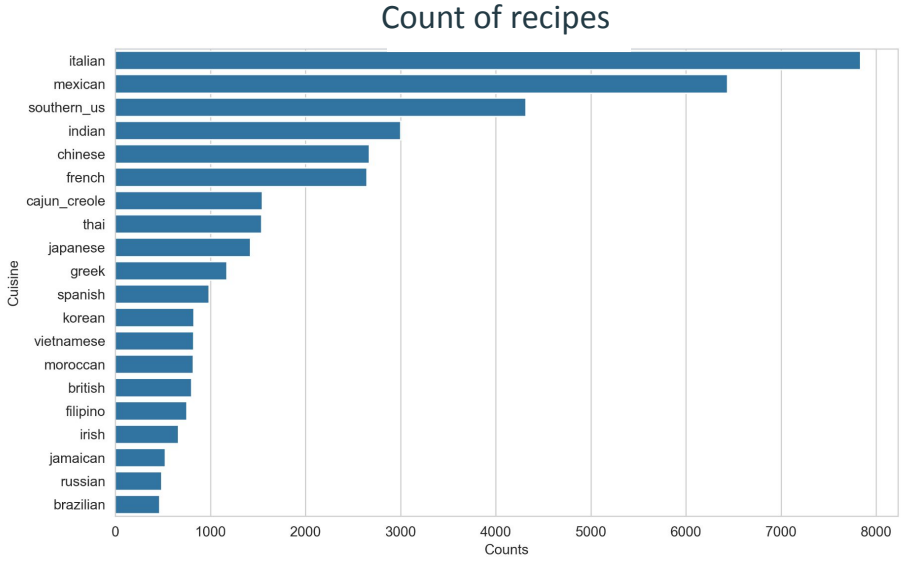

	id	cuisine	ingredients
0	10259	greek	[romaine lettuce, black olives, grape tomatoes...
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, g...
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, g...
3	22213	indian	[water, vegetable oil, wheat, salt]
4	13162	indian	[black pepper, shallots, cornflour, cayenne pe...

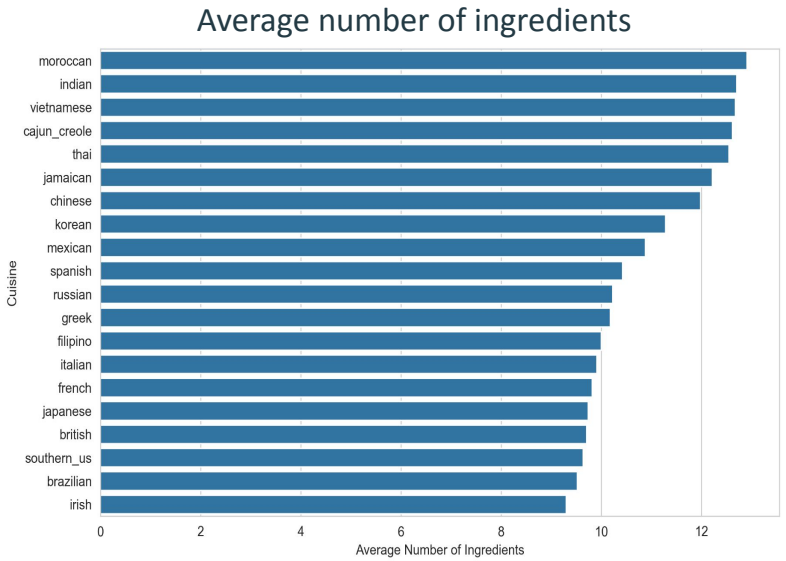

```
['romaine lettuce',
 'black olives',
 'grape tomatoes',
 'garlic',
 'pepper',
 'purple onion',
 'seasoning',
 'garbanzo beans',
 'feta cheese crumbles']
```



```
'romaine lettuce black olives grape tomatoes garlic pepper purple onion seasoning garbanzo beans feta cheese crumbles'
```

EDA & Visualization





Classification Models Comparison

Classifier Model (with CountVectorizer)	**Test Accuracy Score**		
	unigram	**unigram and bigram**	**bigrams**
Logistic Regression	**0.78**	**0.77**	**0.74**
K-Nearest Neighbors	0.63	0.51	0.26
Random Forest Classifier	0.75	0.73	0.68
AdaBoost Classifier	0.55	0.55	0.51
Support Vector Classification (SVC)	0.77	0.77	0.70

Model Results & Improvement

The best classification model is Logistic Regression with CountVectorizer default settings (unigram):

- Train accuracy score: 0.86
- Test accuracy score: 0.78

Model Improvements:

- GridsearchCV to optimize model parameters
- Feature engineering
    - countvectorizer vs. tfidfvectorizer, unigrams vs bigrams vs…
- Ingredients into individual features
    - Total ingredients, avg ingredients

Conclusion

- LogisticRegression can correctly classify with ~80% accuracy
- Much of the False predictions have to do with similar recipe ingredients
- Recipe *instructions* would be a new way to improve the model