Future Food Customer Needs

Lecture	Data Science summer semester 2020
Participants	Rahman, Safie Ur (2572863), s8sarehm@stud.uni-saarland.de Rizvi ,Syed Taqi Abbas (2577651), s8syrizv@stud.uni-saarland.de Munir, Talha (7003008), mumu00001@stud.uni-saarland.de
Submission date	07-17-20
Chair	Univ.-Prof. Dr.-Ing. Wolfgang Maaß Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Executive Summary

The following report presents an intuitive service to predict the food trends in the coming generation for the sake of portfolio development. The service makes use of the Kaggle Dataset 'Food Coded' and makes two types of interpretations. One is to predict the type of Food e.g. Asian, Italian etc. while also calculating the features that contribute to the prediction as global interpretation. This is an important insight as to let the food producer know which areas to focus for a particular food category. Another is the clustering and creation of the different comfort foods. This tells in specific, the comfort foods preferred by the youth e.g. ice cream, chicken etc.

Local interpretations are also presented which can be used by the food producer to know the changing trends if the model predicts wrong and the rebel features contributing.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Contents

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

List of Abbreviation

SHAP: SHapley Additive Explanations

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

List of Figures

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Introduction

The following report presents the results from a data driven service to the task of finding the food trends in the next generation which can be utilized by the food producers to develop/update their product portfolio. The Food and Beverage industry as reported by Statistica[1] is projected to reach a whopping US$224,815m in 2020 and is expected to show an annual growth rate of 10.6%. Figure 1: Food Revenue through the Years. (Figure 1: Food Revenue through the Years)

Such a massive impact also invites detailed research to predict the future needs and every now and then, research is carried out to fore-see the future trends in general as reported by Fraunhofer ISI[2] or specific to an age bracket as for 13-36 year olds as reported by YPulse[3] among several other researches. It is important to note that the researches that we encountered during our literature survey in this short span of time just laid out the trends that could happen without explaining the reasons or factors contributing to that change. In this study, we would give recommendations considering these underlying causes based on the given data that may establish confidence in the results returned by other researches as mentioned before. We will overview the Data Set in Section 2, Procedure and Analysis in Section 3 and finally present the Results and discuss them in Section 4.

1 Data Set

The two data sets used for the project were available on Kaggle[4]. One[5] lists the food choices and preferences of students of Mercyhurst University and spans over 126 uncleaned and raw responses in 61 columns from students on food choices, nutrition, preferences, childhood favorites, and other miscellaneous information. The second dataset[6] is more generalized and records 288 unique responses in 8 columns from participants from different countries and demographics. There were three types of columns in the first data set. Continuous numeric columns like GPA, encoded categorical columns like Gender and free text columns like comfort food. The categorical columns were already encoded by the data provider.

---

[1] "Food & Beverages - worldwide | Statista Market Forecast." https://www.statista.com/outlook/253/100/food-beverages/worldwide. Accessed 21 Jun. 2020.

[2] "50 trends influencing Europe's food sector by 2035." https://www.isi.fraunhofer.de/content/dam/isi/dokumente/ccv/2019/50-trends-influencing-Europes-food-sector.pdf. Accessed 21 Jun. 2020.

[3] "The 14 Biggest Food Trends Gen Z & Millennials Are ... - YPulse." 1 May. 2019, https://www.ypulse.com/article/2019/05/01/the-14-biggest-food-trends-gen-z-millennial-are-interested-in/. Accessed 21 Jun. 2020.

[4] "Kaggle." https://www.kaggle.com/. Accessed 21 Jun. 2020.

[5] "Food choices | Kaggle." 23 Apr. 2017, https://www.kaggle.com/borapajo/food-choices. Accessed 21 Jun. 2020.

[6] "Food Preferences | Kaggle." 18 Jul. 2019, https://www.kaggle.com/vijayashreer/food-preferences. Accessed 21 Jun. 2020.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

2 Procedure and Analysis

2.1 Individual Approach

The first data set, 'Food choices' was the main focus of the project as it provided more clarity and data points for the task of determining product portfolio for a food producer. Our main tasks are to find a predictive model to maximize the utility of the food producer and find a suitable cluster for an individual to focus on their comfort foods for customized products matching similar customer needs. So before modeling the full data set, we consider one individual row to explain our goals and methods for the task.

An individual data point consists of 61 rows, we will explain some of the salient columns and ignore the others. An example to understand a row is as follows:

A female student with 3.654 GPA. Her comfort foods include chocolate, chips and ice cream. The reason for comfort food is mostly stress, boredom or anger. It is encoded by the data provider in column; comfort_food_reason_coded in which the first reason is used as the main category i.e. stress. She cooks whenever she can, but not very often. She prefers American cuisine and thinks her current diet is unhealthy. She has a part time job and eats out once or twice every week where her favorite cuisine is Italian or French. She exercises every day and her ideal diet is small portions, five or six times in a day, of all important food groups.

2.1.1 Data Explorations

Histograms were used to measure the skewness of the data. It shows where the bulk of the values lie in each column. Scatter plots and correlation matrices were produced to check the different relationships between the columns. For open text columns, word clouds were made using the TFIDF weights or frequency counts in case of comma separated values. Details are added in the code description section.

2.1.2 Data Preparation

All numeric columns were imputed using a mean strategy. For the free text columns, SNLP techniques such as bag of words and tfidf were used to extract useful features for clustering. For tfidf calculations, 'TfidfVectorizer' from the sklearn library was used. Standardization is performed in order to smooth the data. It is also needed because for the prediction algorithm, we need to check the relative importance of each variable with the output variable. For the predictive model, some of the rows were dropped as they belonged to categories with less than 5 rows. .Stratification was used to split data into train set, validation set and test set. Details are added in code description.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

2.1.3 Data Modeling

Many different techniques such as Chi-Square test, correlation matrix, random forest and their combination are applied to select the appropriate features. After feature selection, multiple models were tested to check the performance like Multinomial regression, One Vs All Logistic regression, gradient and Ada Boosting models. Ensemble learning technique was applied on top of these models with best parameters through Grid Search Cross Validation. K Fold Stratified Cross Validation is used for establishing confidence in the results returned.

2.1.4 Clustering

The main goal is to divide the data set into suitable clusters for which the food producers can target customized comfort foods mainly. As the comfort food column is open text, we initially have to cluster it for an usable feature in overall clustering. Comfort foods are mainly given as comma separated items. We use text cleaning, and simple comma tokenization and then the TFIDF matrix is calculated with 1-grams as the values are comma separated and then used as the features for K-Means clustering algorithm. The elbow method is used to determine the suitable amount of clusters i.e. 20 in this case. We label each row with its respective cluster, which results in a categorical data column namely comfort_food_coded. Again we use the elbow method to determine the suitable clusters, which are 14 in this case.

3 Results and Discussion

3.1 Key Features of the Predictive model

Prediction algorithm has been implemented in order to find the most important factors that affect the likeness of each food. The table below describes the top 5 features for each food and how one unit change in these features affects the probability of the likeness of food for the customers.

Italian	Asian	American	Spanish
Income (+0.4)	Tortilla Calories (-0.3)	Coffee (-0.34)	Favorite food (-0.23)
Cuisine (+0.3)	Ethnic Food (+0.26)	Ethnic Food (-0.33)	Comfort food (+0.28)
Tortilla Cuisine (+0.29)	Exercise (+0.26)	Thai Food (-0.25)	On_off_campus (+0.27)
On_off_campus (-0.22)	Comfort-food (-0.19)	Income (-0.21)	Waffle Calories (+0.27)
Exercise (-0.19)	Waffle Calories (-0.18)	Favorite Food (+0.21)	Coffee (+0.23)

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

It is observed from the table that every food has different features that would affect their appeal. The importance of these weights explains how much increase in one unit of any feature would increase the likelihood of the particular food. (Jupyter Cell No. 52)
For example, from this table we can predict that if an income of a student is increased by one bin according to our binning of income data, the probability that student food like Italian food is increased by 40% where probability of American Food is decreased by 21%. Similarly if a student has a better exercise routine, the likelihood of a student choosing Asian food would increase by 26%.

If the data of the customers is already available, food producers could start producing foods by looking at these factors. For example, if the food producer is operating in the particular vicinity and he has some representative data of the students residing there, then based on these factors, he could decide whether to produce expensive Italian food or cheap American food etc.

3.2 Key Features of the Clustering

Simple KNN clustering among a small set of features also provided key insights and trends in the data. (Jupyter Cell No. 44-45). It is observed that Students with low income go for eating out less and tend to cook more. Similarly when eating changes are better, people tend to pay more. Also when parents cook more in the home, students go out less but they tend to pay more when they do. These insights will help food producers to prepare food for particular groups such as they could prepare expensive healthy food as there is a particular group of students who are getting health conscious and paying more while eating. Similarly to cater to a group of students who tend to eat more , food producers could produce less expensive food because these students usually eat out more and pay less on average. Also some home-cooked food could also be made since there exists a cluster of students who eat out more due to their parents cooking less in home so home cooked food could be appealing to them.

After clustering the comfort_food choices, it can be seen that the clusters are quite homogeneous hence the large number of clusters. This is due to the small amount of data and similar choice of the students. The histogram of mean values of each cluster (Refer to the cluster histogram picture here) shows important patterns here. For example, health-conscious female students which are heavily involved in sports have a particular liking of ice cream, flavored yogurt, candy bars etc. Whereas male students with similar quantities prefer pizza, dorritos and burgers as their preferred comfort food. The producers can target the female students with healthy chocolate flavored ice-cream with high vitamins or proteins nutrients but not the male students.

3.3 Outlook

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

This prediction algorithm is also used to track the trends in the data. So if the accuracy of the model drops significantly with the passage of time, it could be inferred that the behavior of the customers have been changed and we need to update our model. (Figure and some explanation about the graph)

Predictive algorithm goes hand in hand with the clustering we have done above as Clustering will be used to understand the different cluster of the students and find distinctive behaviors using multiple attributes within each cluster while predictive algorithm will provide information about each attribute of the student and how one attribute affect the student preference of food.

4 Future Work

4.1 Trend Tracking

We could also design the **self-learning pipeline** which could automatically check whether the trends have been changed based on some evaluation metric like accuracy. If there is change in trend, we could initialize the self-learning pipeline which would retrain the model as well as do the clustering again in order to give the updated behavior of our customer

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Appendix

Code Description

Data Exploration

Histograms were used to measure the skewness of the data. It shows where the bulk of the values lie in each column. For example, in the 'Food choices' data set, 60% of the rows are of Female and the other are Male. The most frequent value in comfort_food_reason is 2 i.e Boredom. The most frequent value in current_diet is also 2 i.e cheap/unhealthy/random/too much. For example, this shows that the individuals in this data set consider their diet unhealthy and would perhaps be open to healthier comfort foods. (Figure 2: Histogram)

Scatter plots and correlation matrices were produced to check the different relationships between the columns. For example, for females on average ideal_diet means less sugar however the most frequent answer is adding vegetable/ fruits etc. (Jupyter Cell.9)

For open text columns, word clouds were made using the TFIDF weights or frequency counts in case of comma separated values. (Py File, GetWordFrequency())

Data Cleaning & Preparation

All numeric columns were imputed using a mean strategy using the 'SimpleImputer' class from the sklearn library. For the free text columns, SNLP techniques such as bag of words and tfidf were used to extract useful features for clustering. For tfidf calculations, 'TfidfVectorizer' from the sklearn library was used.
In the next step, standardization is performed in order to smooth the data. It is also needed because for the prediction algorithm, we need to check the relative importance of each variable with the output variable which requires all the data columns standardized and normalized. This was done using the 'MinMaxScalar' from sklearn library, and as most of the data categorical in nature, it was done without any other issues. For the predictive model, some of the rows were dropped as they belonged to categories with less than 5 rows. It helps in reducing the overfitting of the model. (Jupyter Cell. 2)

Data Splitting

In order to develop the appropriate model according to a real world scenario and avoid the risk of over-fitting, the data was split into train set, validation set and test set. (Jupyter Cell. 42)
Stratification was used in splitting in order to avoid risk of clustering all rows of one type of pattern/class in one set and other types of rows in the second set. It also provides representative sampling of all the classes in every set. (Jupyter Cell. 42)

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Feature Selection

Many different techniques were used for feature selections, they are as follows:

1. As most of the columns were categorical in nature, Chi-Square test was performed to check the relative importance with the target variable. (Jupyter Cell. 48)
2. Categorical data is ordinal in nature so correlation was also used to determine the pattern which gave good insights for the project (Jupyter Cell. 47)
3. Random Forest was used for the feature selection as well because it performs well for limited dataset. (Jupyter Cell. 43)
4. Features appearing both in Chi-Square and Random Forest were also used to perform the modeling part. (Jupyter Cell. 49)

Features selected in all these techniques are shown in the jupyter notebook on the relevant cell.

Data Modeling, Hyper parameter tuning & optimization

The training set as gathered from the previous steps was subjected to Linear Classification (Multinomial and OneVsAll), Gradient and Ada Boosting models in a Grid Search with multiple attributes in the parameter grid to select the hyper parameters that best fit the data respectively. (Jupyter Cells. 51, 53-54)

The best parameters sought out were retained and used on the test set to get accuracies. Furthermore, the instances of the above models with best parameters found were passed to 'Ensemble Voting Classifier' which did a ‘soft’ classification to get the final model for the data. (Jupyter Cell. 55)

The resultant model was then subjected to K Fold Stratified Cross Validation. This took the combined set (training and testing) in k-folds and cross-validated the model on the shuffled set to be sure that the selected model is not over fitted to a specific chunk of data.

The standard deviation that results from the K Fold Cross Validation approves of the fact the resultant model is generalized on the whole data and not over-fits. (Figure 3: K Fold Stratified Cross Validation)

Data Evaluation

For evaluation purposes, a confusion matrix is generated that lists down the precision, recall, f1-score, support, average, macro and weighted average values for the multi classes ‘Italian/French/Greek’, ‘Spanish/Mexican’, ‘Asian/Chinese/Thai/Nepal/Indian’ and ‘American.’

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Moreover, we also generate the lifts for the multi-classes relevant to baselines. (Jupyter Cell. 60).
Following the lift calculation, as our model is for a food producer who might not be able to believe in the insights of our model, we have attempted to make our model interpretable by using SHAP plots.

For a particular class and a particular instance, the service tells the features that contributed in a positive/negative way for a particular outcome. In this way, if the outcome is wrong, the food producer can know, what are the features that contributed to that and in how much capacity?
(Figure 5: SHAP Plot for All instances of Italian, French, Greek)
(Figure 6: SHAP Plot for one instance of Italian, French, Greek)

This information is pretty crucial and can be used to know the changing trends and gather insights for future food production.

Clustering

The main goal is to divide the data set into suitable clusters for which the food producers can target customized comfort foods mainly. As the comfort food column is open text, we initially have to cluster it for a usable feature in overall clustering. Comfort foods are mainly given as comma separated items. We use text cleaning, and simple comma tokenization.
The tfidf matrix is calculated with 1-grams as the values are comma separated and then used as the features for KMeans clustering algorithm. The elbow method is used to determine the suitable amount of clusters i.e. 20 in this case.
We label each row with its respective cluster, which results in a categorical data column namely comfort_food_coded.
For the main clusters, we use the following columns:
Gender,calories_day,comfort_food_coded,comfort_food_reasons_coded,cook,diet_current_coded,eating_changes_coded,eating_changes_coded1,eating_out,ethnic_food,exercise,fav_cuisine_coded,fruit_day,healthy_feeling,ideal_diet_coded,nutritional_check,sports,vitamins and comfort_food_coded.
Again we use the elbow method to determine the suitable clusters, which are 14 in this case.

(Figure 8: Elbow Method to Determine Number of Clusters for Comfort Food
Figure 9: Elbow Method to Determine Number of All Clusters
Figure 10: Word Cloud for Comfort Foods)

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Bibliography

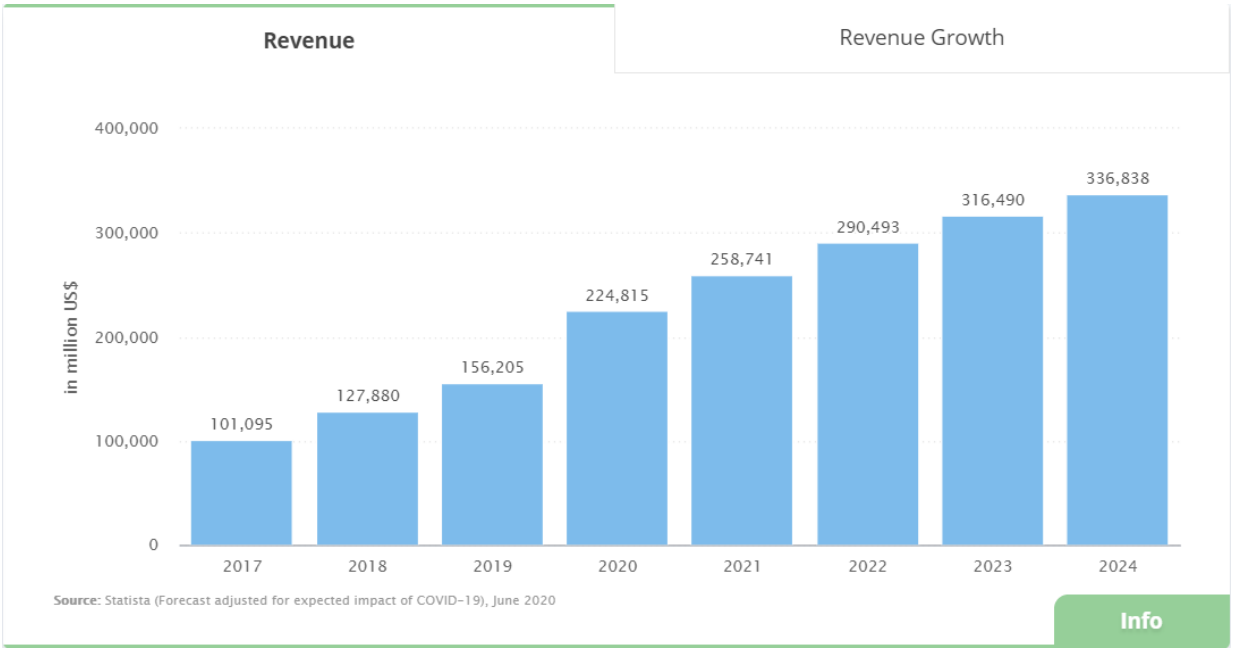

Figure 1: Food Revenue through the Years

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Figure 2: Histogram

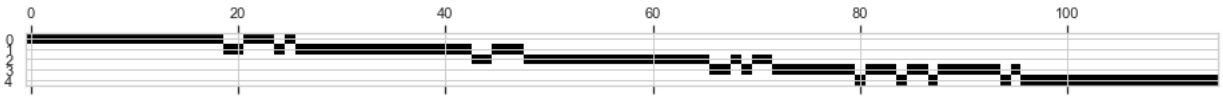

Figure 3: K Fold Stratified Cross Validation

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

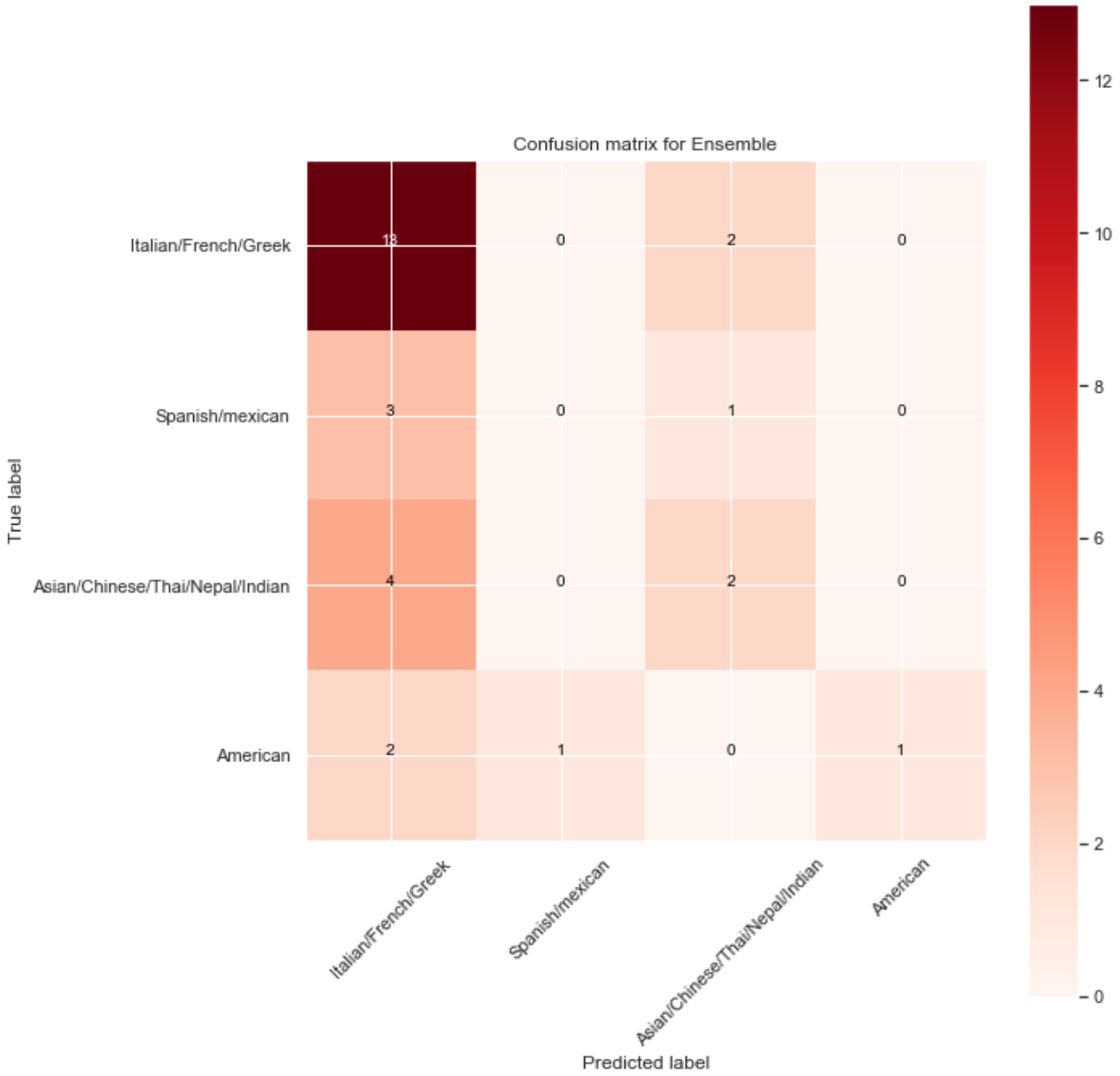

Figure 4: Multi Class Confusion Matrix

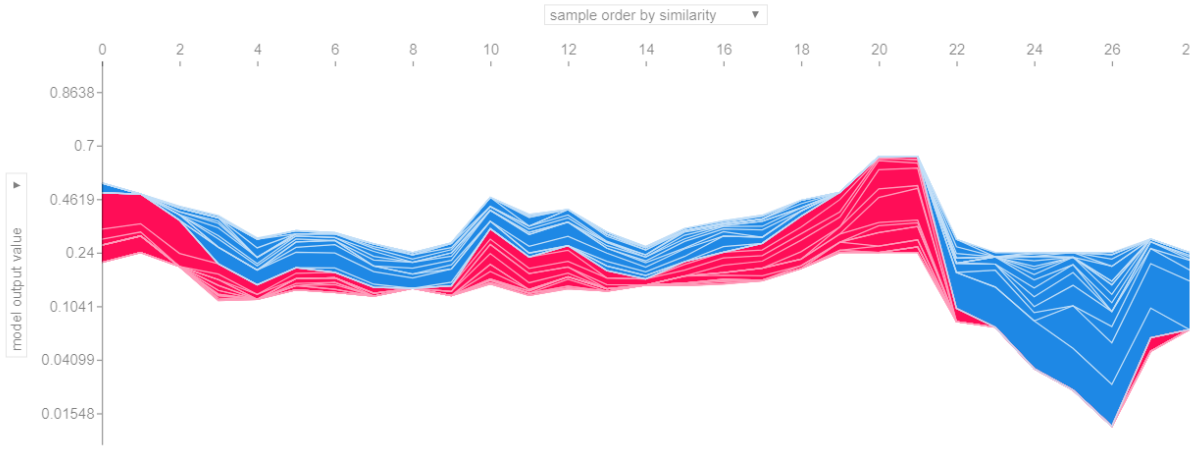

Figure 5: SHAP Plot for All instances of Italian, French, Greek

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

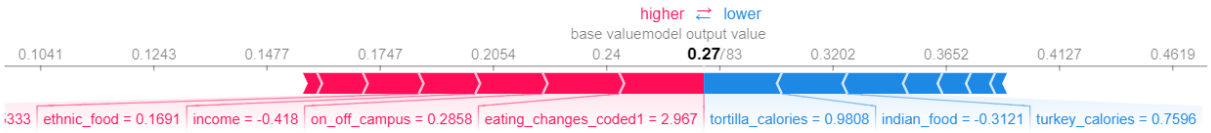

Figure 6: SHAP Plot for one instance of Italian, French, Greek

cluster	income	eating_out	cook	data_index
0.0	4.807551	2.863636	2.636364	70.938776
1.0	4.564516	2.500000	2.960867	57.606061

cluster	eating_changes_coded	eating_out	pay_meal_out	data_index
0.0	1.575000	2.600000	3.375	63.035294
1.0	1.769231	2.807692	3.500	64.000000

cluster	calories_day	eating_out	pay_meal_out	data_index
0.0	3.053267	2.585366	3.365854	62.402299
1.0	3.084513	2.875000	3.541667	66.035714

cluster	eating_out	parents_cook	data_index
0.0	2.60	1.500000	60.662162
1.0	2.75	1.666667	68.024390

cluster	income	parents_cook	data_index
0.0	4.564516	1.596774	57.606061
1.0	4.807551	1.500000	70.938776

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

cluster	eating_changes_coded	cuisine	data_index
0.0	1.659794	1.368270	61.471698
1.0	1.222222	1.419753	84.666667

Figure 7: Some Trend Clusters

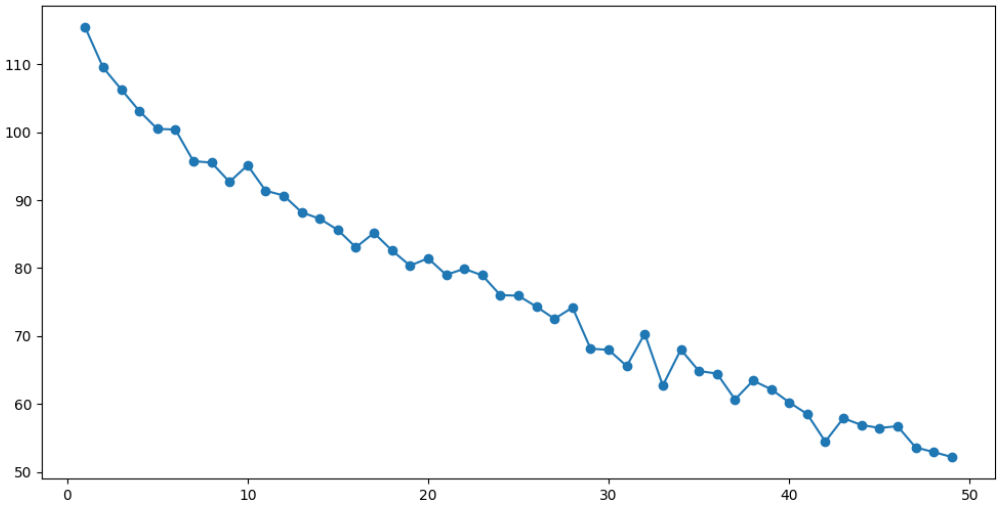

Figure 8: Elbow Method to Determine Number of Clusters for Comfort Food

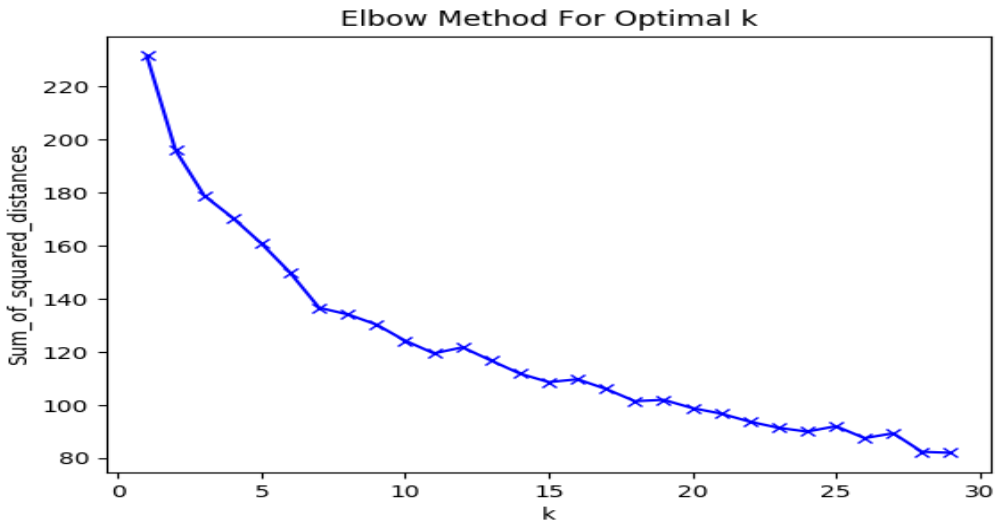

Figure 9: Elbow Method to Determine Number of All Clusters

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Figure 10: Word Cloud for Comfort Foods

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Declaration of Authorship

I affirm that I have produced the work independently, that I have not used any aids other than those specified and that I have clearly marked all literal or analogous reproductions as such.

Location, Date: Saarbrucken, July 17, 2020

Rahman, Safie Ur

Rizvi, Syed Taqi Abbas

Munir, Talha

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736