4/21/2021

# Foundation Project - 1

AMPBA Batch 15

SUBMISSION BY GROUP 10:
Deep Kamal Singh (12020053)
Kshitij Sharma (12020062)
Mohua Sinha (12020015)
Nidhi Srivastava (12020078)
Soumaya Ranjan (12020060)

## Table of Contents

## 1    Background

Food outlets are everywhere, whether it is online or offline, serving mouth-watering food has always been a trending business. If you look at the analytics being used in this food business it's negligible. Menu designs with attractive food images to get the attention of the customers, getting reviews, chatbots, etc are some of the basic technology-based solutions that are available globally today.

## 2    Objective

Track the customer's tastes and preferences based on seasons and moods, by using historical data along with personalized survey data. People may be allergic to specific foods or smells which may result in a very serious problem if neglected. You have to gather as many details as possible from primary and secondary data sources available and develop a Personalized Recommendation Model for Foodies.

### 2.1    Data Sources
- Public forums (Menus)
- Survey data (for personalized preferences)

## 3    Business Problem

### 3.1    Business Understanding: Defining the business problem

#### 3.1.1    Background

In the current scenario, consumers have multiple options for food both in terms of Dining out outlets as well as delivery options. To select an outlet for dining out /delivery options, consumers have options to browse through the individual outlet's menu on their website or to browse the food aggregator platforms to compare the outlets in terms of Ratings, safe and hygiene, offers, distance etc. However, very limited information is available in terms of relating consumers preferences and taste of food based on seasons/ moods.

#### 3.1.2    Business problem
- Present food recommendation engines are not effective
- Significant churn on the account of limited/sub-optimal choices to customers, impacting future sales and retention

#### 3.1.3    Business Objective
- Maximize revenue by providing the most preferred and likeable food options

#### 3.1.4    Business Constraints
- Minimize consumer fatigue by providing most optimal choices
- Minimize the searching time for the preferred choices

## 4    Data requirements

### 4.1    Data attributes/Features required

To gather the historical information and potentially determine people's taste and preferences, we would need to collect the data around the following attributes:

| Attribute/features | How will it help? |
|---|---|
| Food outlet location | Geographic region along with review date will help to determine the seasonality |
| Cuisines/Menu information | Type of cuisines provided by the food outlet |
| Popular dishes | Help to determine the preferences for the dishes for the outlet |
| People Say This Place Is Known For | Help to determine the reasons for visiting the outlet and possible mood |
| Events | Help to determine the mood on the day by linking with the review date |
| Ratings/Reviews | To determine the likeability of a dish and likeability of the outlet |
| Review date | +2 days can be considered as the possible visit date and can help to determine the seasonality impact as well as mood by relating to events around those dates |
| Reviewer info | To determine the demographics of the reviewer, any allergies, food preferences |

### 4.2    Data sources

To capture the above attributes, we are planning to leverage following data sources:

| Data Sources | Data attributes to be captured | Potential challenges |
|---|---|---|
| Swiggy | Information related to outlets:<br>1) Location, Menu<br>2) Cuisines & Popular Dishes<br>3) Outlet known for/Events | Getting the required data through scraping |
| Zomato | 2) Reviews - Dish related info, positive/negative reviews, date of reviews (linking to season /mood) | |
| Dine out | | |
| Social Media Platforms- | Scraping reviews from FB, Twitter for bad/good reviews, cuisine preferences | Unstructured data |
| Personalized Survey | Focus will be to capture the following:<br>1) Demographics<br>2) Any particular allergy<br>3) Food preferences | Number of respondents will be limited |
| Food Blogs | Reviews (positive/negative, cuisine preferences based along with recipes | Unstructured data |

**5      Data collection process**

The data collection process would be done through two modes:
- Unstructured data through <u>Web scraping</u> on websites, online aggregator platforms etc. This will be done for a geography (say North or South) and for one/two year data to capture the seasonality aspect along with reviews and preferences
- Structured data through <u>Personalized survey</u>:  This will be done through a pan-India survey and would be targeting 200+ respondents to capture the food preferences along with the demographic information of the respondents

**5.1      Challenges In Data collection**

- Sourcing the data which is not owned by Business is major challenge – even if the data is arranged from external sources, the reliability and accuracy of data remains a big challenge
- Sourcing data from trusted sources has its own cost and we need to define "discovery" cost and continuous data sourcing cost before we can envisage upon viability of our model

**6      Data understanding with data summary**

Based on the Dish based rating/review data and the customer personalized data, we will do EDA using Multivariate analysis in Python and will do test of linear regression.

**6.1      Data Cleaning and Curation**
- Identifying and monitor for errors in data
- Process formation and preparation of SOP for continuous feed
- Duplicate, redundant and non-usable data identification and removal
- Detailed Data Analysis and process-oriented data structuring
- Inter process communication

**6.2      Data Matrices**

**6.2.1      Customer food preference indices**
- Nutritional compatibility matrix
- Food hypersensitivity reactions matrix
- Flavor preference matrix

**6.2.2      Food Dish composition matrix**
- Ingredient's list – composition
- Complimentary compatibility
- Seasonal/weather compatibility score – humidity, temperature data
- <u>Customer ratings</u> and order history – likelihood pairs

**6.2.3      Historical Data collection**
Gathering customer personalized data for last 2 years through existing survey data. If unavailable to collect it, will collect current data of customer personal preferences using survey forms.

**7      Implementation Model - future strategy**
**Base Idea:** The seasonal, taste-based, and nutritional data will be used as reference data to map with the customer personalized data to create an appropriate training dataset for our recommendation model

1) To build a Recommendation system, will follow MCDA – Multi criteria decision analysis matrix match
2) This will help to figure out top driving factors for a customer's food preference
3) Public surveys to provide data to determine the preferences based on taste, allergic and health conditions
4) Capture the preferences from food reviews from social platforms data.
5) Perform matrix matching of the ingredients of each dish, based on our recipe reference data
6) Further learning in this direction led us to form the data points or features, which we will feed as input to our model
7) Our target is to achieve prediction score (say >~0.90) from the Multinomial Naïve Bayes model,
8) Based on the prediction score, will increase our training dataset, if required, to further tune our model

**8      References:**
- An overview of recommender systems in the healthy food domain | SpringerLink
- A Food Recommender System Considering Nutritional Information and User Preferences | IEEE Journals & Magazine | IEEE Xplore
- Data Cleaning Steps