Dev Zaveri & Joseph Porrino

CMPSC 445

Janghoon Yang

22 April 2024

GitHub Project Link: https://github.com/BabyKangaroo117/Food\_Effect\_Covid

# **Project: Analysis of Food Consumption and COVID-19 Cases**

## **Project Goals**

The project aims to delve into the potential correlation between dietary patterns and the prevalence of COVID-19 cases worldwide. The dietary constraints are based on a per country bases. There are 20 dietary features with percentage values of food quantity based on the total food intake of the country. The output value is the percentage of the population that is presumed to have covid. The significance of this project will be to find the foods that contribute most to confirmed covid cases. Just because there is a correlation, doesn’t mean there is a causation, so the results will be interpretable as to why certain foods contribute more than others. This endeavor encompasses several key objectives:

**Data Preprocessing**: Initial steps involve cleaning and preprocessing the dataset to handle missing values and outliers effectively. The first step was to replace string data with numerical values. The undernourished column had string values of “<2.5” which we replaced with the float 2.4. Next we replaced null values with an imputed data value. We used KNN to compute a value based on the 3 nearest neighbors of the missing value. This will give a more accurate representation than using mean because mean can skew the value. The next step was to remove columns that had majority values of 0. These features are sparse and don’t contribute anything to predicting the target value.

**Exploratory Data Analysis (EDA)**: Through EDA, we gained insights into the distribution and correlation among different food consumption features. We constructed a histogram to see the spread of values among each feature. Most features had majority grouping around a few values. This is a problem during training because it leads a loss in the model’s ability to capture a full range of patterns. This will end up causing the model to be overfitted to the training data. The next analysis tool we used was Pearson correlation coefficient. At most the best features had a moderate correlation to the target value. Most features had a weak correlation to the target value. To analyze collinearity, we constructed a heat map based on the correlation matrix. Vegetal products had a strong collinearity with animal products, milk, and animal fats. Some features had moderate collinearity which is in the range of 0.40 to 0.60. Most features had a weak correlation among each other. Using this data we could get a better understanding of how our model would perform.

**Machine Learning Model Building**: Our primary focus lies in constructing and assessing various machine learning models to forecast COVID-19 confirmed cases based on dietary intake features. The first model we trained was a linear regression model. It ended up performing poorly with an r-squared value of 0.295. Since linear regression couldn’t properly capture the patterns in the data, we moved to training a gradient boosting regressor. GBR is better at capturing nonlinear relationships by combining multiple weak learning decision trees to capture various signals in the data. This model ended up performing substantially better with an r-squared value of 0.65. The next model we trained was a XGBoost model. XGBoost follows the same principles of gradient boosting regressor, except that it has controls for L1 and L2 regularization. Using these hyperparameters we can try to control overfitting which we assumed might occur based on our exploratory data analysis. We found that increasing L1 regularization (alpha) from 0 to 1 and L2 regularization (lambda) from 1 to 1.5, provided the best results. This both increases sparsity by driving some feature coefficients to 0 and decreases complexity by minimizing the size of large coefficients. We also increased the learning rate to 0.15 since we are dealing with a small number of datapoints. The model ended up with an r-squared value of 0.756. This is a big improvement from the GBR model and we can account this to dealing with overfitting. The last model we trained was an SVR model. SVR is a simpler model than XGBoost, but has several kernels that can find patterns in regression data. We used GridSearchCV to determine the best parameters for the model. We selected three parameters for controlling regularization, three parameters to control the fit of rbf and three kernels. After running 100 kfolds on the SVR model using GridSearchCV, the best parameters were {'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}. We created a model with these parameters and ran 50 kfolds. We used kfolds to cross validate our data because we are dealing with a small data pool. Kfolds ensures that each datapoint is used for both training and validation. Since the SVR model is a simple model, this will help us to see the stability and reliability of the model using different data values. We ended up having very mixed results after running the kfolds. Some models performed exceptionally well with r-squared values as high as 0.994 while other models performed terribly with r-squared values in the negatives. This means the model is unable to capture any meaningful information about the relationship between independent and dependent variables. Since the SVR model was so unstable, we decided to use the XGBoost model as our final model. This model is able to capture small non-linear features that other models cannot. It also can account for overfitting.

**Identification of Relevant Dietary Factors**: Ultimately, we strived to pinpoint the most significant dietary factors contributing to the prediction of COVID-19 cases. The importance of this project lies in addressing the relative neglect of dietary patterns amidst the COVID-19 pandemic's extensive research. While considerable attention has been devoted to understanding the virus's transmission and treatment, the potential impact of dietary habits remains underexplored. After training the XGBoost model, we plotted the score of each feature to see how much it contributed to predicting the target value. The top five contributing features were alcoholic beverages, fish and seafood, oil crops, eggs, and tree nuts. This data doesn’t necessarily have a causation that eating more or less of these foods causes covid. What it can explain is that countries with higher covid cases consume more of these foods. It would require scientific research to determine if these foods have an effect on catching covid

**Installation and Usage Instructions**: The first step to run our program is to clone the repo to your local machine. Next, follow the below steps in the IDE of your choosing.

In the terminal:

1. pip install poetry
2. poetry install
3. poetry shell (can be skipped if your computer is not enabled to run scripts)
4. cd GUI
5. poetry run python user\_interface.py

This will start the tkinter app. To use the app move the sliders until the food percentage is at 100%. The model will predict the covid confirmed value with the set parameters.

**Code Structure**: Our preprocessing, EDA, and model training were carried out in a jupyter notebook. The file is located in the Train\_Model folder of the repo. The following were the steps taken to construct our model.

1. **Upload dataset**
2. **Fill in null values** : We decided to replace null values with KNN imputed values because our data set was small and we didn’t want to lose any more samples.
3. **Replace non numerical values**
4. **Remove features with majority 0’s**
5. **Explore data through use of a barplot and histogram:** We constructed a barplot to see the comparison of confirmed covid cases based on country. This was we can see if there is a spread among countries.
6. **Use Pearson correlation coefficient and a heatmap to assess collinearity:** It was important to check the relationship between features to be able to mitigate collinearity. The heatmap showed us that several features had a high collinearity which we could adjust using the regularization of the XGBoost model.
7. **Split data into x and y sets**
8. **Construct models:** We constructed several models that could prediction a regression problem.
9. **Train models**
10. **Assess model performance:** We used r-square, mean square error and explained variance to assess model performance.
11. **Pickle best model**
12. **Construct tkinter application**
13. **Enter data from user input**
14. **Convert user input to shape for model input**
15. **Make prediction**

**Functionalities and Test Results**: Our best performing model was the XGBoost model. This model was able to capture enough weak signals so that the model was more generalizable. The r-squared value was 0.72 with an explained variance score of 0.85. The explained variance score helps to confirm that the dependent variables (features) are predicting the target value. We tested the model with several variations of hyperparameters and found that these were the best:

(objective='reg:squarederror', n\_estimators=70, learning\_rate=0.20, max\_depth=4, reg\_alpha=1.1, reg\_lambda=1.4)

**Conclusion:**

In conclusion, this project successfully reveals potential relationship between dietary patterns and COVID-19 cases through machine learning techniques. Despite limitations regarding the dataset size, the project demonstrates the practical application of data preprocessing, exploratory data analysis, and machine learning model selection. Future endeavors could expand the dataset, explore alternative models, and delve deeper into the mechanisms linking specific dietary components to COVID-19 susceptibility or severity. To fully understand why there is correlation to the food groups revealed in this, would require scientific research in biology and virology to find a potential link to confirmed covid cases. This study helps to narrow potential foods for researchers to study.