**Using Machine Learning to Reduce Food Wastage and Fight Hunger**

1. **Project Description:**

The aim of the project is to use machine learning techniques such as regression to study the patterns of surplus food production in the U.S., build a prediction model to estimate food wastage, and explore various ways to reduce it, thus fighting hunger and starvation.

1. **Background:**

With so many conflicts and wars going on around the world, the worst thing that can happen to humanity is when people die of hunger, starvation, and famine. While everyone requires balanced Food to lead a healthy life, a recent study found that about 44 million people in the U.S. are food insecure. This includes 5 million children who do not have access to nutritious meals. This is a much bigger problem in many developing and underdeveloped countries worldwide. According to the Food Security and Nutrition report, about 783 million people faced hunger in 2022, representing almost one-tenth of the world's population and an increase of 122 million from 2019. (Why Should We Care about Food Waste?, n.d.)

1. **Business Problem:**

In the U.S., despite many people facing hunger, we also have another challenging problem due to excess food production and wastage. It is estimated that about 30-40 percent of the Food we buy ends up in landfills, costing the economy about $161 billion annually. (Nations, n.d.) This comprises time, water, labor, and money spent on farming, transportation, storage, and food distribution, only to waste them. The impacts also extend to climate changes and environmental impacts. According to the U.S. Environmental Protection Agency, food waste in landfills is the largest source of methane emissions that can be controlled and avoided.

On the other end, impoverished people are the most affected by rising inflation and increasing food prices. Safe and healthy Food can be donated, and millions of starving people can be saved. ***The business problem addressed in this project is to bridge the gaps between two significant issues that we have by reducing food wastage and redirecting excess food to those in need, thus reducing hunger.*** This project is intended for non-profit and charitable organizations. (Feeding America, 2023).

1. **Data Explanation:**

Five datasets were used in this project, as described below. The primary datasets were taken from refed.org (ReFED, 2022), while the supplement datasets were taken from the census website. (U.S. Census Bureau, 2020)

1. **US\_State\_Food\_Surplus\_Detail:**

This Dataset, available in CSV format, contains data about the surplus food production in the U.S. from 2010 to 2022. Each row represents the excess output in tons, the amount wasted, donated, etc., for various food categories. It has about 560k rows and 28 columns.

1. **US\_State\_Food\_Surplus\_Summary:**

This dataset is like the Food surplus dataset but contains the Summary data. The data is captured for each state in the U.S. and includes 187k rows and 29 columns.

Other supplemental datasets used in the project are listed below and contain data about the homeless Population, Poverty rates, and U.S. state Population from 2020 to 2023. These datasets were used for data comparison and visualizations.

1. **Homeless\_Populations\_by\_State.xlsx**
2. **US\_Poverty\_Rates.xlsx**
3. **US\_Population.xlsx**
4. **Methods and Analysis:**

The Data exploration, Analysis, and Modelling in this project were done in **R**. As the datasets are in CSV and Excel formats, read\_csv and read\_excel methods in base R were used to load them into Dataframes.

* 1. **Cleansing *of the Data:***

The structure of the data frames was then inspected to check if all columns were loaded with expected datatypes. The first few rows of the data frames were then inspected to check for data formatting errors. During this phase, issues with the currency and other numeric columns were found, such as Dollar signs in currency columns and Commas in Number fields. The text columns had many formatting issues, such as inconsistent capitalization and empty spaces. Reusable functions were used to clean up all these discrepancies in the data. All the extra spaces were trimmed to make the data consistent, and the data in text columns were converted to upper case.

A screenshot of a computer

Description automatically generated

*Table 1: Table showing missing values in the Surplus Food dataset*

In the next step, the datasets were checked for Nulls and duplicates. Table 1 shows the number of missing rows in each column of the Dataset. Upon inspection of a few rows, it was found that the data was missing in all mandatory columns, such as meals wasted, Food donated, Food wasted, etc. So, they were dropped. Nulls in other numeric and character columns were imputed as zeros and N.A., respectively. Also, rows with duplicates in all columns were dropped as well. These steps were performed for all the datasets used in the project. Finally, the structure of the Dataset and a few sample rows from the Dataset were inspected to ensure all updates were made.

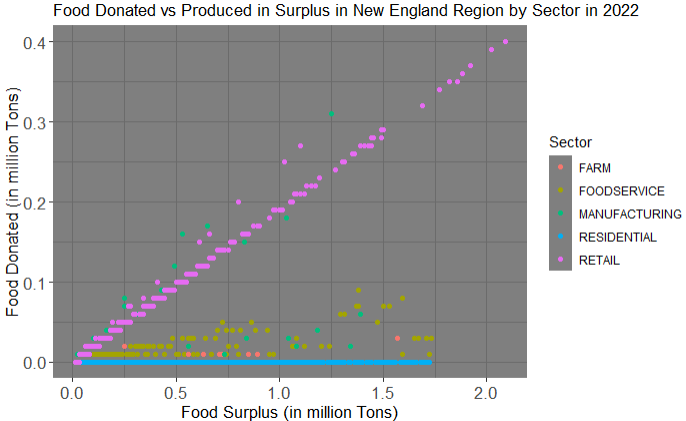
* 1. **Data *Exploration:***

In the data exploration phase, the unique values in the factor fields were compared between the Food Surplus dataset and the Solutions Dataset to identify any discrepancies with the data formatting. This is to avoid data losses while joining the data frames.

A graph with red and yellow dots

Description automatically generated

*Figure 1: Scatter Plot of Food Donated vs Surplus by Population*

The distribution of the numeric data, such as meals wasted, donated, and unused, was explored using scatter plots, histograms, and boxplots. Figure 1 shows the relationship between food donations and surplus in the New England region. The plot indicates a strong positive relation between the two variables in some heavily populated states, but the pattern is inconsistent enough to draw any conclusions. This has been resolved in Figure 2, where the data is plotted based on the food sector, which indicates a strong positive relation between the Food donated and Surplus in Retail but not so much in the other sectors, with almost no relation between the two variables in Residential. 

*Figure 2: Scatter Plot of Food Donated vs Surplus by Sector*

In Figures 3a and 3b, the wasted meal column is plotted on a log scale as the data was heavily skewed from the original scale. The plots indicate skew in all four southwestern states which requires further analysis.

A graph of a graph

Description automatically generated with medium confidence

*Figure 3a: Density plot of Meals wasted in Log scale*

A graph of meals wasted in south west

Description automatically generated

*Figure 3b: Histogram of Meals wasted in Log scale*

Boxplots were plotted for each sector to analyze the distribution of the data on the meals wasted column, indicating the presence of outliers in the food service and Retail sectors (Figure 4). Also, the plot shows that the farm sector wasted the most Food in 2022 in New England compared to other industries.

A graph of different colored squares

Description automatically generated

*Figure 4: Box plot distribution of Meals wasted by sector*

Examining the count plot in Figure 5 indicates that there are more occurrences of Retail data in the Dataset, at least for the New England region compared to the rest, with the Farm sector having only 540 occurrences.

A graph of a number of people

Description automatically generated with medium confidence

*Figure 5: Count plot distribution of sector in the New England region*

* 1. **Research *Questions:***

The below mentioned Research questions were explored in the project:

1. **Which Midwestern state wasted the most Food in 2022, and by how much?** A graph of the states

   Description automatically generated

*Figure 6: Meals wasted by Midwestern states*

**Observation**: Figure 6 indicates Illinois and Ohio wasted almost 350k meals in 2022 and were ranked the top. The Red line indicates the median number of meals wasted; five states were above the median. South and North Dakotas wasted the least amount of Food, most likely due to the lower Population in these states.

1. **What was the U.S. surplus food production trend over the years?**

**Observation**: Figure 7 indicates that the residential sector produced the most excess food while the retail sector produced the least. Residential production spiked in 2020 during the COVID-19 pandemic, while food services, such as restaurants and catering, took a dip when people stopped going to the restaurants. The trend was almost flat for other sectors.

A graph of a food production

Description automatically generated

*Figure 7: Line plot of Surplus production over*

1. **What was the most wasted Farm food in 2022?**

**Observation**: As per Figure 8, the most wasted farm food in 2022 was Lettuce, followed by Apples and Watermelons.

A graph of food in different colors

Description automatically generated with medium confidence

*Figure 8: Wasted Farm food in 2022*

1. **Which sector donated the least amount of Food?**

**Observation:**  Figure 9 indicates that the Residential sector did not donate any food in the New England region. The retail industry donated the most Food.

A graph of food donations

Description automatically generated

*Figure 9: Donated Food by Sector in 2022*

1. **Which state wasted the most Food, and which sectors in 2022?**

**Observation:** The most populated states, such as California, Florida, Texas, and New York, wasted the most Food in 2022. Interestingly, the residential sector wastes a lot of Food in these states, and they never donate any food.

A graph of food wasted in 2022

Description automatically generated

*Figure 10: Food wasted by states and sector*

1. **Which Region wasted most Food unharvested?**

**Observation:** Pacific states waste the most Food due to non-harvesting. Overall, most food waste is from Pacific states in both the residential and farm sectors, which waste the most.

A pie chart with numbers and text with Crust in the background

Description automatically generated

*Figure 11: Unharvested Food by Region*

1. **Which produce was the most grown in Surplus in Texas Farms in 2022?**

**Observation:** Watermelons, Bananas, and Potatoes were the most grown produce in surplus in Texas.

A colorful circle with numbers and text

Description automatically generated

*Figure 12: Excess-grown produce in Texas farms*

1. **Which state produced the most surplus Potatoes in each Region in 2022?**

**Observation:** According to Figure 13, Idaho produced the most surplus Potatoes in 2022.

A chart of a variety of potatoes

Description automatically generated

*Figure 13: Surplus Potatoes Production by Region and State*

1. **Which Midwest state had the highest poverty rate and how much food was Wasted compared in 2022?**

**Observation:** The bubble plot in Figure 14 indicates that Michigan had the Midwest's highest poverty rate in 2022. The bubble size in the plot indicates the amount of Food donated. Illinois and Ohio wasted most of their Food and donated the most compared to other Midwestern states.

A graph with colorful circles

Description automatically generated

*Figure 14: Bubble plot of Donated Food vs Wasted Food*

1. **Which state wasted the most meals per person in retail?**

**Observation:** Figure 15 indicates that Wyoming, Vermont, and Alaska wasted the most meals per person in 2022.

A map of the united states

Description automatically generated

*Figure 15: Meals wasted per person by States in 2022*

* 1. **Data *Preparation:***

Based on the observations in the Analysis phase, there were many outliers in the column Meals\_wasted. Further analysis was performed for other regions, and similar observations were made. Most outliers were in the Farm data, so they were not excluded from the Dataset. In this phase, the data from all sources, such as Food Surplus, Surplus Summary, Unemployment, Homelessness, and U.S. Population, were combined based on the common columns. The combined Dataset had about 35 columns and 540k rows.

* 1. **Feature Selection and Extraction:**

As not all features can contribute to the Model, the correlation between the features was calculated, and the highly correlated features were excluded. The Correlation plot(Figure 16) includes some of the Numeric features that were compared with the Target variable Meals wasted to identify the highly correlated features.

A graph of a graph with red and orange squares

Description automatically generated with medium confidence

*Figure 16: Importance of Features for Model Building*

A Linear model was built only on the Numeric features, and those with p-values higher than 0.05 were discarded. These feature selection techniques reduced the numeric features by half, from 30 to 15. Figure 17 shows each numeric feature's importance in predicting the target variable meals wasted.

A graph with colorful lines

Description automatically generated

*Figure 17: Importance of Features for Model Building*

Per Figure 17, some features like U.S. dollar surplus, tons\_supply, and Population contribute little to predicting the number of meals wasted. So, Principal Component Analysis was implemented to extract the features that can explain most of the variance in the target variable.

A screenshot of a computer

Description automatically generated

*Figure 18a: Summary of PCA Analysis*

A graph with numbers and a line

Description automatically generated

*Figure 18b: PCA Analysis plot*

Figures 18a and 18b show the summary and plot of PCA analysis. Six dimensions contain up to 90% of the information in the original Dataset, and the remaining nine can be discarded.

* 1. ***Model* Building*:***

The original Dataset was split into Train and test sets with a ratio of 80:20. Then, the numeric features were standardized on each Train and test set before applying the PCA. This was done to avoid data leakage from Train into test sets. Also, the factors were converted into Dummy values before fitting the Model to convert all features into Numeric. The Target variable predicted by the Model is the number of Meals wasted, a continuous numeric variable. Linear Regression models were built using Multiple Linear regression, Caret, vTreat, and Recipe packages. RMSE (Root Mean Square Error) was chosen as the Error metric, and R-squared values of the Model were computed to understand the variability explained by the Model.

Also, Regularization methods were used to avoid overfitting in the model and to reduce the sampling error by reducing the variance. Also, the regularization methods penalize the coefficients involved in the linear regression equation. In this project, Ridge, Lasso, and Elastic Net models were trained on the dataset. The models were tested for various values of Lambda and the one lambda with the least amount of Error was chosen to test the model.

1. **Conclusion:**

The four linear regression models were tested for different numbers of PCA parameters, and the results were recorded. Based on the findings, *the vTreat Model with 8 PCA parameters* was identified as the best choice as it had a higher R-squared(0.80) than other models and much lower Root mean squared values(197).

A graph of a graph showing the value of a model

Description automatically generated

*Figure 19: Comparison of R-squared values of Linear Models*

A graph with lines and numbers

Description automatically generated

*Figure 20: Comparison of RMSE values of Linear Models*

However, the models trained without PCA returned better results. As shown in table2, most of the models had the R-squared value of 90% and RMSE of about 191. Hence for this Project, either the ***Simple or Caret Model or a Regularized model such as Elastic Net can be chosen for deployment***.

A screenshot of a computer

Description automatically generated

Table 2: Regression Model Results (without PCA)

1. **Assumptions:**

Some of the assumptions made while working on the project are:

* The data was not biased and was collected from independent sources. As the primary datasets were taken from refed.org, an independent non-profit organization, the data was assumed to be unbiased to any specific sector, subsector, or Region.
* The target variable of the Model – Meals wasted was assumed to have a linear relationship with the independent variables. This was also verified by validating the relationship between some features and the target variable.
* The features used to build the Model were assumed not to be highly correlated with each other to avoid Multicollinearity. This was also validated by plotting the correlation table of the features.

1. **Limitations:**

The target feature of the Model that predicts the number of meals wasted does not factor into the freshness aspect of the food. For instance, the Model can predict the number of cans of tuna that will be wasted and redistributed, but it may not be able to differentiate the Food that expires in a month versus a year. It is up to the people at the distribution center to decide which Food should be donated first.

Though the Model can estimate the number of meals that can be donated or diverted to nearby distribution centers, it does not factor in the cost involved in the storage and distribution. For instance, storing seafood or frozen food costs more than storing grains or bread that does not require refrigeration.

1. **Challenges:**

The Model does not differentiate between healthy Food, such as fresh fruits and vegetables, and Junk food, such as Chips and fried Food. Maintaining a balance between the two can pose a challenge. For instance, based on the Model, if excess fresh produce is reduced, it may lead to an oversupply of non-healthier options. As we don’t have control over the Food being received at donation centers, the oversupply of certain types of Food can create ethical challenges.

1. **Future Uses:**

The scope of this project is to estimate the number of meals wasted; however, it can be expanded further to provide alternate solutions to reduce wastage, such as improving food distribution, Optimizing the harvest, Intelligent rerouting, etc.

1. **Recommendations:**

The Model can expand further by including the time component, such as year and month, to predict food wastage for a specific period of the year. This can be particularly useful in predicting food wastage based on seasonality and taking measures in advance. For instance, if we can expect the number of watermelons wasted during summer, measures can be taken months in advance to redistribute it to other regions or reduce the supply.

1. **Implementation plan:**

Before deploying the Model in production, it must be tested in production-like simulation environments to assess its performance. Also, the enhancements mentioned in the recommendations section, such as including a time component, must be added to make it production ready. A user interface or a mobile app can be built that takes the user input, such as the state, sector, subsector, food type, and season, and the Model can predict the amount of Food that may be wasted for that category. Also, solutions to reduce food wastage can be added as an enhancement.

1. **Ethical Concerns:**

Though the project intends to reduce the surplus food production and increase food donations to the poor and hungry, they are only meant to complement each other and not to cause conflicts. For instance, reducing food production by too much can substantially affect the food supply, reducing the amount of Food being donated, which completely defeats the project's purpose.

Though donating Food by reducing food wastage is a great cause, care must be taken to refrain from donating expired or near-expiry Food. Local laws must be followed, and proper licenses must be secured to avoid conflicts with local government policies.

Reducing food production can significantly impact the local economies and families that rely solely on agriculture. Hence, other solutions, such as redistributing the Food and reducing the transit time to transport Food, must be considered before reducing the food supply.

1. **References:**

Feeding America. (2023). Hunger in America. Feedingamerica.org; Feeding America. <https://www.feedingamerica.org/hunger-in-america>

ReFED. (2022). Food Waste Problem | ReFED. Refed.org. <https://refed.org/food-waste/the-problem>

‌U.S. Census Bureau. (2020). Data. Census.gov. <https://www.census.gov/data.html>

Food Waste in America in 2020: Statistics & Facts. (2020). Recycle Track Systems. <https://www.rts.com/resources/guides/food-waste-america/>

Why should we care about food waste? (n.d.). Www.usda.gov. <https://www.usda.gov/foodlossandwaste/why#:~:text=Resources%20could%20be%20conserved%20for>

Nations, U. (n.d.). Food. United Nations. <https://www.un.org/en/global-issues/food#:~:text=Food%20security%20and%20nutrition%20situation%20remains%20dire%20in%202022&text=According%20to%20the%202023%20edition>

**Appendix :**

Homeless population in the U.S.

The dataset about homelessness includes various details such as total beds year around, participation rates of the year around beds, total units with children, dedicated veteran beds, etc. As there is no clear indication of the number of homeless people in each state, for simplicity, the total number of year-round beds in 2022 was considered for analysis in this project.