Predicting the quantity of food loss in a food supply chain

Maha Mapara and Minhail Khan

15th October, 2020

1 Abstract

Our research problem was about how food loss can be minimized using machine learning. To do this, we created a multivariate regression model to predict food loss quantity using features like geographic location, stage of supply chain loss occurred in, cause of food loss, the year loss was recorded in. As the features were all string variables, we had to encode them into categorical variables, remove missing data and irrelevant columns. While our model was not a good predictor of food loss quantity, our data exploration led to insights on which geographic location had the highest loss for which crops and what stage of the supply chain.

2 Introduction

An estimated 1.3 billion metric tonnes of food is wasted globally, every year. That is 1/3rd of all food produced for human consumption.[4] This has great implications for global hunger and malnutrition, but it also has catastrophic impacts on the environment.

When edible items are discarded, it's not just food that is wasted. Consider all the resources required to bring food from the farm to your table: water for irrigation, land for planting, fuel for powering harvest and transport vehicles. 70% of the world's fresh water is used for agriculture, 28% of the world's agricultural area is used to produce food that is ultimately lost or wasted every year. The carbon footprint of food waste is 3.3 billion tonnes of CO2 equivalent per year. Discarded waste that rots in landfills gives off methane, a potent greenhouse gas 25x more efficient at trapping heat than CO2.

When a bunch of bananas falls off a truck or restaurant owners fill their rubbish bins with uneaten meals, all those resources are essentially wasted right along with the food. From farm to fork, food is lost or wasted at every step of the journey. Food loss takes place on the farm and through supply chains, due to suboptimal farming methods, poor storage and inflexible distribution or buying practices. Sometimes it is something as simple as a bag not being strong enough to hold its contents, and food like rice leaking out of small tears. We're also

generating huge amounts of food waste at the point of consumption, either in shops, restaurants, or at home. Unclear labelling, portion sizes which are too big and simply not eating everything before it goes off all contribute.[8]

By using machine learning approaches[1], we can optimize supply chains[7] and forecast food waste. In this project, we are predicting the percentage of food loss that will occur to understand in which country, for which crop, in what part of the food chain and the reason for the loss to see how food loss can be minimized. By reducing food waste, green house gas emissions can also be reduced.

Our project is focused on predicting the quantity of food loss in order to better understand how the food loss can be minimized.

3 Related work

3.1 Walmart's Eden

Food waste is a global concern and leading grocery franchises have particularly started working to reduce food wastage. Walmart is now using machine learning to reduce food waste. In the past Walmart used manual inspections to assess quality of shipment products but it was highly inefficient. Now Walmart has been using Eden which is a machine learning algorithm that can be used to scan produce to access its quality and freshness. Eden was trained on 1 million old produce photos. With the development of Eden, Walmart saved almost 86 million dollars through reduced food waste, in the 6 months after its launch. [2]

3.2 Applying AI in Food Industry To Streamline Supply Chain Workflow

A blog article from Radio Studio, "Applying AI in Food Industry To Streamline Supply Chain Workflow," discusses a forecasting approach to the problem of running out of raw material during business hours and simultaneously preserving freshness of food. It recommends an application that can be used on the real time basis and with the given data, predicts if supply of a particular commodity will last for the entire day and ultimately sends recommendations to the supplier. [6]

3.3 Deep Learning for Classifying Food Waste

A research paper from the Zurich University of Applied Sciences, "Deep Learning for Classifying Food Waste," uses deep learning to classify food waste in half a million images captured by cameras installed on top of food waste bins. Deep learning to classify food waste in half a million images captured by cameras installed on top of food waste bins. Their designed deep neural network classifies this food waste for every time food is thrown in the waste bins. [3]

4 Data

The Food and Agricultural Organization (FAO) of the United Nations has the world's largest food loss and waste database. The organization conducted an extensive review of literature in the public domain which gathered data and information from almost 500 publications, reports, and studies from various sources (including from organizations like the World Bank, GIZ, IFPRI, and more). We used data from this database.

The data set included information about the country the food loss occurred in, food loss per clean products, percentage loss of food quantity, activity that led to food loss, cause of food loss, crop type, and the measured item percentage. The data set includes data on food loss and waste from 1997 to 2017, from 38 countries. [5]

5 Methods

5.1 Cleaning, processing and exploring the data

1. The original data set had almost 17,000 rows and 22 columns. The total missing data points were 173,740. Moreover, at this stage in the project, our predictor variable was cause of food loss and it had 15,828 missing values. For this reason, we decided to remove all rows where cause of loss was missing data. Then we removed columns that were irrelevant to our project, for example the method of data collection, tag for the data collection, url and references.
Additionally, we removed columns that would have been relevant to our project but had 50 percent or more missing data. These were columns like units of food product, period of storage of the food etc.

2. We also narrowed the scope of our project by choosing to include countries that had 20 or more data points. This decreased the number of countries in the data set from 38 to 18 (countries kept were China, Benin, Ethiopia, Ghana, India, Iran, Pakistan, Nepal, Kenya, South Korea, Malawi, Bangladesh, New Zealand, Nigeria, Rwanda, United Kingdom, Tanzania, Philippines). The dimension of our data after all these changes was 935 x 6.

3. The features that we decided to use for the modeling were country, type of crop, stage in supply chain that food was lost, reported cause of food loss, and the year it occurred. Except for the year, all the data in these features were categorical but in strings. We encoded these to be numerical categories that we could use in our model. The year feature was also numerically categorized.

4 a. Once our data was cleaned and processed, we explored it through some plots and descriptive statistics. Figure 1 shows the summary for the predictor variable, percentage loss of quantity.

	percentage_loss_of_quant
count	935.000000
mean	11.397492
std	12.092827
min	0.000000
25%	2.800000
50%	7.000000
75%	15.300000
max	90.500000

Figure 1: Summary for the predictor variable

4 b. We also looked at the occurrences of each category in our features. Figure 2 shows the counts for each category in our location/stage of supply chain feature.

```
#for location/stage of supply chain
fw_final['fsc_location1'].value_counts()
```

```
Farm                196
Storage             152
Retail              125
Wholesale           113
Transport            64
Processing           59
Harvest              52
WholeSupplyChain     51
Export               28
ParameterEstimate    27
Distribution         10
Packaging             9
Grading               6
Trader                6
Traders               5
-                     1
Pre-Harvest           1
Name: fsc_location1, dtype: int64
```

Figure 2: Counts for each stage of the supply chain where food loss occurred

4 c. Scatter plots were created to visualize the relationship between the percentage loss of quantity (predictor) and each of the features (country, crop, cause of loss, stage of supply chain, time point). These figures (fig 3-6) can be seen in the results and evaluation section.

5.2 Splitting the data

The data was split into training, dev and test sets. The training set contained 60 percent of the data, and the dev and test sets each contained 20 percent of the data.

5.3 Modeling

To predict percentage loss of quantity (a continuous value), we used a multivariate linear regression model with country, crop, cause of loss, stage of supply chain, and time point as the explanatory variables (features).

6 Results and evaluation

6.1 Exploratory plots to see the relationship between the predictor and independent variables

The dot plot for Loss of quantity(%) vs Country (figure 3) shows us which countries have the highest food loss. We can see that Ghana, Benin, China, Tanzania and Pakistan are some of the countries with high food loss values.
Note: As the data available for these countries was not equal in number, this plot may not be an accurate representation of the food loss occurring in these countries.

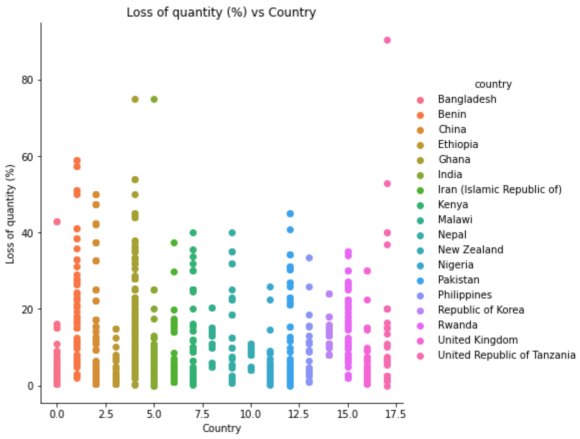


Figure 3: Loss of quantity(%) vs Country

The dot plot for Loss of quantity(%) vs Country (colored by crop) (figure 4 on the next page) shows the highest loss by crop for all the countries. We can see that Pakistan has high loss for wheat, China has high loss for Cabbages and Kenya has high loss for potatoes.

The dot plot for Loss of quantity(%) vs location/stage in supply chain (colored by country) (figure 5 on the next page) shows which stage in the supply chain sees high losses for which country. The wholesale, storage, processing, farm and retail stages have the highest food loss values. Some of the countries with high losses in these stages are Tanzania (in storage and retail), Ghana (in farming), Pakistan (in harvest) and China (in wholesale and storage).

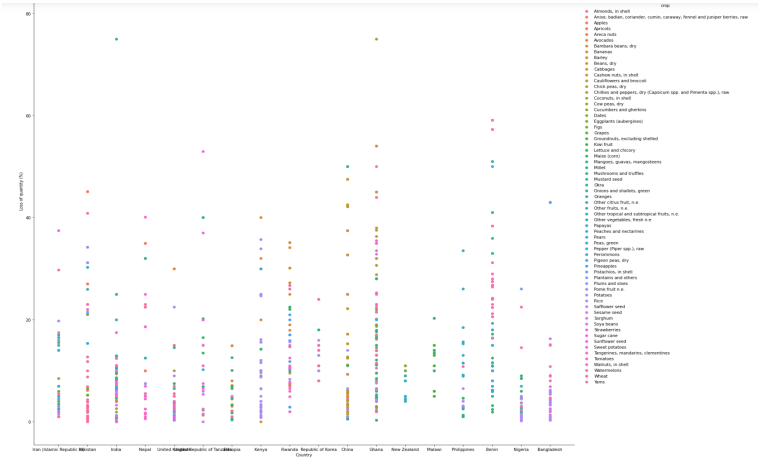


Figure 4: Loss of quantity(%) vs Country (by crop)

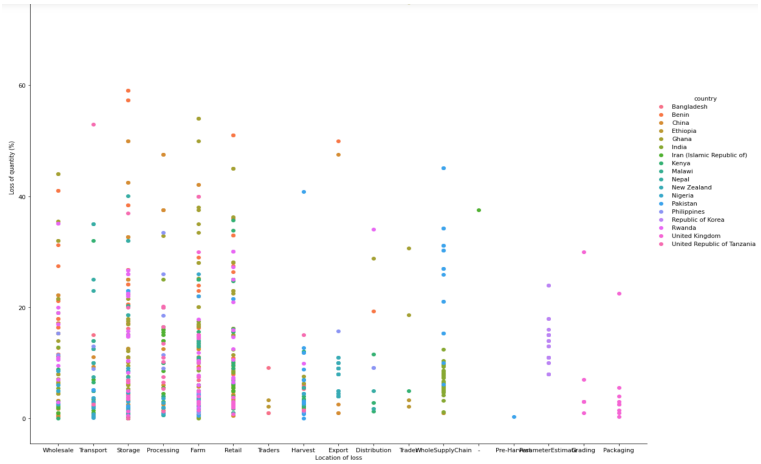


Figure 5: Loss of quantity(%) vs Stage/location in supply chain (by country)

The dot plot for Loss of quantity(%) vs location/stage in supply chain (colored by crop) (figure 6 on the next page) shows which stage in the supply chain sees high losses for which crop. There are high losses in the harvest stage for wheat and for potatoes in the transport and retail stages.

Note: We created other plots for the Loss of quantity(%) vs the time point/year and cause of loss. But these were difficult to analyze and display properly which is why we have left them out of our analyses.

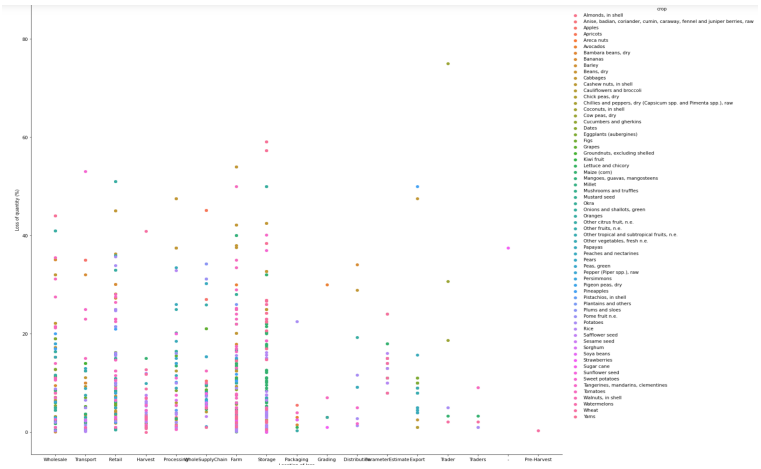


Figure 6: Loss of quantity(%) vs Stage/location in supply chain (by crop)

6.2 Evaluating the regression results

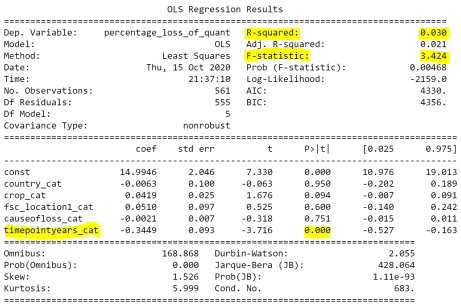
The regression outputs (figures 7-9) were used to evaluate the multivariate regression model. Figures 7-9 display many statistics and we have highlighted the 3 we used to evaluate the model. The 3 statistics are:

1. R-squared: it tells us how well our model fits the data,
2. F-statistic: indicates whether our linear regression model provides a better fit to the data than a model that contains no independent variables,
3. p-value: for each term the p-value tests the null hypothesis that the coefficient is equal to zero (no effect). Typically, we use the coefficient p-values to determine which terms to keep in the regression model.

Figure 7 (next page) is the output for the model on the train set. We can see that the R-squared and F-statistic values are very low. However, one of the features, timepointyears_cat, which the categorical variable for the year food loss occurred in is a statistically significant predictor of percentage loss of quantity. None of the other features were statistically significant.

In figure 8 (next page) we see how the model performed on the validation set. The R-squared and F-statistic values are lower than that for the train set and the feature, timepointyears_cat, is no longer statistically significant.

Due to lack of time and computational challenges we were not able to improve this through one hot encoding or introducing interaction terms. Thus, we decided to move forward and evaluate the model performance on the test set. In figure 9 (next page) we can see that the R-squared and F-statistic values are better than in both figures 7 and 8, but they are still low. Moreover, none of the features were statistically significant.



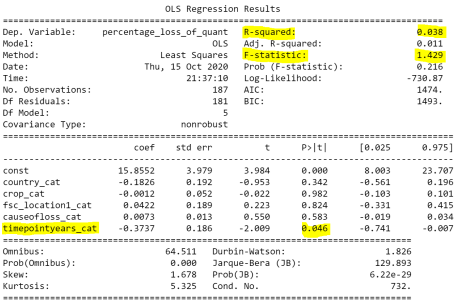
OLS Regression Results

Dep. Variable:	percentage_loss_of_quant	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.021
Method:	Least Squares	F-statistic:	3.424
Date:	Thu, 15 Oct 2020	Prob (F-statistic):	0.00468
Time:	21:37:10	Log-Likelihood:	-2159.0
No. Observations:	561	AIC:	4330.
Df Residuals:	555	BIC:	4356.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	14.9946	2.046	7.330	0.000	10.976	19.013
country_cat	-0.0063	0.100	-0.063	0.950	-0.202	0.189
crop_cat	0.0419	0.025	1.676	0.094	-0.007	0.091
fsc_location1_cat	0.0510	0.097	0.525	0.600	-0.140	0.242
causeofloss_cat	-0.0021	0.007	-0.318	0.751	-0.015	0.011
timepointyears_cat	-0.3449	0.093	-3.716	0.000	-0.527	-0.163

Omnibus:	168.868	Durbin-Watson:	2.055
Prob(Omnibus):	0.000	Jarque-Bera (JB):	428.064
Skew:	1.526	Prob(JB):	1.11e-93
Kurtosis:	5.999	Cond. No.	683.

Figure 7: Regression output for regression on the train set



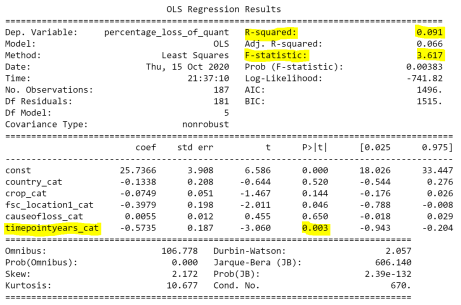
OLS Regression Results

Dep. Variable:	percentage_loss_of_quant	R-squared:	0.038
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	1.429
Date:	Thu, 15 Oct 2020	Prob (F-statistic):	0.216
Time:	21:37:10	Log-Likelihood:	-730.87
No. Observations:	187	AIC:	1474.
Df Residuals:	181	BIC:	1493.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.8552	3.979	3.984	0.000	8.003	23.707
country_cat	-0.1826	0.192	-0.953	0.342	-0.561	0.196
crop_cat	-0.0012	0.052	-0.022	0.982	-0.103	0.101
fsc_location1_cat	0.0422	0.189	0.223	0.824	-0.331	0.415
causeofloss_cat	0.0073	0.013	0.550	0.583	-0.019	0.034
timepointyears_cat	-0.3737	0.186	-2.009	0.046	-0.741	-0.007

Omnibus:	64.511	Durbin-Watson:	1.826
Prob(Omnibus):	0.000	Jarque-Bera (JB):	129.893
Skew:	1.678	Prob(JB):	6.22e-29
Kurtosis:	5.325	Cond. No.	732.

Figure 8: Regression output for regression on the test set



OLS Regression Results

Dep. Variable:	percentage_loss_of_quant	R-squared:	0.091
Model:	OLS	Adj. R-squared:	0.066
Method:	Least Squares	F-statistic:	3.617
Date:	Thu, 15 Oct 2020	Prob (F-statistic):	0.00383
Time:	21:37:10	Log-Likelihood:	-741.82
No. Observations:	187	AIC:	1496.
Df Residuals:	181	BIC:	1515.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	25.7366	3.908	6.586	0.000	18.026	33.447
country_cat	-0.1338	0.208	-0.644	0.520	-0.544	0.276
crop_cat	-0.0749	0.051	-1.467	0.144	-0.176	0.026
fsc_location1_cat	-0.3979	0.198	-2.011	0.046	-0.788	-0.008
causeofloss_cat	0.0055	0.012	0.455	0.650	-0.018	0.029
timepointyears_cat	-0.5735	0.187	-3.060	0.003	-0.943	-0.204

Omnibus:	106.778	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	606.140
Skew:	2.172	Prob(JB):	2.39e-132
Kurtosis:	10.677	Cond. No.	670.

Figure 9: Regression output for regression on the test set

7 Conclusions

7.1 Conclusions for exploratory plots (Figures 3-6)

Through the dot plots (figures 3-6) we were able to see trends emerge on how food loss for crop type, stage of supply chain and country are related to each other. An example for this is for food loss in Pakistan. In figure 3, we see that Pakistan had comparably higher losses than other countries (the maximum loss was about 48%). We then saw in figure 4 that Pakistan reported high food losses for the wheat crop. In figure 5, we see that Pakistan is one of the countries with multiple losses during the harvest, processing and transport stage but most notably in the harvest stage. Figure 6 then shows high losses for wheat during harvest.

From these four plots we can infer that Pakistan is facing the issue of food loss for wheat during the harvest stage of the supply chain. Similarly, we can make inferences on the specificity for food loss in other countries.

7.2 Conclusions for the regression outputs (Figures 7-9)

Based on the R-squared (0.091), F-statistic(3.617) and p-values (no feature was significant) for the regression output for the test set, we can conclude that our multivariate regression model is not a good predictor of percentage quantity of loss. The R-squared and F-statistic values indicate that the model does not fit the data well, and that the features have no impact on predicting the percentage quantity of loss.

7.3 Future directions

The features used were numerically categorical but with one hot encoding the regression results may improve. After one hot encoding there will be more features in the data set and that will make feature expansion easier. Feature expansion will likely improve how well the model will fit the data (a better R-squared value).

Another idea for such work can be to predict a categorical feature like cause of loss using multiclass classification instead of using percentage quantity of loss as the outcome.

7.4 Importance of this work

Work like this is useful for food loss forecasting and to make accurate predictions of where high food loss occurs to minimize it in a particular country, for a particular crop, and stage of the supply chain.

7.5 Lessons from this project

We learned that food loss and waste is a major problem not only in terms of global hunger and malnutrition, but it also has devastating environmental impacts.

On the coding side, we gained exposure to data cleaning, processing and visualization in python. This has improved our coding skills in the language and helped us understand what we should be looking at when starting a data-driven project in the future.

References

[1] Eloy Hontoria Alberto Garre, Mari Carmen Ruiz. Application of machine learning to support production planning of a food industry in the context of waste generation under uncertainty. 2020.

[2] Shane Strumwasser Felipe Caro Ally Kleinman, Keely Schneider. Eden: a new technology to reduce food waste in walmart's supply chain. 2018.

[3] Hans Gelke Amin Mazloumian, Matthias Rosenthal. Deep learning for classifying food waste. 2020.

[4] Anonymous Authors. Towards a sustainable food supply chain powered by artificial intelligence. 2018.

[5] Food and Agricultural Organization of the United Nations. Food loss and waste database.

[6] Shyam Purkayastha. Applying ai in food industry to streamline supply chain workflow. 2020.

[7] P.B. Sharma Rajneesh Mahajan, Suresh Garg. An illustration of logistic regression technique: A case of processed food sector. 2014.

[8] WWF. Wasting food is not an option. 2019.