Total Recall (Samuel Boardman, Khola Jamshad, Riku Kurama, Yucong Lei, Shivani Prabala)

Github Repository

Introduction

Recent workforce reductions at the FDA and the subsequent suspension of a quality-control program at its food-testing laboratories have elucidated the need for an efficient contamination tracking of food products. Our aim is to predict food products likely to be recalled using past inspection data and also pathogen related outbreaks.

Our primary company KPIs are as follows:

- Prevention rate of contaminated products reaching consumers
- Cost savings to manufacturers from early warnings and avoided recalls

Dataset

Our primary dataset is of FDA recalls, which we filter for food product type and remove miscellaneous features mostly relating to company descriptions. We append a separate FDA inspection dataset to provide a counterbalance of non-recalled food products. Entries in both datasets are matched using their FEI (FDA assigned identifying numbers), _recall bool_ is used to label any entry in both datasets as recalled (1) and the remaining as not recalled (0). We engineered a count of the cities mentioned as a new feature and one hot encoded all categorical variables. The data was imbalanced, with a lot more non-recalled products present in our combined dataset than recalled products.

For food related outbreaks, we used the CDC database BEAM, which records various disease outbreaks with different pathogen types from 2018 to 2025.

Methods and Results

For predicting recalls, the DummyClassifier is our baseline model which always predicts the majority class, in this case 0 for not recalled. This gives a 95% accuracy due to data being heavily not recalled but the F1 score was close to 0.

We trained a variety of Random Forest (RF) models, which aggregate the predictions made by many decision trees, using cross validation. Because the ramifications of missing a product that is going to be recalled are so severe, our primary metric was the percentage of recalled products that the model could successfully label as such (also known as *recall*). We tried many combinations of random forest hyperparameters; the best combination attained a recall score of around 88%, at the expense of the accuracy dropping to about 54%, indicating close to half of non-recalled products being mislabelled.

We also trained SVM models for classification and optimized them using GridSearchCSV. However, since SVM did not scale well with large data, they ran extremely slowly. We attempted to counter this by reducing the dataset to only companies that had faced recalls before and while

that improved F1 score to 72%, we were concerned that it introduced an unfavorable bias. Therefore, we continued with the unreduced data.

We compared the 4 models using pipelines before and after we optimized them for predicting recalled products using balanced class weights and hyperparameter tuning. We used the F1 score as our metric to take into account both recall metric and precision metric. Pre-tuning, only RF had any success at predicting recalls but post-tuning, both RF and LR performed well though RF was consistently better at predicting recalls from F1 score hitting a high at 25%. Therefore, we chose optimized RF as our final model.

Another task we carried out was the time series analysis for connecting pathogen related outbreaks and food recalls using a naive seasonal model and a VAR model. The naive seasonal model predicts the average recall percentage of the same month over the past 5 years. For the VAR model, we combined the FDA and the BEAM datasets to better predict future recall probabilities; however, this forced us to truncate our available dataset to the timeframe from 2018 to 2025. This VAR model performed worse than the naive model, so we also added some seasonality to the final VAR model by including the variable that records the average recall percentage in the same months over the past three years. According to our cross validation, this VAR model with added seasonality performed roughly as well as the naive seasonal model.

Future Goals

We expect that we can improve our models by including spatial data, climate over time, and reasons for recall from test.csv in our model. Additionally, implementing a regularized VAR model with seasonality appears beneficial since our VAR model has many parameters due the number of variables and lags. Keeping reasons for recall can also give us a better idea of the factors that cause recall which may assist in developing preventative measures.