Reducing Small Business Food Waste: Analyzing Bagel Sales to Predict Inventory Shortages

Molly Lipman[1], Jordan Lueck[2], Mike Pine[3], and Bryce Walsh[4]

[1]*Mechanical Engineering Department, Tufts University. Email: molly.lipman@tufts.edu*
[2]*Mechanical Engineering Department, Tufts University. Email: jordan.lueck@tufts.edu*
[3]*Computer Science Department, Tufts University. Email: michael.pine@tufts.edu*
[4]*Computer Science Department, Tufts University. Email: bryce.walsh@tufts.edu*

Abstract

Food waste in food and beverage industries continues to be a prominent sustainability issue facing society. This is perhaps most evident with small businesses, whose operating margins often do not allow for advanced inventory management software and services. In this paper we describe various data science techniques that can be employed by small businesses dealing in food or food service to clean and analyze their sales data. Applying such methods can help these companies better predict sale quantities of their products in order to reduce inventory shortages and minimize excess food production. We demonstrate the application of these techniques through a case study on a newly-opened bagel store in the Greater Boston area. Learning models discussed include lasso regression, support vector machines, random forest and extremely random forest classification. We conclude by explaining how to interpret our model results and apply findings to business practices on a manageable scale.

Keywords: food waste, inventory management, machine learning, data science, data cleaning, random forest

1 Introduction

Roughly 1.3 billion tons of edible food is lost or wasted each year, equal to about one third of all food produced for human consumption (Gustavsson *et al.* 2011). This wasted food not only causes social, ethical, and economic issues, but also has a significant negative environmental impact. A case study of six Swedish supermarkets in 2014 found that over the course of 4 years, a total of 1570 tons of fresh food was wasted, causing a calculated carbon footprint of 2500 tons CO2 equivalent (Scholz, Eriksson, and Strid 2015). If we take this same ratio of 2500tCO2e/1570t food wasted and apply it to total global food waste, we get a figure of roughly 2 billion tons of CO2 equivalent contributed to climate change each year from producing, transporting, and disposing of food that is never consumed.

A major contributor to global food waste on the distribution side of the supply chain is inaccurate sales forecasting by retailers (Mena, Adenso-Diaz, and Yurt 2011). If a store poorly estimates how much of a particular item will be sold they run the risk of either suffering economic losses due to not being able to meet demand or contributing to the above problem of food waste. Successfully predicting sales, however, can help a business prevent under- and over-stocking, thus improving both their economic and environmental sustainability (Tsoumakas 2019). We focus on small business sales prediction in this paper, as many small businesses currently employ only human intuition when deciding how much of a certain item to produce or stock, and thus have much to gain from accurate sales forecasting.

Given the multitude of factors that influence small business sales, creating accurate sales models is non-trivial. In this paper, we explore different models in order to predict the sales of bagels for a small newly-opened bagel store in the greater Boston Area. Because the business

Data Science For Sustainability (2020)

Corresponding author
Molly Lipman

hadn't been open a year at the time of the analysis, the project was confined by a small amount of data and an inability to look at seasonal or long term trends. Additionally, because of the business' popularity, they sold out frequently which limits the data's ability to characterize true potential.

1.1 Relevant Works

Food sales prediction is an area of study that has been extensively researched. Many statistical techniques have been applied to this task. Seasonal ARIMA with external variables has been used to predict daily banana sales at a retail outlet (Arunraj and Ahrens 2015), a hybrid moving average system has been used to predict wholesale food sales on a weekly basis (Žliobaitė, Bakker, and Pechenizkiy 2012), and an LSTM network has been successful in using weather data to predict drink sales in a Japanese supermarket chain (Liu and Ichise 2017). Zioblaite's publication used 7 categories of predictors: events, weather, seasonality, price, substitution and cannibalisation, product characteristics, and number of customer visits. Within those categories, they looked at more specific predictors like school vacations or holidays within the event category. Based on the findings of this publication, we explored using holidays, school vacation, air temperature, precipitation, day of the week, and special visitors as predictors. Because of our limited data sources, it was not feasible to include seasonality, price, or substitution and cannibalisation. We did not have any available data on the type of visitors that came, but we did include a feature that kept track of time since the business was featured in a news source, which was intended to be a proxi-variable to model a special customer coming in because of the news source.

In addition, many of the models and algorithms used in these papers were applied to predicting sales at retail or wholesale companies which have been around for a long time and have a vast amount of data available. Several also predict on a weekly rather than daily basis.

By contrast, we look to apply these predictors to a small business scenario by using different models and time resolution of data. For many small businesses that deal in fresh food (those that make fresh made baked goods, for example) weekly predictions are unhelpful as goods must be made in the morning or the night before and sold by the end of the day or else go to waste. Furthermore, small food businesses have a relatively smaller customer base than larger retail and wholesale food stores and thus are forced to deal with a smaller pool of data and often higher variability in daily sales. We attempt to address these issues in our case study with a local bagel company which first opened in October of last year.

2 Methods

2.1 Data Collection & Cleaning

The data came directly from the small business' internal data collection. The software used to collect that information is also used for ticketing during store hours, so all items and modifiers (e.g. 'small' regarding a coffee or 'poppy' regarding a bagel) were included. The main step of cleaning the data was to compile bagel totals for each time step. Originally, this was done with a daily resolution but was then adjusted to an hourly resolution in order to increase the quantity of data. The next cleaning process was to remove outliers or data that were not relevant to the analysis. Any transaction that did not occur during normal store operations was excluded. This was mainly transactions before the business opened on September 13th, 2019, transactions in the evening after the business had closed for the day, and transactions that took place on days the store is not open (Monday and Tuesday). These transactions were likely in the original data set because they were being used to test the software that was being used to compile the transactions.

2.2 Determining Features

The primary method of obtaining possible predictors was to talk to the owners of the small business. With experience running the business, they had anecdotal evidence of the importance of certain factors. Before deciding the quantity of bagels to make for each day they take day of the week, temperature, precipitation (rain or snow), holidays, professional sports games, and school schedules into consideration. Additionally, time since the business was featured in a publication was added. This was incorporated with a number describing days since the last publication. The importance of weather, school schedule (Kirchner *et al.* 2015), holidays, and specialty customer visits (Arunraj and Ahrens 2015) have all been shown to be important in other works examining how to predict sales.

2.3 Model Selection

After determining predictors and cleaning the data, it is then important to choose models that can capture the relationships in the data to best predict for this supervised application. We tested multiple models to not only convey how a model performs but also provide context in relation to others. The four main models implemented were ordinary least squares lasso regression, support vector machines, normal random forests, and extremely random forests.

Lasso Regression

There were two main reasons why lasso regression was utilized. The first is that lasso regression is very simple, fast, and easy to interpret. This means that if it is able to capture the relationship then it should be used and further more expensive models can be deemed unnecessary. The more likely outcome is that lasso regression will not be robust enough to model the data, but instead can act as a good benchmark to show comparisons to the more complex models. Lasso regression is a variant of ordinary least squares regression that introduces a regularization term. It specifically uses L1 regularization which takes the absolute value of the magnitude of the coefficients multiplies it by a constant term and adds that product as a penalty term. As the regularization constant increases the penalty term increases and then subsequent learned coefficients will go to zero.

Support Vector Machines

The second model trained in this experiment was SVMs (support vector machines). SVMs are another linear model with respect to the weights learned and have been used in other papers to predict sales (Vélez *et al.* 2020). One major benefit of SVMs is that they are much more flexible and able to model non-linear boundaries between data points with the kernel trick which allows for more powerful regression. However a drawback of this model is that it is not robust to outliers and can be heavily influenced by a few data points. One of the ways SVMs are altered to account for this is the soft-margin which allows for outliers, specifically ones that inhibit a seperating hyperplane, to be assigned a scaled penalty but not completely affect placements of support vectors. This is called regularization and in order to achieve the optimal SVM model it is important to tune this hyperparameter. The hyperparameters being tuned in this model as a result are the regularization term and the tolerance term used to control penalty of terms far from the decision boundary. This model was tuned using a grid search along with 5-fold cross validation and then averaging across folds to measure performance for each value of the hyperparameter. The kernel function used was the radial basis function (RBF) kernel, because it is the most commonly used and was used in similar applications to predict sales (Vélez *et al.* 2020).

Random Forest

The third model used was random forests. random forests are an ensemble model that aggregates many decision trees in order to control for overfitting and obtain better generalizability across the whole dataset (Zhang *et al.* 2019). The reason to use a decision tree is high interpretability, easy visualizations, and good performance with features that are correlated. However, what makes a random forest difficult to interpret is that it isn't a single tree where you can see each decision (branch) but instead an ensemble of thousands of trees. This ultimately is the main reason why aggregating these decision trees produces better predictions. Because like decisions trees, random forest performs well with correlated features, it was decided to be used in this experiment. The features selected such as school schedule and weather are not completely independent and this model would do well when splitting handling these features. It also can be used to do feature selection and identifying feature importance, in case some features are not contributing to the performance or even cause worse performance.

Extremely Random Forest

The fourth model implemented was a more specific type of random forest. This extremely random forest was used to better handle overfitting. Unlike normal random forests, an extremely random forest does not find an optimal split based on the random subset of features, it makes a random split given a random subset of features. The reason this model was chosen was because this experiment does have a significantly limited dataset and as a result it was important to take measures in the event of overfitting.

Ultimately random forest was used for two main reasons. It handles correlated features better than linear models, and it can be used to determine feature importance. Due to the increased complexity of the model and ability for it to overfit to the already limited dataset, hyperparameter tuning was done. The main hyperparameters vary with random forest models are the number of features considered at each junction and the number of trees in the forest. Limiting the number of features considered at each junction helps prevent overfitting, by randomly choosing which ones to consider. A similar grid search cross validation pipeline was created to find the optimal values. This was done for both the normal random forest as well as the extremely random forest.

These four types of models were used to observe relationships existing in the data and compare with each other.

3 Results & Analysis

In order to evaluate the performance of each model independently and in relation to each other, an error metric was needed. The one used was MSE (mean squared error). The reason for this is because unlike with MAE (mean absolute error), large outlier errors have a greater influence on the MSE. When predicting bagel sales, it is important to note that worse predictions for a given day waste more resources and lead to less profits. The company is better equipped to handle small and distributed daily over- or under-productions of bagels than a single day of seriously erroneous sales predictions. It only seemed reasonable therefore to use MSE when evaluating model performance.

The first model evaluated was lasso regression. Lasso regression only requires hyperparameter tuning for the regularization term. This model selection can be seen in figure 1. Once model selection was performed, the optimal model was run on the test set. The performance is located in Table 1 along with the other models' performance.

The next model attempted was SVMs (support vector machines). Specifically kernelized support vector machines were used on the training data. The different hyperparameters tuned were the

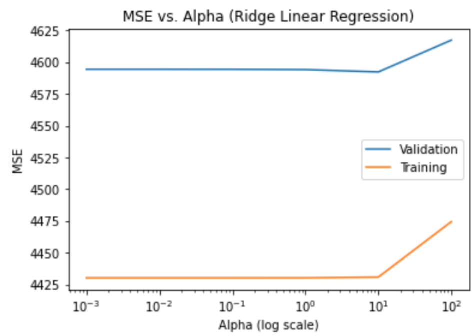
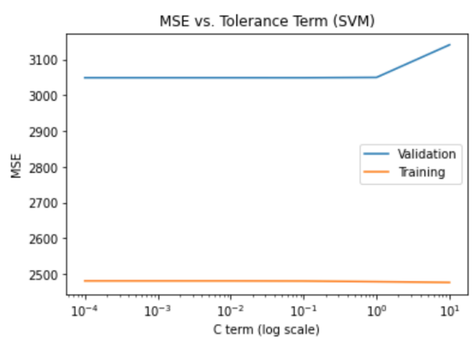
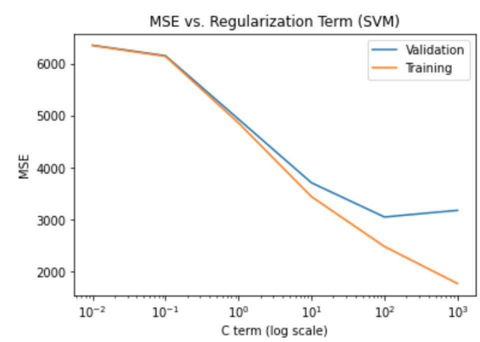


Figure 1. Hyperparameter tuning for Lasso Linear Regression

tolerance for the soft margin in penalizing data ϵ, and the regularization term C. Both the training and validation errors are averaged across the 5 folds and plotted in figures 2a and 2b respectively.



(a) Support vector machines with 5 fold Cross Validation averaged across folds over each value in a Grid Search of ϵ values

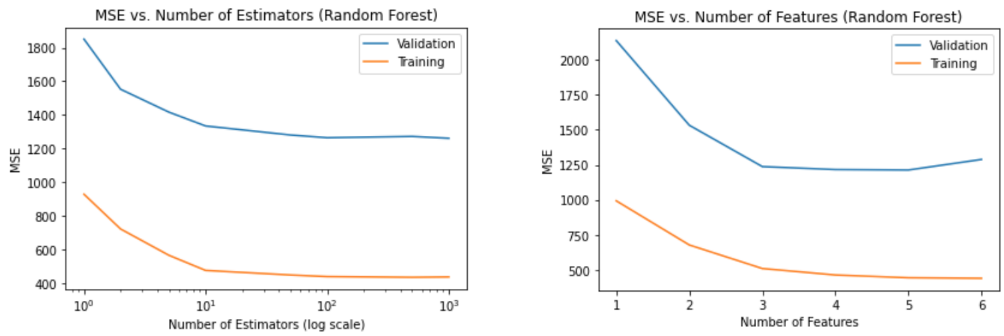


(b) Support vector machines with 5 fold Cross Validation averaged across folds over each value in a Grid Search of C values

Figure 2. Plots of Hyperparameter Tuning for Kernelized SVMs

The optimal values were found to be C equals 100 and ϵ equal to any value less than 10 with the best performance having a validation set mean squared error around 3000. The C value was therefore incorporated into the tuned SVM model, while ϵ was kept as its default value.

In contrast to the linear models, random forest was implemented with the hope that it would perform the best due to the potentially correlated features. The number of features considered at each junction and the number of trees in the forest were both grid searched in order to identify the optimal values of each. Figure 3a and 3b convey the behavior across the searched values for both the validation and training sets against a varying number of estimators (trees) and the maximum number of features considered, respectively. Both were analyzed using 5-fold cross validation.



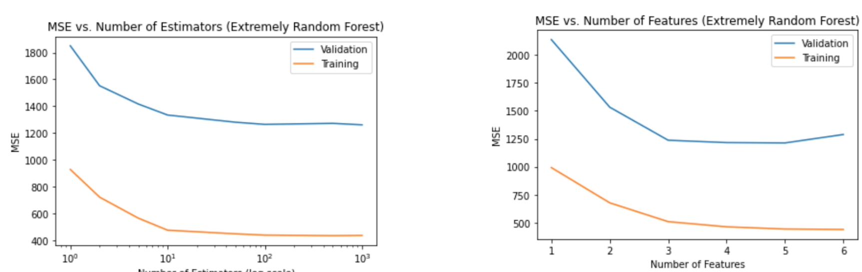
(a) Random forest with Cross validation to find the optimal number of trees in the forest

(b) Random forest with Cross validation to find the optimal number of features considered at each junction

Figure 3. Plots of Hyperparameter Tuning for random forest

The optimal number of trees was 100 and the optimal number of features considered at each junction was five. When six features were considered in random forest the validation error increased while the training set continued to decrease indicating at that point random forest was overfitting to the training set. After 100 trees, it appears that the number of trees added does not lead to better performance. Adding more trees would significantly increase the computation cost without a noticeable change in the error. That is why 100 appears to be the optimal number of trees where each tree provides utility to the model performance.

Due to the limited dataset, extremely random forests were also used to gauge performance and address overfitting concerns. What was found was that overfitting did not seem to be as large a concern as expected. The extremely random forests performed similar to the random forests. The hyperparameters tuned were the same as the random forest models. Figure 4a and 4b show the number of estimators and features respectively for a grid search over both the validation and training sets.

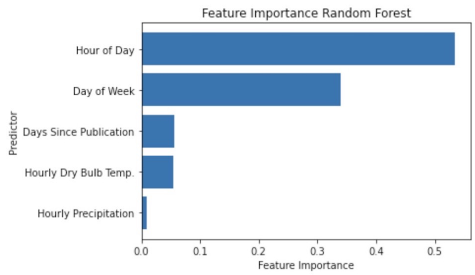


(a) Extremely random forest with Cross validation to find the optimal number of trees in the forest

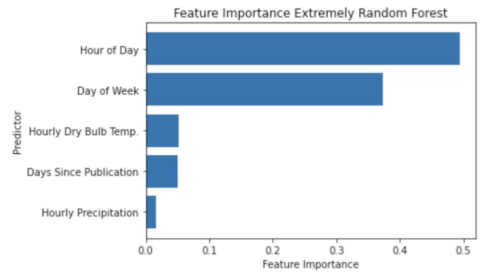
(b) Extremely random forest with Cross validation to find the optimal number of features considered at each junction

Figure 4. Plots of Hyperparameter Tuning for extremely random forest

Figure 4 is nearly identical to Figure 3 conveying that the performance is similar. While this makes it difficult to predict which model will perform better on the test set, because trees in extremely random forest have much less correlation it is expected they have a larger bias and will perform worse than normal random forest.



(a) Random forest Feature Importance.



(b) Extremely random forest Feature Importance.

Figure 5. Plots of Feature Importance for normal random forest and extremely random forest. These values represent the fraction of data points whose prediction was correlated with each feature.

The feature importance charts indicate that hourly precipitation had the smallest impact, and the most important features were hour in the day and day of the week. This aligns with intuition as it seems likely that certain days and times would see more bagel sales than others.

Once there was confidence that the models were not significantly overfitting to the training set, the optimal hyperparameter configurations were run on the test set and these values are reported in Table 1. The training set here included both the training and validation sets used for cross validation, consisting of roughly 90% of all data points. This was done to provide as much data as possible to improve the prediction results on the test set.

Model	Training MSE	Test MSE
Ridge Linear Regression	4446.50	6767.56
Support Vector Machines	2442.31	3031.71
Random Forest	447.32	935.61
Extremely Random Forest	703.08	988.13

Table 1. Summary of Model Performance using MSE as the error metric

Table 1 demonstrates that even the best model, random forest, was unable to achieve a reasonably low MSE. Put in context, an MSE of 935 means that the model can predict bagel sales per hour within 30 bagels of the true value. This, translated to a daily prediction, indicates that the number of bagels in a five hour work day would be off by ±150 bagels. An analysis of the bagel store's average daily sales shows as few as 450 bagels are sold on some weekdays, and as many as 1100 over the weekend. Thus a prediction error of ±150 bagels represents anywhere between 14% and 33% of expected daily sales. Given that there is a temporal bias in our data, this is a best case prediction margin and is not likely narrow enough so as to provide a reliable estimate of how many bagels will actually be sold. An interesting note is that extremely random forests have a training error much closer to the test error most likely due to its inherent high bias low variance. Based on these results, it would make sense to use the extremely random forest as the better model because the normal random forest performance on the training set is not as good an estimate of how well the model will perform on the test set.

4 Discussion

Unfortunately, a prediction accurate to within 150 bagels per day is not close enough to be useful to this particular business. Even though they do not keep track of exactly how many bagels are left over each day the business owners expressed that 150 is more than they have left on a typical day . This process, however, can be useful for future analysis that the business can do with larger amounts of data. Specifically, having multiple full calendar years could allow the analysis to take seasonal trends into consideration. Because the business' first year of operations contained the outbreak of the COVID-19 virus, multiple years of data would be necessary to see unaffected seasonal trends. The business restructured its entire operation in order to adjust to the global pandemic, therefore sales during that time would not have been able to contribute much insight into normal sale trends.

We have provided the business with our script for cleaning transaction data into a usable format. Given that they were previously not using their data at all, this can still be a meaningful contribution in allowing them to use their data to grow their business.

4.1 Further Work

It is potentially viable for these methods to be used in this context once more data is obtained. We have asked the business to start keeping track of how many bagels are left over on days when they don't sell out and to continue to log transaction data. Once the company has a full year of data or more it would be worth revisiting this project with similar methods. A time-series analysis could also be useful once there is enough data to run analysis at the daily level.

5 Acknowledgements

We want to thank Professor Deborah Sunter for her instruction and guidance throughout both this research and the course as a whole. Prof. Sunter provided invaluable input and resources in the initial stages of the study, and continued to support our work through completion. We'd also like to thank the other students in Prof. Sunter's course for their recommendations and feedback at multiple stages of the study's development. Finally, we'd like to the thank the employees and management of the Boston-based bagel store that served as the source of the data analyzed in this paper.

References

Arunraj, N. S., and D. Ahrens. 2015. "A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting." *International Journal of Production Economics* 170:321–335. ISSN: 0925-5273. doi:https://doi.org/10.1016/j.ijpe.2015.09.039. http://www.sciencedirect.com/science/article/pii/S0925527315003783.

Gustavsson, J., C. Cederberg, U. Sonesson, R. Van Otterdijk, and A. Meybeck. 2011. *Global food losses and food waste.*

Kirchner, T. R., A. C. Villanti, J. Cantrell, A. Anesetti-Rothermel, O. Ganz, K. P. Conway, D. M. Vallone, and D. B. Abrams. 2015. "Tobacco retail outlet advertising practices and proximity to schools, parks and public housing affect Synar underage sales violations in Washington, DC." *Tobacco Control* 24 (e1): e52–e58. ISSN: 0964-4563. doi:10.1136/tobaccocontrol-2013-051239. eprint: https://tobaccocontrol.bmj.com/content/24/e1/e52.full.pdf. https://tobaccocontrol.bmj.com/content/24/e1/e52.

Liu, X., and R. Ichise. 2017. "Food Sales Prediction with Meteorological Data — A Case Study of a Japanese Chain Supermarket" [in en]. In *Data Mining and Big Data,* edited by Y. Tan, H. Takagi, and Y. Shi, 93–104. Lecture Notes in Computer Science. Cham: Springer International Publishing. ISBN: 978-3-319-61845-6. doi:10.1007/978-3-319-61845-6_10.

Mena, C., B. Adenso-Diaz, and O. Yurt. 2011. "The causes of food waste in the supplier–retailer interface: Evidences from the UK and Spain." *Resources, Conservation and Recycling* 55 (6): 648–658.

Scholz, K., M. Eriksson, and I. Strid. 2015. "Carbon footprint of supermarket food waste." *Resources, Conservation and Recycling* 94:56–65.

Tsoumakas, G. 2019. "A survey of machine learning techniques for food sales prediction" [in en]. *Artificial Intelligence Review* 52, no. 1 (June): 441–447. ISSN: 0269-2821, 1573-7462, accessed April 30, 2020. doi:10.1007/s10462-018-9637-z. http://link.springer.com/10.1007/s10462-018-9637-z.

Vélez, D., A. Ayuso, C. Perales-González, and J. T. Rodríguez. 2020. "Churn and Net Promoter Score forecasting for business decision-making through a new stepwise regression methodology." *Knowledge-Based Systems* 196:105762. ISSN: 0950-7051. doi:https://doi.org/10.1016/j.knosys.2020.105762. http://www.sciencedirect.com/science/article/pii/S0950705120301684.

Zhang, Y., X. Wu, C. Gu, and Y. Xie. 2019. "Predict Future Sales using Ensembled Random Forests." *CoRR* abs/1904.09031. arXiv: 1904.09031. http://arxiv.org/abs/1904.09031.

Žliobaitė, I., J. Bakker, and M. Pechenizkiy. 2012. "Beating the baseline prediction in food sales: How intelligent an intelligent predictor is?" [In en]. *Expert Systems with Applications* 39, no. 1 (January): 806–815. ISSN: 0957-4174, accessed May 1, 2020. doi:10.1016/j.eswa.2011.07.078. http://www.sciencedirect.com/science/article/pii/S0957417411010402.