Elizabeth McHugh
Math 342W / 650.4
Prof. Adam Kapelner
21 March 2021

Local Café Daily Sales in Brooklyn Is Somewhat Predictable

In New York City, it is incredibly common to walk into a locally owned, small food service establishment to find one of two scenarios—either there are several unoccupied employees standing around, chatting about the weather or condemning the MTA, or else there are no more than two employees running around like crazy, trying to juggle three tasks at once while hoping you'll stick around long enough to order and pay for the cup of coffee they know you came in for. This is particularly the case in residential, commuter neighborhoods, where business seems to wax and wane from day to day with no easily discernable pattern. It would be easy for a local café or restaurant owner to be led to believe, after a reasonable amount of time for observation, that such business must truly be unpredictable, and that they will just have to absorb the costly consequences that come with the inability to predict the next day's business to any degree of accuracy. However, this essay will propose the feasibility of building a predictive model of future business at a local café based upon a discussion of basic data science principles which utilize easily collectable data, showing that such business actually is, at least partially, predictable.

Introduction: Need for Prediction

According to the City of New York's "Small Business First" report, eighty-nine percent of businesses in New York City are considered very small, that is, businesses with fewer than twenty employees. The report also notes that food service industry in New York City is almost entirely composed of very small businesses (City of New York, 2014, p. 9).

In a company of more than a hundred employees (that is, not a “small business”), having one extra employee on the clock or being one employee short in a given shift on a given day may not have huge consequences for the business owners. However, in a business with less than twenty employees (a “very small” business), inaccurately predicting the number of employees needed for a given shift can hold much higher stakes. This is particularly true in the food service industry, where customers tend to come in waves and often do not have an extended time to wait for a cup of coffee on the way to work or to pick up their office’s lunch order while they’re on break. In the case of small local coffee shops and cafes: One employee too many, and the business may lose a large percentage of daily net sales to the extra employee’s salary; one employee too few, and the business risks losing sales from customers who are unwilling to wait an extra ten minutes to place an order or else losing customers completely when their desired orders are unable to be met in a timely manner. It is easy to see that the magnitude of loss experienced by very small local restaurants, cafes and coffee shops becomes even greater, the smaller the business is.

Having an available model which could not only predict future business but also do so using data which the business owners could easily collect and store could have a major positive impact on a business owner’s ability to properly appropriate employee shifts, thus maximizing net profits (that is, the amount the business takes in, after discounting all business-related expenses, including employee salaries) as well as net sales (that is, the total amount of sales after discounting any returns, refunds or customer credits). One important note to mention is that such a predictive model must be highly individualized to the specific local eatery, as, for instance, two nearly identical local coffee shops, located only a block apart may easily have completely different levels of business on a given day, under nearly identical conditions. Thus, this essay will focus on one specific small local café, which will be referred to as G's Café from here on, located in a dense, residential immigrant neighborhood in Brooklyn, New York. However, the same modeling and prediction process should apply to any such small, local eatery, though the factors in consideration will undoubtedly differ.

Building a Model for G's Café: The Basics

Before we discuss the specifics of building a predictive model for G's Café, we must make sure to begin with a common understanding of the terms and notation we will be frequenting. To begin with, though the term has already been used multiple times in the introduction, what exactly is a "model" as discussed in this essay? A *model* is simply an approximation to some real *phenomena*, or natural occurrence. In the case of G's Café, a *mathematical model*--a mathematical object which quantifies observations of some phenomena--is what is needed. Also, what is *prediction*? This seems to be a common term, but when it comes to modeling, prediction is the ability to state, in advance, what outcome of a given phenomenon should be expected, given a specific set of factors.

As in any model, the true causes (called *causal drivers*) of the phenomena we wish to predict are unknown, as is the function which can accurately predict the phenomena. Nevertheless, we denote the causal drivers as $z_1, z_2, \dots, z_n$, and the true function which would produce a perfect model for our phenomenon we denote as $y = t(z_1, \dots, z_n)$, where *y* is the phenomena being predicted, itself. While it is impossible to find $t(z_1, \dots, z_n)$, it *is* possible to approximate t by finding a model which will fit our phenomena well enough to be able to predict future occurrences of y to a reasonable degree of accuracy. This model, which we will call

$g(x_1, \ldots, x_p)$ will take in a list of factors denoted $x_1, x_2, \ldots, x_p$, where p denotes the number of factors whose relationships the model will take into account. While $g(x_1, \ldots, x_p)$ approximates the function t to a reasonable degree, there is a degree of error inherent in any model. This error will be discussed later, as it presents.

Building a Model for G's Café: The Set Up

In the case of G's Café, the phenomenon to be modelled is the amount of business on any given day. While the amount of business could be measured in various terms, including the number of transactions, the number of customers, the number of deliveries, or the daily gross or net sales, the chosen measure of "amount of business" will be defined as $y \coloneqq$ $daily\ gross\ sales\ in\ USD.$ Daily gross sales at G's Café is measured as the sum of all incoming sales transactions, whether via cash, gift card, or electronic payment, in US dollars, completed within a given calendar day. While this metric, previously, was not very accurate for G's Café, as the cash and card sales were recorded on different schedules, and sometimes not recorded properly at all, G's Café's current records (within the past year) are consistently recorded on a daily basis, thus allowing for accurate metrics of the phenomena being measured. For the purposes of G's Café, where the price of an item approximately correlates to the amount of labor required and several individual sales transactions are often entered into the register as one transaction, gross sales is a more appropriate metric than that of the number of customers served or other measures considered. The set of outputs for our predictive model is $y \in \mathcal{Y} = \{\mathbb{R}\}$. (In reality, this should likely be the set of non-negative real numbers, as 0 gross sales is possible, whereas any day with negative gross sales recorded should be excluded from the data.)

One of the most important decisions in modeling to predict daily sales at G’s Café is *feature* selection, or choosing which factors to use to represent the causal drivers of business. Since G’s Café is a very small business, there is not much room for time-costly data collection or data analysis which relies upon collection of additional customer data, so the factors taken into consideration must be readily available or easy to quickly compile. G’s Café’s owners have their own suspicions about what drives their sales. They believe that the true $z_i's$ which affect daily sales include marketing communications, social media, the weather, and whether or not schools are in session. Most of these suspected causal drivers are quite easy to approximate and measure and have readily accessible quantified data. However, there are also other factors not mentioned which are known to affect daily sales at businesses in general, such as business hours, days of the week, month of the year, and holidays.

Undoubtedly, there are countless other factors (a large set of causal drivers) which feed into the daily sales at G’s Café but are impossible to know, or if possible to know impossible to measure. As an example of a causal driver which could be known but not accurately measured, there is little to know doubt that the amount of a customer’s hunger or thirst affects daily sales, as does the attitude and demeanor of the one taking their order; however, neither of these drivers could reasonably be quantified, even if they were known, as the effect and any attempt and measurement would be quite subjective and without clear, consistent metrics.

As well, it should be noted that some causal drivers, even if they were known and measured and trained may not be practical for prediction, as the future value of the factor is impossible to know. Even if the exact effect of a tornado on daily sales was known and accounted for in the model, it would be impossible to use “tornado” as a factor in prediction, since today one cannot say with certainty that there will be a tornado during business hours next

Monday. Thus, some factors which might be useful for training a good model may not be useful for predicting.

While finding the causal driver of any given sale would be impossible, we will attempt to approximate them the best we can with the most accessible data possible. We will do so by selecting the *factors*, or variables which approximate or attempt to measure the causal drivers, which will be used to train a model to predict future sales. It is time to define and quantify the specific factors x_i which will be used in the model for G’s Café. So:

Let $x_1 =: number\ of\ social\ media\ views\ in\ the\ previous\ 24\ hours$ *prior to closing*. This metric might not be the easiest to measure and slightly unreasonable for many small local cafes or restaurants, however G’s Café’s owners check this metric via their social media accounts’ dashboards daily at closing, so this data is readily available, already quantified, and easy to record.

Let $x_2 := marketing\ communication\ in\ 48\ hours\ prior$. x_2 will be quantified as a binary factor, with 1 denoting that some marketing communication was sent in the prior 48 hours. This metric might seem odd. However, for G’s Café, “marketing communications” consist of newsletters, coupon codes, or special holiday notes emailed to customers who have signed up for G’s Café’s mailing list, and the owners have noted that both customers presenting coupon codes from marketing communications and traffic to their online ordering platform from the emailed communications occur almost exclusively in the first two days after sending marketing emails.

Let $x_3 \coloneqq the\ number\ of\ hours\ before\ noon\ that\ G's\ Cafe\ opened$. And, let $x_4 \coloneqq the\ number\ of\ hours\ after\ noon\ that\ G's\ Cafe\ closed$. Since G’s Café has varied business hours, depending on various factors, it is important to capture this data in the model we seek. While it would be simpler to quantify the effect of business hours as simply the number of hours of operation in a given day, this would leave out very important components of the business—in fact it would leave out the very nature of the business as a small café with few employees. If, due to staffing considerations, modified holiday operations, or simply variations in days of the week, G's were to open for business two hours later one Monday (say, at 10 am) than it did the next Monday (at 8 am), it is reasonable to assume that the café missed a large portion of the breakfast and coffee crowd on the first Monday which it was able to capture on the next Monday, thus lowering overall sales for the day. This effect undoubtedly needs to be captured in any model seeking to predict future sales. Using noon as a reference time is mostly arbitrary, but since G's always opens before noon and closes after noon, such metric would seem to allow for a positive association between x_3 and y , and G's Café is dedicated to spreading positivity (it is in their mission statement).

Let $x_5 \coloneqq day\ of\ the\ week$. As with most local businesses in this residential area, Friday through Sunday afternoon are the busiest days of the week.

Let $x_6 \coloneqq month\ of\ the\ year.$

Let $x_7 \coloneqq average\ of\ high\ and\ low\ temperatures\ for\ the\ day$. It should be noted that, to an extent, x_6 and x_7 should be expected to be colinear. However, it is expected that only using one of the two features would not take into account the effect of occurrences such as an unseasonably warm November, when there would be more local foot traffic than usual, or an unseasonably cold spring, when coffee sales might be higher in number but less profitable than the higher grossing smoothies sales which typical begin in mid-Spring. Thus, it is reasonable to attempt modeling with both factors, and in the case that it is found that both factors are not needed, the most costly factor to the data recorder or business (x_7) should be dropped, as x_6 is a factor which would be recorded regardless of whether or not anyone was concerned with modeling the business' sales.

Let $x_8 := holiday$, where "holiday" is defined in this case as: any holiday for which schools or government organizations are closed; any major Muslim, Jewish, or Christian religious holidays (chosen specifically given the demographics of the neighborhood and G's typical customer base); and Chinese New Year. Here, $x_8 = 1$ if the day is a holiday with positive impact on business, $x_8 = -1$ if the day is a holiday with negative impact on business, and $x_8 = 0$ otherwise. G's Café's owners suspect such holidays affect sales, but are unsure of to what extent, so this factor is included.

Finally, let $x_9 \coloneqq number\ of\ days\ since\ opening$. This metric should allow for variation of sales due to the movement of time. For instance, over time, inflation happens, new return business potentially increases, old business potentially decreases, and other time-sensitive variations occur which could be accounted for via this metric.

Now that the factors to be used in our model have been defined, it can be seen that there are $p = 9$ features under consideration in this model. At present, G's Café has one year's worth of accurate, consistent sales data (n = 365) to use for training purposes, as their record keeping practices were initially rough, like many small businesses in the community. (However, given time and a little more attention to data collection, in particular proper records of total daily sales, a better model could conceivably be developed.) Since the number of features in consideration is still small in comparison to the year's worth of data (n = 365) currently available to be used in model development, *overfitting* (that is, having so many features that the model cannot be properly optimized to accurately predict future sales) should not be an issue. *Underfitting* (or not having enough features to accurately predict future sales) in this case would be more of a concern, as it would be possible for there to be important factors which are measurable and easily attainable which have not yet been taken into account in the selected features. The specific features chosen were chosen based on the goal of not only predicting futures sales more accurately but also the goal of not inundating the business owners with data collection which would be too difficult to obtain or too time-costly to record, so only the major factors currently thought to influence sales are included. However, within these features, there are certainly interactions which would be worth exploring that may add to the predictive power of the model. For instance, holiday social media posts with special advertising surrounding holidays may have an effect on the amount of business occurring on the holiday.

Thus, the model we would like to seek is $y = f(x_1, \dots, x_9) + \delta$, where f is the best representative model of the influences of the chosen features on y and δ is a measure of *error due to ignorance* (that is, δ is a measurement of the information not contained within the features $x_1, \dots, x_9$). Unfortunately, as with t and the z_i's, f is another function which we will be unable to find, this time due to errors in the learning process we employ.

Building a Model for G's Café: The Data and Model

The goal of building a predictive model, in our case, is for the owners of G's Café to be able to predict the business at G's Café at least one week in advance in order to facilitate proper scheduling and maximize business efficiency at G's Café, and ultimately to maximize net profits for the business. Thus, to be useful, the model we seek must be able to allow the owners to predict business more accurately than taking a simple average of daily sales over the historical data, which we would define as our null model, $g_0(\boldsymbol{x})$. In addition, though subjective and not concretely measurable, the model should also have better predictive power than the owners' "best guess" as to what business will be like the next week. Ideally, the model should provide the owners insight into their future business which would complement their own "business instinct" and allow for a much more complete understanding of the sales they should expect in the near future. In order to do so, an appropriate model selection must be made, based upon the data set which we are given.

The model we wish to build will employee *supervised learning*, a form of learning from data which uses well defined historical datasets and user specified algorithms to produce a predictive model (function) for a desired phenomenon. In the case of G's Café, a dataset of well labelled historical data for the selected factors will be paired with an algorithm to produce a function used to predict future daily sales at G's Café.

In order to build a model, we must first have data with which to work. In the current case of G's Café, as described above, the historical data, which will be used to train and test our model, will only consist of the 365 days' worth of data they currently possess with adequately recorded sales data from the past year. In general, *training data* is a set of vectors of the x_i's for instance of historical data which we have. That is: $\mathbb{D} = \{< \boldsymbol{x_i}, y_i >: i \in \{1, \dots, n\}\}$, where each $\boldsymbol{x}_i = \left[x_1\ x_2\ x_3 \dots x_p\right]$ is a vector of values of the factors x_i and y_i is the response value corresponding to the ith record of historical data. In the case of G's Café, specifically, $\boldsymbol{x}_{\mathrm{i}}$ is the vector of the eight feature values for the ith day of historical data, and y_i is the daily sales for the corresponding ith day of data.

Before moving on, there is the question as to how the desired historical data should be obtained for G's Café. While G's Café's owners currently maintain an Excel spreadsheet of the response metric of daily sales, along with the day of the week (x_5), month of the year (x_6), and opening and closing times (x_3, x_4) as exported from their in-store POS, this data would need to be cleaned (for example, removing excess "features" such as day of the month and year as well as net expenses which are also present). To this data must be added the remaining features: The number of daily social media views (x_1) is available for export from the store's social metrics account, maintained by a third party. Marketing communications (x_2) are integrated with and sent through their POS system. These are infrequent, so manually entering this information according to the defined metric would not be overly costly, if needed, but the date and time of sent marketing communications is clearly and accurately recorded and easy to access. Accessing the average of high and low temperatures for each day likely would be the most time-consuming metric to obtain, though if a data table of historical daily high and low temperatures for Brooklyn, NY is not readily found via a quick Google search, then the desired data may be obtained in a clean format from the NOAA's Climate Data Tables (https://www.climate.gov/maps-data/dataset/past-weather-zip-code-data-table). Holidays (x_8) would need to be entered, and assigned values, manually. Finally, days since opening (x_9) could be easily obtained and added based on current business records. Of these, obtaining temperatures would be the most costly, time-wise, to obtain, though apart from the amount paid to the data collector for their time, there would be no additional monetary to the business in order to obtain the desired data set.

Now, what model training *algorithm*, or set of steps taken to optimization model parameters for a given hypothesis set, to use must decided upon. We should take the nature of our data into account when considering which algorithm is best suited to output a response in our response space. We have classification algorithms such as the SVM and k-Nearest Neighbors which would be candidates if our response space was discrete and we also have regression algorithms, such as the Ordinary Least Squares algorithm which will produce responses in a continuous space. Of these options, the Ordinary Least Squares algorithm seems to be the most logical choice, as our response space $\mathcal{Y}$, the set of positive real numbers, is continuous.

Thus, the hypothesis set, or the set of potential functions $h(\boldsymbol{x})$, we will be working with is $\mathcal{H} = \{\boldsymbol{w} \cdot \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^{10}\}$, where the weights $\boldsymbol{w}$ will be determined by the OLS algorithm. That is, $\mathcal{H}$ is the set of all lines in $\mathbb{R}^{10}$ over which our algorithm will search to find the line which best fits our data set. Within the hypothesis set exists the absolute best fitting line to model daily sales based upon the given features (that is, the line which most closely approximates f). We will call this line h^*. As this hypothesis set is limited to purely linear functions, and there is no reason to believe that gross sales are precisely a linear function of chosen features, $h^* \neq f$. The error introduced by the difference between h^* and f is referred to as the *misspecification error*, which is error introduced into the model by not having the complex function f in the hypothesis set.

As mentioned, the algorithm, $\mathcal{A}(\mathbb{D}, \mathcal{H})$, which we employ to find the desired model $g(\boldsymbol{x})$ will be that of the *ordinary least squares* (OLS) algorithm in the case where p = 9. Here, $g(\boldsymbol{x}) = \hat{\boldsymbol{y}} = X\boldsymbol{b}$, where $X = [\boldsymbol{1}_n\ \boldsymbol{x}_{\cdot 1} \cdots \boldsymbol{x}_{\cdot 9}]$ and $\boldsymbol{b} = [b_0\ b_1\ b_2 \cdots b_9]$, with $b_i = \underset{w_i \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. [Note that, while there are only * features, we must append the 1-vector to $\boldsymbol{x}$ in order to account for an intercept term of the lines of which the algorithm will search.] The algorithm will return $\boldsymbol{b}$, the vector of weights of the best line which can be fitted based on the chosen features and historical data which is available. It is important to note that, almost certainly, the weights chosen by our algorithm will not be quite those of h^*, due to what is called the *estimation error*.

Thus, once we have obtained $\boldsymbol{b}$, we now have our sought out predictive model, $g(\boldsymbol{x})$ for our phenomenon.

Building a Model for G’s Café: Validation, Metrics, and Other Post-Model Concerns

With $g(\boldsymbol{x})$ theoretically in hand, it would be easy to declare victory and call it a day. Unfortunately, in the discussion of employing the OLS algorithm, there are some important issues which we did not take into consideration, but will now. Having a model is great, but we still need to know if our model will be useful. Unfortunately, without being able to randomly pick up a new data set to test the predictive power of our daily sales model, we must devise a new plan for testing our model. In order to have data on which to test the model, we will split our data, before employing any algorithms, into a test set and a training set. We will train the model with the training data set and then use the test set to *validate* (or test) our model once it has been trained. Finally, we will employee the model using the full data set in order to train our final model.

The question arises as to how to split the data set. In our case, the data set is fairly small, so we will split our data by randomly splitting our data set into five groups, and selecting one of those five groups as our test data (n = 73, 20%) and keeping the other four groups (n = 292, 80%) as our training data. This split should allow for a fair representation from most factors, except perhaps that of holidays, in which it is possible that a training (or test) set may end up with no holidays included. We could get around this by splitting the data again, similarly, and running several different instances of our training in order to select the “best” model of the instances, giving a higher likelihood that all factors will have a chance to be trained into the final model.

With our data now split, and a final model developed, we are able to compute a measure of error for our model we will do so by testing the same metrics on our training data (in sample) and our test data (out-of-sample). In the case of the OLS model for G’s Café, we will use a metric called the Root Mean Squared Error (RMSE), which is calculated as follows: $RMSE = \sqrt{\left(\frac{1}{n} * \sum_i \left(\mathbf{y}_{\mathrm{i}} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}}\right)^{\mathbf{2}} \right)}$. This error metric will give us an interpretable metric of the average error in the model in the units of our response. In this case, the RMSE will tell us, on average, how many dollars the predicted daily sales vary from the actual daily sales in our data set. This is an important metric, since the goal of our predictive model is, ultimately, to help the owners of G’s Café decide how many employees they should schedule on a given day. If they schedule an unnecessary employee, then then are paying approximately $100 (or about 20% of their daily profits) unnecessarily. If they schedule one too few employees for the day, they could easily lose nearly the same amount in lost orders. So, a “good” predictive model, based on past sales and the goals of modeling, should produce a prediction with average error of no more than $50. (Note that average daily sales are currently in the $500-$750 range, so this error would allow for an error of about 10% - 15% of their daily sales.) We would compute this metric both in sample and out-of-sample, but the out-of-sample error is that with which we are more concerned, as it gives a more reasonable estimation of the performance we should expect from our model when utilizing it on new data.

A final consideration is the actual implementation of the prediction model. Once we have the model, it is important to note whether the owners of G’s Café will be using the model to predict on values of the factors contained within the data used to train the model (*interpolation*) or whether they will be using the model to predict using values of factors which are outside of the range of the data used to train the model (*extrapolation*). In the case of most of the factors considered, the owners will clearly be extrapolating. Likely, temperatures, days of the week, months of the year, and marketing communications, and holidays will remain within the range of data used for training. However, it is also like that social media views may be increasing and business hours may change if a decision is made to open earlier or close later than in the past. Most certainly, the number of days since opening (the time metric) will be outside of the range seen before, as it is monotone increasing. In this case, there is a good chance that the owners will be extrapolating when they employ the model in order to predict future business. However, as each of these changes is likely quite near to the range of the training data, and our model is a linear model, the model we create, assuming that the RMSE threshold is met, should be good enough to hold the predictive power desired by the owners.

Building a Model for G's Café: Limitations and Error Discussion

There are limitations to this OLS predictive model of G's Café's daily sales which must be discussed before we conclude. In addition, it is important to note what errors are reasonably expected to be seen and how these errors may be minimized in order to create a better model.

When it comes to the limitations of the proposed model for predicting future sales at G's Café, there are some which are quite significant. One such limitation is the amount of data currently available. As mentioned, due to inadequacies of prior reporting of daily sales, G's Café only has approximately one year's worth of accurate data with which to train a model. This would certainly not be adequate in order to model the effects of certain factors on daily sales, such as holidays. This is one known limitation within the data used to train this model.

One major limitations of this model is that it assumes that the phenomenon is stationary. That is, the model's predictive ability can only really be measured if what drives sales today is the same as what drives sales tomorrow. In the case of G’s Café, or likely food service in general, the driving forces of daily sales are likely to be dynamic. Food and beverage trends come and go, new forms of promotions and marketing come to be, people may some day tire of Instagram pictures of the latest coffee invention or creative sandwich and not take the same cues to order food while scrolling through. Many things could occur, such as a global pandemic, which may interrupt the way in which people go about their daily life which would completely change the driving forces behind how, where, and why they buy their morning coffee or take lunch break. In these cases, over time, it is not reasonable to believe that the model will remain stationary, thus introducing the need to frequently create new models to predict future business. While the OLS model proposed here may be an okay model and allow for more predictability in business for the time being, over time, the same model would likely prove poorer and poorer predictive capability.

However, there are other related limitations to this modeling proposal from a practical standpoint. As G’s Café is a relatively new business, they are still in the stage of experimenting with various means of attracting customers. Many such experimental efforts likely have a temporary effect on business which cannot be accounted for in this model (and likely should not be, even if they could), unless perhaps one were to add a binary factor for “experimental business booster” or something of that nature. For instance, passing out menus on the corner every so often or sending a complimentary package of drinks and pastries to local offices when business is slow may have give an immediate boost to sales for a few days or weeks, but could prove to add more noise to the model if a factor of “handed out menus”, for instance, was added during the modeling process. This is an area in which more thought and consideration could be placed.

In regards to sources of error introduced in a previous section:

Currently, there is a limit on the amount of data which G's Café has collected. Since a year's worth of data would likely not reveal clear patterns, or perhaps any pattern, in relation to the effect of the month of the year or holidays on daily sales, having a high *estimation error* at this point in time would not be surprising. This error could be reduced over time by keeping accurate daily sales data and then employing the OLS algorithm in the future based upon the accumulated data. This would allow for a better fitting predictive line, over time. This is obviously not an immediate solution, but would all for future improvement of the predictive model without changing other factors of the model.

Note also that the model we seek by employing the OLS algorithm assumes that the chosen features, or some variation of the chosen features, will provide us with a linear model. It doesn't seem reasonable to expect daily sales to be a perfect linear combination of the factors selected (or any other set of factors which could be selected). As such, the model misspecification error introduced by the use of the OLS algorithm on the chosen factors could potentially be fairly large. However, since there is some variation from mean sales on a daily basis, and some of the factors selected, such as month of the year, average daily temperature, and adverse weather events are likely to be somewhat colinear, it is likely that the OLS model chosen here would be a better model than the simple average of historical daily sales. It is also likely that this model, based on numbers and not human intuition, would have more predictive power than G's Café owners' best guess when it comes to the goal of predicting daily sales for the next week or two.

One way to improve upon the model trained would be by creating new factors out of the currently chosen factors. For instance, perhaps social media posting has an exponential relationship to daily sales which could be accounted for by incorporating an exponential factor of social media posts into the feature space. Other relationships such as squared factors or products of factors could be explored in order to develop a more fine-tuned model which more closely models the relationships of given factors to the truly best fit line h^*. Doing so could potentially coerce the data to become more linear, thus reducing the error within the model, as well, and creating a better predictive model. However, caution must be emphasized when adding new factors in order to avoid either creating dependent factors or overfitting the model by adding too many factors.

Final Considerations

While this is one proposed model for one local café in the Brooklyn neighborhood of Sunset Park, it is reasonable to expect that a similar model building process could be employed for other small local food service establishments with varied factors taken into consideration. However, while this is reasonable, this model should be able to at least somewhat predict future business at G's Café because the owners have well collected data, though perhaps not enough of it, and accurately recorded historic sales. However, this is *not* the case for many local establishments, who only record and report portions of their daily income or else keep cash records only on weekly or monthly bases. For such establishments, it is not reasonable to expect to build a decent predictive model for future daily sales, as the predictions could be no better than the training data. Also, it might be better to implement other forms of *machine learning* (learning from data in which the computer specifies the algorithm to employ) methods, such as using deep learning, in order to deal with the dynamic nature of business in the food service industry. However, the author is just now discovering such and can take no position on learning algorithms outside of their scope of knowledge.

Conclusion

While it would be quite easy for the owners of small, local eateries in residential neighborhoods to believe that business "has a mind of its own" and is utterly unpredictable, as can be seen by the example of model building using the OLS algorithm for predicting future daily sales at G's Café contained in this essay, predicting future business at such very small businesses, to a certain extent, is possible using the methods of data science. It seems reasonable, based on the potentially available training data and factors considered, that a predictive model could be built for G's Café capable of predicting future daily sales with an average error of no more than $50. However, the model would need to be retrained with new cumulative data on a relatively frequent basis in order to increase the amount of data on which the model is training as well as to take into account potential growth (or decline) of daily sales over time. That is, developing a predictive model is somewhat possible, as long as the business owners are willing to invest in maintaining clear, honest business records over a long enough period of time to employ a learning algorithm to develop a predictive model. Otherwise, any model would be, at best, underfit; at worst, a waste of time.

References

City of New York. (2015). *Small business first*. https://www1.nyc.gov/assets/smallbizfirst/downloads/pdf/small-business-first-report.pdf.