PREDICTIVE ANALYTICS IN FORECASTING

A Senior Project submitted to
the Faculty of California Polytechnic State University,
San Luis Obispo

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Science in Industrial Engineering

by
Matthew Suarez
ABSTRACT

Predicting future demand can be of tremendous help to businesses in scheduling and allocating appropriate amounts of material and labor. The more accurate these predictions are, the more the business will save money by matching supply with demand as closely as possible. The approach for an accurate forecast, and the goal of this project, involves using data analytics techniques on past historical sales data. Working with Campus Dining, a year's worth of their daily sales data will be analyzed and ultimately used for the end result of both an accurate forecasting technique and a way to display the results in a user friendly manner. The feasibility and effectiveness of doing so will be determined at the end of this project.
ACKNOWLEDGMENTS

Special acknowledgments to all of my teachers of whom have helped me, guided me, and aspired me to reach higher

And to my family for their boundless love and support
# TABLE OF CONTENTS

LIST OF FIGURES---------------------------------------------------------------5

I. Introduction---------------------------------------------------------------6

II. Literature Review---------------------------------------------------------8

III. Design (or Theory)-------------------------------------------------------12

IV. Methodology---------------------------------------------------------------19

V. Results------------------------------------------------------------------20

VI. Conclusion---------------------------------------------------------------24

BIBLIOGRAPHY----------------------------------------------------------------26
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Burrito Bowls sold January</td>
<td>12</td>
</tr>
<tr>
<td>2. Weekday ANOVA</td>
<td>13</td>
</tr>
<tr>
<td>3. Seasonal Decomposition</td>
<td>14</td>
</tr>
<tr>
<td>4. ARIMA Model</td>
<td>15</td>
</tr>
<tr>
<td>5. UI Layout</td>
<td>18</td>
</tr>
<tr>
<td>6. Economics Error Comparison</td>
<td>21</td>
</tr>
<tr>
<td>7. Economics Dollar loss Comparison</td>
<td>22</td>
</tr>
<tr>
<td>8. March-May Trend</td>
<td>25</td>
</tr>
</tbody>
</table>
I. Introduction

The subject of this report is the use of data analytics techniques for more accurate forecasting.

Ia. Background/problem

Campus Dining currently does not employ quantitative forecasting techniques opting instead for the subjective approach relying on intuitive judgement and opinions. Not employing quantitative techniques means campus dining will very often over or under supply labor and food items. While the subjective approach may be alright for “landing in the ballpark”, it relies too heavily on someone with much experience and even then, is not fine tuned to be consistently accurate. As someone who has personally worked at Campus Dining, I've witnessed this firsthand as daytime rushes were inadequately stocked or too many people were scheduled to come in on a low-demand day.

The purpose of this study is to utilize analytics techniques to aid campus dining in forecasting demand and designing a system to display the forecast in a user-friendly manner. The scope of the project is limited to Campus Dining’s PICOS venue. The main objectives are as follows:

- Design an accurate forecasting algorithm for individual PICOS menu items
- Design a UI that will take a desired day to be forecasted and output the predicted quantities of material needed to stock up on for PICOS
Ib. Solution approach

The solution approach will consist of the following steps:

- Clean and interpret four years worth of PICOS sales data in R
  - This involves using R to store the data. R will be used as the programming language of choice in this project to manipulate, store, and analyze data.
- Determine the variables that are statistically significant in determining demand of various items at PICOS through ANOVA and decomposition.
  - This means using k-clustering and separating out variables that have the most effect on the hypothetical chosen day to be predicted. For example, if one wanted to predict next Wednesday’s demand, would it be better to use the prior two weeks or prior fourteen wednesdays only?
- Experiment with various forecasting techniques using the determined variables such as machine learning and regression to find the most accurate method.
  - After the most statistically significant variables for each weekday determined, various techniques could be used to actually make the forecast. These techniques will be used and compared against each other by using them to predict the most recent month.
- Design a UI using the R Shiny web application service
  - Shiny is a web application service that will take R code and turn it into a visually appealing webpage. The point of this is to make it easy and intuitive for someone with no statistical background to find out what they will need on a given day.
II. Background and Literature Reviews

Data analytics has three main branches of study: Descriptive, Predictive, and prescriptive. Descriptive is used in nearly every type of analysis (mean, mode, median, standard deviation, histograms, etc) to better understand a data set. Predictive involves using data to find trends and predict future outcomes. Prescriptive is considered the hardest and involves both predictive and descriptive. It is essentially doing both, predicting all possible future outcomes and then understanding those outcomes in the context of what you should do for the purpose of prescribing a solution or path. This project will mainly involve predictive as the goal is to predict future values accurately, and will use descriptive tools as well. It is important to keep in mind the limits of prediction analytics- that it will only show you the future but not how you should respond to it like prescriptive nor will it show you the present situation clearly as a full-scale descriptive analysis will. Nevertheless, predictive has an important role to play, namely in forecasting future demand and supply needs. The following are typical examples of predictive analytics in action in the world:

- Large online retailers directing customers to items they’re most likely to buy
- Universities predicting what kinds of students will choose to enroll at their school
- Research groups in healthcare using predictive analytics to better classify and treat various forms of cancer.
- Airlines increasing customer satisfaction and revenue by more accurately predicting the number of passengers who won’t show up.

As these examples show, predictive analytics can be of great use in various industries by helping businesses optimize existing processes, identify hidden opportunities, understand customer behavior and anticipate problems before they arise.
Further evidence of the usefulness of predictive analytics is written by Joe F. Hair Jr. who writes, “Data mining and predictive analytics are increasingly popular because of the substantial contributions they can make in converting information to knowledge.” (Hair, 1) Hair goes on to describe the specific role of predictive analytics in marketing in his paper, *Knowledge creation in marketing: the role of predictive analytics*. According to his findings, in the future, we can expect predictive analytics to be used more and more in all fields to enhance the ability to understand and predict future developments. Additionally, he mentions two types of data which will be examined- structured (numbers) and unstructured (images and text). Overall, Hair does a good job of explaining the uses of predictive data analytics in other fields and a good examination of the roles it will play.

Matthew Waller’s publication, *Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management*, also details this subject but with more focus on supply chain. In this paper, further consideration is spent on what will be made obsolete. Additionally, it looks into what skills will be needed by supply chain management data scientists. Those skills it goes over include quantitative skills, management theory, and the ability to visualize data to others. Overall, a very insightful publication.

“Creating analytic models is both art and science.” (Eckerson, 14)

Galit and Koppius’s publication, *Predictive Analytics in Information Systems Research*, details some of the technical details when considering machine learning. Within predictive analytics there are two techniques to building a model when utilizing machine learning: Supervised learning and unsupervised learning. Supervised learning is the process of creating predictive models using a set of historical data which itself holds what is wanted to be predicted. For example, if you want to predict which days are likely to have a high demand (like in this project), you use the past sales data in order to
train a model to identify the characteristics of the high-demand days. Techniques involved in supervised learning include regression, grouping, and time-series analysis. Regression is simply using past values to predict future values, the focus of this project, and is commonly used in forecasting and variance analysis. Time-series analysis is similar to regression and involves analyzing data in relation to time to extract meaningful statistics and other characteristics about the data. Classification techniques identifies the group new data point should be in determined by its characteristics. Unsupervised Learning does not use previously known results to train its models, unlike supervised. It instead tries to find patterns and relationships existing within that particular data set without trying to predict a future value. It is commonly used to find clusters (groups of similar characteristics) within a data set.

Another publication, this one by Eckerson - *Extending the Value of Your Data Warehousing Investment*- goes over some additional technical concepts. He goes over the idea of splitting data when creating your model and the specifics of doing so. One half is the training model while the other half is the test set to test the strength of your model to see how well it predicts. For the final step, it needs to be validated in real time by comparing it to live data. It's important to note, this is a very iterative process. According to Eckerson, “Most analysts identify and test many combinations of variables to see which have the most impact.” (Eckerson, 14) Finding the high impact variables means identifying significant trends in the data and consulting those with experience in the field to hone in on key variables. And then a variety of algorithms are tested to see what works best in predicting the test set from the model set. Eckerson explains these in detail quite sufficiently.

More onto experimental techniques, Mike West details the use of bayesian statistics in forecasting in his article *Bayesian Forecasting*. Bayesian statistics is different from frequentist statistics in that it incorporates a more subjective approach. Namely, it incorporates a guess as to what one might think a future value will be, called
the prior. The advantage of doing so is that a more natural understanding can be obtained as it fits in line with how most people interpret probability, unlike frequentist which can be quite convoluted. Onto forecasting, Mike West details how bayesian is incorporated into time series. The effectiveness will of course be affected by the prior, so an educated guess is important- new hires wouldn't be able to do this for example. But the advantages are better interpretation and communication of the results. Mike West does a good job going over the bayesian techniques and the advantages of those techniques.

Another experimental technique, detailed by Billy Williams in a publication entitled *Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results*, analyzes the forecasting of vehicle traffic. Specifically, it ‘presents the theoretical basis for modeling univariate traffic condition data streams as seasonal autoregressive integrated moving average processes.’ So, in other words, it is asking if using the specific moving average is good for modeling traffic. According to their conclusion- using actual traffic data- it is. It is a very insightful study and method.

Onto the visualization components of such analytics is the paper, *Forecasting Hotspots—A Predictive Analytics Approach*, goes into detail about the visualization of ‘hotspots’ or spots in the data with a high incidence of events. In order to facilitate that type of forecasting, the authors of this paper have created a predictive visual analytics toolkit that provides analysts with linked spatiotemporal and statistical analytic views. Basically, the system models spatiotemporal events (the hotspots) by using two methods: kernel density estimation and seasonal trend decomposition by loess smoothing for temporal predictions. The idea is to help analysts perform hypothesis testing much faster to plan around and allocate resources to these hotspots.
III. Theory and Design

This section will illustrate some of the technical aspects of the project.

Figure 1: Quantity of Burrito Bowls sold by day through January

Figure 1. illustrates the variation of sales over the month of January. Figure 8 (page 24) shows the variation of sales since March. Both of these graphs can be used to get a good look and start making some initial observations on the data. This for example, seems to show Saturdays as being consistently the lowest of all other days, while Tuesdays and Thursdays seem to be the highest. Additionally, we can see the general trend over a longer period of time on Figure 8, which, while has some extreme dips for breaks, seems to indicate a general increasing trend. To make the claim that Saturdays are actually statistically significantly lower than other days, an analysis of variation (ANOVA) was conducted on each weekday. Figure 2 is the results of said ANOVA over the month of January, which takes the difference of the mean sales of each day and compares these groups for statistical significance. From this, we can conclude that the means of Mondays-Thursdays is actually statistically significantly lower...
different from the means of Fridays and Sundays which is also statistically significantly
different from the means of Saturdays. Therefore, it appears there are three main
groups- Mondays through Thursdays in one, Fridays and Sundays in another, and then
finally Saturdays alone in the last. Sundays being statistically significantly higher than
Saturdays is indeed odd. The reason may be because PICOS is situated next to
Chick-fil-a, which happens to be open on Saturdays but not on Sundays- meaning
PICOS has far less competition on Sunday. If Chick-fil-a were open on Sundays as well,
PICOS Sunday sales would most likely resemble current Saturday sales. Ultimately,
these groupings help to understand the data and can be used to improve the accuracy
of a potential forecast via linear regression.

![Figure 2: ANOVA of weekdays on quantity burrito bowls sold](image)

Figure 3 depicts what is called a decomposition of the data. Basically, it looks at
the overall trend line and separates the seasonal component from the underlying
trendline. From this we can get a sense of how much the season affects the sales, and
how much sales are increasing at a steady rate—(due in part to increasing

numbers of cal poly students). This underlying trend line has an overall increasing value which is to be expected. Each dip corresponds to the weekends—specifically saturdays—and the rises are on Monday’s-thursdays, with slight dips on Tuesdays, which corresponds to our findings from the ANOVA. The final block of Fig 3 gives us the remainders or the difference between the seasonal and underlying trend line at each point. Using the remainder values, we can obtain the values needed to make an arima model which is discussed and shown below.
Figure 4 shows off said arima model. Arima models (Autoregressive integrated moving average) have some unique uses- namely that they are best with ‘stationary data’ and predicting dailies (non-stationary would resemble stocks). So anything involving complex seasonality with a relatively consistent underlying trendline is where this type of model works best. Since our data fits these criteria (found from the decomposition), an arima model was chosen. For the purpose of the model, the frequency chosen is seven (best for daily forecasts)- which treats each week as a season instead of each month. Different frequencies were tested (daily and monthly) however monthly required too much data to be accurate (compiling each month as a data point requires many months, more than the amount on hand) and daily failed to capture the overall underlying trend accurately.

Sales data for a campus dining facility has unique complications as well- namely breaks, such as Spring and Winter break, where sales dropped to zero. Because of this, the time series would be ‘broken’ in the sense that having gaps of zero sales from an otherwise independent event would affect future predictions in ways they shouldn’t.
Holidays and other events would also prove to be an issue, but single days can be estimated or cut out with little harm- multiple weeks however is a bit more difficult. Getting past this is a common and difficult issue in forecasting, but one that is made dramatically more difficult due to the long stretches of zero sales unique to PICO’s location on a college campus. Four possible solutions were looked at:

- **Solution 1: keep them at zero**
  - This solution is to ignore it and keep them in at zero which harms accuracy. Generally this can be done if there aren’t many instances, but would be worse with the more zero sales days included.

- **Solution 2: Cut them out entirely**
  - Cutting everything out and essentially “gluing” the pre-break sales to post-break sales is another possibility, but this harms the forecast as well from giving a less accurate underlying seasonal trendline. There will be a ‘jump’ in the trend that would have been otherwise smooth with time where there shouldn’t be with this approach.

- **Solution 3: Interpolate approximate values**
  - This approach involves plugging in rough estimations on what these points would have been if there was no spring break and school continued on as normal so to speak. This of course relies on accurate estimations in the first place.

- **Solution 4: Start the time series after the problem area**
  - Alternatively, if enough data points occur after the problem area, the forecast could instead simply start after the break. This way no awkward estimations have to be made or cuts in the middle of the time series. This of course will forfeit any data prior to the break, but generally the more recent data is what's important.
Ultimately, solution 4 was tested to have the best effect on forecasting accuracy and was chosen as the way to deal with this unique problem.

Additionally, a simple linear regression model was tested as well. These models involve approximating one straight line that best resembles the trend. Unfortunately, these do not capture the cyclicality of daily fluctuations well, and thus is not expected to forecast accurately.

Another approach sometimes used in forecasting is a machine learning model. These models use pattern learning algorithms to analyze vast amounts of relevant data in order to determine significances, discover unique patterns and make future predictions. Since machine learning algorithms ‘think’ in a unique manner, possible complex connections can be made between data groups that would be undetectable via traditional methods. While this kind of model may be the most accurate, it also requires a tremendous amount of data to be accurate and was therefore determined to be unfeasible with available data. Possible data sets would need to be obtained- such as rain, temperature, population data, etc (anything that could remotely affect sales) in order for the algorithm to churn through it all and find the connections. This approach is considered the most technologically advanced and perhaps most accurate, but is not commonly used.
Figure 5: UI screenshot

Figure 5 is a picture of the UI designed to display the forecast model. The UI component of this project had some unique specifications from the client:

- Simple and easy to use
- Clear, with pictures displaying the items as well as the trend prediction line
- A drop-down calendar selection tool for the forecast day input

Following these specifications, the user will input a day through the aforementioned calendar tool and select an item (one of three for PICOS’s main items) and it will display the amount predicted to be sold as well as the ingredient totals needed.
Measuring the accuracy of a forecast model involves comparing forecasted values to actual sales values. This can be done by forecasting the future and waiting for said day to come for comparison, or by splitting the data into a test group and forecast group. Since the model was working off of a fewer number of data points (because the model starts after spring break), forecasted values went approximately two weeks into May. After May went by, the comparisons between those live data points and the forecasted values could be made. There are many different measures used for this comparison—such as ME, RMSE, MAE, MPE, MAPE, and MASE. Since this data involves points much greater than zero, MAPE (mean absolute percent error) was chosen as the main accuracy measure. This measure is also the most commonly used in the industry for forecasting accuracy. Generally, a MAPE score of less than 20% is considered ok but a score of less than 10% is considered excellent. The results will be discussed in the following section.

As for testing the display of the UI, ideally some user-interface tests would be implemented to gauge its design. However, since the client has made some clear instructions on its specifications, this would be limited. Regardless, unofficial testing of limited sample size (sending it out to a few coworkers/family) proved that it was indeed simple and easy to use and informative. Overall, five people rated it in two different categories on a 1-10 point scale—ease of use and aesthetic. The average score was a 10 for ease of use and 9 for aesthetic. Overall, I'd say this indicates a well designed UI to fit the client's criteria.
V. Results and Discussion

The accuracy test of the arima model using the MAPE method and compared to the accuracy of two other base forecasting methods. The first- Naive- is the baseline forecasting method (i.e. using yesterday's sales as tomorrow's forecast) and is commonly used as a benchmark for comparison purposes. The second is the seasonal naive model, which is another standard forecasting method that uses the first value of the past season as your forecast. From the ANOVA results, each weekday group is considered it's own season for this method. The results are as follows:

- Naive for the data produced a MAPE of 19.9%
- Seasonal Naive produced MAPE of 6.6%
- Arima produced MAPE of 8.8%
- Simple Linear regression>20%

From these results, we can see that even the naive approach produces a result less than 20% which beats out a base simple regression approach. This would make sense as the linear regression would fail to capture the seasonality, while the naive would capture it at least partially by always using recent data. However, what comes as a surprise is that the seasonal naive actually beats out the arima ever so slightly. The reasoning on why this is the case may have to due with PICOS’s customers- students- which are subject to peculiarities that normal students are not. Namely, students abide by a daily schedule that changes everyday. Because of this, the day has more of an impact than anything else and can be treated as its own unique season along with other days in it’s group from the ANOVA. So while the seasonal naive is not usually expected that much better than the regular naive, this is one such case where it would make sense for it to be. The arima model might be slightly less accurate due to it putting too much weight behind other seasonalities, which ultimately don’t have as much of a statistical significance. Since the seasonal naive was the best approach, it will function
as the selected model in the UI, which is also good design for it’s simplicity and ease of understanding.

Economic Justification

<table>
<thead>
<tr>
<th>May</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasonal Naive</td>
<td>338</td>
<td>337</td>
<td>339</td>
<td>298</td>
<td>258</td>
<td>179</td>
<td>212</td>
<td>-</td>
</tr>
<tr>
<td>Original Forecast (their method)</td>
<td>320</td>
<td>315</td>
<td>320</td>
<td>300</td>
<td>280</td>
<td>160</td>
<td>230</td>
<td>-</td>
</tr>
<tr>
<td>Actual Sales</td>
<td>317</td>
<td>351</td>
<td>333</td>
<td>325</td>
<td>181</td>
<td>178</td>
<td>178</td>
<td>-</td>
</tr>
<tr>
<td>Seasonal Naive Error</td>
<td>21</td>
<td>-14</td>
<td>6</td>
<td>-27</td>
<td>77</td>
<td>1</td>
<td>34</td>
<td>-</td>
</tr>
<tr>
<td>Original Error</td>
<td>3</td>
<td>-36</td>
<td>-13</td>
<td>-25</td>
<td>99</td>
<td>-18</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6: Economics of the forecast

Figure 6 shows and compares the seasonal naive forecasts values to the actual forecasts made (albeit a rough approximation) on burrito bowl sales. The error is how off each forecast was in terms of burritos. If the number is negative, then the forecast lowballed and if it’s positive then it went over. The cost of materials of each burrito is roughly $3.00, thus in every instance where a sale was missed due to lacking the necessary ingredients is a ‘loss’ of money equivalent to the missed profit and each instance where excess burrito ingredients were not sold for that day and thrown away is a ‘loss’ of its material cost. There may be other soft costs associated with having unhappy customers created from missing a sale as well, but these are harder to quantify and are left out of this analysis. The following table adds up all the instances each error was negative for both forecasts and multiplies that by the cost of the burrito missed out on ($7.00). When the error is positive, the only money lost out on is the
materials cost and thus multiplies that by the material cost ($3.00). A comparison of the loss of both methods can then be made.

<table>
<thead>
<tr>
<th></th>
<th>Sum of the 'negatives'</th>
<th>Negative Costs</th>
<th>Sum of the 'positives'</th>
<th>Positive Costs</th>
<th>Total Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasonal Naive</td>
<td>41</td>
<td>$287</td>
<td>139</td>
<td>$417</td>
<td>$704</td>
</tr>
<tr>
<td>Original</td>
<td>92</td>
<td>$644</td>
<td>154</td>
<td>$462</td>
<td>$1106</td>
</tr>
</tbody>
</table>

Figure 7: Economic Comparison

From Figure 7, we can compare total costs each forecasting system has, and see that the seasonal naive forecasting model would produce a net $402 in savings per week for burritos. Since the system forecasts PICOS's three main items, we can roughly expect a $1,206 savings per week ($402 per item). Assuming roughly 50 work weeks total, that is roughly $60,300 in total yearly savings. The server costs of using the UI is $9 per month for 100 hours of use per month. This is a very minor cost that adds value by allowing the on floor employees to easily see and identify what they'll need for a given day without relying on the supervisor- and translate into possible time savings.

For this project, there are major limitations and assumptions that need to be discussed. The first and foremost is this algorithm only works well for “regular” days (i.e. Cal Poly events are not accounted for). Holidays are easier to take into account (zero sales) vs an event that increases or decreases sales compared to the norm because while holidays are consistent year to year and have a clear representation in the data, events can go by unnoticed and change year to year. Ideally, the algorithm could flag forecast dates with an event scheduled and look back at past values for that event, but without data going more than a year back it is hard to get this estimate.

The second biggest assumption is concerned with PICO’s menu items. Pico's items from their sales data can contain different ingredients. For example, while data
records the ‘burrito’ sales, it does not distinguish chicken burrito’s from steak burritos, presumably because they cost the same. So an assumption of 50/50 between ingredient types or an equal spread is made (which is also how it currently is) despite there actually being no evidence steak and chicken burrito’s sell in equal quantities or that every burrito bowl contains the same ingredients every time (customers can customize certain things on the menu). This is a question that remains unanswered unless a change to point of sales data recordings is made.
VI. Conclusions

This project began with campus dining wanted a forecasting system and algorithm. To summarize, the following main steps were completed:

○ Took in a year's worth of sales data
○ Organized, cleaned, and analyzed the patterns in the data
○ Tested various forecast algorithms against real sales
○ Created a UI to display the forecast
○ Measured the approximate amount of money saved by using the forecast system

The results were interesting as the seasonal naive approach was determined to be the best. This is in part due to its accuracy as well as its simplicity and ease of use. The visual UI uses this as its model in calculating the forecasted values, and itself adds value to the company by allowing on-floor employees to stock up on what they need in a quick and efficient manner. This UI was developed in accordance to the client's design specifications, which features a drop-down calendar for day selection, a visual representation of the forecasted curve, and predicted sale values for both the main item and the ingredients it consists of. Some major limitations were noted, namely that abnormal events aren't accounted for and some assumptions about specific ingredients were made. The economic analysis determined large savings in using the system.

Over the course of the project I've personally learned some important things. Namely the importance of clarifying and identifying what data you'll need as early as possible. I've also learned some things about the technical aspects of forecasting such as forecasting normal days is not nearly as valuable as forecasting the abnormal days- (which is harder to do, but something I believe I could do with the right data). This is because abnormal days are where the majority of losses happen from inadequate stock. In addition, forecasting daily is considerably harder than forecasting weekly or...
monthly as daily fluctuations tend to be harder to predict. Also, a major lesson to be gleaned is that complex data doesn't always need a complex algorithm.

![Burritos sold over time](image)

**Figure 8: Burrito's sold March-May**
Works Cited


