**Milestone 3: White Paper**

Jamie Tran

Bellevue University

DSC680 Applied Data Science

Professor Amirfarrokh Iranitalab

**Business Problem:**

The business problem we will be tackling for this project is trying to determine the hiring targets for a local pizza parlor. Often times fast food restaurants hire young seasonal talent in order to preserve profits. Small businesses such as these cannot afford to go through the trouble of training talent that ultimately won’t be needed and will only be employed for a short period of time. This will slow down productivity and throw away a portion of their product taking on unneeded labor. This model will help forecast the future sales going into the next year in order to hire just the right number of employees throughout the year according to the results.

**Background/History:**

According to a binwise.com 60% of restaurants fail within the first year and 80% of restaurants fail within the first 5 years. With those statistics the outlook of the food industry seems quite grim. Our model will help those who have made it past that first year and give their business the best shot they can have at surpassing that five-year mark. In the service industry there are many things that cannot be controlled but the number of staff the company keeps employed is one of them.

**Data Explanation (Data Prep/Data Dictionary/etc.):**

For the following research I gathered my dataset from Kaggle.com. My dataset contains sales information from a pizza parlor every day in 2015. This data frame consists of 48,619 rows and 12 columns letting us the know the date and time each order was placed, how many pizzas were ordered, the size of the pizza, what type of pizzas were ordered, etc. I will start by converting the order\_date column to datetime using pandas. Making sure this is correctly formatted is essential to ensure our model can be run and with accurate results. After the order\_date column was converted I then set the datetime column as the index for the dataset. Once that has been completed, I removed all columns other than the total\_price column. I had initially anticipated using most of the columns in our research, however after a great amount of trial and error the additional characteristics were not compatible with our SARIMAX model. Lastly, I resampled the data so values would be grouped together by one single date and the total prices for each day would be totaled instead of having an individual row for each sale that occurred throughout the year.

**Methods:**

For my research I have chosen to perform Seasonal Autoregressive Integrated Moving Average + Exogenous Variables (SARIMAX) model using python. This forecasting/time series analysis model was selected for our business problem because we were looking to forecast sales data according to how business might change seasonally to help the business owners determine their hiring targets throughout the year according to our projected number of sales. The order date column and total price column will be essential in using this model.

**Analysis:**

As I would with any model, I will begin this analysis by utilizing exploratory data analysis to familiarize myself with the data and undercover any underlying trends. Through this analysis the following illustrations were created:

A graph of blue lines

Description automatically generated with medium confidence

The first graph shows us the total prices of orders placed throughout the year. From this chart we can see that most orders placed land between $25-$40 with larger spikes taking place in the last three months of the year and the beginning of summer in June.

A bar graph with different colored bars

Description automatically generated with medium confidence

This second graph lets us know that classic pizza is the most popular pizza which doesn’t come as a surprise. However, it was suprising to see that all other pizza types are really not that far behind. A barcode with numbers

Description automatically generated

The third chart shows us the number of pizzas ordered throughout the year. We can see people generally order a max of two pizzas with the exception of the last three months of the year and summer time. Similar pattern to what the first illustration had shown us.

A graph of a pizza size

Description automatically generated with medium confidence

This last chart shows us large and medium pizzas are most frequently ordered amongst the pizza parlor’s customers. Since extra large and double XL pizza are rarely ordered this tells us there may be larger gathering that prompt those orders.

After the EDA I performed the Augmented Dickey-Fuller (ADF) test to determine whether or not the data was stationary to evaluate the d value for our analysis. Luckily, since our p-value was 0.0035 which is less than 0.05 we know that our data is stationary. This means that no further differencing was required and because of that we know our d value in our model will be 0. Moving on to determing out p and q values I plotted the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for our prepped data frame, the outputs are listed below.

A graph with blue dots and numbers

Description automatically generated

The following graph will help us find our q value. We can see that there is only one large spike that is significantly greater than the mid line. This lets us know that our q value will equal one.

A graph with blue dots and numbers

Description automatically generated

The partial autocorrelation graph will help us find our p value. Similar to the last illustration we see there is only one large spike that is significantly greater than the mid line. Again, because of this we know that our p value will equal one.

Now that we are confident in our p, d, q values we will use these numbers to define our SARIMAX model and fit it. From running our model we were able to produce the following results.

A screenshot of a computer

Description automatically generated

After reviewing these results and taking a shot at ARIMA and ARIMAX models in addition to the SARIMAX model, we can see that our specfic data may not be the best fit for our models. With the SARIMAX in particular we know that there are some fitting issues with the Log Likelihood value being negative and our AIC/BIC values being relatively high. From these results we are also able to see in the coef values are quite large, which tells us the variance between values can vary greatly.

Using the model that was built I generated the forecasted sales for the pizza parlor in 2016. From those forcasted values we produced three values, the lowest total price, highest total price, and mean total price for each day of the following year. With our SARIMAX results being far less confident than I would have liked, having these three values gives the businesses a range to work with when determining their hiring targets. I have the following results plotted below.

A graph showing a red line

Description automatically generated

**Conclusion:**

At this time, I certainly do not believe this model is ready to be rolled out to the public. While we were able to deliver a forecasted range to the pizza parlor it is far less centralized than I would like when offering services to small business. The end goal is to be able to help small business cut on costs and not waste their time with under or over hiring. Our current results could help them in a general sense however, I want to have the integrity to deliver stronger results than the one given. I had a lot of trouble utilizing this specific data set with the ARMIA, ARMIAX, and SARIMAX models. While I could try and troubleshoot even further or try out different models, ultimately this data set may have not been the compatible with the goals I was trying to achieve with this research.

**Assumptions:**

A few assumptions we are making with this analysis is that this dataset is complete and accurate. We are forecasting next year’s sales data under the assumption that there were no technical issues or mishaps that led to unrecorded sales. We are also assuming that this past year worth of sales is their baseline of sales received. Meaning that there were no extraordinary events that resulted in an influx or major decrease in sales.

**Limitations:**

A limitation that we currently face with this model is it can only forecast the future anticipated volume of the pizza shop. It does not give a distinct number of people that need to be present for each day. Although our intention with this research is help businesses evaluate their hiring targets, they can really use this analysis for a multitude of different purposes. After receiving the results, it is up to the owners to evaluate the staffing necessary because on sales information.

**Challenges:**

A challenge of this analysis with the given data frame is that we only have one year’s worth of data to work off of. Having a bigger dataset would help us better account for the ordinary swings of business in the food industry. For example, outdoor gatherings will always spike in warmer seasons such as spring and summer so that is a somewhat constant variable that we can count on to boost sales. However, maybe one year the pizza parlor partners up to do a few fundraisers that would cause abnormal spikes in sales and slight decrease in profits for the amount sold. Another scenario could be the pizza parlor is next to a movie theater and a long-awaited movie finally releases, this would also affect sales on an irregular basis. Events such as these would be more likely to show on long-term data making our results even more accurate.

**Future Uses/Additional Assumptions:**

In the future I can see this model be used as service that can be offered to businesses to help them better manage many aspects of their business, not just hiring targets. Restaurants for example can use this data to do long term planning with their business ventures. For example, they can use this projection to decide whether or not they can handle opening up a second location, if the business will use up enough product to order in bulk and possibly switch distributors, etc.

**Recommendations:**

To build upon this model we can make it more interactive. If we go the route of making this accessible through an online resource, we can ask for the client’s input on more detailed information. For instance, when the businesses are entering in their sales data we can prompt them with a few questions. Such as how many employees are required for x amount of sales per day and their feedback on their purpose for wanting this data. With the first question we can code the website to do the math on how many employees are required based on their input. This way we can print a nice specified number of hiring targets for the user throughout the year. As for the second question this could help us expand the services on our website to possibly add additional services.

**Implementation plan:**

I see this model being implemented through local businesses offering their analysis as a service to other nearby establishments. This could also be made more publicly available by giving everyone access to this tool online. Similar to TurboTax, businesses could plug in their sales information from the last year to have this report ran for them. The goal of this is to help give businesses the resources they need to succeed in the long run.

**Ethical Assessment:**

Some ethical considerations we need to consider before rolling out this model is making sure our model is as reliable as it can be and the transparency behind letting our clients know that actual results may differ from the forecast. The reality is there will always be unknown variable that cannot be accounted for because they simply do not exist yet. Businesses using this service are likely too small to have an internal forecasting team and do not have the data science knowledge to perform this analysis themselves. With that being said they are putting a lot of their trust in their given results to run their business. If they blindly follow our results, (which should be fairly accurately to the information they have given us prior) then they could strongly believe that we had a hand in ruining their business and coming off as untrustworthy scammers. Before engaging with any business these precautions must be thoroughly understood.

**Ten Questions:**

1. Is forecast really that helpful, it could always end up being wrong?
   * It is true that the forecast could always end up being incorrect. There are many factors that can affect sales that simply do not exist yet. Our goal is to give these businesses a starting point/reference values. These results can always be reevaluated as more data is collected or various circumstances arise.
2. Will this be able to work on a brand-new business?
   * This model does require at least a year’s worth of sales data. So while it cannot aid business on day 1, it will be able to assist them once they have made it past the one year milestone.
3. Can I use this even if my business has been open for a long time?
   * Absolutely! This model would thrive better and produce more accurate results with the more data you have.
4. Wouldn’t internal forecasters be able to get this done for them?
   * Unless you are a part of a mass corporate chain, most businesses are too small to hire an internal forecasting team and lack the skills themselves.
5. Why doesn’t it tell us exactly how many people to hire?
   * The model in its current state does not give exact hiring targets because this varies widely by business unit and sector. For example, an ice cream shop that simply scoops prepared ice cream for customers would need a lot less people per dollar in sales than a full-size restaurant that serves a whole menu of different entrees alongside a bar.
6. Will I have to enter in my sales data into the model manually?
   * This model will not require you enter each day’s sales data one by one. You will just need to upload a csv file with your data and you will be good to go!
7. Will the forecast be accurate for multiple years?
   * We recommend entered updated data each year for the most accurate results.
8. How often can I do this?
   * You can technically do this as often as you would like. However, if you are running this model each day there might not be much of a change in your results. Performing this analysis quarterly, bi-annually, or annually would be best.
9. What do I need to include in my data?
   * At the bare minimum you will need the date and total sales for the day. However, if you want to include other numerical values such as the amount of product sold per day that would work just fine as well.
10. How long does the results take to process?
    * Depending on the size of your data and the capacity of your computer the results will be processed in a matter of seconds to a few minutes at the latest.

**References:**

Mdismielhossenabir. “Pizza Sales Dataset.” *Kaggle*, Kaggle, 10 Jan. 2024, www.kaggle.com/code/mdismielhossenabir/pizza-sales-dataset.

Krimmel, Matthew. “Why Do Restaurants Fail? Restaurant Failure Rate Statistics.” *Complete Bar Inventory Software System*, 10 July 2024, home.binwise.com/blog/restaurant-failure-rate.

**Appendix:**

First 10 rows of the prepared dataset:

total\_price

datetime

2015-01-01 2713.85

2015-01-02 2731.90

2015-01-03 2662.40

2015-01-04 1755.45

2015-01-05 2065.95

2015-01-06 2428.95

2015-01-07 2202.20

2015-01-08 2838.35

2015-01-09 2127.35

2015-01-10 2463.95

First 5 rows and last 5 rows of all the forecasted values:

2016-01-01 2312.432825

2016-01-02 2204.408910

2016-01-03 2259.018505

2016-01-04 2070.313906

2016-01-05 2269.973173

...

2016-12-26 2132.671170

2016-12-27 2024.647255

2016-12-28 2079.256850

2016-12-29 1890.552252

2016-12-30 2090.211518

Freq: D, Name: predicted\_mean, Length: 365, dtype: float64

lower total\_price upper total\_price

2016-01-01 1287.554339 3337.311311

2016-01-02 1179.529577 3229.288244

2016-01-03 1234.139163 3283.897847

2016-01-04 1045.434563 3095.193250

2016-01-05 1245.093829 3294.852517

... ... ...

2016-12-26 1062.990829 3202.351511

2016-12-27 954.964446 3094.330064

2016-12-28 1009.574015 3148.939685

2016-12-29 820.869412 2960.235091

2016-12-30 1020.528674 3159.894362

[365 rows x 2 columns]