Predictive Sales Analytics of My Family's Business

Parsa Keyvani

March 2022

Contents

1 Introduction

In this research project I use data analytics tools to find sales insights and predict the sales performance of my family's business. The company gathers the data in a daily time-series fashion where each row corresponds to an order being placed. The data include 12 datasets: 6 datasets for sales data and another 6 datasets for returns data. The data values are in Farsi language and dates are in Jalali format. The dependent variable to be studied is Total Sales and the independent variables that might help with obtaining better forecasts are:

- **The Buying Power:** The buying power of the consumer is largely dependent on their income. The amount of the product which the consumer is willing and able to purchase also depends on the type of goods. Since school and office supplies are considered normal goods, a consumer will buy more if his income increases.
- **The Number of Customers (Lagged):** The number of customers positively affect the sales of a company since more consumers suggest more demand for the company's products.
- **Discounts:** a price discount provides a monetary gain that incentivizes consumers to purchase the product. Consumers also perceive a higher level of savings for a product when a higher price discount is provided. Therefore, a higher price discount leads to higher sales.
- **Sale Return (Lagged):** as total sales increase, sales return also increases because as sales increase, there is a higher possibility of returns.

Note: due to the presence of large missing values and abnormalities for some of the independent variables, the project did not use all the desired independent variables for analysis.

The methods employed for forecasting the time series data are exponential smoothing (SNAIVE), ARIMA, and multiple regression with seasonal decomposition, and piecewise_linear model. The exponential smoothing is used to indicate the baseline model performance, and other models are used to predict future sales.

1.1 Importance of Demand Forecasting for a Stationery Business

Demand forecasting is the process of estimating the future sales of a particular product or service to customers over a defined period, using historical sales data to make informed business decisions (Intuit, 2022). It could either be carried out as a bottom-up approach where judgmental approaches are used or conducted using advanced methods that statisticians have developed. General forecasting approaches include judgmental,

experimental, relational/causal, and time-series approaches. Moreover, demand forecasting will contribute to the following enhancements in my family's business.

1. **Improved Inventory Management** Demand for the company's stationery products exhibits seasonality. Hence accurate demand forecasting will help prevent large stocks in the inventory. Utilizing demand forecasting in this project will also reduce the risk of damages or losses of products in the inventory.

2. **Improved Planning and Control** A good demand forecast helps in the better allocation of resources. For instance, it allows better planning of the supply of raw materials and other inputs, price, and promotional activities. It will control the number of internal costs such as storage cost and wastage and external costs such as loss of customer perception, opportunity cost, loss of market share.

3. **Improved future decision making** Demand forecasting will allow a better understanding of the consumers' behavior and will allow decision-makers to respond in advance to unfavorable situations and prepare a more robust and accurate growth plan.

2 Literature Review

2.1 Market Demand for Stationery Products

The market for stationery products is constantly growing, especially with the rise of technology, where consumers now seek convenient one-stop-shop options instead of selling one-product-category (Aurmanarom, 2010). However, a study has reported that the global demand for stationery products during 2016 and 2020 grew sluggishly as laptops, desktops, and smartphones replaced traditional stationery items (Market Research Company, 2022). Although short-term market growth does not seem promising, a recent market report conducted by Market Research Company forecasted the global stationery market to reach $30 billion from its 2021 value at $24 billion. Therefore, the long-term projections suggest steady demand from educational institutes and workplaces. Several studies were conducted to explain the long-term increase in sales demand:

1. Rising educational programs can fuel demand for stationery products. Rising inclination towards higher education and government initiatives boost the educational sector and direct the market on a positive trend.
2. The demand for stationery products is likely to remain high across the corporate sector due to its increasing scale of businesses and its necessity to maintain records regarding employees, turnover,

profits, and other crucial aspects. Hence, the demand for products, such as file boxes, folders, binders, clipboards, and other similar products for documentation is likely to sustain; however, increased reliance on digital records may restrict the growth of the stationery market.

3. Customization in design, personalized printing, and comprehensive marketing strategies bridge the gap between the manufacturers and the consumers, which opens new doors for market expansion.
4. The proportion of young population is projected to increase with an estimated 4 percent growth from 2020 to 2035, which suggests a steady increase in demand for stationery products (Frey, 2021).

The studies expected growth to remain mainly concentrated across the developing countries, particularly in Asia. Countries such as India and China are extensively investing in primary and secondary educational programs to meet the population's requirements (Market Research Company, 2022). China's stationary market grew significantly over the past years and reached $18 billion in 2019, and according to recent studies, the market is predicted to remain among the fastest-growing regional markets (Market Research Company, 2022). Similarly, Iran's stationery market is also growing. The stationery market reached $1 billion in 2017, with over 40% of the market being held by Iranian producers and the rest being imported mainly from China and Germany (Financial Tribune, 2017). Chinese products alone hold a 50% share of Iran's stationery imports as the variety, and low prices of Chinese stationery have made them popular among both importers and consumers (Financial Tribune, 2017).

2.2 Stationery Products Sale Forecast

For forecasting stationery products sales, relatively fewer studies are conducted. Most studies have focused on time series models, and predictions of the stationary product sales were made with support vector regression and methods alike. Economic indicators used for improved forecasting accuracy of stationery products sales include gross domestic product (GDP), consumer price index (CPI), interest rate, and unemployment rate. Other indicators such as the number of consumers, the price of related goods, product discounts, and the cost of the products are also used for forecasting the future demand for stationery products. The most relevant literature relating to forecasting product sales of my family's business is the "*Applicability of Forecasting Models and Techniques for Stationery Business: A Case Study from Sri Lanka*" conducted by Dewmini Danushika Illeperuma and Thashika Rupasinghe from the University of Kelaniya. This paper presents a demand forecasting methodology for a stationery company in Sri Lanka. The data is collected and prepared as a monthly time series from April 2008 until April 2013. The demand for stationery products exhibits strong seasonality and cyclicality. Furthermore, the literature utilizes a combination of judgmental

methods, quantitative methods, and Artificial Intelligence methods as the literature proved to have generated higher forecasting accuracy. Time series forecasting methods used in the literature are single exponential smoothing, Croston's method, Moving average, Additive Winter, and ARIMA. Methods based on judgment are unaided judgment, prediction markets, and Delphi. Methods used for quantitative methods are extrapolation, quantitative analogies, and rule-based forecasting. Moreover, using all three methods, they built a model that produced a relatively low mean absolute percentage error (MAPE) that forecasts the sales of drawing books (Dewmini et al., 2013).

3 Data and Time Series Characteristics

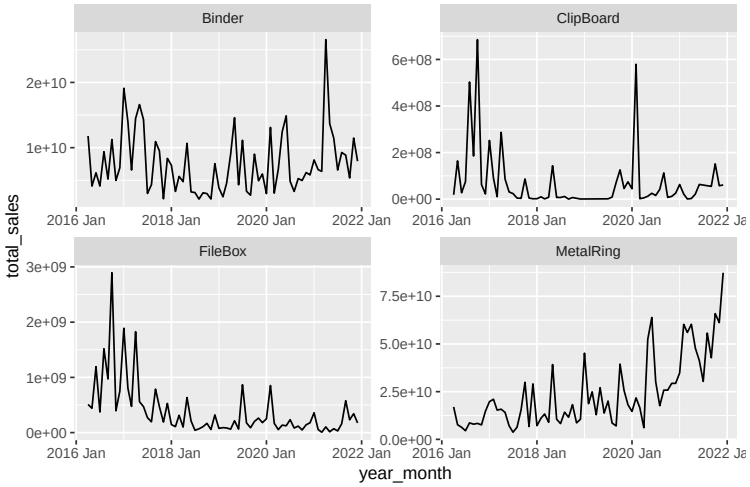
3.1 Data Collection and Preprocessing

The data are extracted directly from the company and include 45 different variables from which we choose six total variables in this study. There are a total of 12 datasets which include 6 datasets for sales of the past 6 years (2016-2021) and another 6 datasets for the sales returns for the past 6 years. The data is recorded daily where each row indicates a purchase made by a customer. This is not ideal as we want each row to represent total sales made by customers in a day. We corrected this by summing the values of sales and returns datasets in each day so that each row would indicate one distinct day. Moreover, the variables of interest were translated from Farsi to English and the dates were also converted from Jalili to Gregorian. Additionally, in the raw data, the company's product names column includes more than 100 unique names because each product was differentiated by its color. This is not of our interest and to fix this for our analysis, we grouped and combined all the products and their respective colors into one product. After the modifications the 12 datasets were merged and converted from daily to monthly data. Our dependent variable is the company's historical sales. And independent variables that will support our analysis and forecasting are: sales returned and the number of distinct customers. The company manufactures and sells four different products: binders, metal rings that are placed in the binders, clip boards, and file boxes. The variables of interest are prepared to give historical values of each product.

Graph 1 below shows the total sales with respect to each product. We can see that the product that generates the most revenue for the company is Metal rings. The graphs indicate strong seasonality in all the products. Additionally, there seems to be two cyclicalities in clip board sales: one in 2016 and another in 2020 with a sharp and unusual increase in sales. Similarly, there is a cyclicality in the sales of binders in 2022 with an unusal abropt increase in sales. Moreover, two products show clear trend. Metal ring sales have a clear

upward trend and file box sales have a clear downward trend since 2016. For binder sales and clip board sales the graphs do not seem to show a clear trend.

Graph 1: Total Sales with Respect to Each Product

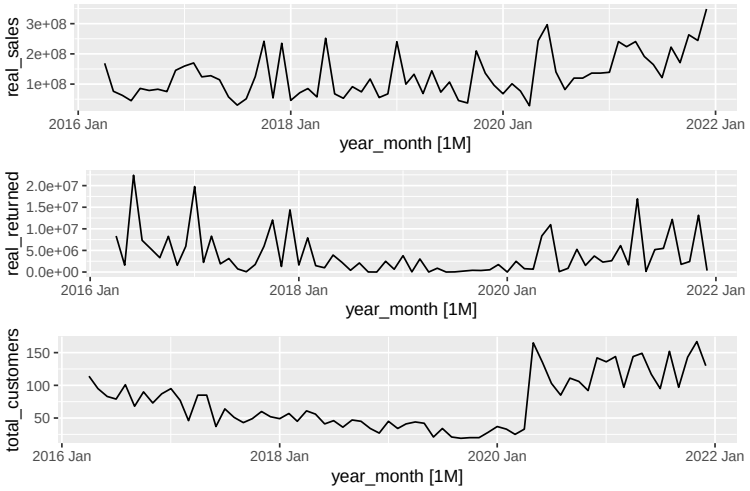


Note: Due to our time constraint, further analysis and forecasting in this project will examine metal ring sales as it generates the most revenue and therefore is the most important product for the company.

3.2 Inflation Adjustment

Total sales and returned sales are measured in nominal Rials (Iranian currency). Since inflation is often a significant component of apparent growth, it is important to adjust these variables for inflation to uncover the real values. Additionally, deflation may stabilize the variance of random or seasonal fluctuations and highlight cyclical patterns in the data (Nau, 2020). Deflation is accomplished by dividing a monetary time series by a price index. Our monetary time series are metal ring sales and metal ring sales returned, and the price index is Iran's Consumer Price Index (CPI). CPI for Iran is collected from the Federal Reserve Bank of St. Louis (World Bank, 1960). The inflation adjustment has been applied to all the indicated variables. Graph 2 shows the adjusted for inflation values. After inflation adjustment, we can see that the trend for the variables has been to some extent decreased, which signifies that the remaining trend is real growth. Furthermore, the seasonal and cyclical patterns and the sales' magnitude stand out more clearly when displayed in real terms.

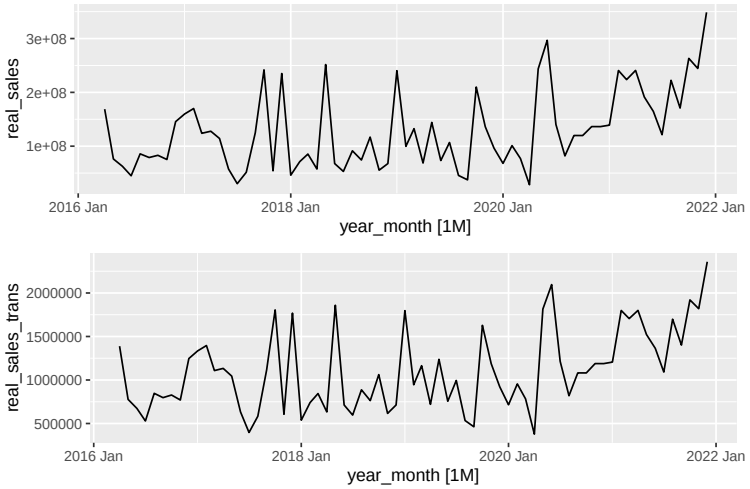
Graph 2: Selected Variables' Historical Values with Respect to Metal Ring



3.3 Box-Cox Transformation

The main use of variable transformation is to reduce changing variability and skewness and other distributional features that complicate analysis. Since our dependent variable as shown in Graph 2 have changing variablility over time, it is nessecery to correct this. Box-Cox transformation method is used to do so. For real sales, $\lambda_{optimal} = 0.730$, which suggest that square root plus linear transformation would be the best transformation method. Therefore, the transformed series is shown in Graph 3. The graph shows that changing variability was reduced in real sales transformation, but it is not completely eliminated.

Graph 3: Before and After Real Sales Transformation Adjustment



3.4 Time Series Decomposition

We use time series decomposition to better understand the series. The main objectives for a decomposition is to estimate seasonal effects that can be used to create and present seasonally adjusted values to better identify trends in the series. Since we already used Box_Cox transformation, we use additive decomposition on the transformed variables, which is computed below.

$$y_t = S_t + T_t + R_t$$

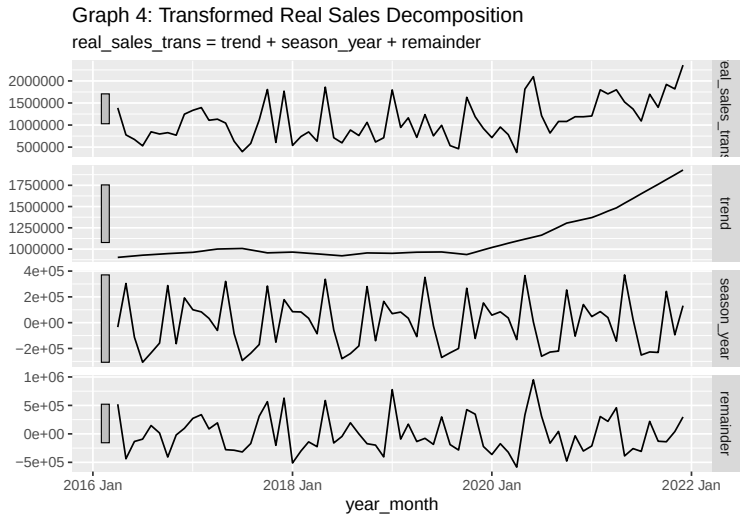
Where y_t = data at period t

, T_t = trend-cycle component at period t]

, S_t = seasonal component at period t

, R_t = remainder component at period t

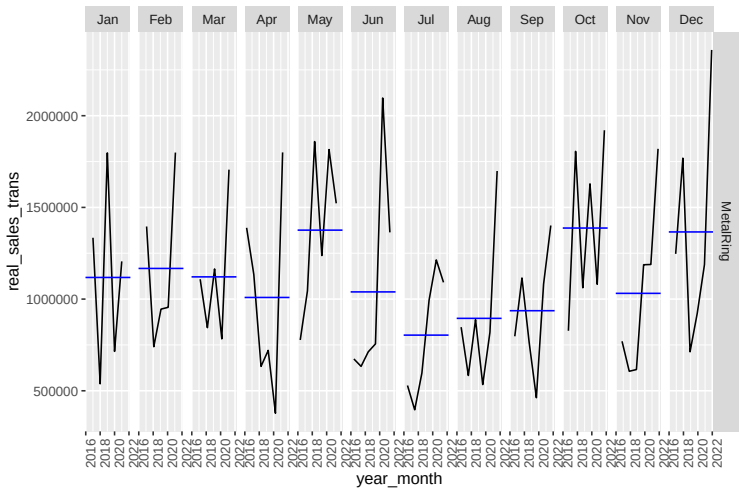
After decomposition, Graph 4 clearly dipicts an upward trend beginning from the late 2019, a strong seasonality, and some cyclicality.



3.5 Seasonality and Autocorrelation

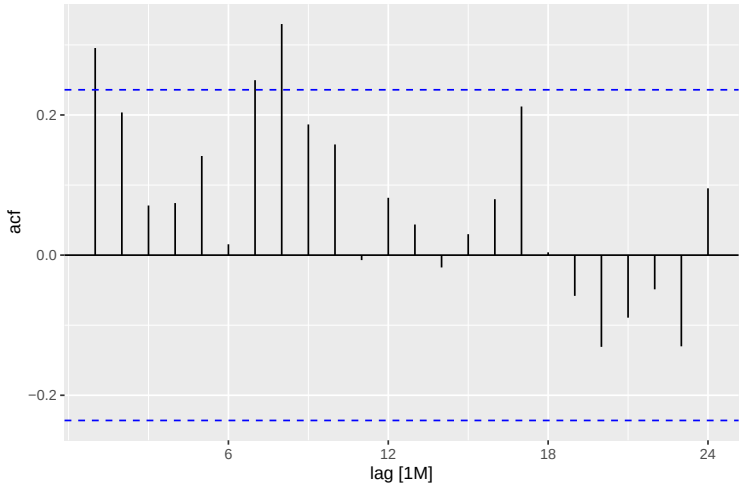
Seasonality Observations: gg_subseries in Graph 5 confirm our observation of the presence of strong seasonality and trend in the variables. Furthermore, the plot indicate that sales were highest in May, October, and December and lowest in July of each year.

Graph 5: Transformed Real Sales Seasonal Plot



Autocorrelation Observations: Graph 6 shows a particularly high correlation between current values and their first, seventh, and eighth lags, which denotes strong positive autocorrelation and hints strong seasonality at those lags. Additionally, the autocorrelations for small lags tend to be somewhat large and positive, which signifies some trend in the data. Moreover, the autocorrelation seems to get smaller in larger lags, which implies that each observation is more positively associated with its recent past.

Graph 6: Transformed Real Sales Correlogram



4 Application of Forecasting Methods

In this section, we attempt to forecast metal rings' transformed real sales by using various forecasting models. The forecasting methods implemented are SNAIVE, ARIMA, Piecewise Linear model, and multiple regression models. Furthermore, the train test split ratio used in the project is 80:20. Since our series

is strongly seasonal, our strategy is to forecast seasonal and non-seasonal components separately and then combine these forecasts to obtain forecasts for y_t, which is the transformed real sales of metal rings. Formally,

$$\hat{y}_t = \hat{S}_t + \hat{A}_t$$

$\hat{y}_t$ obtains forecasts of transformed real sales as the sum of forecasts of seasonal component ($\hat{S}_t$) and forecasts of seasonally adjusted component ($\hat{A}_t$).

Model 1: The SNAIVE model is used as our baseline model and models y_t directly. This model is chosen to be compared with the performance of more sophisticated models. The SNAIVE model is also chosen because our series exhibits strong seasonality.

Model 2: For this model, we forecast y_t by forecasting S_t and A_t separately by using decomposition_model() function. By default, this function uses SNAIVE() to forecast the seasonal part (S_t). And to fit a model for A_t, it uses an ARIMA framework.

Model 3: This model doesn't forecast S_t and A_t separately, rather it models y_t directly using ARIMA() with transformed_real_sales as its argument, which is y_t.

Model 4: this model forecasts y_t by forecasting S_t and A_t separately by using decomposition_model() function. By default, this function uses SNAIVE() to forecast the seasonal part (S_t). model 4 uses a regression framework to fit a model for A_t. Since our dependent variable (Transformed Real Sales) has some linear trend we have included a trend variable such as:

$$A_t = \beta_0 + \beta_1 t + \beta_2 \text{realReturned} + \beta_3 \text{totalCustomers} + e_t$$

Model 5: Similar to Model 4, this model also uses decomposition_model() function to forecast St and uses a regression framework to fit a model for At. The difference between the two models is that this model also includes the first lag of the two independent variables in the regression shown below.

$$A_t = \beta_0 + \beta_1 t + \beta_2 \text{realReturned} + \beta_3 \text{lagRealReturned} + \beta_4 \text{totalCustomers} + \beta_5 \text{lagTotalCustomers} + e_t$$

Model 6 and 7: Model 6 uses a linear regression model with the main purpose of comparing with Model 7, which uses a piecewise linear regression model. As shown in Graph 4, we can see that the data has a clear upward trend starting from October of 2019 onwards. We placed a kink in October 2019 and used a

piecewise linear model to better capture this behavior. Model 6 is as shown below:

$$\text{transformedRealSales}_t = \beta_0 + \beta_1 \text{tLin}$$

And Model 7 is as shown below:

$$\text{transformedRealSales}_t = \beta_0 + \beta_1 \text{tLin} + \beta_2 \text{t19Oct}$$

4.1 Model Accuracy Analysis

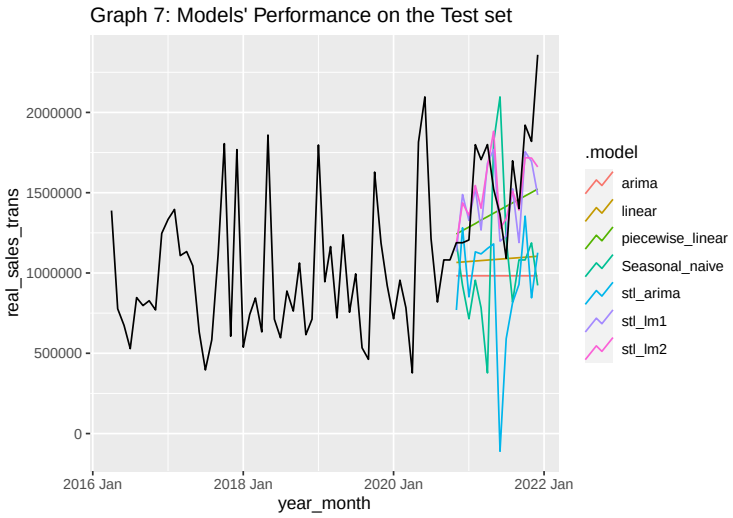


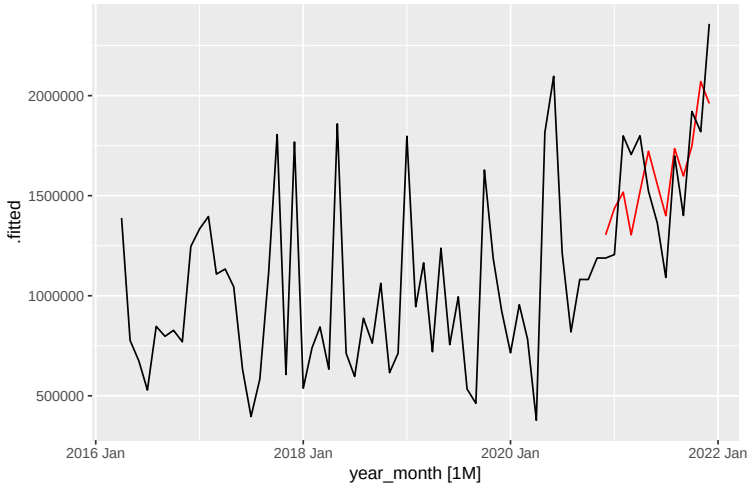
Table 1: Accuracy Results

.model	.type	ME	RMSE	MAE	MPE	MAPE
arima	Test	593550	686892	593550	34.649	34.65
linear	Test	491106	596546	491106	27.931	27.93
piecewise_linear	Test	193006	358036	281847	8.707	16.36
Seasonal_naive	Test	493259	782741	657255	26.640	38.66
stl_arima	Test	644739	748815	658033	40.116	41.23
stl_lm1	Test	127905	313469	243289	5.551	14.74
stl_lm2	Test	69933	271601	214000	1.713	13.12

According to table 1, the best-performing model is stl_lm2, the 5th model in section 4 analysis. The RMSE value for this model is the lowest, and the MAPE value shows only around a 13 percent error rate,

which indicates that the model is relatively accurate. Overall, we can say that multiple regression models performed better than other models. Additionally, we can confirm that the piecewise linear model performed significantly better than the linear model, suggesting that placing a kink in October 2019 resulted in a more accurate prediction. Lastly, we can see that both Arima models performed worse than our baseline model (seasonal_naive), indicating low prediction accuracy.

Graph 8: Best Performing Model on the Test Set



5 Conclusion

The goal of this paper was to predict my family's business product sales to improve the company's sales performance. By analyzing and preparing the company's data, producing visualizations, and building various forecasting models to predict the products' future sales, we were able to gain valuable insights. Additionally, we were able to build a model with 87 percent prediction accuracy that is 24 percent more accurate than the baseline model performance. This model will give the company a better understanding of the company's sales historical pattern and will improve business decisions. In the future, I plan on expanding this project to give more insights about the company's sales and customers. Additionally, I plan to include more independent variables in my analyses, such as the nation's GDP and product sales discounts. I also plan to use the Vector Autoregression model (VAR) and other Machine Learning models to forecast the sales and come up with an even more accurate model. Lastly, I plan on expanding my analyses to all four products.

6 Bibliography

[1] Frey, William H. 2021. "What the 2020 Census Will Reveal about America: Stagnating Growth, an Aging Population, and Youthful Diversity." Brookings (blog). January 11, 2021. https://www.brookings.edu/research/what-the-2020-census-will-reveal-about-america-stagnating-growth-an-aging-population-and-youthful-diversity/.

[2] Aurmanarom, C. (2010). An exploration of the impact of brand personality on consumer buying intentions toward specialist stationery products across age groups. https://ro.ecu.edu.au/theses_hons/1235

[3] "Stationery Market Analysis." Market Research Company, n.d. Accessed February 28, 2022. https://www.factmr.com/report/stationery-market/toc.

[4] "Stationery Producers Reclaiming Iran's $1b Market." Financial Tribune, September 11, 2017. https://financialtribune.com/articles/economy-domestic-economy/72197/stationery-producers-reclaiming-iran-s-1b-market.

[5] Danushika Dewmini, Thashika Illeperuma, and Thashika Rupasinghe, "Applicability of Forecasting Models and Techniques for Stationery Business: A Case Study from Sri Lanka," International Journal of Engineering Research 2 (November 1, 2013): 2319–6890.

[6] Inc, Intuit. n.d. "What Is Demand Forecasting?" Accessed March 1, 2022. https://www.tradegecko.com/ebooks/demand-forecasting.

[7] Nau, Robert. "Inflation Adjustment." Fuqua School of Business, Duke University, 18 Aug. 2020, https://faculty.fuqua.duke.edu/~rnau/Decision411_2007/411infla.htm.

[8] World Bank, "Consumer Price Index for Islamic Republic of Iran," FRED, Federal Reserve Bank of St. Louis (FRED, Federal Reserve Bank of St. Louis, January 1, 1960), https://fred.stlouisfed.org/series/DDOE01IRA086NWDB.