Sales Data Analysis and Forecasting Report

The CEO has a mandate of expanding the company internationally next year. To accomplish this goal, he has determined the company's workforce will need to expand by 20% and marketing spend will need to increase by 15%. Annual revenue must be projected to grow by at least 15% for the next 12 months to accommodate the increase in costs and time it will take for marketing to take effect and the international expansion to be fully realized. With a high degree of accuracy, what will the company's projected revenue be next year; and does projected revenue support the CEO's expansion plans?

E-commerce stores have exploded in popularity over the last 2 decades. Having a forecast on next year's sales is extremely beneficial to a business. It gives executives information on how fast/slow to grow the company, manage inventory to minimize waste, when to test new products, etc. The goal is to build a model that can accurately predict future sales which will allow the executive team to make a data driven business decision as to whether or not to follow through with expansion plans that are pivotal to the company's growth and success. This should be done with a high degree of accuracy (low margin of error).

Our data wasn't too dirty this time but did require cleaning. First I checked for duplicates and found no duplicate rows in the data. Order Date & Ship Date were converted to Datetime so we can use the data in EDA. The Sales column was renamed to Sales Revenue so as not to be ambiguous with the number of sales. Sales Revenue was then rounded to two decimal places. Postal Code was renamed to Zip Code as that is more commonly used. The only null values we had were in the Zip Code column so nulls were filled as 99999 as this is not a real zip code. The Zip Code column was then converted to strings then the decimal point and trailing 0 were dropped. After further inspection, some zip codes had only 4 digits which I determined was because the leading 0's were dropped because they were floats. A leading 0 was added to these 4 digit zip codes to correct this.

All other columns were inspected to ensure there were not similar errors. In doing so, I found that all values in the Country column were United States so this column was dropped as it provided no further value in our analysis.

Our Sales Revenue column was then statistically inspected and the max value seemed to be a very large outlier. I plotted the distribution and confirmed this. To ensure this observation was not an error to be dropped, I filtered for only Sales Revenue values over 1,000 and found that there were 462 instances. These were plotted as a boxplot and as there were quite a few observations outside 1.5*IQR, I determined that these values are not in error, they are simply large orders.

I then looked at the distribution of orders with Sales Revenue less than or equal to 1,000 and found that almost half of all orders in our data are $50 or below in sales revenue.

Turning to look at Sales Revenue over time, we can see we have sales data spanning 5 years so the initial plot was too cluttered to be that useful. After aggregating Sales Revenue per month, this showed we have some seasonality in our sales but our sales overall are increasing over time.
Lastly, we exported the cleaned data to a CSV file.

Sales Trends and Patterns

When we analyze monthly sales revenue over time, the patterns in our sales become more apparent. We can see that we do have seasonality in our sales with a peak in sales revenue coming in November and December with a sharp dropoff in January and February. This is confirmed by the bar plot showing all sales over time by month.

We also can see that, despite the seasonality, our sales overall are increasing over time at a growth rate of 22.62% over the last 5 years. Looking at the most recent years, it appears that sales are growing at a faster rate of 51.42%.

Insight 1) The fastest way to grow revenue would be to address our lowest sales months of January and February. To increase sales in these months without cannibalizing holiday sales, it would be beneficial to assess the sales volume left in customer's carts after the holidays and aggressively deploying email campaigns to push those customers over the finish line. We could also release new products in January to drive new sales.

Customer Analysis

The consumer segment clearly brings in the most revenue, with home office accounting for the least revenue.

We've also identified our top 20 customers by revenue and top 20 customers by total orders. This is great to know as our power users are also the most likely to refer.

Next we calculated the Recency, Frequency, and Monetary Value (RFM) of our customers. We can see that we have a right skew distribution for all 3 metrics with most customers most recent order coming within the last 50 days, 10 being the most common number or orders a customer makes, with most being between 7-14 orders, and the majority of customers monetary value being $3,000 or less. We also see a moderate correlation between Monetary Value and Frequency.

Insight 2) We can use targeted ad campaigns aimed at our top customers and offer them referral bonuses and exclusive discounts. To take it a step further to reward these customers' loyalty, we can explore a brand ambassador program to incentivize these customers to drive sales for us.

Geographical Analysis

The West region narrowly edged out the East region as the region accounting for the most Total Revenue with Central and South regions accounting for the least Total Revenue.

The three states with the highest total revenue were California, New York, and Texas, while the three states with the lowest revenue were North Dakota, West Virginia, and Maine.

The three cities with the highest revenue were New York City, Los Angeles, and Seattle.

Insight 3) In analyzing state and city data, we find that although New York produces the highest sales revenue of any city, California produces more sales revenue than the state of New York due to 3 California cities being in the top 10 of sales revenue. I would recommend deploying more marketing dollars in California and New York City only to maximize ROI of the increased ad spend.

Product Analysis

Our product analysis shows that the Canon imageCLASS 2200 Advanced Copier is by far our top selling product by sales revenue.

On the whole, the technology category is our highest revenue source.

Phones and chairs are the sub-categories that account for the highest share of total revenue.

Insight 4) Although the Technology segment does account for more sales revenue than the other categories, it isn't by much. They are all within $150,000 of each other. I'd recommend we analyze our costs of goods sold per segment first and only then deploy additional resources toward the segment that has the highest profit margin.

Revenue Analysis

When looking at only orders greater than $ 1,000 in sales revenue, the boxplot shows us that with as many outliers above 1.5 * the inner quartile range as we have, we just being a number of large orders in our data rather than there being some kind of error.

When looking at orders of 1,000 in sales revenue or less, we can now more easily see that almost half of all orders are $50 or below in sales revenue.

Insight 5) The majority of our orders are smaller orders under $50. Depending on the

direction the company wants to go in, we should either:
a) Double down on this low revenue purchase behavior of our customers by carrying more products in this price range, or

b) Assess the profit margins on our more expensive products and adjust pricing to increase total order size.

Modeling

After training 3 different models and forecasting the next 12 months of sales revenue, these are our results:

The Prophet Model predicts a 25.8% increase in sales revenue next year.

The Holt-Winters Exponential Smoothing Model predicts a 39.43% increase in sales revenue next year.

The SARIMA Model predicts a 25.09% increase in sales revenue next year

It is very encouraging that 2 of the models were within 1% of each other in their predictions. For our use case in using these predictions to make the determination of whether or not to proceed with such a large capital expenditure in expanding the company's territory, we would want to take a very conservative approach.

To this end, we will use the lowest prediction to decide whether it is prudent to proceed with the expansion or not. The SARIMA Model had the lowest prediction; an increase of 25.09% in sales revenue next year. This is well above our required threshold of 15% annual growth, so the data shows we can proceed with the expansion.