‘

**Technical Report**

**Sales Data Descriptive and Predictive Analysis**

Shannon School of Business, Cape Breton University

2022 Winter Predictive Modeling/Analytics (MGSC-5125-11)

Professor: Dr. Alejandro Vital Soto

March 30, 2022

Submitted By:

Anand Kumar (0253413)

Jasleen Kour (0254437)

Bindu Priya Lingala (0246499)

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Page Number** |
| 1 | Executive Summary | 3 |
| 2 | Introduction | 4 |
| 3 | Project Description | 5 |
| 4 | Data Description | 6 |
| 6 | Conclusions | 16 |
| 7 | References | 17 |
| 7 | Appendixes | 18 |
|  |  |  |

**Executive Summary**

**Project objective:**

This project was conducted to help answer some key business decisions for the future to ascertain which regions and items would generate high revenues and what key factors play a role in its increase or decrease and to check for relationship among various sales parameters.

**Methodlogy:**

For our sales predictive analytics we conducted our analsyis on MS-Excel using Data Analtyics toolpack on the 5000 sales records. The sales records were a time series data ranging from 2010 to 2017. We identified the key fields and formed the key business queries. The data fields were carefully examined to check for any missing or null values. To prepare forecast, some additional calculated fields were also included in the dataset like Lead estimation.

For our predictive analytics we used regression models like multiple regression, forecasting Models like Moving average, Simple Exponential, Holt-winters additive model and also conducted hypothesis testing to predict the revenue estimates for the future pertaining to different regions and different items.

**Conculsion:**

The majority of business revenue is generated from the sales of the household items and office suppiles and least from the fruits and meat. The revenue estimates for the Sub-Saharan Africa show a slight decline for the future. The revenue figures are directly impacted from the amount of units of an item and unit cost of each item alongwith sales channel and number of lead days.

# **Introduction**

Rapid development of Integrated systems has led to emergence of large amounts of data which is a huge challenge for organizations to manage and analyze. The analysis of sales data is particularly important for organizations as it provides basis for key business decisions. The sales data provides an insight into how the company is performing in terms of products sold, customer sales, unit costs and total revenue generated. Sales data retrieval is a big challenge not only because of the voluminous amount of data that is generated but also because of various attributes which comprise the data.

The sales data plays an important role in making business decisions pertaining to production, operations, logistics and marketing. It helps the organizations to forecast their revenues and consequently improve their performance in areas of concern by highlighting the possible causes for it. As the nature of the sales data becomes complex, more advanced analytical techniques are required to extract information and make predictions about the future.

The analyzing the historical sales data records for making predictions about the future is the end goal of the project

# **Project Description**

This project is based on sales data of a company which export the various items to different geographical locations around the world. The dataset is uniform in nature, doesn’t have any missing values. Although it is covered as per different regions parameter such as, Asia, Australia and Oceania, Central America and Caribbean, Europe, Middle East and North Africa, North America, Sub-Saharan Africa. Under these regions they are covering 186 countries. The dataset is about 12 types of items which has been distributed around this network of countries. All the orders have been taken by just two modes wither offline or online and order priority is of 4 types mentioned as L, M, H, and C. Now, the order dates attribute consists of dates spread from year 2010 till 2017 and shipping date is just some days past the order date for any respective order. Each order has its unique order ID. All these orders are recorded with the data of units ordered, unit prices, unit cost, and total cost, total revenues generated and total profit. So, all this comprised of 14 different column and 5000 rows of rich data to be analyzed. This data was taken from the data.world resource.

We meticulously examined the data and fund that it has very diverse content to apply different types of methodologies and find certain link in the data which will eventually help us to predict revenue forecasts for the future. The goal of the project is to perform descriptive analytics and predict revenue figures for the future along while studying the patterns observed in the data. For our analysis we first re-grouped the dataset as per continents and not region. The reason to that is, the regions which were mentioned in the data were too vague to consider and didn’t reflect any set boundaries. Additionally, as this is a big dataset, we took this approach to minimize the regions prior to starting the work. This helped us to understand the data better as per geography which we use generally wo segment different regions. From all the recorded sales we found that the data for sales occurred in 2017 was too available till the last month but till July instead. For this flaw, we have not used the sales data of 2017 in most of the case studies. To gain better insights for our data analysis we framed some key business questions which are as follows:

Q1. Analyze the region-wise revenues to forecast estimates about future for region with highest revenue generation. Which forecasting model gives better forecast results.?

Q2. Identify the factors that led to decrease in revenue for North American region? Check if any multicollinearity exists within the different sales parameters for the region.

Q3. Apply the KNN to predict the sales channel for an item with given lead time, unit price and unit cost for an item listed in Sales data for Canada.

Q4. Predict the Sales (units sold) for upcoming quarters across all continents.

Q5. From the available data of orders per continent, try to find the independence of the group of count of order dates for each year.

# **Data/ Description of Data**

The dataset was collected from the Data.world website as it had comprehensive dataset with around 5000 sales records. The dataset is uniform in nature, doesn’t have any missing values. Although it is covered as per different regions parameter such as, Asia, Australia and Oceania, Central America and Caribbean, Europe, Middle East and North Africa, North America, Sub-Saharan Africa. Under these regions they are covering 186 countries. The data was collected from year 2010 to 2017. The data has around 14 fields which comprises of around 5000 records .

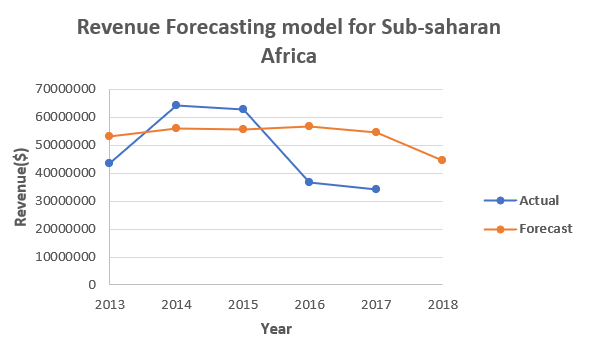
Description of the fields is given below;

|  |  |  |  |
| --- | --- | --- | --- |
| **SNO** | **Field/Variable** | **Datatype** | **Definition** |
| 1 | Region | string | Name of each region where sales were recorded |
| 2 | country | string | Sales figures by country |
| 3 | Item category | string | List of items sold |
| 4 | Order date | Date/Time | The date at which order was made |
| 5 | Shipping date | Date/Time | The date at which the order was executed. |
| 6 | Unit cost | Number (Decimal) | The cost of producing a single unit of an item |
| 7 | Unit price | Number (Decimal) | The price at which a single unit of an item is sold. |
| 8 | Order priority | string | The priority list for item order execution. |
| 9 | Order\_ID | string | The unique ID for each order. |
| 10 | Units sold | Number (decimal) | Total units of an item sold. |
| 11 | Total revenue | Number (decimal) | Total revenue generated over the years. |
| 12 | Total profit | Number (decimal) | Total profit generated over the years. |

**Q1:** For the region-wise analysis, the revenue figures were filtered on the basis of region demonstrating cumulative revenue figures from 2010 to 2017 using pivot tables. The referenced fields used for this analysis were region, total revenue and item category. The pivot table identified Sub-Saharan Africa as the region with the highest cumulative revenue calculated as $400329148.31. For conducting further analysis, we identified household items and office supplies that have contributed to high revenues across the regions for the company.

Using scatter plot, we visualized the revenue figures for Sub-Saharan Africa to identify any trend or seasonality. As per the scatter plot, no specific trend or visible seasonality exists in the total revenue generated till 2017. Thus, moving average and exponential smoothing were used to predict the revenue for year 2018.

**Forecasting using moving average method**: The moving average method was applied to the revenue figures of Sub-Saharan Africa for Household items to identify the underlying direction in which the revenue figures are changing.

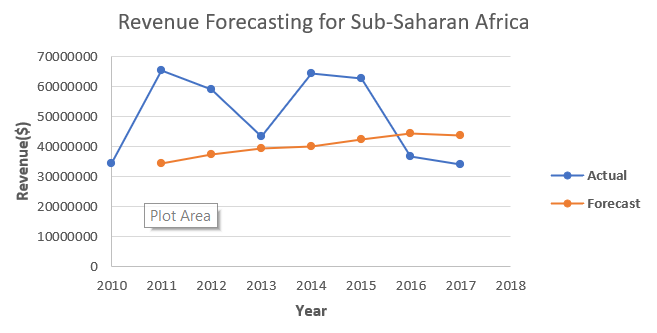


*Figure 1 using Simple moving Average Forecasting model*

The simple moving average model is computed on the basis of recent ‘k’ observation wherein the value of k=3(assumed since larger values result in smoother forecast). For checking the efficiency of our Forecasting model we calculated various error metrics pertaining to the forecasted values.

**Forecasting using exponential Smoothing:** Exponential smoothing was also conducted on the same data using actual revenue figures of Household items for Sub-Saharan Africa.

For better computation of forecast the value of has been assumed.



*Figure 2 using Exponential Smoothing Forecasting Model*

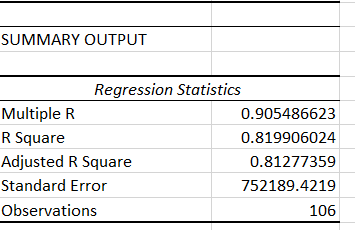
Various metrics were also calculated to make a comparison of the two forecasting models. The value of the Mean Absolute Percentage error (MAPE) for the both the models are quite identical for moving average model and Exponential smoothing model of 31.96% and 30.32% respectively which means that both the models are equally appropriate in forecasting the revenue for 2018. So, the company can use any of two models to forecast the data for next year. For MAD and MAPE errors following formulas were referred;

Mean absolute deviation Error (MAD)=

Mean Absolute Percentage Error (MAPE)=

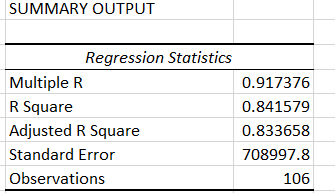
**Q2:** North America was identified as the region with the lowest revenue generated over the years from 2010 to 2017. To identify which factors were significant in the drop in total revenue generation we used three referenced fields in the data namely, unit cost, units sold and sales channel. Another field that was used is lead estimation by using the order date values and shipped date values.

To determine which factors are significant, regression Method from the data analysis tool in MS-Excel was performed on data using total revenue values as the dependent variable and unit cost, units sold, sales channel and lead estimation as the independent variables. The following table shows the key values for regression model.



*Figure 3 Regression Model for North America with all variables*

The value of R represents the multiple correlation coefficient as 0.905 which indicates that revenue figures are highly predictable by other dependent variables. Also, the R Square value of 0.81 suggests that around 81.8% of variation in revenue can be explained by the changes in dependent variables. But as in the figure 3 except units sold and unit cost none of the dependent variable are significant (p-value >0.5). So, we removed them from our model one by one to access if the any interaction among other dependent variables is significant which shows interaction for all dependent variables is significant (p>0.05).

**

*Figure 4 Regression model for North America with only significant variables*

We can clearly see that value of R and adjusted r square reflect changes after interaction was also included in the regression model to analyze the total revenue figures for North America as the adjusted R-square value has increased from 0.81 to 0.83 stating that regression model has improved. Thus, no dependent variable could be removed from the regression model as all have impact in sales revenue across North America. To predict the revenue unit cost, units sold, sales channel and lead estimation have to be considered.

**Q3:** As one of the key Business decisions is to predict the best sales channel for sale of items across a country. For our analysis we selected Canada sales data for all items collectively. The referenced fields selected were unit cost, unit price and lead estimation days. Firstly we normalized the values for unit cost , unit price and lead estimation using z-score normalizing method wherein the mean average and standard deviation were used. For taking a decision on sales channel we used K-Nearest Neighbor Method in MS-Excel for which given data is taken as training data and the Decision variable is the testing data.

For the training model part all of the data is used and for the hyperparameter k, we will use the k=5, since the dataset is of the appropriate size. For our testing model we used a query as sales entry with unit cost, unit price and lead estimation values as $250, $550 and $18 respectively.

For the prediction step, we calculated the distance between the new observation which is the query in this case and all the observations in the training dataset. As for the k=5, the shortest distances calculated are as above which are identified as Record 3, 6,2,8 and 4

using Euclidean Distance excel formula = SQRT(SUMXMY2)

Thus, as per the majority decision we can conclude that the stated query for an item with unit cost as $250 , unit price as $550 and lead estimation days as 18 , the appropriate sales channel decision is selling the product online.

Q4. From the available dataset, we have inputs for units sold as per various countries and group of countries which can be re-grouped as continents. This would help us analyze the data much better with existing geography and segregate the data for units sold for a certain period. Moreover, the data has been categorized as per quarters and not year. This has been done so we can bring more meaning from the numbers as year’s data doesn’t give much flexibility to practice different methods and variations in the predictions. Although, as per the line charts on time series for the units sold gave us a clear picture of the seasonal accent, trend present in the data, we still checked it statistically and found that by using moving average, exponential smoothing the error metrics were high and decided to not move forward with that approach. Finally, by look of the eye we found that the data did have seasonality and trend on time series plot for this selected data. Therefore, the method Holt-Winter’s Additive model with seasonality and trend seemed to be an ideal choice for such data. Initial values for Level, Trend and Seasonality are calculated using the following formulae’s:

Level: Text

Description automatically generated

Trend: Text

Description automatically generated with medium confidence

Seasonality: A picture containing logo

Description automatically generated

Since the quarters under first four years (2010 - 2013) complete a full cycle of increase and decrease of units sold, they were chosen as the reference data and forecast was implement on next quarters of years (2014 - 2017). As we did not have the entire data for the year 2017, the forecasts of the year serve the purpose of the performance of this statistical model and calculations. For forecasted numbers formula used is:

Forecast: A picture containing clock, watch

Description automatically generated

As we proceed, we need to update the values of all the factors used for forecasting in the following manner:

Level component: 

Trend component: 

Seasonality component: 

To continue to forecast the units sold beyond k periods/quarters of seven years of data (or 28 quarters) which is last point of observed data. we use updated forecast method as below:



As the entire predictive model was formulated in link to the values of Level, Trend, and Seasonality component after several trial of altering the values, we found values of 0.1, 0.8, and 0.3 respectively to be the close fit for the component values. We also included calculations of Root Mean Squared Error to be the error metric in this statistical model to help us understand the difference of Observed and Calculated forecast values of units sold. This helped us understand what values for components included in this study best reveals the forecasts.

A picture containing text, clock, watch

Description automatically generated

**Q5:** Across the entire geographical area where the company has its root, they have stored the data for all the order dates each year and tried to find if there is any significant difference in number of orders received each year per continent. An important point to notice here is, we have not considered the data for the year “2017” because the data is not complete and only present till month of July. Instead, we only consider the data from year “2010” to “2016” and present our findings. First the data has been segmented year-wise for each continent mentioning total count of order dates (which is summation of all the days each country received an order). The data has been grouped year-wise (considered independent) and analyzed for Null Hypothesis that mean of each group is equal.

HO : µ1 = µ2 = µ3 = …. = µn

H1 : at least one mean is different from others

We have used ANOVA- Single Factor method from Data Analysis tool in Microsoft Excel to conclude the results of our Null Hypothesis. This tool is available in Data Analysis toolkit under Data. As shown in the ANOVA summary, we have in total 7 groups with Count, Sum, Average and, Variance parameters. This gives us a brief knowledge of the descriptive spread for the groups. ANOVA stands for Analysis of Variance used to check if there is any statistical difference between the means of the all the chosen independent groups. As given in the results, the calculations for “between groups” signifies that we cannot reject the Null Hypothesis because the calculated F-statistics value < F-critical value (0.0078 < 2.4453) and p-value is greater than the selected alpha value (0.9999 > 0.05)

# **Conclusion**

Overall, the revenue generation for most of the regions has been largely impacted by the sales of Household items and Office supplies. The key regions to look out for in a forecast would be Sub-Saharan Africa and North America. The revenue figures are dependent on various factors like unit cost, units sold, lead time and sales channel used. The forecast for coming years shows a slight decline from the previous year figures. The suitable sales channel decision can be made considering the unit price , unit cost and number of days an item order is executed and shipped.

# **References**

[1] Business Analytics, 3rd Edition by James R. Evans, Pearson 2019

[2] 5000 Sales Records by Jonathan Major, Data.World 2019

*https://data.world/bobmajor/sales/workspace/file?filename=5000+Sales+Records.csv*

# **Appendixes**