**PAPER 1:**

**Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management**

**By: Matthew A. Waller1 and Stanley E. Fawcett2**

Source: https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12010

Year: 2013

This paper illuminates the myriad of opportunities for research where supply chain management (SCM) intersects with data science, predictive analytics, and big data, collectively referred to as DPB . Highlights growing combination of resources, tools, and applications has deep implications in the field of supply chain management (SCM), presenting a doozy of an opportunity and a challenge to our field .

Over the past few years, research has been done trying to understand the DPB’s implications for research and education in business logistics and SCM which reveals that with the vast and growing amount of data, domain knowledge and analysis cannot be separated.

To identify and solve problems with SCM we need a Broad awareness of many different methods of estimation and sampling, understanding application of qualitative and quantitative methods of forecasting, Numerical methods of optimization, capital budgeting and determining opportunity cost.

The paper defines  *‘SCM data science is the application of quantitative and qualitative methods from a variety of disciplines in combination with SCM theory to solve relevant SCM problems and predict outcomes, taking into account data quality and availability issues’*.

The paper identifies that lack of domain knowledge might lead to False positives.

False positive arises as there is huge number of attributes which lead to correlation that do not really exist.The problem is that as the number of variables increases—that is, as the use of a theoretical data mining proliferates—the chances of false positives increases exponentially which lead to which result in wasted time and money.

Sheds light on use predictive analysis (Both qualitative and quantitative) in logistic & SCM.

It calls for forecasting , optimization of implementable solutions, Applied probability along with application anchoring and framing effects from psychology and data mining as methods to improve supply chain design , storage of inventory and competitiveness.

Stresses for a need for data capture at multiple points in the supply chain and consumer sentiment, these data must be analyzed and quantified which can help in optimizing Forecasting, Inventory management, Transportation management, Human resources for Users, retailers and manufacturers in various capacities.

PAPER 2

**Leveraging Insights from “Buy-Online Pickup-in-Store” data to Improve On-Shelf Availability**

Source: Google scholar. http://matthewalanham.com/

By: Pranav Saboo, Sachin U. Arakeri, Shantam D. Mogali, Sushree S. Patra, Zaid Ahmed, Matthew A. Lanham , Krannert School of Management, Purdue University, West Lafayette, Indiana

Year: 2020

How does it relate to our problem statement?

* The dataset was collected from a national retailer which consists of online transactions from January 2019 to December 2019 for 3,547 products across 246 stores. This is very similar to the dataset we have chosen, and has almost similar columns.
* The paper describes a study that is based on the motivation to reduce lost sales opportunities
* by improving on-shelf availability (OSA), subsequently improving the overall revenue and
* profitability of the retail store. This is one of the principal ideas that we had too, to predict demand vs supply of stock.

Assumptions:

The primary focus of the paper is to predict if a stock is OOS(Out of stock) or not OOS. Since the data is skewed towards not OOS, techniques like upsampling have been used to make the classification more accurate.

Main claims:

* Dimensionality reduction was performed to remove features that were highly correlated to avoid redundancy, risk of overfitting, and segments that were extremely skewed observed increment in F1 score on average, by 60%.
* Engineered features showed significant improvement on model performance.
* To overcome the computational challenge, leveraged all the cores of the server by exploiting multiprocessing capabilities.
* Random forest to be the best performer, works great on large datasets and also assists in extracting variable importance.

Takeaway:

* Several approaches and detailed information regarding data cleaning, preprocessing, upsampling, target transformation etc have been clearly elucidated.
* The clear representation of the process of arriving at a good solution to the problem statement gives a fair idea of the path to be followed to come up with an accurate model for our dataset.
* Several good models have been compared, giving insights on models and key paths that we too could pursue and improve to solve the problem statement.
* The paper also led us to several other informative research papers and articles of interest in the demand forecasting domain

PAPER 3

**Big data analytics and demand forecasting in supply chains: A conceptual analysis**

Erik Hofmann, Emanuel Rutschmann, "Big data analytics and demand forecasting in supply chains: a conceptual analysis", International Journal of Logistics Management, The, <https://doi.org/10.1108/IJLM-04-2017-0088>

Year: 2020

How does it relate to our problem statement?

* This paper aims to examine how demand forecasting can be made more accurate by employing data analysis methods.

Retail chains have an immense amount of data on their customers, like recent purchases and items they commonly buy together. Most retailers do use this data to provide more relevant deals or recommendations to the customer in order to increase sales. However, by analyzing the relation of forecasting methods to big data analytics as outlined in this paper, insights can be provided to enterprises to apply in their operational, tactical, or strategic demand plans.

Assumptions:

* The various forecasting methods used are not likely to disappear in the near future, but will rather be improved and adapted.
* that specific analytic types fit “better” with certain types of forecasting methods and that only alignment between the input, scope, and method ensures appropriate analysis of data to support decision-making for demand planning and forecasting.

Main claims:

* First, the improved forecasting method could directly replace the conventional methods.
* Second, the improvement of the method could enable it to be employed in a different situation.
* Third, the improvement could lead to a shift in methods in which an improved method replaces another method.
* The effect of price changes could be modeled when sufficient data on analog changes in the past becomes available. Moreover, complex time-series can be predicted more accurately with causal methods than with time-series.

Takeaway:

* In most situations, forecasting methods will benefit from BDA.
* An exemption are causal models created with AA, which introduced a new way of forecasting that will replace other methods. Causal models use recognized patterns and domain knowledge to make predictions as analyzing the past is always a necessary first step to discover hidden patterns.
* Three main forecast situations that benefit from BDA have been identified. Strategic long-term estimation using self-service EA, tactical medium-term forecast situations using DAA, PDAA, and EX, and operational forecasts which can benefit from improved time-series methods.

Other References, EDA summary

Paper 4 :

**Increasing Supply-Chain Visibility with Rule-Based RFID Data Analysis**

The paper describes the tool's core concepts include rule-based analysis techniques and a map-based representation interface. With these features, it lets users visualize the supply-chain structure, together with performance metrics, and detect problematic hot spots.

Collection of data is done through RFID technology that can automatically generate event data that digitally describes how physical entities, such as single items or pallets, move through supply-chain processes. With this continuous stream of data we can analyze it with data mining and other techniques to detect and locate inefficiencies in the chain.

It sheds light on Supply Chain Visualizer Concept which aims to reduce the complexity with analysis techniques while still keeping the user in the decision-making loop.

To begin, we use a shipping and receiving model and the notion of shipping, receiving, and internal events to dynamically build up the supply-chain model combined with the concept of our rule-based analysis.

The rule-based analysis, preprocess all incoming event data in two steps. First, it splits up grouped events into single events and creates chronologically ordered linked lists for each individual item . Second, we calculate time differences (based on eventTime), distances (based on the geo coordinates translation of Location), and the resulting movement velocities (based on eventTime and Location) for each pair of preceding and following events.

The analysis starts at each root of a product flow and checks each of the currently enabled rules for each time ordered pair of events denoted as ei and ei + 1. Every time a pair of events fails a check, the analysis engine generates a new inconsistency event within the location of ei. The following rules represent checks for basic supply-chain conditions that we developed during discussions with industry experts.

Paper contained other visualisation techniques Map-Based View and Performance metrics.

TABLE I  
EDA Summary

| 1 | Number of rows | 180519 |
| --- | --- | --- |
| 2 | Number of columns | 53 |
| 3 | Number of categorical columns | 23 |
| 4 | Number of numerical columns | 20 |
| 5 | Number of columns with significant NULL values | 3 |
| 6 | Number of irrelevant columns (intuitive) | 12 |

* Number of rows - 180519
* Number of columns- 53
* Number of categorical columns - 23
* Number of numerical columns -20

Data Cleaning and Feature Selection - drop columns that are unimportant

* Columns with too many null values
  + Order Zipcode 155679
  + Product Description 180519
  + Customer Zipcode 3
* Irrelevant columns
  + Customer email, password
  + Redundant columns due to address splitting

Scope to Split the dataset into different domains - such as transactions, geographical locations etc before modelling, so as to obtain a model specific to our problem statement and reduce unnecessary computation.

Correlation Insights

Some interesting insights observed from the correlation matrix after performing pearson's correlations are:

* Benefit as well as sales of a product depend on Days of shipment
* Discount and discount rate are correlated to the department of the products
* Sales is directly correlated to price, and discount as well as days of shipment
* Similarly profit per order
* Product Status -whether its in stock or not - depends on sales and price

Project Roles (Tentative)

1. Split Visualisation & Detailed Literature Review into broad categories, allot and explore individually.
2. Since the dataset has lot of attributes we will see if PCA will help us get better correlations
3. Explore separate approaches to solve the problem statement and compare.

Content of PHASE 1 report:

**Abstract**

Over the last decade, with the onset of the internet, the outlook of retail and businesses has changed drastically, and multichannel (i.e, offline and online) retail has become extremely popular. It becomes essential to bring about coordination between technologies. However, the massive upsurge in both quantity as well as diversity of data have resulted in datasets that are no longer manageable manually or even by conventional management tools. The motivation for this study is to come up with unique implementations to maximise profits of an online retail store, by analysing the Supply Chain Management (SCM), forecasting demand using the massive datasets from previous sales in order to facilitate effective replenishment, ensuring sustainable development and subsequently predicting the time it takes for orders to reach customers. This research has been built on multiple other papers that have been done in the area, putting together theoretical concepts of SCM and predictive analysis.

Key words : Data Analysis, Supply Chain Management, Demand Forecasting, Trend Analysis, Regression

**Introduction**

Today, E-commerce and Online Retail have become extremely popular, and more so during recent times and through the pandemic. Products are now widely accessible in the remotest of locations, and this has led to a massive surge in sales. Rising competition among various e-retail providers has also upped the importance of customer satisfaction, which depends prominently on appealing prices and timely delivery, which in turn depend on availability of stock. Furthemore, with the rising need of the world moving towards sustainable development, minimising loss and overstocking in warehouses becomes one of the primary points that can be acted upon.

The rise in online transactions and also the use of artificial intelligence and IOT(methods such as RFID) in retail locations has resulted in the availability of several new and potentially invaluable datasets, and new methods of data science, primarily predictive analytics, have been developed. The fundamental process of a supply chain however, remains unchanged. Most research today is focused on bringing together a combination of integrated supply chain management and data science concepts, and this is often referred to as *supply chain predictive analysis* and sometimes as *DPB (Data Science, Predictions and Big Data).*

*A definition of SCM predictive analysis : SCM predictive analysis is the application of quantitative and qualitative methods from a variety of disciplines in combination with SCM theory to solve relevant SCM problems and predict outcomes, taking into account data quality and availability issues. [1]*

This paper aims to use concepts from existing research and optimize the final stages of a supply chain by predicting upcoming sales of products in a multi-purpose online store, and in turn use it to predict an optimized time of delivery (TOD).

**[~1 page] What have others done to solve it - critique others' approach and cite the work**

The advent of technology has made it clear that data analytics and successful businesses go hand in hand. Over the years, research has been done trying to understand implications of research and analysis in business logistics, revealing that with the tremendous amount of data being generated every day, domain knowledge and analysis cannot be separated. More specifically, as the number of variables increases - i.e, as the use of a theoretical data mining proliferates, the chances of false positives increases exponentially, which results in wasted resources.

The use of both qualitative and quantitative predictive analysis calls for data captured at multiple points in the supply chain as well as consumer sentiment, which can be used in forecasting inventory,transport and manpower, designing solutions that use applied probability along with effective data mining, and in optimization of implementable solutions. [1] One study gives interesting insights about direct and indirect relationships between the supply chain flow, customer services and finance, and talks about the importance of an integrated supply chain strategy where customer service is the yield of the entire system, achieved by synchronizing the requirements of the final customer and reaching a balance between high customer service and costs. [4] Decision Variables and Performance measures deeply affect the accuracy of the problem at hand. [5]

Efficient Supply Chain Management (SCM) has been an area of interest for several years, with a significant number of works in academic literature, each focusing on approaches ranging from regression to deep learning methods and neural networks. Data Preprocessing, Dimensionality reduction, Effective Sampling and Feature engineering are highlighted to be important steps in developing an effective model. Data gathered in the field is often found to be skewed, and techniques such as upsampling has been used effectively in some studies to improve accuracy of models, along with methods like target transformation. [2] In [6], the products were classified into four categories in order to be able to analyze the correlation of the sale of the products by channels, and then pre-processed to remove the outliers identified through statistical methods.

The problem of Demand Forecasting in order to predict Stock replacement has been approached in several ways by different people. In [6], the author describes combination of K-means clustering and ANNs (NAR and NARX) , and the model has been validated using the by the R2 coefficient of determination to observe the adjustment of the forecast equation and the Mean Squared Error (MSE) to analyze the accuracy measure, which revealed that NARX presented a good performance for all the products.

Another paper [2] compares the use of Logistic Regression, Random Forests and AdaBoost to first classify products into categories, and then compare their F scores, to find that random forests work best in product classification. The results are then used to test different models on different categories to specifically predict Out Of Stock(OOS) products. This process is already accounted for in our dataset, as categories are recorded too.

Bibliography

[1]Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management By: Matthew A. Waller and Stanley E. Fawcett,

University of Arkansas

Weber State University

2013

[2]Leveraging Insights from “Buy-Online Pickup-in-Store” data to Improve On-Shelf Availability By: Pranav Saboo, Sachin U. Arakeri, Shantam D. Mogali, Sushree S. Patra, Zaid Ahmed, Matthew A. Lanham

Purdue University

[3]Big data analytics and demand forecasting in supply chains: A conceptual analysis

Erik Hofmann, Emanuel Rutschmann

University of St. Gallen

Deloitte Consulting AG

2018

[4]The effects of an integrative supply chain strategy on customer service and financial performance: an analysis of direct versus indirect relationships Shawnee K. Vickery, Jayanth Jayaram, Cornelia Droge, Roger Calantone

Michigan State University

University of South Carolina

2003

[5]Supply Chain Design and Analysis:Models and Methods By: Benita M. Beamon

University of Washington

1998

[6]Towards a predictive Approach for Omni Channel Retailing in Supply Chains By: Marina Meireles Perieira, Enzo Morosini Frazzon

Federal University of Santa Catarina

2019

[7]Increasing Supply-Chain Visibility with Rule-Based RFID Data Analysis By:Alexander Ilic, Thomas Andersen, and Florian Michahelles

ETH Zurich

2009

M. M. Pereira and E. M. Frazzon, “Towards a Predictive Approach for Omni-channel Retailing Supply Chains,” *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 844–850, Jan. 2019.

E. Hofmann and E. Rutschmann, “Big data analytics and demand forecasting in supply chains: a conceptual analysis,” *The International Journal of Logistics Management*, vol. 29, no. 2, pp. 739–766, Jan. 2018.

M. A. Waller and S. E. Fawcett, “Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management,” *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, Jun. 2013.

A. Ilic, T. Andersen, and F. Michahelles, “Increasing Supply-Chain Visibility with Rule-Based RFID Data Analysis,” *IEEE Internet Comput.*, vol. 13, no. 1, pp. 31–38, Jan. 2009.

S. K. Vickery, J. Jayaram, C. Droge, and R. Calantone, “The effects of an integrative supply chain strategy on customer service and financial performance: an analysis of direct versus indirect relationships,” *J. Oper. Manage.*, vol. 21, no. 5, pp. 523–539, Dec. 2003.

B. M. Beamon, “Supply chain design and analysis:: Models and methods,” *Int. J. Prod. Econ.*, vol. 55, no. 3, pp. 281–294, Aug. 1998.

P. Saboo, S. U. Arakeri, S. D. Mogali, S. S. Patra, Z. Ahmed, and M. A. Lanham, “Leveraging Insights from ‘Buy-Online Pickup-in-Store’ data to Improve On-Shelf Availability,” [Online]. Available: http://matthewalanham.com/Students/2020/ICDATA20\_Leveraging%20Insights%20from%20Buy-Online%20Pickup-in-Store%20data%20to%20Improve%20On-Shelf%20Availability.pdf.

<https://www.kaggle.com/sukanthen/e-commerce-multi-output-models-project-cse07>

**[~0.5-1 page] Proposed problem statement with the specific issue you intend to address**

The scope for optimisation of fields such as transportation, inventory management and human resources etc. is immense, and especially so with the availability of valuable datasets. The data set that has been used in this research consists of a DataSet of Supply Chains used by the company DataCo Global describes Provisioning , Production , Sales , Commercial Distribution.

Supply chain data has a vast scope for optimisations in various fields such as transportation, inventory management and human resources.

This project concentrates on Inventory management from the perspective of the distributor, and customer satisfaction.

Aim: To perform statistical analysis and visualisation on product popularity, profit, sales and supply, and draw insights to relate various attributes in the dataset.

Some probable predictions that are expected from this dataset are:

* Category wise demand forecasting of products.
* Predicting the shipping days by finding a meaningful relation between shipping days and other attributes such as location, product prices and demand/sales.
* The dataset can be divided into frames that explain location, product, order details and history of the customer, to be able to find hidden patterns between these different sets of attributes. This would contribute to a knowledge base for further optimisations.

Further, this project aims to produce visualisations and statistical models in such that the statistics behind it are abstracted and the real world meaning is effectively conveyed.

**[~1-2 pages] How is your approach (or the type of problem you are looking at) different from what has already been done? (or if you are attempting to improve upon someone else's work, explain in what way it distinguishes itself from what has been reported)**

For companies with perishable stock and for fast paced e-commerce it is important to provide impeccable customer service. In order to keep up with the ever increasing demands, a company has to make the best out of available customer purchase data.

The objective is to meaningfully use present data to forecast the demand for each of the departments so that the provider is always one step ahead of the customer.

Through EDA, it was found that the dataset has as many as 23 categorical variables whose effect is not visible while using Pearson's Correlation. So, it was concluded that Feature Engineering would be extremely valuable in order to extract maximum valid information for the predictive model (Feature Selection) .

Forecasting the future demand for products of each department based on purchase history,delivery history and location of the warehouse , is to be using statistical methods for forecast , Trend Projection and visualisation.

The time series data includes as many as 20,000 rows for sales of 2 months - October and November.

One notebook on Kaggle [8] describes a Multi-output Regression model with target Features - Days for shipping (real) and Days for shipment (scheduled). This project also aims to experiment with the Regression model so as to find attributes of the Dataset that predict the actual shipping rate effectively and how Demand and Supply fit into the model.

Since the dataset has 53 attributes, experiments on the regression model are expected to be done through:

1. *Attribute selection*

The attributes can be broadly classified into those that explain the delivery details and those that explain the product. To find what combination of these attributes would predict the shipping delays better.

1. *Dimensionality reduction [PCA].*

Content of PHASE 2 Report