Capstone Project
Fall 2022

Data Science for Supply Chain Analysis

University at Buffalo
School of Engineering and Applied Sciences

By: Venkata Sesha Aditya Mantri Pragada

1. Abstract

The process of industrialization and increasing number of customers has made process management and balance of the supply and demand requests as essential part of any organisation. Supply chain procedures involve managing customer requests and ensuring steady progress in any product or service providing company. This essay on Data Science for Supply Chain Management demonstrates how data science is being implemented in today's evolving corporate industry to promote technological growth. Data science provides efficient problem-solving tools such as machine learning models and database management system to efficiently manage, generate and use data. From manual data entry to automated cloud-based data management, data science offers multiple methods for data handling, data warehousing and data pipeline construction. To better understand the implementation of data science tools in Supply Chain Management research paper study, hands on analysis and case study analysis in IKEA are well explained in this essay as per literature review practices implemented in scientific journals. All approaches aim to convey the underlying aspects of data science in supply chain practices of modern companies. Introduction section establishes need for data science and ongoing processes in supply chain that use data science. Background describes how supply chain management efficiency and functionality improves using data science methods and tools. Methods section involves studying data science methods from the perspective of research paper, hands on analysis and case study of a real-life company "IKEA". Towards the conclusion section, relevance of data science methods and techniques to solve real world supply chain problems is clearly conveyed and all the learnings in this essay are summarized briefly.

2. Introduction

Oracle a software technology solution provider and inventor of world's first autonomous database named "Oracle" defines supply chain fundamentally as "management of flow of goods, data and finances related to a product or service from procurement stage to the final delivery". Over the times industries have evolved from small scale industries to large scale industries based on their profits and increase in customer demands for new products. In the process of growth, the companies began to generate multiple inputs of data from machines, people and measuring instruments. To store the ever-increasing data that has been produced computers have proved to be an efficient tool. Evolution in modern computing has taken place from Charles Babbage's computing machine to IBM Watson of recent times to use Artificial Intelligence in answering complex calculations. To accommodate the huge amount of data, modern computing has come up with cloud-based software and platforms to provide storage, retrieval and multiple services for the data of many organisations. To deal with complex structure of organisations in an industry and how data flows between the organisations many SaaS platforms are used. Implementation of supply chain management is crucial for data transfer between large organisations. However, computers have limitations of their own in processing multiple requests. This situation can be better understood by the phrase: "Computers are fast but not infinitely fast and computer memory is inexpensive but not free." These limitations and load distributions of an organisation are managed by data science tools and algorithms. Algorithms are crucial to data science as many machine learning libraries in Python language involve usage of data structures with logics explained using searching, sorting, divide and conquer and merging algorithms.

The above explanation helps us to understand why data science is an essential element in modern day supply chain management as both of them share a common task of managing and directing the workload in large organisation by implementing software technologies under a set of guidelines defined by specific organisations.

3. Background

Implementation of Data science to promote Supply chain management functions promotes domain integration. In the modern era of computers and automation, data science has found many applications in supply chain processes to automate computing, financial management, order planning and inventory control. Efficient inventory control through forecasting is crucial to ensure a supply and demand balance between the client and customers. Warehousing for inventory raw material in mechanical production

companies helps to prevent shortage of manufacturing materials. Forecasting models developed in R Studio are used to predict quantity of raw material required to manufacture a product. This quantity prediction is important for reducing material wastage and improve shelf life of the manufactured product.

Establishing the points and goals of a discussion makes it easy and clear to convey the ideas and approaches that need to be communicated to the audience. By establishing the background, key concepts and methodologies implemented in supply chain can be better understood. In the research paper written by Jayashankar M Swaminathan, the term Supply chain management has been defined as: "Supply chain is a set of entities involved in design of a product or service from the designing stage to the consumption stage." The term supply chain was coined by Keith Oliver a consultant at Booz Allen Hamilton. The end-to-end process of supply chain includes processes such as "procurement, planning, forecast, consumption and distribution. "The supply chain cycle between a supplier and customer involves elements named: Demand and Supply planning, raw material sourcing, manufacturing and operations, logistics for distribution and order-based customer care. There are various flows that occur through above elements named: "Supply Information flow depends on Cost, capacity and product design. "Demand Information flow" consists of customer requirement, Orders), "material flow" is functionally dependent on new components and "reverse material flow" is a function of returned/recycled products. When mentioning above terms and practically implementing them, cash flow associated with each process can occur in both the directions between the supplier and customer. Hence, anticipating the customer demand in advance requires forecasting. To understand a cash flow mechanism the income sources and expense values with their sources must be mentioned in the supplier's balance sheet.

Identifying income drains and minimising expenditure will require statistical analysis and data collection from all organisations involved in the supply chain cycle. To perform these tasks data science offers practical technologies and procedures to successfully construct data pipelines enabling data and insight transfer. Through the aid of data science, the supply chain life cycle gains an increase in processing speed and ensures efficient decision-making processes. Data science will help supply chain by performing functions of data collection, data cleaning, data analysis and Machine Learning Model predictions. Database management tools like "MySQL and PostgreSQL" are used for efficient management of data and generating functional dependencies between the relations in an existing database. By integrating programming languages like Python with PostgreSQL Relational Database Management systems (RDBMS) can be built to analyse and relate progress of a product across different stages in a supply chain life cycle. In the modern-day application to cater to large amounts of data cloud services and SaaS (Software as a Service) platforms have proven to be helpful tools. AWS and google cloud offer storage as well as online data processing tools. AWS and Snowflake have made construction and management of data pipelines across multiple platforms an easy task. This large amount of information the is generated across multiple organisations by different instruments is referred to as Big Data. Apache, Spark and Hadoop offer big data handling systems. Hadoop Distributed File System (HDFS) operates using data nodes which ensure replication and more secure and fast handling of large information. So Big Data handling in Supply chain cycles is being made efficient using tools like, AWS, Hadoop and Python. Further methods involved in solving Supply chain industry problems can be better understood when studied through case studies, journals and the ongoing works of companies and organisations around the world. The implementation of above methods using data science techniques and methodology is explained in "Methods" section.

4.0 Methods

The methods of implementation of data science ideas in supply chain can be studied through research papers, personal case study experience and ongoing implementations in companies. Studying implementation of data science techniques in Supply Chain Management will make understanding of the topic easy to understand and implement.

4.1 Financial Risk Analysis of Data-Driven Food Supply Chains of USA

The financial risk management model under the Internet financial model is proposed using data science methodologies. During the cash flow between multiple entities of the internet supply, the amount of cash flow is bound to change. If the change of cash flow indicates a decrease this might pose a risk to all entities in the supply chain being observed. To predict this risk in advance, counter measures of the risk are stated and the financial risk model is constructed based on multivariate piecewise regression analysis. Fuzzy decision method is used to analyse the risk assessment of the internet supply chain.

The key element of a financial risk model is "Financial risk index system". Correlation value of parameters helps to decide a suitable variable to determine the correct choice of hyper parameters for the model training. Correlation values are represented in heatmaps. Heatmap is a representation of the correlation matrix with values ranging from 0 to 1. The parameters when fit into a model will determine the risk index system. This risk index system is specific to the enterprise being studied. To construct the index system, selection of indicator parameters needs to be done as per the financial goal of an enterprise. The types of indices that are used in supply chain of an internet enterprise are" Liquidity index, Asset management capability index, Indebtedness index and Profitability indicators."

In the scientific paper authored by Qi Feng Yang, Yingying Wang and Yidong Ren, the method of piecewise regression analysis is combined with statistical analysis and comprehensive decision making to perform a financial risk model building. The financial model discussed in the paper uses mean to understand the structure and profit change rate of an Internet supply chain. The value of variance in this model shows how much the structure of the prediction varies. Mean and variance calculated using this model are used as parameter for the fuzzy decision-making model to give a quantitative measure of the financial risk associated with an internet supply chain. Apart from the implementation, the underlying mechanism of a model is a statistical function.

The empirical analysis performed in the research revealed that volume ratio of market value as 0.201 and correlation value as 2.283. The data analysis in the research paper is performed using Excel 2007 and SPSS 19.0 in combination with MATLAB mathematical programming. Inferring from the discussion of the author, mathematical implementation of data science techniques in MATLAB helped in the analyses of financial risk model. Using the correlation value and volume ratio, the financial risk model predicted that risk decreases year by year. Using this information more funds were allocated to start new ventures to help in the growth of an upcoming enterprise. Overall, the financial model discussed in this paper has good piecewise fit for financial risk management and management of supply chain under the internet financial model. In addition to the existing analysis, KNN(K-Nearest-Neighbours) can be implemented to identify median of the volume ratio values. KNN is an unsupervised machine learning model used for improving accuracy of model training process. This integration will make the existing model more efficient and increase test accuracy. Thus, the model developed in this paper using financial modelling proved to be effective in reducing the financial risk associated with the supply chain.

4.2 Risk Analysis of Supply Chain in Banking System

Banking systems require more cash inflow than cash outflow to be functioning and expanding business to larger customers. Cash Inflow and Cash Outflow are two major factors that impact how well the supply chain between bank and multiple customers keeps building. Cash inflow of a bank is earned through Customer income, security/checking accounts and repaid loans. Cash outflow from the bank is caused by Online payments, Mortgages collected, Credit card transactions and sanctioned loans. Banks need to predict the loan interest rates as per the customer demand to increase their business profit. Keeping Cash inflow greater than Cash Outflow is crucial to the financial growth of a bank.

In the case study below a dataset of an upcoming bank is studied using Cash inflow and Cash Outflow. To refine this study, customers who take healthy loans are identified. Customers who have Income, Security Account and Mortgage greater than their expenses through online, loans and Credit card are considered healthy customers. In this manner the percentage of healthy loans being approved must be kept in check to

ensure continued financial growth. To better understand the cashflow between bank and customers as two large organisations in a Supply chain following dataset from Kaggle is selected. The dataset has columns named 'ID', 'Age', 'Experience', 'Income', 'ZIP Code', 'Family', 'CC Average', 'Education', 'Mortgage', 'Personal Loan', 'Securities Account', 'CD Account', 'Online', 'Credit Card'. As per the above inferences, columns describing cash inflow are Income, Securities account and Mortgage. Cash outflow from the bank occurs due to "Personal Loan, Online and Credit Card". Keeping these points in mind data is analysed in a Jupyter Notebook using Python Scripting Language.

Data is imported from a CSV (Comma Separated Value) file using pandas' library and stored as a dataframe. Null entries are removed to reduce errors. Columns with text values and irrelevant to cashflow calculation are removed. Net cash inflow is calculated as sum of Income, Security accounts and mortgage received by bank. Net cash outflow is calculated as sum of loans, online and credit card payments. The cash inflow value is 369695 $ and cash outflow value is 286948 $. Observation study revealed that cash inflow to the bank is greater than cash outflow. Hence, the existing Bank has a positive cashflow and growth is possible. To refine the analysis every transaction performed by individual customer is studied and the customers with healthy loans are identified.

To identify healthy loans as per the filter condition, cash inflow and cash outflow values are calculated for each customer. Customers with cash inflow greater than cash outflow are filtered from the data source. Out of the total 5000 customers in the bank dataset, 3537 customers have cash inflow greater than cash outflow. Cash inflow being greater than cash outflow indicates the customer's potential to repay the loan. So, the percentage of healthy loans approved by the bank are 70.74%. While studying the cash inflow system to the bank, 56.55% inflow was from customer income and 43.3% income was from Mortgages. Securities account and CD account contributed 0.08% and 0.04% respectively. This means that in future campaigns the bank needs to improve their management and usage of Securities and CD account. Also, the bank has income and Mortgages as the largest contributors to cashflow. To ensure good financial health, banks need to branch out their business and start growing their Securities account contribution as well. Otherwise, the maintenance of the Securities account might turn out to be a financial risk in the near future.

In the programming aspect, Python dataframe helped in easy categorization of cash inflow sources and identifying the primary contributors. Observing the banking system as a whole might not give a complete picture of all customers. Matplotlib library in Python provides pie charts to develop visuals to understand contribution split up for each cash inflow category. Balance sheet analysis is implemented in Supply chain to identify healthy growth of the banking system thus improving future business prospects and increasing profits. So, in depth analysis using data collection, cleaning and processing will help to identify healthy loans that are being approved. Through this hands-on experience, the healthy loans are identified and the bank should take initiatives and campaigns in future to further increase their inflow sources and percentage of healthy loans approved. Future scope of this hands-on case study can be extended to include effect of customer's family members involved with the same bank and categorisation of loans of based on reason for approval. This analysis emphasizes on understanding multiple reasons behind loan approval systems in banks. On the whole this analysis explained multiple contributors in a supply chain and how they affect the banking system.

4.3 Supply Chain Analysis in IKEA

IKEA offers a variety of furniture products and household solutions. It has a large variety of global producers and consumer audience. To understand the company process in a better manner, a case study has been conducted by T. Sund and M. Asplund as cited and explained in the following sentences. Being a large organisation IKEA has many workflows to cater to a large variety of customer needs. The following elements are used in the case study to describe elements of each workflow. Product is manufactured and refined for sale. Order means a request for the manufacturing of multiple products. Shipment refers to the transportation of the manufactured product it specifically refers to when and where the product is located. Consignment refers to the place from which customer has requested the product. Pallet is a flat structure

used to transport products. The above discussed terms are used as validators for analysing the shipment tracking in IKEA products. The case study performed by the author devised a product level tracking system that goes beyond current systems to generate product level information to improve traceability. By improving traceability, the case study aims to provide users with permissions to approve a product's origin and shipping routes. This idea is a customer centric idea as it helps in improving the availability of multiple shipping options for users.

To perform data collection and data storage for this new project System design is implemented using JavaScript and Node.js as the environment. The environment consists of six components which are described as follows. Client application interacts with users to perform data collection. Data encryption and signed transactions are performed here. Controller is present between the client and validator to schedule transactions and perform load balancing. IFPS (Interactive Financial planning system) it is an off-chain file storage connected to a network of nodes. Validator uses a software named "Quorum" to validate transactions and make decisions to process orders to shipment stage. Smart contracts consist of object contract with object information and Ownership registry that controls access and ownership to the previous object contracts. To perform this case study Amazon was chosen as a cloud environment. Elastic Cloud compute also known as EC2 was the choice of cloud environment for the test. EC allows to setup multiple virtual machines through an API. Environment configuration and setup are automated for each order using Terraform and purpose-built scripts. Through loads generated by client orders, 10 orders were processed per minute. This was made possible initialising purpose-built script for each test case from the environment.

At the end of this case study both benefits and future improvements are identified and stated as follows, the bottlenecks identified are that, terraform cannot handle more than ten requests per minute for shipment stage of the supply chain. Network Latency can also be a factor that impacted the performance. It was also found that in an organisation such as IKEA to improve their performance, IKEA needs to partition their processing into multiple subsystems. During epidemic of COVID-19 it is essential for the sever to be able to handle large number of requests as the number of online customers increased exponentially. Inferring from these observations, the case study is a prototype which needs further upgrade through load balancing severs. The benefits were that each processing request was solved despite high processing time. Also, there was no failure of client-side environment developed by JavaScript and Node.js.

Supply chain trouble shooting organisation like IKEA includes procedures such as expired load, document mismatch, damage to product, logistic malfunction and payment issues. The case study discussed above provides a suitable methodology to address these customer requests. Data collection and analysis during these stages of the supply chain cycle will help in locating and correcting these errors. Machine Learning algorithms are implemented in Supply chain to solve tasks such as "Demand-Supply match making, Pricing and Incentives and User segmentation based on functions". Deep learning and Neural Networks are used to improve accuracies of the existing models used to perform these functions. Continuous update and training of training and testing datasets of a model is crucial to performance improvement of a model.

5. Conclusion

In this Capstone stone Data science in Supply chain has been studied from three different perspectives. The three perspectives covered are namely "Research paper, Hands on analysis and Case study on an Industrial Supply chain process". On the whole, these processes helped to understand how different Machine Learning models are implemented in Supply chain and how insight van be drawn from the results. The first research paper discussed is based on Supply chain model of the internet to analyse financial risk. This model helped me to understand how parameters of choice influence efficiency of the model and how financial risk is being forecasted. Implementation of MATLAB and Excel formulas helped to perform the data cleaning and data processing stages of the analysis. The statistical function implemented as a Machine Learning model helped to speed up the processing of a large number of entries involved in the research paper. In the second explanation a case study is implemented based on Kaggle dataset of a loans approved in a bank. The concept of balance sheet is understood as sum of all cash inflow sources compared with sum of all cash outflow

sources. Dataset taken from Kaggle was cleaned and processed to remove errors before starting analysis procedure. In the analysis phase cash inflow of bank turned out to be greater than cash outflow. So, the bank had an overall good financial health. Upon examination of each customer, only 70% turned out to be healthy loans. Accordingly future campaign suggestions were proposed. This hands-on analysis also identified loan and mortgage as primary contributors to the bank's cash inflow. This analysis helped me to practically calculate, measure and perform data science methodologies in the context of banking system analysis. The future scope of the analysis can be extended to providing dashboard analysis and reporting to a real time client to help in decision making processes for bank campaigns. In the last discussion the case study analysis of IKEA is considered. The supply chain elements involved in IKEA production workflow were listed out and their functions are studied. The case study of interest explored a methodology following data science technologies to improve the traceability of products. The case study helped to improve the efficiency of supply chain in IKEA. The case study also provided user with multiple ways of determining their transport route to deliver the shipment.

Summing up, the three ideas of understanding supply chain management provide a good literature review and process description of ongoing supply chain methodologies. This essay provides a detailed step by step understanding of Data science applications in Supply chain. It is crucial to stay up to date with the ongoing phase of Artificial Intelligence and Automation of Industrial processes. The knowledge of Artificial Intelligence has made Supply chain management more efficient and improved product quality of large organisations such as IKEA. Through data science tools Supply chains can predict customer demands in advance and ensure faster delivery mechanisms between customer and client. The value of Research articles and Industrial Engineering processes in the field of supply chain is highly noteworthy. Hands on experience and continuous learning will be the driving forces for innovation and growth in the field of Supply chain management through data science. In this manner integrated approaches across multiple domains of the industry will help in the innovation process of new products and services for future development and technological advances.

Word count: 3927

6. References

[1] Supply Chain Management (SCM) | Oracle. (n.d.). https://www.oracle.com/scm/

[2] Lu, L. X., & Swaminathan, J. M. (2015, January 1). Supply Chain Management. Elsevier eBooks. https://doi.org/10.1016/b978-0-08-097086-8.73032-7

[3] S. Benzidia, N. Makaoui, and O. Bentahar, "The impact of big data analytics and artificial intelligence on green supply chain process integration and hospital environmental performance," Technological Forecasting and Social Change, Apr. 01, 2021. [Online]. Available: https://doi.org/10.1016/j.techfore.2020.120557

[4] "Bank_Loan_modelling," Kaggle, Aug. 29, 2018. [Online]. Available: https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling/data

[5] Vmantrip, "GitHub - vmantrip/Bank-Loan-analysis: Healthy loans are identified. Balance sheet analysis is performed to understand cashflow.," GitHub. [Online]. Available: https://github.com/vmantrip/Bank-Loan-analysis

[6] T. Sund, C. Lööf, S. Nadjm-Tehrani, and M. Asplund, "Blockchain-based event processing in supply chains—A case study at IKEA," Robotics and Computer-Integrated Manufacturing, Oct. 01, 2020. [Online]. Available: https://doi.org/10.1016/j.rcim.2020.101971

[7] "Trends of Data Science and Applications," SpringerLink. [Online]. Available: https://link.springer.com/book/10.1007/978-981-33-6815-6