EAS 504: Applications of Data Science – Industrial Overview – Spring 2023

-Lecture by ManojKumar Rangasamy Kannadasan

Name: Tananki Ranga Sai Saran Rohit

UB ID: 50441793

UB Email: rangasai@buffalo.edu

Ques 1 : Describe the market sector or sub-space covered in this lecture :

The market sector or sub-space covered in this lecture is Role of Data Science in eCommerce. Data science is a multidisciplinary field that uses scientific methods,processes, algorithms and systems to extract knowledge and insights from structured and unstructured data . Data science is now a days offered as a product to its customers, that's how data science is wide spread in the industry today. Data science plays a crucial role in eCommerce as it allows businesses to leverage the power of data to gain insights and make data-driven decisions.Data science helps Empowering management to make better decisions , Directing actions based on trends-which in turn help to define goals,because data driven decisions are always better than decisions.Data scientists are often treated as trusted advisiors to the management ,as they help to take better data driven decision for the wellbeing of the organization. Data science can help eCommerce businesses optimize their inventory management by predicting demand and ensuring that products are stocked in the right quantities at the right time. E-commerce companies use data science to analyze customer behavior and recommend products that best match their interests and past purchases. This can increase customer satisfaction and improve conversion rates.

Ques 2 : What data science related skills and technologies are commonly used in this sector?

E commerce Companies who are interested to solve their business needs ,they develop state of art solutions that typically integrate descriptive, predictive and prescriptive elements. Data science also helps when dealing with large and complex data in various forms (audio,text,video etc.) and multidisciplinary approach. Search, SEO, Ads/Marketing, Structured Data are various departments. There are different moralities of search like text search, Faceted search , image search, voice search, conversational Search , Recommendations . As we all know , e – commerce companies have a lot of products ,which implies it has to store and process lots and lots of data .So, E-commerce companies use big data technologies such as Hadoop and Spark to process and analyze large volumes of data. E-commerce companies use statistical analysis to analyze customer behavior, sales trends, and other related data. This requires the use of statistical methods such as regression analysis, hypothesis testing and time series analysis. E-commerce companies use machine learning algorithms to analyze large data sets and extract insights. This includes techniques such as clustering, decision trees and neural networks. Deep Semantic Similarity Model (DSSM) is a neural network model used to measure the semantic similarity between two texts. DSSM is trained using a large corpus of text data and uses a technique called pairwise ranking loss to optimize the model parameters. E-commerce companies use cloud computing platforms such as Amazon Web Services ,GCP and Microsoft Azure to store and process data, such that they can scale quickly as per the growing business demands. Understanding user behavior and expectations, Domain expertise. All these data science skills and technologies mentioned above necessitates a multi-disciplinary team helps in solving business challenges in this sector.

Ques 3 : How are data and computing related methods used in typical workflows in this sector? Illustrate with an example.

The applications of information data and computing related techniques are best explained using an example that how an e-commerce platform works . The online shopping platform initially collects information about customer behavior such as browsing history, search queries, purchase history and product reviews. The collected data is cleaned and preprocessed, such as removing irrelevant data, handling missing values, and correcting errors and this collected data is then used to analyze the data and identify patterns using various machine learning and data mining techniques. For example, clustering algorithms can be used to group customers according to their behavior, while associative rule mining , market basket analysis can be used to identify common products. Based on the insights gained from analyzing the data, the e-commerce platform creates a recommendation engine that can provide customers with personalized product recommendations. This engine uses machine learning algorithms such as collaborative filtering, matrix factors or deep learning to recommend products that the customer is likely to be interested in. E-commerce companies use statistical analysis to analyze customer behavior, sales trends, and other related data. This requires the use of statistical methods such as regression analysis, hypothesis testing and time series analysis. E-commerce companies use machine learning algorithms to analyze large

data sets and extract insights. This includes techniques such as clustering, decision trees and neural networks . DSSM is trained using a large corpus of text data and uses a technique called pairwise ranking loss to optimize the model parameters. Convolutional Latent Serantic Model (CLSM) is a type of neural network architecture used for text classification and information retrieval tasks. It is a variant of the convolutional neural network (CNN) that has been adapted for processing text data. Implicit feedback helps avoiding bias during data collection. Fastcat is an open-source library that enables fast and memory-efficient large-scale, high-dimensional data aggregation. It is designed for computationally efficient processing of large data and is particularly useful for processing multidimensional data such as images and videos.

Ques 4 : What are the data science related challenges one might encounter in this domain?

Some of the data science challenges that can be encountered are data variability - different types of data , quantity - data size, speed - processing speed where data can often be complex and difficult to process and data can contain missing values, which are some of the obstacles that can be encountered in ecommerce industry. The first step of any data science project is finding and collecting necessary data assets. As we all know , e – commerce companies have a lot of products ,which implies it has to store and process lots and lots of data . However, the availability of suitable data is still one of the most common challenges that organizations and data scientists face as many times data needs to be pulled from different servers and different systems which usually have different formats , and consolidating data from lots of disparate and semi-structured sources is a complex process and this directly impacts their ability to build robust ML models. Data prearation and cleaning often helps ,but then data scientists spend a lot of time processing and preparing data so that it’s consistent and structured enough to be analyzed. E-commerce platforms require real-time data processing to provide timely recommendations, detect fraud and manage inventory.Also most of the time, data consistency and data authenticity can be a challenge as data may not always be available in the correct format or may be stored in outdated systems.

Ques 5 : What do you find interesting about the nature of data science opportunities in this domain?

Increasing demand in the use of data is exponentially growing from past few years. For instance, the amount of data in 2010, which was being created every two days, and in 2021 it was being created just for every 40 minutes. The amount of data ,which is being created by applications, sensors ,electronic devices, virtual assistants , smart phones, self driving cars , AI devices etc… Similarly, in the field of e-commerce. One interesting opportunity in e-commerce data science is the vast amount of data generated from various sources such as customer events, website traffic, social media and advertising campaigns. E-commerce platforms can provide customers with highly personalized recommendations and experiences. Data science techniques can be used to

analyze customer behavior and preferences to provide personalized recommendations and campaigns that improve customer satisfaction and increase sales. Also, E-commerce platforms can provide customers with highly personalized recommendations and experiences. Data science techniques can be used to analyze customer behavior and preferences to provide personalized recommendations and campaigns that improve customer satisfaction and increase sales.

(i) Please discuss how sellers and buyers may need different data features in an e-commerce platform such as e-Bay.

Sellers need detailed product information to build a list and attract buyers, while buyers need accurate product information to make informed purchasing decisions. Therefore, sellers can request access to features such as product classification, product descriptions, images and specifications, while buyers can request information such as product reviews, ratings and comparisons. Similarly, sellers need customer information to personalize marketing and customer service efforts, while buyers need customer information to build trust and make informed purchasing decisions . Sellers, can request access to features like shipping counters, shipping labels, and delivery tracking, while buyers can request information like shipping costs, delivery times, and tracking updates. While, buyers need shipping and delivery information to estimate delivery times and choose the best shipping options. Therefore, sellers can request access to features such as customer profiles, order history and feedback ratings, while buyers can request information such as seller ratings, reviews and feedback.

(ii). Describe briefly the algorithmic steps involved in query correction as described in the lecture.

After the query is issued, a candidate generation process takes place in which a large number of candidates are generated to determine whether the query is valid or not. Then, two different models are generated- language model and error model and then we combine the two model and , we come up with the rankings for all candidates which is generated. Then we choose the top rank query and choose it as corrected query. The algorithm then ranks the candidate queries based on how likely they are to be the correct query. This is typically done by comparing the candidate queries to a dictionary or language model to calculate their probability or likelihood . Language models are trained on large amounts of text data, such as books, articles, or web pages, and are used in a variety of natural language processing tasks, such as speech recognition, machine translation, and text generation. The error model is used in combination with a language model to generate candidate corrected queries that are likely to be the correct versions of the user's query. In, many cases , in the candidate generation part itself, we decide whether if query correction is needed or not.