Name: Sahithi Dodda
Person Number: 50441731
UB Email: sahithid@buffalo.edu

Assignment 5

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

Lecture by ManojKumar Rangasamy Kannadasan – Role Of Data Science in eCommerce

Q1): Describe the market sector or sub-space covered in this lecture.

Ans: Role of Data Science in eCommerce is the market sector or sub-space covered in this lecture by ManojKumar Kannadasan. Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. The goal of data science is to find patterns, trends, and relationships in data to make informed decisions, solve problems, and increase business value. By using data-driven techniques, organizations can make better decisions, optimize their operations, and stay ahead of their competitors. Data science has multiple roles in ecommerce businesses. E-commerce companies use data science in various ways to improve their operations and enhance the customer experience. It can help analyze customer behavior and preferences to extract insights which can help ecommerce businesses make informed decisions about various aspects of their operations, such as product recommendations, inventory management, pricing strategies, and customer segmentation. By analyzing customer behavior, preferences, and purchase patterns, data scientists can create personalized shopping experiences for customers, improving sales, customer satisfaction and loyalty. By analyzing sales data, it can forecast demand and optimize inventory levels to reduce costs associated with overstocking or stock-outs. Data science in E-commerce also helps in Identification, refining of target audiences, Recruiting the right talent for the organization. In addition, data science can help determine the optimal pricing strategy by analyzing pricing data and competitor pricing to increase sales and revenue. Data science can segment customers based on their behavior, preferences, and demographics, which can assist in creating targeted marketing campaigns and promotions. Furthermore, it can identify patterns of fraudulent activity by analyzing transaction data, leading to minimizing fraud losses and protecting customer data. Overall, data science is critical in helping ecommerce businesses make informed decisions based on data, leading to better customer experiences, increased revenue, and reduced risks.

Q2): What data science related skills and technologies are commonly used in this sector?

Ans: Data science plays a critical role in the ecommerce sector, and professionals in this field require a diverse set of skills and technologies to effectively extract insights and optimize

business operations. Among the most common skills used are Natural Language Processing (NLP) which helps to enhance content to facilitate simpler and more effective product purchases for customers and Computer vision, which is used to extract information from images. To make predictions about customer behavior and optimize pricing, data scientists need experience with Python, R and popular frameworks like TensorFlow, Keras, and scikit-learn. They must also have experience in supervised and unsupervised learning techniques. Large datasets generated by ecommerce businesses require the use of big data technologies like Hadoop, Spark, and NoSQL databases for processing and analysis. Effective data visualization tools such as Tableau, Power BI, or Python libraries like Matplotlib or Seaborn are critical for communicating insights to stakeholders. Also, strong foundation in statistics, data visualization, and manipulation is also required to extract insights from large datasets. Knowledge in Cloud computing platforms such as AWS, Azure, and GCP which provide scalable and secure solutions for storing and processing data is also necessary, and recommendation systems can be used to target users with the best product recommendations. By creating a personalized user profile based on customers search history, clicks, and browsing time, data scientists can provide the best product recommendations to encourage users to buy more products, ultimately resulting in increased profits for the company. To rank questions, deep neural network models such as Deep Semantic Similarity Model (DSSM) and Convolutional Latent Semantic Model (CLSM) are utilized, but training and optimization can be time-consuming. However, data scientists can use the FASTcat model to improve training efficiency, as it is feedforward neural network that uses n-grams as input.

Q3): How are data and computing related methods used in typical workflows in this sector? Illustrate with an example.

Ans: In the ecommerce sector, data and computing related methods are used extensively to streamline processes, optimize marketing efforts, and gain insights into customer behavior. One common example of this is the use of data analytics to track and analyze customer data. For example, consider an ecommerce company that sells clothing and accessories online. By analyzing customer data, they may identify that a certain demographic (such as women aged 25-35) tends to purchase a certain type of product (such as dresses). Using this insight, the company can then tailor their marketing efforts to this demographic by promoting relevant products on their website and in their email marketing campaigns. Another example in this field where data and computing-related methods are utilized is the prevalence of misspelled words in user search queries. Misspellings can result in no search results being returned, frustrating users and hindering sales. To address this problem, spell check engines are commonly used. However, given the large amounts of data involved in ecommerce workflows, these spell check engines must be capable of processing gigabytes or even terabytes of data in real time. This requires balancing speed, accuracy, and memory footprint when selecting a model for spell checking. To create an error model for spell checking, probability theory is used, such as the Bayes Theorem. This involves calculating the probability of finding certain words given the presence of other words in a given search query. The resulting language model can then be used to determine the likelihood of misspelled words given the correct spelling, improving the overall accuracy of the spell check engine. In summary, data and computing methods play a critical role in addressing the problem

of misspellings in ecommerce workflows, requiring the use of sophisticated language models and probabilistic techniques to ensure accurate and efficient spell checking.

Q4): What are the data science related challenges one might encounter in this domain?

Ans: There are several data science related challenges that one might encounter in the ecommerce domain. For instance, In the ecommerce domain, recommendation systems face several challenges, including precision of the recommendations, data quality and cleanliness, personalization, scalability, and multilingual support. The precision of a recommender system is critical to its effectiveness and must be efficient, fast, and accurate in providing helpful recommendations to users. However, data quality can be messy and incomplete, leading to inaccurate recommendations. Personalization can also be challenging to provide, particularly as ecommerce platforms grow and user preferences change over time. Additionally, scalability can become an issue as the amount of data increases and multilingual support is crucial in ecommerce, particularly as companies operate in multiple countries with different languages. One Example for the challenges faced by recommendation systems in the ecommerce domain is Netflix's recommendation system. Netflix uses data science techniques and algorithms to recommend TV shows and movies to its users, based on their viewing history, ratings, and other factors such as genre preferences. However, Netflix faces challenges in ensuring the precision of its recommendations. For example, some users may watch a show or movie just because it's popular, which can skew their viewing history and result in inaccurate recommendations. Additionally, new releases can have a large impact on the system's accuracy, as they may not have enough data yet to accurately recommend them to users. To overcome these challenges, Netflix continuously improves its algorithms, incorporating new data points and machine learning models. Overall, the success of recommendation systems in ecommerce depends on their ability to overcome these challenges and provide accurate, helpful recommendations to users.

Q5): What do you find interesting about the nature of data science opportunities in this domain?

Ans: One interesting aspect of data science in the ecommerce sector is the abundance of data available for analysis, providing a chance for data scientists to discover insights that could potentially enhance a company's growth. Data science can be applied to optimize pricing strategies, identify opportunities for upselling and cross-selling, and detect fraudulent activities. Furthermore, data science can improve the user experience by providing personalized recommendations and enhancing the user interface. For instance, by analyzing a user's browsing history, purchase history, and other relevant data points, ecommerce companies can create personalized shopping experiences. They can also optimize the user interface, making it more intuitive and user-friendly. A real-life example of the potential of data science in ecommerce is Alibaba Group's success, one of the largest ecommerce companies globally. Alibaba has invested extensively in data science, using it to optimize product search, personalize recommendations, and improve supply chain management. By analyzing user data and product attributes, Alibaba's recommendation system has achieved an accuracy rate of over 90%, significantly enhancing the user experience. Additionally, Alibaba uses data science to optimize pricing strategies, allowing it to offer competitive prices and attract more customers. Another example is Amazon's use of

data science to optimize supply chain management, reducing delivery times and improving efficiency. By analyzing data from its warehouses and delivery routes, Amazon has reduced delivery times to as low as one hour for select products, providing a better experience for its customers. Ultimately, data science presents significant opportunities for ecommerce companies to improve their operations and enhance the user experience. Investing in data science can bring significant benefits to these companies.

(i) Please discuss how sellers and buyers may need different data features in an e-commerce platform such as e-Bay. (10 pts of the 80 C+R points in the rubric)

Ans: E-commerce platforms like eBay need to cater to the specific data requirements of both buyers and sellers to ensure a positive user experience. While buyers demand filters, placeholders, customized adverts, and a basket to hold all the products they wish to purchase, sellers require data on user interactions and various transactions that they interacted with on product pages. This helps sellers to distinguish between users who are interested in their items and those who are just browsing, and to target interested users with more products that are like the ones they interacted with, resulting in greater sales. Moreover, it is important to note that buyers require access to specific data points to aid them in their decision-making process. This includes features such as a page displaying the most popular products, search filters to eliminate unsuitable products, a shopping cart to hold selected items, and a section for personalized advertisements. Additionally, the provision of daily deals can entice buyers to purchase items they may have only been mildly interested in before. These data elements are essential for buyers to make informed decisions and feel satisfied with their purchases on an e-commerce platform like eBay. eBay's recommendation system uses data science to provide personalized recommendations to buyers based on their browsing and purchase history, while sellers can use data science to optimize their product listings, pricing strategies, and supply chain management. Ultimately, catering to the specific data requirements of both buyers and sellers is critical for e-commerce platforms to succeed and remain competitive in the market.

(ii) Describe briefly the algorithmic steps involved in query correction as described in the lecture. (10 pts of the 80 C+R points in the rubric)

Ans: In an e-commerce platform like eBay, there is a vast amount of user-generated content in the form of item level information when buyers and sellers are typing queries. To improve the user experience, spell correction is necessary for both buyers and sellers as they may make mistakes while typing their queries. However, performing spell correction for millions of queries and users is a challenging task that involves several steps. First, when a query is issued, a candidate generation process takes place to generate a lot of candidates to verify if it is a valid query or not. The mathematical formulation involves identifying the probability of the candidate that should be recommended given the query that the user typed. This is done using Bayes' theorem, which splits the probability into two factors: the probability of the candidate and the probability that the user intended to type the query. Then, two different models - the language model and the error model - are applied to the candidate features generated. The probability of the candidate i.e., the probability of the query occurring and the language model corrects it for

any spelling mistakes. The error model, on the other hand, is trained on a set of known queries and their correct versions to estimate the probability of a given query being an error based on the probability of observing the query given the correct query. Here, the probability that the user intended to type the query is estimated by the error model. Once the queries are processed through both models, they are given to the ranker function, where they are ranked based on precision. Finally, the top-ranked query is selected and presented as the corrected query. In summary, spell correction is an essential feature in an e-commerce platform, and it involves several steps, including candidate generation, language model, error model, and ranker function. By using Bayes' theorem, the probability of the corrected query is determined by considering the probability of the candidate and the probability that the user intended to type the query.

(iii) Also, answer the following multiple-choice questions: You can list the question number and the letter corresponding to the correct choice as Answer in your report, (2x5 = 10 pts of the 80 C+R points in the rubric)

Ans:

Q1) C
Q2) B
Q3) C
Q4) B
Q5) D