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

Assignment 8

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

Lecture by Matthew Nagowski and Eric Hanson – Data Science in the Financial Sector

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

Ans: This lecture will examine the market sector or sub-space of how data science and analytics are used in Commercial banking. The banking industry mobilizes public funds and turns them into productive investments. This enables people and corporations to invest in their future while also stimulating economic growth. The banking sector is important to the economy because it mobilizes funds, creates credit, provides payment services, manages risk, and promotes economic stability. Credit unions are non-profit financial cooperatives that provide their members with banking services. Their members, who are often individuals with a shared link, own and manage them. Data Science is applied in commercial banking to detect and prevent fraud. Sophisticated analytics and machine learning algorithms are used to detect irregularities in transaction data and highlight probable fraud. Data science is being utilized to create tailored marketing initiatives that will attract and keep customers. Customer preferences, spending habits, and behavior are identified using analytics techniques, allowing banks to offer customized goods and services.

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

Ans: Commercial banks recognize that data, sophisticated analytics, and artificial intelligence (AI) are potentially valuable tools for negotiating a challenging market marked by squeezed margins, high costs to serve, and disruptive competition in important areas like as small and medium companies. The capacity to process, analyze, and generate insights from enormous amounts of data is referred to as data analytics. Commercial banks employ data analytics to acquire knowledge about consumer habits and market circumstances by analyzing customer data, market trends, and risk measures. Commercial banks rely on analytics to obtain insights into client behavior, manage risk, and make sound business choices. Commercial banks must be able to evaluate and generate insights from vast amounts of data to remain competitive and fulfill the changing demands of their clients. Analytics allows banks to use data to produce value for both consumers and shareholders. SAS and Stata are used in commercial banking for data analysis to examine massive amounts of data and uncover emerging trends and changing patterns. Banks utilize these technologies to obtain insights into consumer behavior and market circumstances by analyzing customer data, market trends, and financial indicators. The commercial banking

industry landscape evolves daily owing to a multitude of reasons. Macroeconomic instability, geopolitical upheaval, inflation, conflict, rising interest rates, and supply chain disruption are some of these variables. SAS and Stata are used to create visualizations and analyses that give data on organization performance. Banks utilize these technologies to track key performance metrics, consumer behavior, and market trends.

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

Ans: During the loan approval procedure, data and computing-related approaches are utilized to gather, clean, analyze, model, and assess data. These techniques allow loan officers to make data-driven judgments and manage risks more efficiently, which is crucial for the bank's financial health. The loan officer simulates several situations and predicts the chance of default using modeling and simulation techniques such as Monte Carlo simulation. The loan officer, for example, may mimic the impact of fluctuations in interest rates or market circumstances on the applicant's capacity to repay the loan. The loan officer decides whether to accept or refuse the loan based on the data analysis and modeling findings. The loan officer may also utilize visualization tools like Tableau to produce representations that assist them explain the outcomes of the investigation to stakeholders. DTW works by warping, or extending and compressing, one sequence's time axis to match the other while reducing the disparity between the two. It is important in applications such as voice recognition, gesture recognition, and biological signal analysis since it allows for the comparison of patterns with different speeds, accelerations, and noise. DTW comes in multiple flavors, including global and local DTW, and may be tailored with different factors such as the kind of distance metric used and the window size for local DTW. It is a sophisticated tool for evaluating time series data, with applications in finance, healthcare, and voice recognition. DTW may be used to detect fraud trends and abnormalities in financial transactions. It may be used to compare transactional temporal sequences and detect those that differ considerably from usual patterns. It can, for example, be used to detect credit card fraud by comparing a card's transaction history with the cardholder's regular spending pattern. DTW may also be used to detect insider trading by comparing individual trading patterns to those of the market as a whole.

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

Ans: Commercial banking data comprises a significant quantity of sensitive customer data, such as personal information, government identification numbers, and credit card information. Data scientists must ensure that this information remains private and secure, and that it is not vulnerable to theft or hacking. As we have seen in the past, several financial sector websites and applications have been hacked, exposing critical information such as credit card details and dumping them on the dark web. This will have a significant impact on businesses since it will erode consumer trust and have an influence on the organization's operations. To fully realize the promise of data science, it is critical to ensure data quality, compliance with legal standards, and create models that are visible and straightforward. Commercial banking works with a vast volume of data, which necessitates the automation and integration of data science techniques. Data science models must be linked into existing business operations, and automated methods for

data processing, model training, and deployment must be created. While data science approaches can bring significant insights, they might be difficult for non-experts to grasp. It is critical to ensure that models are subject to interpretation, and transparent to stakeholders, especially when it comes to regulatory compliance.

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

Ans: Throughout the last several years, there has been an exponential increase in the usage of data in online transactions, mobile phone apps, including sales transactions, customer interactions, online banking, bill payments, statement production, and more.... For example, in 2010, data was generated every two days, however in 2021, data was generated just every 40 minutes. Data scientists may utilize this data to uncover insights and trends that can assist drive business decisions and improve performance. Data science may aid in this process by evaluating vast amounts of transactional data and discovering trends and abnormalities that may suggest fraudulent conduct. Banks can identify possibly fraudulent transactions for additional examination or rejection by utilizing machine learning algorithms to detect odd activities. By automating manual operations, minimizing mistakes and redundancies, and optimizing workflows, data science may help banks enhance their operational efficiency. This can assist banks in lowering expenses, improving overall performance, and improving the client experience. By designing and deploying data science-driven apps, data scientists may also assist banks in improving their fraud detection and prevention systems. Natural language processing algorithms, for example, may be used to evaluate text-based data sources like emails and social media in order to detect possible fraud threats and warn institutions to take action.

(i) According to the lecture, what are the types of technical and business questions that are considered to evaluate the validity of a model for a banking application? (10 pts of the 80 C+R points in the rubric)

Ans: In the commercial banking industry, model validation ensures that the mathematical and statistical models used by banks to make key decisions are accurate, dependable, and devoid of mistakes or biases. Model validation is an important aspect of the risk management process in banking since it aids in the identification and mitigation of possible hazards connected with the usage of models. The following are the questions that are asked during model validation: What are the model's limitations? What is the quality of the data? What kind of data is used to create the model? What do we expect in terms of error forecasting? What kind of testings were done during the development process? want to do some back testing? What is a sensitivity analysis? What are the major assumptions that were backed up by research and evidence? Scope of stability testing?

(ii) Describe how the clustering model meets the business purposes of M&T Bank, and what characteristics of the bank's portfolio were being captured. (10 pts of the 80 C+R points in the rubric)

Ans: The clustering approach consists of two steps, with some control at each stage. The first is the computation of distance, which contains two algorithm enhancements to lower computing costs. The entire time warp is constrained by the size of the window. DTW is a method for comparing time series data that varies in length and speed. It entails determining the distance between two time series by warping them in time to bring them closer together. The DTW clustering method is a two-step procedure with some control at each stage. The first step is to compute the distance between each time series in the data set and each other time series. Two algorithm modifications are applied to lower the computing expenses involved in this stage. The second stage is to optimize the window size, which is a parameter that controls the amount of warping that can occur between two time series. The RPN (relationship pricing net) count is used by the bank to determine how balances behave. The M&T bank divides balance behavior into 25 clusters, which are then divided into 5 general behaviors. A "hill" cluster indicates that funds have left the bank due to better market rates or possibilities elsewhere. If a cluster is reducing, it signifies that balances have left the bank, but their history may indicate the possibility of future high balances. If the cluster expands, it indicates that balances have risen independent of growth markets or prices and are deemed price insensitive. This is how the features of the M&T bank's portfolio are recorded.

(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) D
Q2) C
Q3) D
Q4) C
Q5) A