EAS 504 - Applications of Data Science : Industry Overview - Assignment 8

Name : Aboli Rawanhate

UB Number : 50374341

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

In this lecture by Matthew Nagowski and Eric Hanson, the market sector or sub-space covered was the Financial sector. The lecture was about the issues in the financial sector and the need of data science in that, how data science has solved some financial sector problems are explained with examples. Data is being maintained in this sector over the years, this is used for analytics to solve problems in various sectors of business solutions for financial sectors, such as Loss Forecasting and Credit Risk Management, Interest Rate & Liquidity, Risk Capital Planning, Business Unit Analytical Support, etc. M&T Bank also uses analytics and modeling to inform business choices and meet increased regulatory requirements.

- To keep track of transactions, credit ratings, and other financial metrics without experiencing any delays, risk management has been introduced.
- Advanced Machine Learning algorithms are used for Fraud Detection.
- Companies may acquire insight into customer behavior in real-time with the aid of data scientists and real-time analytics to make better strategic business decisions.
- Machine learning is used to extract business intelligence in order to acquire insight into consumers and their behavior.

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

Machine learning and data science has successfully solved many financial problems. Some of the examples of data science skills and technologies used in this sector is:

- Data analytics is primarily about communicating stories to finance leaders for making decisions.
- Tableau, Qlikview for data visualization and sophisticated reporting to business customers
- SAS, SQL, and Tableau are used to offer transaction-level information to attribute changes in deposit balances.
- Excel is utilized for rapid ad-hoc analysis and reporting
- To assure appropriate financing, Stata, a generalized linear model, and SSRS are used to track and anticipate the intra-monthly volatility of deposits.
- Geometrically lag To analyze the demand for mortgage loans in different economic environments, regression models, SQL, and SAS are employed
- SQL and SSRS are used for data reporting and management
- Python, SAS, Stata and R are used for econometric model estimation, analytical manipulation and model execution
- New deposit demand elasticity models are supported by SQL, SAS, and non-linear regression.
- For advanced reporting and data visualization, Qlikview is used
- For documentation, latex is used

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

There are various data and computing methods used in Typical workflow of the financial sector. We will see data science involved in treasury division workflow.
The Treasury Division is divided into many sections, each of which makes considerable use of machine learning and data science techniques.
These stages include stress testing, capital adequacy, ALCO, IRR, and liquidity risk; financial analysis and planning; product management; modeling functions; balance sheet strategies; and data, model support, and project management activities. Multiple modeling functions are employed across the bank, and there is a centralized modeling department that supports all of the different products customers of different companies for generating predictive models, building content analytics, and supporting risk management decision-making and research. Everything is done depending on client behavior. As a result, everything is unrelated to consumers' failure to meet their obligations.
The consumer model and group, as well as the commercial model and group, are focused on consumer and commercial borrowers' credit practices. Following that, they focus on behavior such as our deposit with propensity for our clients open and close deposit account fee income revenue streams to the bank, the desire for new loans to appear on the bank's balance sheet, and the liquidity visibility of deposits on bank accounts.

Q4. What are the data science related challenges one might encounter in this domain?
The use of models carries a risk since they might not accurately reflect reality. Because the financial industry is one of the most sensitive businesses, the risks are enormous, making it difficult to deliver outcomes that are error-free. Models that are correctly created and utilized are preferable, while models that continue to work over time are also desirable.

- Additional risk emerges as a consequence of model development constraints/limitations result in a poor model
- Bad design choices (e.g., statistical approach, segmentation) and inappropriate business assumptions results in inaccurate results
- Business assumptions that are incorrect lead to wrong decisions
- If the data is not sufficient or does not reflect the complete population, the assumptions and interpretation will be incorrect
- Poor data quality, such as a large number of missing or inaccurate data points will make models train on values which are not relevant and hence it will generate a lot of validation errors
- Model implementation errors
- Model can be used for a purpose for which it was not designed and hence assumptions don't match with the problem statement being used for producing incorrect results

Q5. What do you find interesting about the nature of data science opportunities in this domain?
Data Scientists have a variety of options with the financial industry being one of the most important businesses. Some of the things that I found interesting are follows:

Behavioral modeling is used to inform bank strategy

- Pricing strategy optimization by price elasticity and balance flow modeling
- To drive retention initiatives, identify early warning signals (silent attrition, balance migration to direct banks)

- Hadoop-based transactional analysis (internal ACH data, external ACH data such as MX).
- Clustering is utilized in the financial industry to drive business insights. Dynamic Temporal Warp organizes time patterns by shifting each one on top of the other and comparing the results.
- The similarity of things is described by the dendrogram of findings, and comparable objects are clustered together. During the procedure, the analyst chooses the number of clusters, while the algorithm determines the cluster division point.
- The business benefits from advanced analytics, machine learning and modeling includes: Customer Lifetime Value (align marketing, service price, customer experience, network, and capabilities). Prospecting or sales targeting strategies like next best product, network analysis, prioritizing customer outreach, RbROE, expansion of opportunities are optimized. Automation can help to improve process efficiency. Advanced data visualization and data modeling applications, support for joint ML projects with universities. Risk transfer; risk quantification

Q6. 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?

The technical and business questions that are considered to evaluate the validity of the model for a banking application:

Technical questions - Goal is to understand the model and know the risks

- It's possible that there's a flaw in the model itself (e.g. Poor data, inappropriate assumptions)
- Deficiencies in the model building process (e.g. weakness in the variable selection, model testing was not rigorous enough or the wrong tests were done)

Business questions -

- Is the model sound from a conceptual standpoint? Is the chosen approach likely to be effective in light of the model's business goal? What other approaches were taken into account?
- Is the model accurate in capturing the distinctive features of a bank's portfolio?
- Is it a good representation of how the business actually operates?
- Are major assumptions backed up by research and evidence?
- What kind of data did you have to work with to create the model? What is the data's quality?
- What are the model's limitations? What might we expect in terms of forecast errors?
- What kind of testing was done during the development process?
- Are there controls in place when the model is run to ensure that the correct input data is used and that the model is the approved version?
- Is everything listed above well-documented?

Q7. Describe how the clustering model meets the business purposes of M&T Bank, and what characteristics of the bank's portfolio were being captured

Clustering model which is used to find groups in data points and to develop a paradigm or typology. It is used in M&T to drive business insights. Deposit behavior responds to changes in incentives over time, which aids in the development of deposit behavior typologies. The Dynamic Time Warp (DTW) algorithm for grouping time series patterns is used to solve this problem. DTW algorithms shift Time by a specific

amount of observations or time period, depending on how you measure your data, to see if we can line them up to reduce the distance between them to determine if they are connected. This approach employed the hierarchical clustering method, which employs dendrograms to characterize behavior based on distance.
It calculates the distance of any 2 data points clustered together and then continues until we break off more branches from the street until we're left with just one cluster or one observation. There are methods that can be used to optimize the number of clusters. DTW finds the best match between two sequences by extending and compressing parts of each. Distance computation and clustering are two phases in the process. The distance matrix must be computed. Two algorithm modifications have been explored to reduce computing costs: 1. The window size, for starters, restricts the entire "time warp."2. We use a modified version of the fundamental DTW method to improve the temporal optimization of this window size. If and when accuracy is enhanced, it utilizes the Euclidean distance. DTW lower bound is the name for this version. The final dendrogram was created using a Hierarchical (dendrogram) technique with Ward's D statistic. Ward's D - seeks to uncover tight patterns of behavior inside a single cluster while minimizing overall within cluster variation. Over 25 clusters were compared, and clearer trends emerged when the average behavior was simplified. All of the patterns were eventually grouped into five categories: increasing, decreasing, hill, volatile, and seasonal.

Q8. MCQs:

Q1. D
Q2. C
Q3. D
Q4. C
Q5. A