Data Science in the Financial Sector
(Talk by : Matthew Nagowski and Eric Hanson)

Introduction

With the rise of capabilities in the field of Data science and modeling its use has popularized in many domains and fields. One such domain is the banking sector where Data science and modeling play a vital role in commercial banking for Treasury division. Using these tools we can go deeper to answer questions regarding risks, opportunities and make informed decisions about making better use of a bank's financial resources. Clustering – an algorithm / technique used in machine learning to group similar things together is used in the commercial banking or Treasury division to cluster commercial deposit relationships. The similar groups are clustered together and their behavior such as deposit relationships can be studied. The obtained information can then be used to strategies pricing, reduce customer churn and plan various schemes for the customers in the future. It can also be helpful for identification of deposit relationships that are at risk of leaving the bank or who have potential to grow. The report discusses the clustering model and how it is used to improve business of commercial banks (M&T Bank). The report also answers important questions such as skills related to this domain and opportunities of using data science in modeling for the treasury division of M&T bank.

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

The market sector of the sub-space covered in the report is the Treasury division of M&T bank which lies under the commercial banking sector. It manages deposits, investment, loan and using the bank's financial resources responsibly. The treasury professionals use tools and techniques to manage the opportunities, risks and understand the customer behavior using data science and modeling.

The market sector also uses **clustering techniques**; an algorithm / technique used in machine learning to group similar things together is used in the **commercial banking or Treasury division to cluster commercial deposit relationships**. This can prove very useful for deciding the strategies for pricing, customer churn and identifying the relationship between deposit and risk of leaving the bank or having potential growth. This lecture also discusses the inaccurate callibarations done by banks which resulted in the great depression in the year of 2008 and also talks about new reformed techniques and tools which banks and finance sector use currently in order to make wise and fruitful decisions for the banks.

The lecture also talks about developing behavioral topologies, Methodologies for Clustering Analysis & DTW and how it is used to find patterns in the customer behavior using the data and answer some deep questions which can help banks make wise business decisions. It also talks about stakeholders responsibility, including Treasury professionals at commercial banks, researchers in the field of data science and modeling.

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

The domain of commercial banking needs fundamental domain knowledge to understand the working of this sector and to understand its intersection with the society. But, using data science we can make useful inferences for making financial decisions. These skills include :

1. **Excel** : As most of the data is stored in excel format, knowledge of excel is handy while dealing with this type of data.
2. **Statistics / econometrics** : for understanding the distributions and mathematical model of the financial sector.
3. **Regression** : Multiple linear regression, logistic regression, panel data methods
4. Algorithmic thinking (e.g. Object oriented programming).
5. **Programming** skills in **Python, Java, Javascript, Processing**
6. **Statistical programming and machine learning using Python Pandas, R, SAS, etc.**
7. **Data structures** (**SQL**), discrete math.
8. **Data visualization** : making the data understandable for the stakeholder is a key element in representation of data.
9. **Communication skills** : Communication with stakeholders, teams and presentation skills also play vital roles.

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

The workflow of the data and computing is very similar to how it's set up in the usual data science cycle. The data is collected from the customer when he wishes to receive services from the bank. This data includes personal information, marital status, personal finances etc. An exploratory data analysis is done on this dataset to get information about the customers. In the lecture they have mentioned about the segregation of customers together. This can be done using clustering so that customers with similar behavior can be grouped together and a strategy can be implemented to either offer them services or avoid customer churns. The lecture briefly discussed Clustering Analysis and DTW (Dynamic Time Warp).

The Dynamic Time Warp stretches and compresses the two sequences to find an optimal match. Unlike the regular clustering analysis where it only takes into consideration the Euclidean distance between two points. It was also observed that the DTW gives a more efficient result which can be later used for analysis and building models based on these observations. It’s a **two step process where step 1 is the calculation of the distance and step 2 is clustering analysis.** This is done in order to have some level of control in each level/step.

In the first step, the window size is limited for getting a “time warp”. This algorithm also uses Euclidean distance if it gives a more accurate result. So, that an optimum result is the first priority. Here, they also deployed a variant of the basic DTW algorithm which is used to optimize the size further. This variant is known as DTW lower bound. **For step 2,** We deployed clustering. The clustering is done in the form of a dendrogram approach using Ward's D statistics which reduces the number of clusters in the graph and using this a final dendrogram is made.

For example : Risk management: It's very important to manage commercial bank activities including the risk of operation, market and credit. A model has been deployed to predict which customer can have a defaulter next month. This model can later be deployed to make better decisions while lending money to a customer.[3]

Risk management workflow

1. Data collection: Personal data is collected having a history of finances, credit score and business plans.
2. Data preparation is done by using EDA to get insights from the data.
3. Model development: Based on the insight a data model can be made which avoids customers who can be potential defaulters.
4. Model validation: It's an important part of the complete life cycle as it makes sure it is working accurately and is reliable.
5. Model deployment: After deployment the model helps banks to make informed decisions about lending money.

In the commercial banking sector, the role of data science and its related methodologies is increasingly critical. Through the application of these tools, banks are able to segment their customer base, enhance fraud detection, improve risk management, and innovate in the creation of new products and services.

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

The commercial banking and financial domain deals with various cross culture challenges. The following challenges are prominent in this domain :

Data Quality : The data quality has been a major concern in the field of data science. As the models are built upon the data, the quality of the data defines the quality of the model. For the majority of time data is isolated and is difficult to access, making the quality vary in consistency making it difficult to do analysis.

Model Interpretability : The model deals with a lot of parameters making its interpretability difficult to the stakeholder of a person having limited domain expertise. Hence, making it less buys from stakeholders.

Model Failure Risk : The models once created can make either Worst or Best decisions which directly lead to downfall of growth of the bank as well as the economy of a nation.

Staying Updated : The ever changing field needs staying updated with most recent financial changes in the economy as well as new machine learning models to enhance the analytical capabilities.

Working with regulations and obtaining deep insights : As this data is personal and cannot be directly used for analysis and modeling. The hidden data needs to be adjusted accordingly so that the model is accurate.

Additionally, it demands knowing the limitations of data science. Most functionalities cannot be solved just by using data science. But a correct use of data science can help us make excellent inference from the data and help us build models which in turn reduce the risk and facilitate opportunities of a commercial bank.

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

Commercial banking is very fascinating as it deals with decision making based on real data and has several implications. Following are the few data science opportunities that particularly interest me :

1. **Data science and its impact on financial decision-making** : By using exploratory data analysis we can get info about customer data and gain a deeper understanding of the possible risk and opportunity faced by the bank. The information inferred can be used to make wise financial decisions enabling useful use of banks financial resources. **Example :** It can be used for forecasting the flow of the cash, assessing the impact of interest rates and price financial products.
2. **Large and Complex Data Challenges :** The bank dataset including customer personal information and transactions can take large values and are complex making it difficult to work with. This inter-relations makes its complex of making data science models. This is an exciting opportunity to learn about complex relationships and learning how to manage large datasets and developing new innovative solutions.
3. **Excellence in visual and presentation skills :** Visuals and representations of complex datasets, its relations and results always excite me and make me think how can the details of the data be shown more effectively so that even a person with limited knowledge about the subject will also understand it. Therefore, efficient communication is essential for data scientists to make an impact in the real commercial banking division.
4. **Solving real life problems :** Managing risk, predicting the customer behavior, cash flow forecast, product price financing etc. are many challenges faced by data science working professionals. The use of data science will enable us to develop solutions to these challenges helping the banks to make better decisions. Example : data science is used to predict customer churn and identification of frauds.
5. **Variety in data :** Professionals working in this domain have access to a variety of data including but not limited to financial data,customer data, and market data. Data science can be used to analyze the data to extract valuable insights. Example : Identification of customer segmentation, behavior understanding of customer and market prediction.

In conclusion, Data science is currently revolutionizing the financial decisions in banks and people with these skills will have a significant impact on the future of the financial industry.

Additional Questions:

(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)

As discussed in the lecture, following are the technical and business questions that needs to be answered for model evaluation and to check its validity for a banking application :

" Technical questions:

- Whether the model is technically good. Along with its efficiency with respect to business methodology and use of alternative methodologies.
- The model approximation is important as well as its ability to capture unique characters of the bank profiles.
- The evidence, anticipation, assumptions supported by brief analysis along with empirical evidence.
- The quality of the data and its availability for making the model.
- Limitations and edge cases of the model?
- It's also important that it has been through various types of testing. Such as performance during development, backtesting and sensitivity and stability of the testing.
- What is the anticipated prediction error?
- Controls monitoring the right input data and whether the model is approved etc.
- Are all the questions well documented? " [1]

"Business questions:

- How accurate are the insights with respect to the business needs ?
- Is it efficiently improving the performance of the bank?
- Easy to use and understanding?
- Reliability of the model and its consistency.
- Whether its deployment and maintenance is cost-efficient?
- Is it meeting the stakeholders' demands and needs?" [1]

In the realm of commercial banking, it's important to consider several key factors beyond initial inquiries when deploying banking models. Firstly, regulatory compliance is crucial, as models must adhere to a wide array of regulations and be thoroughly calibrated for compliance before deployment. Secondly, model risk management is essential to ensure the reliability and accuracy of the model outputs, addressing the inherent risks associated with their performance. Finally, robust model governance is needed, which entails setting clear guidelines and procedures for monitoring and managing the model, including regular updates and performance evaluations. By integrating these elements, a commercial bank can effectively implement and maintain dependable models that align with regulatory standards and operational objectives.

(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)

The clustering models help the bank to understand the behavior of the customer with regards to the commercial deposits. The insights obtained can then be used to **make strategies which are customer centric and which can in turn reduce customer churn**. Using this a detailed analysis and report can be obtained of **potential deposit relationships** which are on the **verge of leaving**. A **customer retention strategy** can be employed in order to keep them as loyal customers. This model can further be used to identify the deposit relationships that have a growing potential using balances of these relationships.

The clustering model of the M&T Banks is used for customer segregation. It is responsible for capturing the size of the deposit relationship, the type of depositor and his relationship history with the bank. Along with that the models also take a note of the current interest rate in the market. The clustering technique is very effective for grouping similar customers together enabling a develop a targeted strategy to employ them.

Example : During analysis it may be found that a certain group is sensitive to high interest rate and might be a potential churn. This can be addressed by deploying personalized interest rates or might have constant interest rates. Later, the bank could develop a retention strategy for this cluster in order to keep them a loyal customer of the bank. This is a valuable tool for the bank for understanding the behavior of customers who deposited an amount, potential risks of leaving the services of the bank and developing strategies to increase the deposit balances. Here are some specific examples of how M&T Bank can use the clustering model to improve its business:

Overall, the clustering model is a powerful tool that can help M&T Bank to improve its business in a number of ways.

(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)

Q1 D
Q2 C
Q3 D
Q4 C
Q5 A

REFERENCES

[1] Lecture video - Data Science in the Financial Sector (Lecture 8: Matthew Nagowski and Eric Hanson)
[2] Lecture slides - Data Science in the Financial Sector
[3] Article - https://activewizards.com/blog/top-9-data-science-use-cases-in-banking/
[4] Article - https://stats.stackexchange.com/questions/131281/dynamic-time-warping-clustering