Assignment II: MIS 64038 Analytics in Practice

---------------------------Vdodda@kent.edu-------------------------

Case 1: Marketing Campaign

This case represents a classic scenario of missing one-minute detail in the whole data mining process which ultimately led to a disaster scenario for a major Canadian bank. The case involved a logistic response model being built by an external supplier (not us) for acquisition of new customers regarding a given bank product of a well-known Canadian bank. The model was built and worked very well when looking at validation results. This model was then implemented and actioned on within a future marketing campaign. During the development process, the tools that were used both generated the solution as well as the validation results. However, during the scoring process, the tool did not automatically generate the score. The user had to take the output equation results from the model development process and generate a scoring routine to score a given list of bank customers. In scoring, the user had to manually create the score by multiplying coefficients with variables. As part of this process, there was also a transformation of this equation to a logistic function. As part of this transformation, the user had to multiply the entire equation by -1. This fact of multiplying the equation by -1 was forgotten by the user when scoring the list of eligible customers. Guess what happened. Names with the highest scores represented the worst names with the opposite scenario happening for the lowest scores. The campaign went out by targeting names with the highest scores which ultimately resulted in horrific results.

When the supplier did the backend against a control random group of names promoted across all model deciles, they flipped the sign the right way to -1 and validated that the model worked quite well. Unfortunately, this did not appease the client's unhappiness as the bulk of their campaign names represented so-called targeted names within the top few deciles but who were in fact the worst names. From a net eligible universe of 500M names, the client ended up losing well in excess of $100M.

Example Answer:

There are at least two areas (phases) of the project that could be linked to the failure of the above project: the commination phase and the operationalization phase. As for the communication phase (phase 5), it seems that the third-party consulting firm failed to clearly communicate and/or emphasize enough on the importance of the steps needed to be taken to use the model's output. Generally, once an analytics model is handed to a team, the development team is responsible to commutate the details of how the model should be used and to train the end users who will be interacting with the model.

In addition, the model did not seem to have been properly operationalized (phase 6), as it required several manual operations which significantly increases the chance of errors. Automation should be used as much as possible to minimize the impact of human's errors. In addition, checks and balances and models' monitoring should be an integral part of the model's operationalization. This scenario might have been prevented if there were checks and balances as part of the implementation process. By checking score distributions as well as the model variable means within the targeted deciles during model development and the current list implementation, this error would have been caught. The user would have noted that significant changes in both score distribution as well as model variable means for the targeted deciles would have occurred between time of model development and the current list scoring run. They then would have investigated this further by checking their coding in further detail and would have caught the omission and corrected it by multiplying the equation by -1. They say that the devil is in the details, but in data mining the devil is in the data.

Case 2: Fraud Detection in Banking

A community bank partnered with an analytics solution provider to develop new fraud detection algorithm for ATM withdrawals. The bank provided historical data and the company trained a model that seemed to provide an acceptable performance when tested on the data. Once implemented, however, the bank faced a major tragedy: the algorithm was too slow in the production environment, and, as such, most ATM withdrawal requests were timed-out and customers were not able to withdraw from their accounts. Discuss which aspects of the project were ignored and which phase(s) of the analytics project, the problem can be associated to?

Answer:

The above project fails in following phases:

- **Operational Phase Failure.**

Operational Phase Failure:

The primary issue in the Banking fraud detection project appears to have been a failure during the project's operational phase. After being built and tested in a sandbox environment, the model did not work as anticipated when deployed in a real production environment. As a result, the algorithm processed ATM withdrawal requests excessively slowly, preventing many customers from accessing their accounts.

There might be several reasons why the model failed in the production environment. One cause is that during the project, differences between the operating and development environments were missed. For example, the solution provider may have generated a model in a different programming language from the one used in the live environment. Alternatively, other differences in hardware or software configurations might have affected the algorithm's performance.

Another option is that the model generation process was not transparent. In other words, the IT team in charge of putting the model into action may not have completely understood how it was built, particularly the data conversions and procedures used. This might have led to errors or omissions while replicating the model in the production environment, adding to the algorithm's poor performance.

Overall, this instance highlights the need for rigorous planning and execution during an analytics project's operational phase. Before models are deployed in a real-world context, they must be properly checked and validated, and any potential differences between the operational and development settings must be considered. Furthermore, maintaining transparency and communication throughout the project is vital to ensuring that all stakeholders understand the project's aims and expectations.

Case 3: Amazon Rekognition

Amazon Rekognition is a cloud-based software as a service (SaaS) computer vision platform that was launched in 2016. It has been sold and used by a number of United States government agencies, including U.S. Immigration and Customs Enforcement (ICE) and Orlando, Florida police, as well as private entities. Rekognition provides a number of computer vision capabilities, which can be divided into two categories: Algorithms that are pre-trained on data collected by Amazon or its partners, and algorithms that a user can train on a custom dataset. In January 2019, MIT researchers published a peer-reviewed study asserting that Rekognition had more difficulty in identifying dark-skinned females than competitors such as IBM and Microsoft. In the study, Rekognition misidentified darker-skinned women as men 31% of the time, but made no mistakes for light-skinned men. The problem, AI researchers and engineers say, is that the vast sets of images the systems have been trained on skew heavily toward white men. In June 2020, Amazon announced it was implementing a one-year moratorium on police use of Rekognition, in response to the George Floyd protest. In May 2021, Amazon announced that they are extending its global ban on police use of its facial recognition software until further notice. Discuss which aspects of the project were ignored and which phase(s) of the analytics project, the problem can be associated to?

Answer:

The above project fails in following phases:

1. **Data preparation phase.**
2. **Operational failure.**

Data preparation phase:

- The misidentification of dark-skinned females by Amazon Rekognition may be traced back to numerous areas of the analytics project that were neglected. One of the primary causes is a lack of diversity and discrimination in the training data sets. The system misinterpreted dark-skinned women because the data used to train the algorithm was heavily skewed toward white men.

- This problem is connected to the data collection and preparation phase of the analytics project. The selection of training data sets is a critical step in developing an effective machine learning model. In this case, it appears that the dataset used to train the algorithm was inadequately diversified and may not have accurately mirrored the population for which it was developed.
- When screening for facial recognition, MIT researchers revealed that dark-skinned women are mistaken for men 31% of the time, whereas light-skinned men are not affected. This is because the model was built mostly with data from white guys. This indicates that the data preparation operation was unsuccessful. To make an educated screening choice, both humans and robots require relevant data.

- Decision-making quality will worsen. Good data preparation is required to build accurate and efficient models that are both valid and trustworthy. The accuracy of any analytical model is directly proportional to the quality of the data into which it is input.

Operational failure:

- Even after the algorithm has been installed, the project's life cycle will go through some phases all the way to deployment. Following the deployment of the algorithm, the team should monitor and measure the model's performance on real-time data during this stage. • Because the distribution of the data we used may change or the link between the model and data may change, the model may fail when we apply new or unknown data on our algorithm.

- Furthermore, it appears that Amazon did not thoroughly test and validate its algorithm before exposing it to users. The issue with misidentification of dark-skinned women should have been discovered during the testing phase of the analytics project. This highlights the need for testing and validation in ensuring that the model works as intended and has no unanticipated implications.
- Finally, the issues with Amazon Rekognition may be linked back to many stages of the analytics project, such as data collection and preparation, ethical concerns about deploying the technology, and testing and validation. It is vital to solve these concerns to construct accurate, fair, and ethical machine learning models.

Case 4: IBM Watson in Healthcare

Some time back, MD Anderson Cancer Center, the largest cancer center in the US, announced that it is going to introduce IBM Watson's computing system into the healthcare system. With the help of Artificial Intelligence, this system was supposed to accelerate the decision-making process of physicians while treating cancer tumors. But IBM Watson turned out to be a failure, as it did not deliver what it promised. It failed to analyze huge volumes of patients' health data and publish studies to offer cancer treatment options. Here, are a few possible reasons why IBM Watson flopped in the healthcare industry, according to the experts. The AI technology that Watson uses is not a problem. The problem is that it is not given enough time to gather quality data and use personalized medicine. IBM launched Watson in a hurry as something that can handle something as complex as healthcare. They were quite aggressive in the marketing of their product, without realizing the importance of making it competent first. Watson was supposed to be launched as a software product, in which oncologists can simply enter their patient data and receive commendable treatment recommendations. This was how IBM advertised its Watson Health, but it failed to deliver this effect. IBM failed to work with the hospitals to ensure the proper functioning of Watson. Another reason for Watson's failure is that IBM used data from its own development partner, MSKCC, to train it. Since the system is trained through the hospital's own data, the results it gave after queries were biased towards the hospital's own cancer treatments. It did not include data from other hospitals and other smaller clinical facilities. While such a trained system can be helpful in treating simple and generic cancer cases, complex ones need a different approach to the approach. Smaller hospitals cannot even access the same methods of treatment as their bigger counterparts. Discuss which aspects of the project were ignored and which phase(s) of the analytics project, the problem can be associated to?

Answer:

The above project fails in following phases:

1. **Discovery Phase.**
2. **Data Preparation phase.**
3. **Model Building.**

Discovery Phase:

- IBM Watson's failure in the healthcare field was caused by several problems, including a lack of time to collect sufficient data and apply tailored therapy. IBM rushed Watson to the market and failed to collaborate with hospitals to ensure its correct operation. This suggests that the data gathering, and preparation phase received insufficient attention and importance.
- To design efficient machine learning models for healthcare, the company must first comprehend and establish the business challenge, then define the project's goal, determine the type of data needed for the project, and ensure that all resources are available to assure the project's success. In the instance of IBM Watson, however, the researchers solely concentrated on their sample data set and utilized the model to evaluate massive amounts of patient data and publish articles to propose cancer therapy choices.
- Furthermore, the program could not deal with the intricate healthcare systems. IBM neglected to collect data from several sources and failed to focus on the project's scope, which would have aided in better modeling. The team also failed to gather project pain points, which might have been utilized to improve the model.
- Finally, the paper emphasizes the need for adequate planning and preparation in the analytics project. To guarantee that machine learning models in the healthcare business are accurate, fair, and ethical, the organization should prioritize data gathering and preparation. Furthermore, the team should collaborate with hospitals to ensure that the models are tailored to their unique needs and pain spots. By solving these issues, machine learning models capable of giving individualized and effective cancer treatment alternatives can be constructed.

Data Preparation phase:

Certainly, the article highlights the significance of data preparation in analytics projects, which is the most critical, iterative, and time-consuming step. During this phase, the company creates an analytic sandbox in which the team works for the life of the project, collecting data from various sources.

In the instance of IBM Watson in the healthcare business, the team failed to collect the desired data, and the system was trained using data from MSKCC, the company's own development partner. Because the model was exclusively trained on data from the hospital, the findings were skewed toward the institution's own cancer therapies.

To fix this issue, the team could have gathered data from various healthcare systems where the model was employed (i.e., the client's healthcare systems) and considered difficult scenarios. IBM might have achieved the desired results by employing a variety of data sets, including their own data set, end-user system data sets, and difficult cases. These data sets included their own data set, end-user system data sets, and complex situations.

Furthermore, IBM Watson's failure in the healthcare field might be ascribed to a lack of different data during the model training phase. The team relied on data from its own development partner, MSKCC, and excluded information from other hospitals and smaller clinical centers. This implies that the dataset used to train the algorithm does not appropriately represent the population it was designed to serve.

Finally, the essay underlines the significance of good data preparation in the analytics project, as well as the requirement for diverse, representative data in the model training phase. By solving these issues, machine learning models that are accurate, fair, and ethical in the healthcare business may be produced.

Model Building:

For starters, the corporation relied on skewed data from its own healthcare system, which resulted in skewed outcomes for their own cancer treatments. This is a typical issue in machine learning when the training data is insufficiently varied to represent the underlying population of interest. The training data used to develop the model should be representative of the population under study and should be free of bias.

Second, analysts did not properly prepare datasets for testing, training, and production. This meant that the little data analyzed and evaluated failed to give answers for complicated cancer instances where other techniques may be required to forecast the remedy. This problem highlights the significance of effective data management, in which the data is cleaned, labeled, and partitioned into multiple groups in order to train, validate, and test the model.

Third, the model deployment step was insufficiently tested and validated to guarantee that the model functions as expected with no unforeseen effects. This is an important phase in the machine learning process since models behave differently in a production setting than in a controlled testing environment. Proper testing and validation guarantee that the model's predictions are accurate, fair, and ethical.

It is critical to overcome these issues to construct successful machine learning models for healthcare. Machine learning models should be developed on varied, representative data and tested and verified in production. Furthermore, data preparation should include adequate cleaning, labeling, and partitioning into separate sets for training, validation, and testing. Finally, it is critical to guarantee that machine learning models forecast accurately, fairly, and ethically.

Case 5: AI for University Admission

The researchers tried to develop a robot Todai, to crack the entrance test for the University of Tokyo. Its one of the tasks that only humans can do with required efficiency but researchers thought they could train machines for this purpose. Unfortunately, the results were opposite to their expectations as AI was not smart enough in understanding the questions. It would be better to introduce a broad spectrum of related information in the robotic system; so, it can answer the questions rightly. Respective members from the National Institute of Information gave their statement about Todai: "It is not good at answering a type of question that requires the ability to grasp the meaning in a broad spectrum". Discuss which aspects of the project were ignored and which phase(s) of the analytics project, the problem can be associated to?

Answer:

Below is the failure in this project:

- **Data preparation.**
- **Model planning.**

The development of a robot named Todai by scientists in order to pass the University of Tokyo admissions exam. While it is often assumed that only humans can pass the test with the required efficiency, the researchers concluded that with enough training and a correct algorithm, computers may also pass the exam.

However, according to the paper, the researchers' algorithm for training the robot finally failed catastrophically, yielding the exact opposite consequence of what was planned. The robot was unable to answer most of the exam questions, prompting the researchers to conclude that their technique was ineffective.

Data preparation:

Data preparation is a critical stage in the building of a successful machine learning model. It entails gathering, cleaning, and preparing data for use in training and testing the algorithm. The accuracy and efficiency of the model are heavily influenced by the quality and quantity of data utilized to train it.

The failure of the Todai robot can be linked to a lack of sufficient data preparation. The robot was created to pass the University of Tokyo admission exam, which needs a thorough mastery of a variety of subjects such as language, mathematics, and science.

One of the Todai robot's significant flaws was its inability to understand the meaning of a wide variety of words. This indicates that the data used to train the model did not cover a wide enough range of words and phrases. As a result, the algorithm misidentified and misunderstood a considerable percentage of terms in the exam.

This problem may have been averted if the scientists and researchers had used a larger range of data in their model training. The more diversified the data, the better the model will be able to learn to detect and interpret distinct words and phrases. Furthermore, integrating more data linked to language and comprehension would have increased the model's performance in comprehending and responding test questions.

Model planning:

It is critical to understand that parallelization refers to the act of breaking down large jobs into smaller, more manageable sub-tasks that may be completed concurrently. The computational complexity of the model may be lowered by using parallelization during the model planning phase, resulting in more efficient data processing and enhanced performance.

Todai's failure to grasp the meaning of a wide variety of words may have resulted from a lack of parallelization during the model development phase. The model may have been able to absorb and grasp a greater range of words and phrases by breaking down the work of language comprehension into smaller, more manageable sub-tasks.

Furthermore, the model's breadth and the unique obstacles it may confront must be considered. Todai's model was created in order to pass the University of Tokyo admission exam, which needs a comprehensive grasp of many courses as well as a high level of language comprehension. The researchers may have neglected crucial elements that led to the model's poor performance by neglecting to address these specific difficulties during the model development phase.

Case 6: Mars Orbiter

In 1999, NASA took a $125 million dollar hit due to the loss of a Mars orbiter. The loss was later attributed to a mix-up in the units of measurement used by Lockheed Martin's engineering team and NASA's internal team-Lockheed was using English units of measurement and NASA was using more conventional metric system measurements. According to an internal review panel at NASA's Jet Propulsion Laboratory, "The loss of the orbiter] was an end-to-end process problem... something went wrong in our system processes in checks and balances that we have that should have caught this and fixed it." Fixing this "end-to-end" process problem likely would have prevented this loss. NASA also blamed Congressional budget constraints for a portion of the error. So, additional funding would have also helped. Discuss which aspects of the project were ignored and which phase(s) of the analytics project, the problem can be associated to?

Answer:

Below is the failure in this project:

- **Discovery phase**
- **Communication phase**
- **Operational phase**

Discovery phase:

The Mars orbiter project's failure can be ascribed to a lack of appropriate project planning, which is the initial part of an analytics project. The team did not appropriately describe the project's goal, clarify the business challenge, or determine the type of data necessary. As a result, there was disagreement over the units of measurement to be utilized, which resulted in the orbiter's loss.

Data Preparation:

The mismatch in the units of measurement employed for the Mars orbiter project was not found during the data processing phase, as previously stated. Cleaning, converting, and assuring the consistency of the data to be utilized for the project are all part of proper data preparation.

Model Building:

Although it is not stated specifically in the case, it may be concluded that the project's failure can also be linked to a lack of effective model construction. The scientists failed to account for the changes in measurement units, resulting in inaccurate computations and the loss of the orbiter. Proper model construction includes selecting the best algorithm, guaranteeing its correctness, and assessing its usefulness.

Communication phase:

In the Mars orbiter mission. This emphasizes the significance of excellent communication throughout the analytics project, particularly when it comes to data-related issues such as measurement units.

Proper communication between the two teams might have aided in identifying the mismatch in measurement units early on, thereby preventing the orbiter's demise. Furthermore, during the communication phase, explicit norms and standards for the use of certain measuring units might have been developed, which would have helped both teams stay on the same page.

As a result, it is critical to have a solid communication plan in place during the analytics project, where all stakeholders are kept up to speed on progress and any inconsistencies or difficulties are handled as soon as possible. Effective communication can assist in avoiding possible disasters and assure the analytics project's success.

Conclusion:

The case studies underscore the importance of being thorough across all phases of an analytics project, from acquiring and preparing data to developing and deploying models. Neglecting any of these stages can lead to severe consequences, such as biased or failed results, as well as financial losses. It is critical to rigorously test and validate analytics solutions on diverse data sets, while ensuring seamless integration with existing workflows and processes. Additionally, managing expectations and setting achievable goals during the project's discovery and definition phase, as well as creating contingency plans to address potential issues, is crucial. Ultimately, the key to success lies in taking a comprehensive and holistic approach that considers all aspects of the project from start to finish.

References:

1. https://www.wired.com/story/the-failure-of-ibm-watson-highlights-ais-limits/
2. https://www.pcworld.com/article/241162/todai_robot_flunks_university_entrance_exam.html
3. https://www.jpl.nasa.gov/news/news.php?feature=7208
4. https://www.cio.com/article/3279464/the-importance-of-data-governance-in-the-age-of-big-data.html
5. https://www.nytimes.com/1999/09/30/us/mars-climate-orbiter-thrusters-firing-cut-mission-short.html
6. https://www.independent.co.uk/life-style/gadgets-and-tech/news/artificial-intelligence-robot-todai-entrance-exam-university-of-tokyo-ai-national-institute-of-a7597506.html
7. https://www.forbes.com/sites/robertszczerba/2017/04/17/why-ai-failed-to-pass-the-university-of-tokyo-entrance-exam/?sh=57a85e0a7de9