Failed Analytics Project Real-World Case

Introduction:

In the early days of human evolution, the exchange of goods and services happened by way of a barter system. As human development began, this eventually moved from exchanging goods and services for the same to exchanging goods and services for money. Moreover, the term "business" emerged, which referred to exchanging goods and services to generate profit.

As humans evolved and technology evolved, businesses' conduct also became. In the early days of business, they involved only the purchase of goods and services directly from one person to another or one organization to another. As technological advancements progressed, goods were purchased from one place to another worldwide.

The advancements in conducting business paved the path for a huge market for any company or organization to create a customer base and generate profits.

In today's scenario, every company or organization aims to increase its market share or customers. By catering to the need of the customer in the earliest possible manner, every company is trying to attract new customers and, at the same time, retain their customers.

The aim to increase market share for every organization required them to gather data and perform various forms of analytics to understand past, present, and future customer behavior.

The concept of business analytics is becoming an important way to study customer needs. Business analytics has different areas of study, namely the study of historical information to assess customer behavior, called "Descriptive analytics." Predictive analytics involves studying historical data and statistical information to predict the future. Another area is Prescriptive analytics, which is the advancement of predictive analytics that uses advanced algorithms to suggest the best course of action for the future.

Organizations must plan any business analytics project to achieve the best possible outcome. A business analytics project requires resources, time, and money. Improperly designed analytics projects can lead to impacting the business immensely. Various steps are outlined to implement any business analytics project effectively. In this discussion, we will look at a company's failed

Abhinav Thupili- athupili@kent.edu (811234108)

Failed Analytics Project Real-World Case

business analytics project and how this can be avoided with steps outlined in the business analytics lifecycle.

Lifecycle of Analytics Project:

Earlier, we stated that to implement an analytics project, they are various steps outlined. Below listed are the steps involved in this lifecycle process:

1. Discovery Phase
2. Data Preparation Phase.
3. Model Planning Phase.
4. Model Building Phase.
5. Communication Phase.
6. Operation Phase.

- **Discovery Phase:**

  The first and foremost aspect of any business analytics project is understanding what they are trying to solve. Figuring out the right problem paves the way to answering the right questions, which helps in a successful analytics project. The discovery phase helps the analytics team to answer this question.

The steps involved in the Discovery phase:

- The organization needs to understand and formulate the business problem.
- They need to lay down all the aspects creating the business problem.
- They need to define and formalize the scope of the problem once all the elements are laid down.
- Within the organization, they are many sources of data available. The team responsible needs to figure out which data is relevant to answer the current business problem.
- The resources required for business analytics project increase exponentially concerning organization and problems to be solved. So, the organization needs to determine the

resources, like people, infrastructure, software, and other aspects required for the project.

- **Data Preparation Phase:**

Once the organization decides on the problem to be solved, we move on to the next step. *Data preparation* is a vital phase that involves collecting, processing, and cleansing the accumulated data.

The steps involved in the Data preparation phase:

- The phase is considered the most iterative and time intensive as huge amounts of data must be collected depending upon the organization and its problem.
- As the organization handles multiple projects in due course of business, the analytics project must be treated separately from the rest of the company's projects. They need to create an analytic sandbox that helps the team work on the project independently without impacting other operations.
- The important task in this phase is to collect all the data from different sources available, which can later be refined as per requirement. No data source should be left out as we do not know which data is useful for the project.
- Once the data is collected, cleaning it and applying necessary transformations are imperative to process for further phases.

- **Model Planning Phase:**

The third phase of the analytic project is the model planning phase can also be called the additional data preparation phase after the initial data preparation phase. At this stage, it is important to establish a relationship between the variables from the data collected. Exploring the relationship between the variables helps to understand the data better. In this phase, the data scientist involved in the project will offer valuable opinions on the data that can help to interpret the data to help solve the problem better. We should also consider the views of all the people working in the analytic sandbox of the project, which could help give different perspectives on the variables of the data.

Failed Analytics Project Real-World Case

Further, in this phase, the analyst determines the methods, techniques, and workflow required for the next model-building phase. The model planning phase is also the last step of preparations before implementing models and testing.

- **Model Building Phase:**

  Once the data is cleaned and the analytics team has decided on the required techniques, we move on to the model-building phase. The cleaned data is further processed and divided into training, testing, and production stages for implementation. The training models are used to understand the data with the techniques fixed by the analyst. This training model will help the algorithm to understand the model. Once the algorithm is trained, it is tested on the test model to check the quality and performance of the data. The process of training and trying the models is essential as this establishes the success of the business objectives. By training and testing the models and evaluating the performance, the team can tune the techniques and performance metrics determined earlier.

- **Communication Phase:**

  In the earlier phase, the results and hypothesis derived from studying the data would have been fuzzy and vague. As in the building phase, the analyst tries to get the best results by refining the techniques and metrics. The team must present clear results in quantifiable terms in the communication phase. The project's major stakeholders are interested in the business value of the project. In summary, the communication phase should be able to do the following:

- The team should be able to identify the key findings.
- Quantify the business value from deriving the findings.
- Develop a narrative to summarize and convey the findings to the stakeholders.

Abhinav Thupili- athupili@kent.edu (811234108)

Failed Analytics Project Real-World Case

- **Operation Phase:**

Once the team finalizes that the results derived are performing well for the given problem statement, it is the final stage of the analytics project. The last step is the operations phase, where the team moves the project from the sandbox to the live environment. At this stage, the team deploys models with live and unseen data to see the model's performance. The data is studied in real-time or batch to understand if the model created can produce similar quantifiable results on the unseen data. The 2017 SAS survey shows that 83% have invested in data-driven projects, but only 33% yielded satisfactory results. Therefore, the operations phase of the analytics project must deliver good results. If the model's performance is unsatisfactory on the live data, the team must return to the earlier phase to modify the data, called model recalibration.

Problem Statement:

The core idea of taking up a business analytics project is to enable businesses to make important decisions. Business analytics help the organization make decisions externally, i.e., to increase market share and build better products for customer satisfaction but also help the organization create projects internally that contribute to the company's overall growth. Since the analytics projects can also help internally, Amazon.com Inc's developed an AI recruiting tool in 2018 to help find the best talent to help the company's growth.

- **Problem:**

  As per a Reuters article in 2018, Amazon aimed to recruit the best talent available in the market with machine-learning algorithms. Amazon is one of the leading e-commerce with a market share of 37.8% just in the United States of America. With such a huge market share, it was evident that the data with the company is in huge amounts, which would help them understand and classify any business problem.

Abhinav Thupili- athupili@kent.edu (811234108)

Failed Analytics Project Real-World Case

- **Results of the Problem:**

As per the article in Reuters in 2018, Amazon created the AI-power recruitment tool to automate job applications by screening applicants' resumes. The data used to train the model for this analytics project was from the resumes submitted to the company over the ten years. The resulting model was a failure as it was a biased model against women.

Analysis of the failure of the Analytics Project:

From the above image, the following analytics can be interpreted as to why Amazon's AI tool failed to succeed:

- The AI tool was fed with data that existed with the company for easier interpretation, but the company should have considered the consequences of the data they used.
- From the above image, it was clear that males dominated most of the industry, holding most of the roles. The AI trained by Amazon followed the same pattern to create a model.
- The data prepared for the model needed to be implemented correctly, which was the problem's root cause.
- Since the Second phase of the analytics project was prepared with incorrect data, the model outcome resulted in considering "Male" as the best outcome for any data fed to it.

Understanding Data Preparation Phase for Problem Statement:

Every analytics project is a huge undertaking prone to failure without proper planning and implementation. The best example is the Amazon AI tool which failed to deliver the results. These mistakes can be avoided by carefully implementing the steps in the lifecycle of an analytics project.

For the current scenario, we need to understand the steps involved in the Data preparation phase. It will help us understand how the mistakes made by Amazon can be avoided. The following can be considered as various stages involved in the Data preparation phase:

1. Collection of Data.
2. Data Cleaning.

Abhinav Thupili- athupili@kent.edu (811234108)

Failed Analytics Project Real-World Case

3. Feature Engineering.

- **Collection of Data:**

Data plays an important role in any analytics model. Data forms a basis for training, validating, and testing the model. Therefore, collecting enough correct data from various sources is very important. Data can only be accurate and complete sometimes. Multiple problems are involved in data collection are:

- Manual data entry is a major form of data collection in most of the industry. However, this is also prone to errors due to human errors.
- Data collected is from various sources from within and outside the organization for data analysis. Therefore, there is no uniform standard for the content and formats of data collected.
- As data is collected from various sources within the organization, there is a high probability of parallel entry of the same data in different forms.
- Data is not always exact or stored in a complete form from every source. Therefore, sometimes data is approximated concerning other values within the data itself. It can lead to misinterpretations in the analysis.
- When data is about measurements, the measurement data is severely prone to errors due to instrument discrepancies or systematic errors.

- **Data cleaning:**

Data cannot be used directly in raw format when collected from different sources. For data analysis, the data needs to be processed and cleaned so that it is suitable to be used and draw quantitative and qualitative conclusions. This process can be called Data cleaning. The main problems that can be faced while cleaning data are:

- Missing values: The missing values can be caused due to equipment malfunction, not considered due to no domain knowledge, not regarded as important at the time of entry, or deleted to keep consistent with remaining data. Missing can be classified as Missing at random (MAR) or not at unexpected (MNAR). If the data is randomly missing, we can conclude and ignore such missing data. However, when

data are not missing at random, it could be biased while completing as such missing values could have potentially high predictive power.

- Outliers: Outliers are data points with extreme values that do not match the range of the data points within the dataset. In terms of data cleaning, outliers can disrupt the conclusions of the data. Understanding and classifying outliers as good and bad is important because some can still affect predicting data analysis conclusions.
- **Feature Engineering:**

Features of a dataset are the attributes of the data that define the data. These features play an important role in a machine algorithm, as they help to infer better conclusions. When a new attribute is added to the dataset or when two or more characteristics are combined under one or two or more attributes are combined to create a new attribute, such process is termed Feature engineering. It is important that feature engineering has a significant predicting power when implemented properly, but it also can cause bad conclusions when executed improperly.

Assumptions for the Problem Statement:

The following assumptions can be drawn from Amazon's failure to implement the Analytics project:

- The main cause of amazon's project failure could be improper planning during the data preparation phase.
- As stated earlier, data collection is very important in an analytics project—the algorithm that Amazon's model used was solely trained on the data of resumes by applicants. It invariably created a basis for the model as the resumes of applicants consisted same and repetitive attributes, which it learned over training and validating.
- Regarding data cleaning, data sourced from applicants' resumes needed to be processed thoroughly. The algorithm failed to assign significance to the applicant's skills or any other attributes as it needed to be properly established.
- Feature engineering of the data was also one of the failed aspects of the project. The team created roughly 500 models based on the data to understand different roles and

locations. As attributes needed to be correctly labeled and modified for the algorithm to understand, the algorithm completely ignored the most significant aspects of the applicants' resumes.

Mitigation of Failures in Amazon's Analytics Project:

There are various ways in which the Amazon analytics project could have become a better prediction model. The below steps can be a few measures that could have been followed to mitigate the failure:

- The model was a binary classification problem of classifying the resumes with yes or no. The model failed as the data provided was from a historical basis, where a wide range of companies, including Amazon, employed "Males" for most of the roles. It occurred as the data was not updated to reflect the current day scenario. Therefore, it is important to sample the data provided to the model in a way it does not create any form of bias.
- The data used by the company needed to be broadened for the model to make an accurate prediction. The team tried to select candidates solely based on resumes, which is not a practical application in real-world scenarios. Therefore, the team should have considered multiple data points for classifying eligible candidates. Sourcing more relevant and quality information on candidates is one way of collecting more data.
- Regarding data cleaning, the team needed to understand that the resumes consisted of repetitive terms that could affect the model's predictive power. The team needed more people with domain knowledge that could have helped identify the data's different aspects.
- Another way to maintain the data quality is to understand the accuracy of the data recorded, the consistency with which the data is used, and the uniqueness of each data point.
- The team also needed to create different attributes that could have helped the model better to identify eligible candidates. The team could have used other data transformations like normalizations, log transformations, or dummy variables to create a better model for the prediction.

Abhinav Thupili- athupili@kent.edu (811234108)

Failed Analytics Project Real-World Case

Conclusion:

We can infer from the above that a company as huge as Amazon can also commit a mistake when implementing a business analytics project. It shows how a business Analytics project is a huge undertaking besides the implementation. Many organizations, like Amazon, need help implementing the analytics project as they need to understand the right that needs to be used at different stages in the lifecycle of an analytics project. These steps that are laid out in this discussion help us understand how a successful analytics project can be planned and executed. The idea of automating candidate selection using AI was an innovative idea by Amazon. However, they needed to understand that resumes may be one of the criteria for choosing a candidate but only one for selecting a candidate. It shows how careful planning and understanding the analytics phases will always yield better for any organization. In conclusion, every organization needs a lot of resources, time, and planning to implement a project successfully. Especially when it comes to analytics projects, the main point to remember is to employ the right resource that helps the idea to be implemented correctly.

References:

1. Amazon scraps secret AI recruiting tool that showed bias against women | Reuters.
2. Comparing Descriptive, Predictive, Prescriptive, and Diagnostic Analytics - insightsoftware.
3. Top Data Preparation Challenges and How to Overcome Them (techtarget.com).
4. 7 Fundamental Steps to Complete a Data Analytics Project (dataiku.com) .
5. Learn How To Overcome Common Mistakes And Errors In Data Analysis (limeproxies.netlify.app).

Abhinav Thupili- athupili@kent.edu (811234108)