Assignment 01

Failed Analytics project of Google- Google Flu Trend

E.R.M.S.C Ekanayake

Introduction to Business Analytics

To find more information and trends in the current business data the process of gathering, processing, and applying statistical models to the current data is simply called business analytics. Most of the time Analytics is done to fulfill the requirements of stakeholders. Therefore, the primary purpose of Business Analytics should be an outcome that can present as a solution for the stakeholder's problems. Business Analytics can Reveal previously unknown information or problems which can be useful to overcome some business challenges, increase efficiency, revenue, reputation, productivity, gain competitive advantage, etc.

When considering the content of Business Analytics, four major types of analytics can be identified as follows.

- Descriptive statistics – identifying patterns and trends by analyzing past data.
- Predictive analytics – based on existing data, making predictions and forecasts.
- Prescriptive analytics – using data for decision-making and making effective business operations.
- Diagnostic analytics – identifying the causes of business problems.

Why Business Analytics is important?

The modern business world is driven by data-based decisions. So, having a better understanding of the data, makes it easy for organizations' operations. By properly following up the business analytics step by step, many advantages can be achieved.

- Improving decision-making - Business analytics provide insight into plenty of information that was hidden before. It makes easy the decision-making process and leads to effective decision-making.

- Finding opportunities - with the help of new data discovered by analytics, businesses can identify new chances for growth and innovation. For instance, finding new target customer trends helps the organization reach them with the customer's expected product.
- Improving competitiveness – by identifying market trends, changes in customers' behavior, market changes, and so on, organizations can adapt them by changing their business approach. So, it provides them a competitive advantage and makes them strong to survive in the rival business world.
- Increasing efficiency – through analytics, organizations can identify their weaknesses, hidden problems, and solutions to overcome those. After streamlining the business process, they can address those business inherited problems.
- Mitigating risk – With the help of business analytics organizations can identify root causes for their business problems, can foresee upcoming issues, and hide information about their customers, market, and so on. It makes the organization enable to take the necessary steps before more issues arise.

Not only the above-mentioned advantages. There are numerous others.

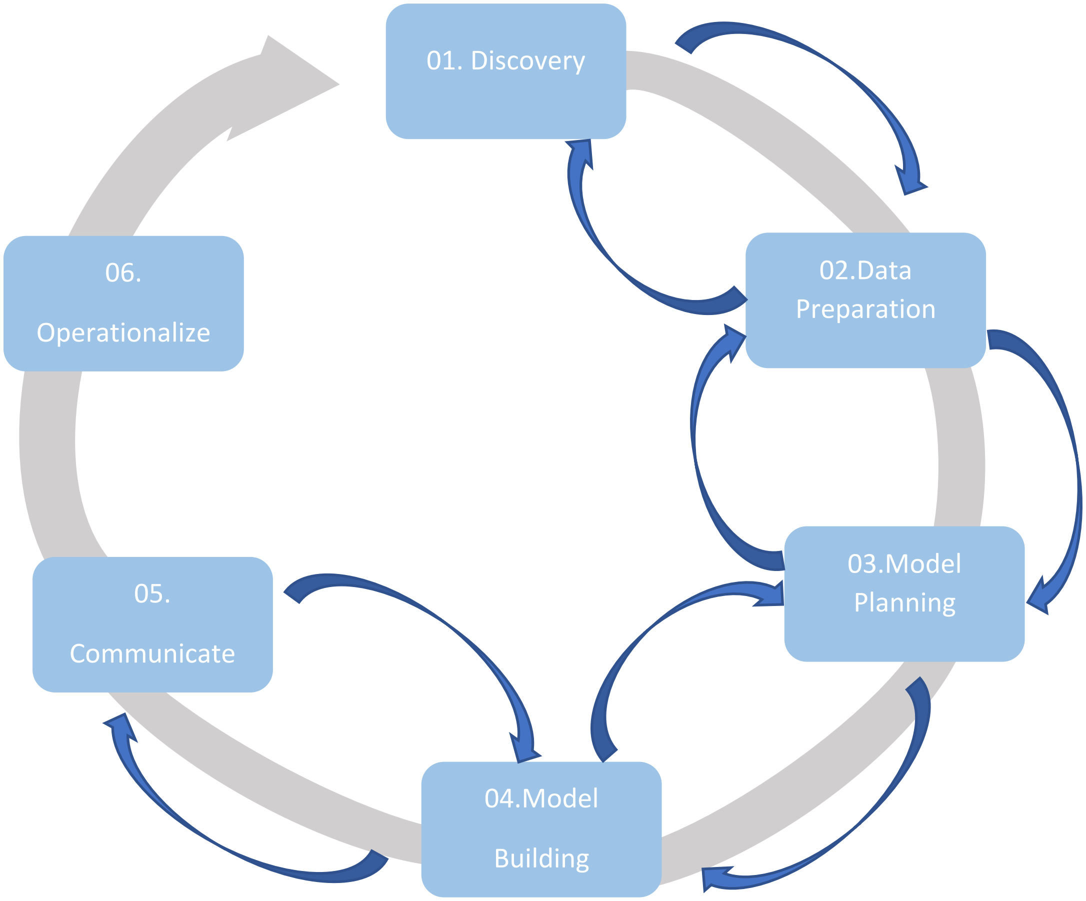
Phases of Business Analytics project

First, Business analytics projects streamline the process under 6 major steps. It creates a structured method to handle data which makes confirm at the end of the project, they achieve the project goals.

This structure provides guidance and support to go through the entire process. If someone messed up with the analytics project, they could clearly identify in which phase they are stuck and what further steps can be taken to overcome the issue. The analytics project life cycle can be identified as follow.

As indicated in the below figure, though it is a straightforward cycle, it can be followed up both forward and backward. For instance, while a data analyst working on the model-building phase, he finds out some drawbacks of the model. So, before completing the model-building process he can go back to the model-planning process and do the necessary changes as required.

Data Analytics Lifecycle



01. Discovery Phase

This phase is important because this is the step that makes the project aligns with the goal of the business and stakeholders' requirements. This provides the roadmap for the entire process. Therefore, getting the help of a team is required including some parties such as Business Analysts, Data Engineers, and Business stakeholders is crucial.

There are some key steps that should be taken during the Discovery phase.

Understand the business domain:

Determining how much business knowledge is needed to complete the project is important. For this can get the help of domain experts who are familiar with the business process. Probably the business analyst or data engineer may be not that much familiar with the business, therefore getting the help of experts is better to understand the business domain. Trying to understand the business capacity is the primary purpose of this step.

Identify key stakeholders:

There may be numerous stakeholders for a specific business. Out of those stakeholders identifying the relevant stakeholders should be done under this step. The best way to do this is to interview or discuss with relevant stakeholders and clearly identify what are their problems and what each stakeholder may expect from the project, and what are their desired outcomes.

Understand the problem:

In this step, it is significant to state the problem. Under this should be addressed some questions such as: What is the current situation of the business, what are the pain points, why solving this is important, and to whom it is important? Here getting the ideas of stakeholders is crucial because they are the people who are affected by the above-mentioned problems. According to the problem, they must determine what needs to be achieved, and what are the success or failure conditions.

Develop hypotheses:

After clearly defining the problem, developing hypotheses can be done by getting the help of domain experts as well as stakeholders. This involves making assumptions about the connection between variables and testing them by using further analysis.

Identify Data sources:

The analysis is driven by data. So, data is the base for the entire project. Therefore, identifying key data sources is significant. In a business domain, there may be different types of data coming from different departments and sections. Out of those, there may be so much sensitive data required for the project. So, deciding how to get that type of data should be decided under this step. Not only from internal sources it may require some data from external sources as well. Finally, they must have a clear scope of the type of data required for the project completion.

Identify resources:

It is important to identify the required resources for the project at the beginning. Assessing available resources technology, tools, data, people, and time provides a clear view of identifying the other required resources.

Learn from the past:

If there are any other similar data analysis projects done in the past, they also must be considered under this phase. If there were any failed projects must be considered reasons for the failure of them.

02. Data Preparation phase

In this phase it is required to gather data, clean, format, preprocess, and so on. Under this phase, they make an analytical sandbox. Here it takes a copy of the cleaned data and loads it into a distinct database. So, analysts can work through this data without making an impact on the production data.

Data collection:

Out of the above-identified data sources, now should gather data. It may be from different databases, spreadsheets, text files, web scraping, and so on.

Data cleaning:

The raw data gathered may be in different forms and structures. So, the raw data may be noisy, incomplete, and inconsistent and need to be prepared and remove different types of errors such as duplicating, missing values, outliers, human errors, and so on.

Data transformation:

The raw data should be transformed into another form, making it easy to analyze by machine learning algorithms. By following up the different methods such as normalizing, feature scaling, and discretization data transformation can be done.

Data integration:

The data got from different sources should be combined and integrated to create one dataset that can be easily used for analysis.

Data sampling:

In some situations, there may be an enormous amount of data which makes hard the analysis. So, it is needed to make some representative samples for analysis.

03. Model planning phase

The third phase of the analytical cycle is model planning which finds an appropriate model for the project. Specifying the project goals, choosing suitable procedures and algorithms, and determining the plan for model improvement and evaluation are done under this phase.

Assess data:

Exploring the data to find the connection between variables. When exploring data here, there may be so many other issues such as data quality problems, data structure errors, missing values, and so on.

Select modeling techniques:

This can be included supervised machine learning, unsupervised machine learning, classification, regression analysis, clustering, or time series analysis. Here it should be considered the availability of data before selecting the modeling techniques.

Identify the modeling approach:

After modeling techniques are selected, next should decide on the modeling approach. Simply it means choosing algorithms, establishing parameters, and determining the suitable training and testing methods.

Create a model development plan:

Here, develops a plan to evaluate the efficiency of the build model. This may include cross-validation techniques, hold-out datasets, and other techniques.

Getting required resources:

It may require some resources such as software, hardware, and people. As identified in the first phase, now gathering those resources is done under this step. As well as data infrastructure such as databases, data warehouses should be set up.

04. Model Building phase

Under this phase, the team creates datasets for testing, training, and production intentions. As identified under the model planning phase it is required to build and executes the model.

Model building:

After the model is selected, it starts to build the model and then trains on the cleaned data.

Model testing:

Then it is needed to measure the validation of the built model. By using suitable validation techniques such as cross-validation the team can evaluate the efficiency and accuracy of the model.

Model implementation:

After getting confirmed of the model's accuracy, the team can deploy the model in the production environment.

Model monitoring and maintenance:

After the model implementation, it is needed to monitor the performance of the model continuously and make sure the model generates the expected results and achieves goals. If not, model updates or improvements should be done.

05. Communicating phase

In the fifth phase, it is required to communicate all the key findings of the project to the stakeholders. Here it is important to present the outcome in an understandable way to the stakeholders. Clearly, it should mention all the success and failure criteria of the project.

Define communication plan:

First, should develop a communication plan which includes intentions, audience, the content of messages, and methods of communicating.

Create reports:

The results of the project can be presented in summarize the report and provide recommendations. As well as through data visualization, the team can convey some complex information by using graphs, charts, and dashboards.

Present the results:

After finalizing everything the team can present the findings to the stakeholder as tailored messages.

06. Operationalization phase

Under this final phase, the team distributes the outcome, code, reports, and technical documents to a big audience. Based on these findings the next step is starting the pilot project.

In this phase, the sandbox data need to be moved into the live environment. Not only that but continuous monitoring is also required to make sure the weather results match the business goals.

If there are any mismatches, the team can go back and do the necessary changes before it implements in real.

All the actions needed to be taken throughout the data analytics lifecycle can be brief as follow:

Discovery	**Data Preparation**	**Model Planning**	**Model Building**	**Communicate Results**	**Operationalize**
Learn Business domain	Data collection	Assess data	Model building	Communication plan	Prepare final documents
Identify key stakeholders	Data cleaning	Select modeling techniques	Model testing	Create reports	Turn Sandbox to live
Define the problem	Data transformation	Identify the modeling approach	Model implementation	Present the results	Implement pilot program
Develop hypotheses	Data integration	Create a model development plan	Model monitoring		
Identify Data sources	Data sampling	Getting required resources			
Identify resources	Data formatting				
Learn from past					

The real-world companies that do Business Analytics and their interesting findings

There are so many companies that are engaged with business analytics projects, and they have become successful in the real world. Out of those, there are a few examples of successful projects and their findings.

Redfin

It is an online real estate brokerage marketplace and they have used analytics to figure out the factors which impact the price of houses. There are some interesting findings of their analytics projects.

- There are so many factors people do consider before they purchase a house. Out of those, they have identified three factors as very crucial. Price, location, and the number of bedrooms and bathrooms are the most considerable factors when compared to the others.
- While the COVID-19 pandemic influencing the business world in a variety of ways, it has impacted housing preferences too. People have begun to look for larger houses with more outdoor space. According to the findings of Redfin analytics, the price of houses with a balcony or backyard soared by 72% in 2020 than the previous year.
- With the help of analytics, Redfin has identified the most suitable day for their online home listing. According to them, the houses listed on Thursday sell faster and can earn extra money than houses listed on the other days of the week. Monday is the worst day for their house listing. If someone lists the house on Thursday, he can sell his house in an average of five days faster and earn for extra $3000.
- The houses which are near Starbucks are worth 96% or more than the other houses. If t there is a Starbucks within a quarter mile, they have more demand than the other houses. As well as some other houses a quarter mile from Trader Joe’s were worth 40% or more than without nearby.

Facebook / Meta

This is a social networking and social media platform owns by American company meta platforms. Because they are dealing with different ethnicities worldwide, they have found some interesting factors through their analytics.

- According to their findings, most Facebook users are willing to adopt particular behaviors which they find out their friends doing at the same time. Simply this is called social proof, and it is a strong driver of behavior.

- Emojis play a considerable role in Facebook posts. For instance, posts with the 😂 emoji get more shares and comments than posts without it.
- The videos on Facebook are very attractive to users. Therefore, Facebook pays attention to video content while it is increasing users' viewing time more. As well as they have found out that there is a high viewing time for short videos. So, Facebook has introduced short reels and stories in September 2021.
- The posts made on weekends have higher engagement than the post on the other days of the week. So, if someone needs to grab more attention for his post, the best is to post it on the weekend.

Walmart

It is one of the world's biggest retail companies which has developed its network globally. It has conducted various analytics projects over years and there are some interesting findings of them.

- Walmart has identified the weather as a main factor that changes the customers' behavior, and it leads to changes in sales patterns. For instance, during a hurricane, customers try to make some personal stocks of emergency supplies. During a heatwave, they buy more air conditioners, fans, and so on.
- Based on the analytics they have noticed that there is a growing trend of customers towards healthier and sustainable products. Customers are attracted to organic, natural products rather than others. Therefore, Walmart is increasing the variety of those products to attract more customers.
- According to the analytics, Walmart customers who use "Clicks and Collect" which means online purchasing and picking up from their convenience store, are more profitable than the customers who shop in-store because they tend to purchase more items and pay more money per transaction.
- Through proper placement of the product on shelves they try to maximize their sales and minimize waste. They try to increase the visibility of the product by increasing the distance between products, placing low-price items a little below the eye level selves, and placing similar products together. For instance, place chips and dips often close to each other.

Introduction to Google

Google is an American multinational company that was founded by two Ph.D. students in 1998. Google provides different services such as Google ads, Google Drive, Gmail, YouTube, Android, and so on. But it is world-famous as the best search engine and it is the most widely used search engine now. Because the Google search engine is fast and user-friendly.

To hold the position of the market leader, they give priority to innovation and research in different projects such as artificial intelligence, renewable energy, self-driving cars, quantum computing, natural language processing, and so on.

Today Google has become a giant company in the world which is operated in over 100 countries worldwide.

Successful Business Analytics projects done by Google

Google Analytics 360 implementation for Coca-Cola

As one of the paid services of Google Analytics, they have incorporated Coca-Cola to launch this project. With getting access to a big range of data and analytics they could develop more customer-oriented programs and marketing campaigns. This project helped Coca-Cola company to identify how to increase traffic for their website, what sources to reach some specific customer segments, and so on. As a result, Coca-Cola could increase its online sales by 20% and reduce the bounce rate by 14%.

Increase user experience for Pizza Hut

This project targeted online sales of Pizza Hut. Google used analytics to identify the areas where online customers are confused and leading to frustration. As well as they found out customer preferences to tailor the Pizza Hut website. With the help of this data, Pizza Hut improved its website and increased the user-friendliness of its website. As well as Pizza Hut streamlined its website targeting mobile devices, smartphones, and tablets to order their food.

Finally, this project was successful resulting 7% increase in online orders and a 48% decrease in customer calls for orders. This caused a further increasement in their customer satisfaction rates and loyalty.

Increasing App user engagement for ASOS

ASOS is an online beauty & fashion store. They offer their product and services through websites and mobile apps. Google analytics followed up on all the data regarding ASOS customers' purchase history, search queries, products they preferred and viewed, etc.

Based on this information and analytics ASOS did some changes to their app. Based on customer preferences and browsing history ASOS introduced personalized product recommendations and added social media sharing features.

As a result of those types of changes on their app, the revenue from the app increased by 30% and it increased by 15% app engagement too.

Google flue trend

This project was launched by google in a collaboration with the Center of Disease Control (CDC) in 2008. It is a web-based application and community data source which used search queries. The main purpose of it was to track the development of influenza-like illnesses worldwide.

Google analyzed the frequency and pattern of data they received through quarries related to a specific geographic area. Through overtime tracking of these trends and patterns, Google attempted to make 10 days early warning alert system for flu outbreaks which makes it easy for public health officials of that region to take action. This program was used in more than 25 countries worldwide.

Google always tried to avoid privacy violations when they gather data. They collected millions of anonymous search queries. The search log included IP addresses that could be used back to trace the region. Google used just machines to run the program and access data without getting any human involvement for this.

In the beginning, Google found out that their project is very successful and Google Trends predictions were almost 97% correct compared to data from CDC. But later on, the reports assured that sometimes the Google Trends predictions are inaccurate. For instance, during the flu outbreak season from 2011 to 2013, it overestimated the flu incidence. The enormous data had led to an overfitted prediction. As well as it entirely missed an enormous outbreak such as the Swine flu in 2009. Finally, Google Flu Trend was discontinued in 2015 after it was found that Google Flu trend has overestimated the outbreak of flu incidents in the United States.

Why google flue trend was failed?

Lack of ability of algorithms:

One of the reasons for this project's failure was the lack of ability of Google algorithms to identify the correct flu. People make flu-related searches and symptoms search for different reasons when they are not actually sick. Due to google directly did not interact with people for data gathering, they were unable to identify which cases are real and which cases are not. As well as some keywords such as “fever”, and “cough” were tracked by Google which led to increasing the flaw over time in their search algorithm.

Lack of transparency:

One of the other main reasons for failure is the lack of transparency in the project. The specific search terms used to make predictions were not revealed. Even the algorithm used for the project was not available to the public. Due to that researchers, and public health officials were unable to realize how this tool made predictions and asses its correctness.

Limited scope:

Google flu trend was totally based on the search quarry data, and it did not consider other factors which influence the predictions. For instance, it did not consider the search behavior of people.

Though this program was implemented in over 25 countries around the world, Google did not consider the search pattern changes in different regions and countries.

Other factors:

There are several other factors that affect flu cases. Weather, vaccination rate, other climate factors and etc. But Google has not considered these factors when they gather data for the project. For instance, a sudden heat wave in a specific area can be caused a fever and it cannot be considered for the data set which is going to predict the flu outbreak around the world.

What are the correction steps to be taken?

The Google Flu trend project was stopped by 2015. If Google going to implement this project again, they have to overcome the previously identified issues and pain points. According to the Business Analytics life cycle, Google should start from the first phase. In the discovery phase and data preparation phase, Google better focus on the following facts.

1. Discovery and Data Preparation

Integrate with other data sources:

One of the mistakes done by Google for this project was not collaborating with other data sources. Therefore, it was hard for them to measure the accuracy of the data they received. To overcome this issue, first, they must identify credible other data sources. This may include data from different social media platforms, electronic well-being records, and other online sources.

Use advanced algorithms:

Before re-launch this project, to identify the most correct data, Google can improve its algorithms to identify the correct cases. By increasing the relevant search terms and or can use combinations of search terms to identify the correct data. Analyzing the trends in online behavior through flu outbreaks and combining search phrases during those periods make this process easy.

After gathering data, Google can filter out relevant data for their project out of their gigantic dataset with the help of improved algorithms, Data filtering can be used to filter out irrelevant and misleading search quarries.

Use sophisticated data analysis techniques:

Here Google can identify the patterns by analyzing the existing data to increase the accuracy of their valuations. After analyzing the existing search quarry data, they can have some sort of understanding regarding user behavior in flu seasons.

Further, to identify some complex patterns in numerous data sets, Google can use advanced data analysis techniques such as Neural networks and Deep learning algorithms. By applying these techniques to Google search quarries, they can identify complex patterns which are unable to detect by traditional statistics methods.

Improve transparency:

One of the other drawbacks of this project was the lack of transparency. Google can provide information regarding the data sources and methods they used to make their estimates. By providing regular updates on the machine algorithms and allowing researchers and other interested parties to access data Google can increase the transparency of the project. This is really important to increase the stakeholders' trustworthiness regarding the project.

2.Build a model and test

Google can build a model to predict the flu. After applying the data analysis techniques, the model can be tested to measure accuracy. By comparing the model predictions with actual flu data from other public sources, google can determine the accuracy of data.

Improve the model:

After testing the model, if there are any improvements to be done, the model can be adjusted again to increase its accuracy of it. This may involve using different machine learning algorithms, incorporating other data sources, or maybe changing the assigned weight of different variables.

3.Validate the model predictions

To make sure the model works properly over time, it is important to get feedback from public health officials. Due to the flu being related to some other demographic factors, other factors which affect the flu, getting feedback from officials can be combined with the analysis.

As well as the model estimates can be validated by using other sources such as data from surveys, laboratory government, etc. Through this, Google can identify the pain points of the model where algorithms do not perform well.

Summary

Google Flu Trends was implemented in 2008 by Google targeting to track and predict the flu outbreak and make warning alerts 14 days early. Google used search queries to collect data and analyzed it. In the beginning, it seemed to be a favorable tool for public health officials as an accurate and quick monitoring method.

Later, the failure of Google flu trend in 2013 to predict accurately the flu, damaged the trustworthiness of this tool. This led to generate some criticism of the methodology and the reliability of the tool.

After failing several times to make accurate predictions caused it destroyed the trustworthiness. Finally, in 2015, Google decided to stop the project. But this project was helpful for the well-being of people worldwide. If it was successful could save the life of a large number of people who die due to flu worldwide and can reduce the number of hospitalizations for flu. If it could be re-launch, there are some corrective actions to be taken and must overcome the previous drawbacks of the project.