Capstone

Gage Bishop Hockerman

Table Of Contents:

What is Data Analysis 2

Problem Definition 3

Data Acquisition 4

Data Cleansing and Preparation 5

Model Development and Analysis 6

Conclusion 7

Sources 8

**What is Data Analysis?**

Data Analysis is the process of working with data to identify useful information that can help us make informed decisions and conclusions. The first step in Data Analysis is to identify the problem and define it to be “solved”. The next step is Data Acquisition- where you collect and find new information to solve the problem that has been identified. After the data has been collected it needs to be thoroughly checked to make sure the appropriate values are in the correct spaces, any outliers have been identified to be handled, and duplicate information has been deleted. Additionally the data needs to be outlined in the proper format. Finally after the data has been collected, combed through, and formatted- we can begin to develop and analyze the data.

What are the benefits of Data Analysis?

Why do we analyze data and what are the benefits of analyzing data? Data Analysis can assist with everything like reducing hazards and optimizing the customer experience. Data Analysis allows companies to optimize and improve the customer experience, make informed decisions, streamline operations, and mitigate risks and fraud.

Optimizing and improving the customer experience is one of the big benefits of data analysis. When applying analytics for the control process, designing, and optimizing business operations it ensures efficiency and effectiveness to fulfill expectations and operational excellence. It allows companies to decrease their risks of damaging brand loyalty and customer experience.

Analysis can help steer business decisions as well as reduce a company's financial losses, allowing them to make informed decisions. Predictive analytics anticipates what may happen in reaction to a change. Prescriptive analytics gives recommendations on how a company should respond to those changes. Additionally, sales data can be collected and analyzed, providing companies with the opportunity to evaluate performances and weather to proceed and apply those changes across the company.

Operational efficiency can be enhanced thanks to data analytics by collecting and analyzing the supply chain- allowing for the identification of potential future problems, revealing delays that may have been unseen, as well as helping companies optimize their inventory levels based on seasonal and other varying patterns.

Risks are everywhere for a business. Data analysis can help businesses identify and evaluate potential risks as well as help with taking preventative measures.

For my capstone, I was focused on optimizing the customer experience for League of Legend viewers. With the idea of fulfilling their expectations and exceeding them for future World Championship matches.

**What is Problem Definition?**

The goal of identifying and defining a problem is to find correlation between data. “Solving” the problem allows us to make informed decisions or draw conclusions backed by data. To identify a problem you need to understand the problem before solving it or developing a solution. It’s beneficial to identify and understand the root-cause(s) of the problem you’re trying to solve, the problem behind the problem. Identify who’s impacted by the problem, define the boundaries of the solution, and identify any constraints to impose on the solution. A problem definition is a statement regarding an area of concern, something to be improved, eliminated, or a question that needs better investigation.

Defining the problem will allow us to create a focused plan and how we can execute said plan. All the information necessary will be taken into account during this step so we can proceed and acquire the data that is needed.

Simple data/simple problems, is when a problem with the data given can simply be solved with a statistical equation because the y axis is a yes or no question. For example, my first capstone problem was correlating a team's win rate with a certain champion picked.

Complex data is composed of unique interchanging data that’s harder to analyze due to the complexity of the statistical equation. The y axis has more than two options to it. My most recent capstone problem is taking the correlation between years of experience and gold gathered per match per player.

**What is Data Acquisition?**

Data Acquisition is the process of gathering raw data that has been acquired to solve the identified problem. Once the Problem Definition has been defined, we need to create a strategy to collect and combine the appropriate data. To do so, we need to determine what data we need. There is Quantitative Data, which can be measured numerically and Qualitative Data, which is descriptive.

Once the type of data necessary has been identified- you need a DMP (Data Management Platform) to identify and aggregate the data from the pulled sources. Once the data has been collected, you can clean the data.

To acquire the data for my capstone I went to Kaggle. Kaggle is a free online resource that houses datasets. Allowing you to browse and download different datasets provided by users. The dataset I chose is a dataset on League Of Legends World Tournament 2022. I exported the dataset to Google Sheets and provided it with a title and description.

**What is Data Cleansing and Preparation?**

Cleaning the data or ‘scrubbing’ it, is very important in making sure that you’re working with high quality data. Making sure that the best data is being used, multiple cleaning tasks need to be completed. Tasks such as:

* Removing errors and duplicates (=IF(OR (Value=Below, Value=Above), "DUPILICATE FOUND","")) - inevitable when collecting data from numerous sources.When collecting from multiple data sources- there's more likelihood that the data may be duplicated or mislabeled. If the data is incorrect the algorithms and outcomes become unreliable, even if they may look correct.
* Removing outliers- the action of getting rid of irrelevant points that could represent measurement errors, data entry or processing errors, or poor sampling.
* Bringing structure to the data- fixing typos, layout issues, making sure the data variables are easy to read, and validating each section's data; this will help us map and manipulate the data easier. Frequently used file formatting such as excel, json, and csv.
* Filling in major gaps- when cleaning, important data may be missing. Once gaps have been identified we can proceed to fill those gaps.
* Scaling of a graph- transforming the data so it fits within the specific scale. Scaling data is based on how far apart the data points are.This will be different model to model.

There is no one correct way to assign tasks in data cleaning because based on the dataset the process will vary. However, it is important to develop a template for yourself while data cleansing to establish an outline of how it can be done correctly and to the best of your ability.

After completing the cleaning tasks, we should be able to answer some basic questions.

* Does the data make sense?
* Does it prove or disprove the outlined problem?
* Can trends be located in the data?

After this, the cleansed data will be developed into a model and analyzed.

**What is Model Development and Analysis?**

Model development is a repetitive process in which many models are derived, tested, and built upon until the desired criteria has been built and met. It allows us to create a visual representation of a whole information system or parts of it- showing connections between data points and structures. The point of model development is to illustrate types of data used and stored within the made system, the relationships among the data types, the way data can be grouped/organized, and the formats and attributes of the data.

Types of data modeling that are more commonly used are linear regression and logistic regression binary.

Linear regression is the approach for modeling the relationship between the scalar response and one or more variables. In the case of one variable it’s called a simple linear regression- if more than one variable is used it’s called multiple linear regression.

In a binary logistic regression the target variable is binary, it can only take 1 or 2 values; 0 or 1. This is the most utilized type of regression model in readmission prediction.

I used these two types in my capstone. I originally used logistic regression for my first capstone problem that focused on a team picking a specific champion and if they won with the champion. One being if the team won and zero being when the team lost.

For my second capstone problem I utilized the linear regression approach. In professional LOL play, what is the correlation between years of experience and gold per minute? The x axis was years of experience while the y axis was average gold per minute.

Conclusion:

Throughout the semester I had the opportunity to work with Derrick and learn directly from him the pipeline that a Data Analyst uses in their work life. Everything from defining a problem, acquiring data to solve the identified problem, cleaning and preparing the data, and lastly, developing the model for analysis. Being able to directly work on a project while learning the step by step process with Derrick allowed me to have a unique and informative experience, solidifying my desire to pursue a career as a Data Analyst.

Thank you so much for working with me all semester Derrick and Tim.

Sources:

A. Areyana, “League of Legends: World Championship 2022,” *Kaggle*, 28-Dec-2022. [Online]. Available: https://www.kaggle.com/datasets/ilyadziamidovich/league-of-legends-world-championship-2022?resource=download. [Accessed: 09-Mar-2023].

Oracle Selixir, “Oracle's Elixir - Lol Esports stats,” *Oracle's Elixir - LoL Esports Stats*, 05-Nov-2022. [Online]. Available: https://oracleselixir.com/stats/players/byTournament/2022%20Season%20World%20Championship%2FMain%20Event. [Accessed: 09-Mar-2023].

L. Portal, “2022 World Championship - Statistics,” *Liquipedia Portal*, 06-Nov-2022. [Online]. Available: https://liquipedia.net/leagueoflegends/World\_Championship/2022/Statistics. [Accessed: 09-Mar-2023].

S. McGahey, “Model Development - an overview | ScienceDirect Topics,” *www.sciencedirect.com*, 2022. https://www.sciencedirect.com/topics/computer-science/model-development#:~:text=Model%20development%20is%20an%20iterative (accessed Mar. 24, 2023).

“Python Machine Learning - Logistic Regression,” *www.w3schools.com*. https://www.w3schools.com/python/python\_ml\_logistic\_regression.asp (accessed Dec. 01, 2022).

M. Stojilkovic, “Logistic Regression in Python – Real Python,” realpython.com, Jan. 13, 2020. https://realpython.com/logistic-regression-python/#math-prerequisites (accessed February 05, 2023).

“Filling the area between lines — Matplotlib 3.7.1 documentation,” matplotlib.org. https://matplotlib.org/stable/gallery/lines\_bars\_and\_markers/fill\_between\_demo.html (accessed March 25, 2023).

“What is Data Modeling? | IBM,” www.ibm.com. https://www.ibm.com/topics/data-modeling (accessed Mar. 13, 2023).

M. Chowdury, “You are being redirected...,” www.analyticsinsight.net, Feb. 26, 2022. https://www.analyticsinsight.net/5-benefits-of-using-data-analysis-in-business-in-2022-and-beyond/ (accessed Mar. 11, 2023).

GeeksforGeeks, “GeeksforGeeks | A computer science portal for geeks,” GeeksforGeeks, 2019. https://www.geeksforgeeks.org/ (accessed Nov. 01, 2023).

M. Fritz and P. Berger, “Improving the User Experience through Practical Data Analytics,” ScienceDirect, 2015. https://www.sciencedirect.com/book/9780128006351/improving-the-user-experience-through-practical-data-analytics (accessed Apr. 25, 2023).