Data Science Methodology

Final Assignment

Which topic did you choose to apply the data science methodology to?

Hospital Data Management

Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer.

You are required to:

1. ***Describe the problem, related to the topic you selected.***
2. ***Phrase the problem as a question to be answered using data.***

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

The main goal of healthcare organizations is to provide quality treatment at reasonable cost. To maintain high standards of patient service, providers must make the right medical decisions. The vast amount of unstructured healthcare data complicates decision-making.

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with.

1. ***Analytic Approach***
2. ***Data Requirements***
3. ***Data Collection***
4. ***Data Understanding and Preparation***
5. ***Modeling and Evaluation***

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

Business Understanding

It forms the basis of an effective solution to the hospital data management. Business partners who need the analytics solution play a critical role in this phase by defining the problem, the project objectives, and the solution requirements from a healthcare organization perspective.

Analytical Approach

The approach is to ask the person asking the question to clarify the most appropriate form or approach.

- Descriptive
  - Current status

- Diagnostic: Statistical Analysis
    - What happened?
    - Why is this happening?
- Predictive: Forecasting
    - What is this trend continue?
    - What will happen next?
- Prescriptive o How do we solve it?
- Daily admissions: Outpatient / Inpatient
- Type of problems
- Departments
- Employees

Data Requirement

This involves identifying the content, formats, and data sources needed for the initial data collection.

- Contents, formats, representations suitable for decision tree classifier
    - One record per patient with columns representing variables
    - Contents covering all aspects of each patient's clinical history
        - Transactional format
        - Transformations required
    - Treatment and diagnosis
    - Hospital inventory management
    - Employees, salaries and other expenditure

Data Collection

At this stage, the data requirements are reviewed and a decision is made as to whether more or less data is required for the collection.

Techniques such as descriptive statistics and visualization can be applied to the dataset to evaluate the content, quality and information of the original data.

The gaps in the data are identified and plans for filling or replacement must be made. Data can be collected from

- Available sources
    - Corporate data warehouse
    - In-patient record system
    - Out-patient record system
    - Inventory management system
    - HR Department o Claim payment system
    - Disease management program information

Data Understanding

Understanding the data includes all activities related to creating the data-set.

The “Understanding Data” section of the Data Science methodology answers the question:

- Is the data you collect representative of the problem you are trying to solve? To understand data, descriptive statistics had to be established in the data columns that would become variables in the model.
- First, these statistics included Hearst, Uni-variate, and statistics for each variable, such as mean, median, minimum, maximum, and standard deviation.
- Second, pairwise correlations have been used to determine the degree of correlation between the linked variables and those that, if any, are highly correlated, meaning that they are essentially redundant, making it only relevant for the modeling.
- Third, the histograms of the variables were examined to understand their distributions.

Histograms are a good way to understand how values or variables are distributed and what kind of data preparation may be needed to make the variable more useful in a model.

For example, if a categorical variable contains too many different values to be meaningful in a model, the histogram can help decide how to consolidate those values.

- Univariate, statistics and histograms are also used to assess the quality of the data. On the basis of the data provided, some values can be recorded or deleted if necessary, e.g. For example, if a particular variable has a lot of missing values.
- The question then arises as to whether “missing” means something. Sometimes a missing value means “no” or “0” (zero), or sometimes simply “we do not know”.
- Or if a variable contains invalid or misleading values; For example, a numeric variable called “age” containing 0 to 100 and 999, where “triple-9” actually means “missing”, will be treated as a valid value unless we have corrected it.
- First, the importance of heart failure was determined based on a primary diagnosis of heart failure. However, the data comprehension study revealed that the initial definition did not cover all expected cases of heart failure due to clinical experience.
- This involved returning to the data collection phase, adding secondary and tertiary diagnoses, and creating a more complete definition of heart failure approval.
- This is just an example of the interactive processes in the methodology. The more you work with the problem and the data, the more you learn and the more the model can be adjusted, which ultimately leads to a better resolution of the problem.

Data Preparation

- With data collection and understanding, data preparation is the slowest phase of a data science project. As a rule, it takes up 70% or 90% of the total project time. By automating certain data collection and preparation processes in the database, this time can be reduced to only 50%.

- This saving of time means that data scientists should focus more on creating models.
- data transformation in the data preparation phase involves putting the data in a state where it may be easier to work.
- To work efficiently with data, missing or invalid values must be changed and duplicates removed to ensure proper formatting of all data.
- Feature engineering is essential when machine learning tools are used to analyze data. When working with text, text analysis steps are required to code the data for manipulation.
- The data scientist must know what he is looking for in his file to answer the question. Text analysis is essential to ensure that the correct groups are defined and that programming does not overlook what is hidden inside.
- The data preparation phase is the basis for the next steps to answer the question. Although this phase may take some time, the results will support the project if done correctly. If this is omitted, the result is not at the same level and can sit on the drawing board.
- Make sure you spend time in this area and use the tools available to automate common steps to speed up data preparation. Pay attention to details in this area.

Data Modelling

Data modeling focuses on the development of descriptive or predictive models. • An example of a descriptive model might be the following: if someone did it, they probably prefer it.

- A predictive model attempts to provide yes / no results or to stop / continue. These models are based on an analytic approach learned either statistically or Machine learning. The Data Scientist will use a training set for predictive modeling.
- A training set is a set of historical data in which the results are already known. The training set serves as an indicator to determine if the model needs to be calibrated.
- At this point, the data scientist will use several algorithms to ensure that the variables involved are really needed.
- The success of data collection, preparation and modeling depends on an understanding of the problem in question and the appropriate analytical approach.
- The data support the answer to the question and the quality of the ingredients in the kitchen is the basis of the result.
- Each step requires constant improvements, adjustments and tweaking to ensure the strength of the result.

Model Evaluation

A model evaluation goes hand in hand with the creation of models. The modeling and evaluation steps are performed iteratively. The evaluation of the model is carried out during the development of the model and before deployment.

- The evaluation evaluates the quality of the model, but also provides the opportunity to determine if it meets the initial requirements. The evaluation answers the question:

- Does the model used really answer the original question or should it be adapted? The evaluation of the model can have two main phases.
- The first phase is the diagnostic measurement phase, which ensures that the model works as intended.

  If the model is predictive, a decision tree can be used to assess whether the response provided by the model matches the original design. This allows areas to be displayed where adjustments are required. If the model is a descriptive model that evaluates the relationships, a set of tests with known results can be applied and the model refined as necessary.
- The second evaluation phase that can be used is the statistical significance test. This type of evaluation can be applied to the model to ensure that the model data is processed and interpreted correctly. This is to avoid a second unnecessary assumption when the answer is revealed.