Topic: CANCER SURVIVAL PREDICTION WITH SYNTHETIC DATA

Members: Hugo Martin Teemus, Robert Israel, Maria Malk

1. Business understanding

Goals

- Background

Lung cancer is a major global health problem, with survival rates dependent on a variety of factors, including the course of treatment. The goal of this research project is to use synthetic medical data to predict 5-year survival rates after the first lung cancer diagnosis. By analyzing synthetically made patient treatment trajectories, we aim to identify key procedures, drugs and other various treatment options that influence survival and develop a predictive model to help make more informed decisions about patient care.

- Business Goals

Prediction accuracy: Develop a high-accuracy prediction model to predict the 5-year survival rate of lung cancer patients based on their treatment courses that consist of drugs given and medical procedures made. Utilize advanced feature selection techniques and pattern finding to identify important elements in survival of the patient.

- Business Success Criteria

Develop a predictive model with the best AUC score possible. Our goal is not set in numbers but we strive for at least 80% accuracy on the test set.

If possible identify a concise list of significant features influencing survival outcomes, that can be used as itself as a part of further research. The publication of this list is based on legal aspects, whether we are allowed to know it, since it needs to be legally approved.

Demonstrate the added value of subsequence analysis in predicting survival rates.

Assessing the situation

- Inventory of Resources

We work with synthetic medical data that replicates the treatment process of lung cancer patients that was generated for this project. For the tool we use algorithms such as PCA, stepwise selection and network science approaches. We incorporate what we have learned into our courses and if possible ask additional information for experts who have worked with this kind of data. Our go to software program is jupyter notebook. No additional hardware needed.

- Requirements, Assumptions, and Constraints

Data is synthetic so no need to worry about ethical standards, but going forward if we apply our code on real life data we need to consider ethical standards and data privacy regulations. There is an assumption that synthetic data accurately represents real-world scenarios. Project results and a poster are due 11.december.

- Risks and Contingencies

The first risk is the quality of synthetic data and whether it matches real life data. It determines if our methods and algorithms are effective on real life data. We would like to make our model as simple as possible to minimize overfitting and limit the usage of computational resources. If overfitting occurs the contingency is to cross-validate and regulate the data. If we run out of computational resources we will ask the University of Tartu to provide more computational resource or simplify our model.

- Terminology

Treatment Trajectory: Sequence of medical procedures a patient undergoes post-lung cancer diagnosis and drugs given.
Feature Selection: Process of identifying the most relevant variables influencing the prediction model, contains procedures and drugs given.
Subsequence Analysis: Examination of frequent patterns within treatment trajectories.
Network Science Approach: Application of graph theory to identify key-player nodes in treatment networks.

- Costs and Benefits

The cost consists of the contribution of our skills, time and knowledge. The benefit is that our work can improve the accuracy of predicting survival rates and identifying important treatment characteristics

Defining Your Data-Mining Goals

- Data-Mining Goals

Better understanding of how different treatments affect people's survival rate for further research. Ideally our model can be implemented on many different medical datasets. We will be making a poster of our work.

- Data-Mining Success Criteria

Since using RandomForrest method the AUCROC was 65% we strive to have a higher AUCROC score for given data then that.

2. Data understanding

Gathering data

Data is given by a domain expert. We are given two datasets as CSV files. One is synthetic data based on prostate cancer trajectories and other synthetic data based on lung cancer trajectories. Additional data can be gathered by research by experts who have a legal permit to conduct this kind of research and have an ethical agreement with the government to do so. For this project no additional synthetic data is not required and if it is needed it is provided by the same domain expert.

Description of data

Data tables have three different values given:

- SUBJECT_ID - unique patient id
- DEFINITION_ID - medical intervention that took place with the patient
- TIME - time in years when the intervention took place

At time 0, each patient was diagnosed with cancer. Some patients have the status "death", which means that the patient died at that time. For all patients, the year before death/last record has been deleted. Looking at the data, if a patient survived we don't have any data about procedures done a year before the last one, it can be calculated as an approximate time. The medical interventions consist of drugs, conditions, measurement, procedures and observations for both tables. These are numbered to differentiate between before named interventions.

Prostate cancer synthetic data

- consist of 4623 different interventions, containing all drugs, conditions, measurement, procedures, observations and death
    - 2089 conditions
    - 368 drugs
    - 1315 measurements
    - 425 observations
    - 425 procedures
    - 61 patients with death as last value
- There are 698 subject described

Lung cancer synthetic data

- consist of 4864 different interventions, containing all drugs, conditions, measurement, procedures, observations and death
    - 2399 conditions
    - 418 drugs
    - 1333 measurements
    - 224 observations
    - 489 procedures
    - 263 patients with death as last value
- There are 727 subjects described

Exploring and validating data

All data in the dataset is complete as far as we can validate using our given knowledge. There are no missing values, so we can use all available information. The dataset is organized by subject ID and also chronologically ordered by time of occurrence within each subject. In cases where the last recorded event is marked as "death", no further information is available for that subject. Therefore, all data up to the time of the last recorded event are considered valid for our analysis.

We concluded that the first step to make the given data available for our model to predict survival rate based on different medical interventions is to vectorise all the interventions based on the survival of the subject. If the subject survived, the intervention was classified as 1 and if the subject did not survive the intervention classification was 0. With that step done we are ready to start our model building.

3. Planning your project

1. Analyzing the test and validation datasets that are given. Deciding on the approach for processing the datasets. Current flow for data processing uses one-hot encoding and vectorization to create one-hot vectors.

- For the features part (X) of the data, each row represents a single patient, each column represents all unique medical interventions. For each patient, the value of the medical intervention is ‘1’ if the intervention was done to the patient, otherwise the value will be ‘0’.
- For the prediction part (y) of the data, we want to predict the values for the column labeled ‘death’, where the value in a row represents if the patient died within 5 years of the intervention or not. Value will be ‘1’ if the patient died within 5 years of the last intervention, otherwise the value will be ‘0’.

  Processed data is used for feature selection.
    - Contribution: Everyone will spend 2-3 hours on getting to know the data or analyzing it and first processing

2. Analyzing possible approaches for feature selection. Analyzing the processed data to find correlations between the medical interventions done and the ‘death’ value of the patient. Trying out different methods on our processed data. Selecting the approach, that selects the features with which the models get the highest accuracy.

- Current methods we’re trying:
    - Analyzing the dataset, to find the features correlated with survival rates by ourselves, e.g. selecting interventions most done to alive and dead patients to filter out noisy data; finding the survival percentage of patients for each intervention.
    - Feeding the processed data into algorithms, that calculate the importance and variance scores for each feature. Currently, it seems PCA finds the best features for predicting survival rates.
    - Contribution: 15-20h , Mostly Hugo and Robert

3. Evaluating different approaches for training models. Testing the models with the selected features from the processed data. Doing hyperparameter tuning to get the model with the best AUC score. Currently, our best AUC score seems to be around 0.85, using PCA and RandomForestClassifier.
    - Contribution: 15-20h, Mostly Hugo and Robert

4. Extract the most critical features from cancer treatment trajectories that effectively indicate whether a patient is likely to survive or not. Then try to visualize the

extracted features from a comprehensive overview of the most crucial “key players” influencing cancer treatment outcomes.

- Contribution: 15-20h, everyone

5. Design the poster, reporting findings. Making a plan to pinpoint the objectives to complete the workflow.
    - Contribution: 10-15h, Mostly Maria