G7 Report

Lung Cancer Survival Prediction with Synthetic Data

Uku Kangur, Jaan Märten Huik

Business understanding

Identifying your business goals

Background

In the developed countries 3 disease groups with highest mortality are cardiovascular diseases, cancers, neurodegenerative diseases and overweight/obesity which is a big risk factor for all previous disease groups. While cardiovascular diseases are more prevalent compared to cancers, modern drugs, more easily implementable screening options and clear disease specific guidelines allow most of the patients to enjoy similar quality of life as unaffected people. Cancers however have a notoriously wide range of effects to patients and the prognosis depends on an uncountable large set of factors like unhealthy habits, genomics, stress, driver mutations of cancer colonie, how advanced the disease is at discovery etc. Wide spectrum of clinical presentation in different patients makes it difficult to adequately assess the prognosis, decide on best available treatment options and drug schemes and preserve the quality of life for patients.

More accurate prognosis would allow patients to plan their life, allow physicians to choose treatment plans more likely to have an effect on cancers and allow Healthcare insurances to allocate already scarce funds to hotspots where they would serve the best outcome. Therefore any model that succeeds with these goals could have a big impact on healthcare systems and have a meaningful effect on individual lives and families as well.

Business goals

Our main goal in developing a model is to predict their mortality in a period of one year – that is to as accurately as possible to classify patients according to available data into two classes 1 or 0, dead or alive. These classifications might allow physicians to initiate more aggressive and toxic drugs in those patients that show increased risk of dying within a year. In contrast, patients who wish not to undergo uncomfortable treatments could plan and enjoy their limited time with palliative treatments.

Business success criteria

Since our end goal is to predict mortality using already available data that includes their survivorship status we can quantify the accuracy of the model quite easily. Therefore models that have higher prediction rates are better.

Assessing your situation

Inventory of resources
Our most valuable assets are two Ferrari-engine brains with a diverse set of domain knowledge. Uku Kangur is PhD student from University of Tartu Institute of Computer Science focusing on spread and mitigation of misinformation which makes him very acquainted with data analysis and model logic. Jaan Märten Huik is a medical student in his final year with beginner-level skills in Python and R. He has gained experience in the oncology department and currently is involved in a genetics and personalized medicine clinic, focusing on genomic data.
We are guided and supervised by Markus Haug — Ph.D student and a junior researcher in Health Informatics who has answers to all of the questions and provided us with synthetic data from Estonian Cancer Registry. Estonian Cancer Registry was started in 1978 and collects increasingly more information on incidence, death and so much more.
Additionally we have two laptops, access to cutting edge LLMs and unlimited access to vast space of internet.

Requirements, assumptions, and constraints
The final deadline for submission of our work is December 11th 2023, and the poster will be presented December 15th. Requirements of the final submission are filled out form with information about the project, access for instructor to code repository.

Risks and contingencies
There are no detectable risks and contingencies, we have made copies of the data, the code is synced and backed up. The only potential mechanism for delay is Jaan having an e.coli infection due to drinking tap water in Kuressaare and Uku slipping on ice and getting hospitalized. If that were to become true, we have no backup plan.

Terminology
- **Cancer Incidence**: number of new cancer cases in a specific population during a given time period. It's often expressed as a rate per 100,000 people. It helps in understanding how widespread cancer is within a particular demographic or geographical area.
- **Cancer Mortality**: number of deaths caused by cancer in a specific population over a certain period. Like incidence, it's often presented as a rate per 100,000 people and indicates the severity or lethality of cancer in that population.
- **Predictive Modeling**: a data-mining technique used to predict future outcomes based on historical data. In the context of cancer research, it can be used to forecast cancer trends or the effectiveness of treatments.
- **Data Normalization**: the process of organizing data in a database to reduce redundancy and improve data integrity. This is crucial in large datasets like those used in medical research to ensure accuracy and consistency.
- **Machine Learning**: a subset of artificial intelligence that provides systems the ability to automatically learn and improve from experience. In cancer research, machine learning algorithms can be used to identify patterns in complex datasets, like genomic data.

- **Data Mining**: the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It's essential in analyzing vast amounts of medical data to extract meaningful insights.
- **SMOTE**: Synthetic Minority Over-sampling TEchnique.
- **AUROC**: an Area Under the Receiver Operating Characteristic.

Costs and benefits

Our hourly rate is 150€. We are both expected to contribute 30 hours to project work, which is 60 man hours, this makes the total cost of man hours 9 000 €.

Defining your data-mining goals

Data-mining goals

The modeling process was started by creating a RandomForrest model as a baseline model. For this so called wide-dataset was created by:

1. Creating features according to interventions and order:
    a. Removing “Death” column from the dataset.
    b. For each unique intervention and its order in the sequence, a separate feature was created.
    c. Each intervention and its sequence number was transformed into a distinct characteristic.
    d. The data was organized in a way that there is one row per patient, with the value being the time associated with each intervention.
2. Normalizing of time intervals for each patient:
    a. Time intervals were normalized for each patient.
    b. Time data was adjusted so that each row (representing a patient's timeline) has times ranging from 0 to 1.
    c. It was ensured that '1' represents the time unit of the last intervention in the patient’s trajectory, excluding death so the model couldn’t cheat by using shorter absolute time.
3. Reintroducing the "Death" column:
    a. The column labeled "death" was reintroduced to the dataset.

Since death was underrepresented in the provided dataset, we used SMOTE to balance the dataset. RandomForrest model (n_estimators=100) was then trained on a balanced dataset to obtain a baseline accuracy and AUROC score.

Data-mining success criteria

We aim to develop a model that equals or surpasses the performance of Markus Haug's initial Random Forest model on the test dataset. To be even more precise, our benchmark for success is achieving an AUROC score of 0.75 or higher, matching the score attained by Markus's model after feature selection.

Data understanding

Gathering data

Outline data requirements

The project requires medical history data from lung cancer patients. As the aim is to predict survival of the patients during the next year based on these trajectories, the data should have at least a few years of medical trajectory data.

Verify data availability

Real medical data is extremely sensitive data under GDPR and acquiring and working with such data would need consent from the patients. Luckily the University of Tartu Health Informatics Group has provided us with similar synthetic data generated from real medical data from lung cancer patients.

Define selection criteria

The data will be given to us in tabular (.csv) format, so that we can easily read and process it with python without any additional parsing from medical documents or databases. For us relevant data would have separated information about subjects, their health trajectories and also the timeline of their trajectories (which in reality is the case as well).

Describing data

The data provided comes in two separate files since they were generated with different synthetic data generation methods. For our use case we will develop our solution on the first dataset and then validate the generalizability of our pipeline using the second dataset. As an example of what the data includes,, we introduce 5 rows from the first dataset:

SUBJECT_ID	DEFINITION_ID	TIME
1	drug_217	0.004807
1	condition_1922	0.008643
1	condition_785	0.027792
1	drug_49	0.032515
1	measurement_132	0.056765

We can see here that we have three columns:

- SUBJECT_ID is the identifier for each patient
- DEFINITION_ID is the medical intervention done with the patient
- TIME is the time of the intervention in years, where time 0 is when the patient got the cancer diagnose

Some of the patients trajectories end with the definition_id of “death”. This is our classification label, which we aim to predict. To no cheat (overfit) regarding the prediction, the last year medical trajectory before death has been removed from the data.

Also when solving the prediction problem, one also has to consider that the two datasets do not have the same definition_ids, so “drug_217” in the first dataset does not correspond to “drug_217” in the second dataset. This means the solution should rather be a universally applicable data pipeline rather than a specifically engineered solution for only one dataset.

Exploring data

The first dataset includes 727 unique subjects, 4864 unique interventions and the range of the time is 0.00000129 - 16.86671224 (so from around 0 to around 16.86 years).

The second dataset includes 698 unique subjects, 4623 unique interventions and the range of the time is 0.00000768 - 18.54682926 (so from around 0 to around 18.55 years).’

In both datasets we observer that there are also specific types of interventions (in addition to death). There are 4 of them and types are:

- measurement - 1332 unique interventions
- condition - 2399 unique interventions
- drug - 418 unique interventions
- procedure - 490 unique interventions

We note that the first dataset has 263 subjects who died and 464 subject who did not. The second dataset has 61 subjects who died and 637 subjects who did not. This means we must also do class balancing on our data before training the model.

Verifying data quality

The data that we have for the task is sufficient to solve the main objective. However there are some limitations to this data such as it being quite noisy - not all medical interventions are relevant and this can make the model focus on the wrong features of the data. This smart feature selection and cleaning is the most important key part to solving this task effectively and efficiently.

Planning

Our general plan was to have weekly meetings to discuss progress and possible concerns and reasons for occurring stagnations.
General steps:

1. Prepare the data for learning and develop the baseline model that all new models are compared to.
2. Analyze the dataset, how many classes of interventions were in the dataset, how many times did they occur etc. Discover and describe existing frequent patterns.

After preparing the dataset for modeling and creating the baseline Random Forest model we moved on to test different strategies:

1. Randomly filter out measurements that don’t seem to impact models’ performance and train the Random Forest model again on the smaller datasets.
2. Perform associate rule mining and consciously remove (or combine) the feature that is occurring together with another feature.

3. Find correlations between features and drop those that are highly correlated. Additionally find correlations between features and Death column to determine highly relevant features.
4. Fine-tune several machine learning models including hyperparameter tuning and cross-fold validation to determine the best model for our data.

The final objective is to sell our shares in the initial public offering (IPO), cash in our unicorn money and retire in the Bahamas.