Cancer Survival Prediction with Synthetic Data

Members: Marten Mathias Jaani, Norman Tolmats, Joosep Lember

Repository link: https://github.com/Norman90844/IDSproject

TASK 2

Business Understanding for Predictive Model Development in Healthcare

Identifying Business Goals

Background: The project is situated within a healthcare data analysis context, focusing on the development of a predictive model for patient survival. Using synthetic data based on real cancer patient trajectories (lung and prostate cancer), the goal is to predict 5-year survival rates. The data includes patient IDs, medical interventions, and the timing of these interventions in years starting from the diagnosis date. Data consists of a lot of definitions from which we should be able to train our model and be able to predict in the future with it.

Business Goals: The primary business objective is to create a predictive model that forecasts patient survival over a 5-year period post-diagnosis. This model aims to identify significant features and vital treatment sequences to assist stakeholders in making informed decisions. To be able to do so we need to filter out the important features like drugs and other conditions that have the most effect on the patients life expectancy.

Business Success Criteria: Success will be measured by the model's accuracy, specifically its ability to classify patient outcomes correctly (survival or death within a 5-year period). The target metric is a high AUC-ROC value, demonstrating the model's ability to distinguish between the two classes effectively. We will try to get the score as high as possible and hope we will succeed.

Assessing the Situation

Inventory of Resources: Resources include two synthetic datasets in CSV format, computational tools for data analysis (like Python, Jupyter Notebooks), and access to real patient data for validation purposes through our instructor. We intend to use one of the datasets to train our model and the other one to test it.

Requirements, Assumptions, and Constraints: The model requires high-quality, relevant features from the datasets that actually have an effect on the patients health outcome. It's assumed that medical interventions are critical predictors of patient survival. Also we can assume there is a lot of noise in the given data as there are many attributes. A major constraint is the inability to share real patient-level data. The synthetic data might not be as accurate as would be the real data.

Risks and Contingencies: Risks include potential inaccuracies in synthetic data and the model's limited applicability to real-world scenarios. Contingency plans involve iterative model refinement and validation against real patient data.

Terminology: Key terms include AUC-ROC, synthetic data, patient trajectory, medical interventions, key players, graphs and survival prediction.

Costs and Benefits: Costs involve time and resources for data processing/cleaning and model development. Benefits include potential advancements in predictive healthcare analytics and contributions to scientific research. Everyone would benefit from it so basically the whole society because it could help a person or its close one to fight against lung and prostate cancer.

Defining Data-Mining Goals

Data-Mining Goals: To develop an effective classification model that can accurately predict patient mortality within a year. The model should be able to handle noisy data and select relevant features. The last is most important because we have so many different features that apply on different times to the patients that it will be hard to predict overly accurately.

Data-Mining Success Criteria: Success will be judged on the model's accuracy, interpretability, and the relevance of the identified features. The ability to generalize findings to real patient data is also crucial. We would call it a success if our model is stable on many different datasets and gives approximately accurate predictions on all of them.

This project has significant implications in healthcare, providing tools for better patient outcome predictions and potentially guiding future research in lung and prostate cancer treatment. The focus will be on creating a model that is both accurate and interpretable, using innovative methods like graph creation and key-player identification for feature selection.

TASK 3

Data Understanding for Predictive Model Development in Healthcare

Gathering Data

Outline Data Requirements: The data required for this project should include patient identifiers, medical intervention details, and the timing of these interventions. It is also essential to have data on patient outcomes, specifically their survival status at the end of the study period.

Verify Data Availability: The available dataset that we plan on using to train the data contains 560,971 rows and 3 columns, encompassing a wide range of SUBJECT_IDs and DEFINITION_IDs over time. This suggests a comprehensive dataset that likely meets the requirements for developing a predictive model. There are 984 patients in that dataset and out of them died 263.

Define Selection Criteria: The selection criteria for data to be included in the model development will be based on the relevance and frequency of medical interventions, the completeness of patient trajectory data, and the accuracy of the survival status information. Using all of these selections we are aiming to get the best result.

Describing Data

The dataset consists of three columns: SUBJECT_ID, DEFINITION_ID, and TIME. SUBJECT_ID is a unique identifier for patients; DEFINITION_ID categorizes the type of medical intervention (drug, measurement, condition, procedure etc), and TIME represents the time in years since diagnosis at which the intervention occurred. The range of SUBJECT_IDs and DEFINITION_IDs indicates a diverse patient population and a variety of medical interventions, spanning a substantial timeframe.

Exploring Data

An initial exploration of the dataset reveals a large number of unique patients (SUBJECT_ID) and medical interventions (DEFINITION_ID). The interventions vary from drugs to conditions and measurements, suggesting a detailed recording of patient treatment trajectories. All of these interventions also have their own unique ID-s depending on what drug or procedure was done. TIME values are fractional, indicating events are recorded with high precision, possibly up to the day of occurrence.

Verifying Data Quality

The data quality can be initially assessed by checking for missing values but it seems that there are no missing values in our given data sets, inconsistencies in the data (like negative time values), and duplicate records. The integrity of SUBJECT_IDs must be confirmed to ensure that they uniquely identify patients throughout the dataset. For DEFINITION_ID, it is crucial to understand whether the interventions are consistently coded and if the definitions remain consistent across the dataset.

The quality verification will also involve ensuring that the 'death' condition is accurately recorded and corresponds to the correct time since diagnosis. Given the context of predicting 5-year survival rates, any discrepancies in recording the death event could significantly impact the model's accuracy.

In conclusion, the dataset appears to be robust and relevant for developing a predictive model for patient survival. The next steps will involve cleaning the data to address any quality issues identified and preparing it for the modeling phase, ensuring that the features selected for the model are the most predictive of the patient outcomes.

TASK 4

Project Plan for Predictive Model Development

1. **Data Preprocessing** (7 hours/team member)
   - Cleaning: Handling missing values, errors, and inconsistencies.
   - Transformation: Encoding categorical variables, etc.
   - Feature Engineering: Deriving new features from existing data.
2. **Exploratory Data Analysis** (5 hours/team member)
   - Statistical Analysis: Understanding distributions and relationships.
   - Visualization: Creating plots to observe patterns and outliers.
   - Hypothesis Testing: Assessing assumptions about data features.
3. **Model Development** (8 hours/team member)
   - Selection: Choosing appropriate machine learning algorithms.
   - Training: Building models using training datasets.
   - Optimization: Tuning hyperparameters to improve performance.
4. **Model Evaluation** (5 hours/team member)
   - Cross-Validation: Assessing model performance with unseen data.
   - Metrics Analysis: Evaluating AUC-ROC and other relevant metrics.
   - Interpretability: Ensuring model decisions are understandable.
5. **Validation and Testing** (3 hours/team member)
   - Real Data Testing: Applying the model to real patient data (via the project sponsor).
   - Performance Assessment: Comparing model outputs against actual outcomes.
   - Iteration: Refining the model based on performance feedback.
6. **Documentation and Reporting** (2 hours/team member)
   - Documentation: Detailing the methodology, code, and findings.
   - Reporting: Summarizing the results and insights for stakeholders.
   - Publishing: Preparing materials for publication or presentation.

Methods and Tools:

- **Python**: For all data manipulation, analysis, and model development tasks.

- **Scikit-learn/TensorFlow/Keras**: For implementing various machine learning models.
- **Pandas/Numpy**: For data manipulation and numerical analysis.
- **Matplotlib/Seaborn**: For data visualization.
- **Jupyter Notebooks**: For code organization and sharing among team members.