DATA-PROJECT-LUNG-CANCER

Henry Marie MONT, Jean-Côme CARISSAN

November 2023

Business Understanding Report

Identifying Our Business Goals:

In this project, our main goal is to predict 5-year survival rates after a lung cancer diagnosis. We're using synthetic medical data to make these predictions more accurate and help in making timely medical decisions. The idea is to improve how we approach patient care for those dealing with lung cancer.

We have two main objectives. Firstly, we want to enhance patient outcomes by providing more accurate survival rate predictions. This should contribute to better decision-making and the creation of personalized treatment plans. Secondly, we aim to make healthcare resource usage more efficient by identifying key features and treatment patterns through advanced data-mining techniques.

Success means achieving significantly better accuracy in predicting 5-year survival rates compared to current models. We also want to see real impacts on clinical decision-making and improvements in resource allocation efficiency in the healthcare system.

Assessing Our Situation:

Looking at what we have, we're using synthetic medical data, making sure we play by the ethical rules. Our team is composed of 2 persons, but we can also rely on the help or tips of Markus HAUG who has proposed the project:

- Henry Marie MONT: Erasmus student from France, studying in his M1 of Computer Science.
- Jean-Côme CARISSAN: Erasmus student from France, studying in his M1 of Computer Science.
- Markus HAUG: Ph.D student in Health Informatics at Tartu University.

In terms of requirements, we're keen on following data privacy regulations and ethical use of synthetic data, but this shouldn't be too hard as we are working on synthetic data. The success of the project depends on how well our algorithms perform and how understandable they are.

Concerning the risks, we are actively scouting the synthetic data to catch any inaccuracies or problem. We should be careful to not over fit our model, especially with medical information which are not really relevant to lung cancer treatment.

To keep everyone on the same page, we've set clear terminology of the words used in the project, this might get expanded later on:

Synthetic Medical Data: Fictitious healthcare information generated to simulate real-world patient data without compromising privacy or breaching ethical standards.

Key Features: Essential variables identified through data-mining techniques, crucial for predicting 5-year survival rates post-lung cancer diagnosis.

Treatment Subsequences: Sequences of medical interventions derived from patient treatment trajectories, indicating patterns influencing survival predictions.

Algorithm Effectiveness: The measure of how well implemented algorithms, such as PCA and Stepwise Selection, perform in identifying crucial features and treatment subsequences.

Model Performance Metrics: Quantitative measures evaluating the accuracy and effectiveness of the predictive model, including metrics such as precision, recall, and F1 score.

Resource Allocation Efficiency: Enhancing the optimal utilization of healthcare resources by identifying key features and treatment patterns, contributing to more effective patient care.

Defining Our Data-Mining Goals:

In the data-mining part, we're focusing on identifying key features and treatment patterns. We will try to use techniques like Principal Component Analysis (PCA), Stepwise Selection, and robust algorithms for this.

Success in data mining is tied to how well these algorithms work. We're looking for clear improvements in model performance metrics, and we hope that Markus HAUG will be able to validate our findings with real clinical input.

In a nutshell, our goal is to use synthetic medical data for more accurate 5-year survival predictions after lung cancer diagnosis. We want to improve patient outcomes and make healthcare resource usage more efficient. Success is all about being more accurate, making a real impact on decisions, and ensuring our methods are practical and ethical.

Data Understanding Report

Gathering Data

The dataset crucial for this project was provided by Markus HAUG. Upon initial loading, it was evident that the key columns `SUBJECT_ID`, `DEFINITION_ID`, and `TIME` are readily available, laying the groundwork for subsequent analyses.

To achieve our project goal of predicting 1-year mortality, a focused approach is essential. The dataset includes various medical interventions and a temporal component (`TIME`). Defining selection criteria becomes imperative, targeting specific medical interventions and time intervals leading up to the 1-year mark.

Describing Data

The dataset, comprising 560,971 entries with a memory usage of 12.8 MB, provides a comprehensive overview of the patient landscape. The `TIME` column, signifying the time in years since the initial cancer diagnosis, exhibits a distribution with a mean of around 4.16 years. Notably, the minimum time is recorded at 0.000001 year, representing the moment of the initial diagnosis, while the maximum extends to 16.87 years.

A more nuanced exploration of the `TIME` distribution reveals stability within the 0 to 7.5 years range. This timeframe captures a substantial portion of patient data, suggesting its clinical significance. However, the presence of outliers beyond 12.5 years necessitates a deeper investigation to discern their clinical context and potential impact on the predictive model.

Verifying Data Quality

Ensuring the quality of the dataset involves rigorous checks for missing values, duplicate entries, and outliers. The good news emerges—no missing values were detected, affirming the completeness of essential information. Similarly, the dataset exhibits integrity with no duplicate entries identified.

The intriguing observation of outliers in the `TIME` distribution, particularly those extending beyond 12.5 years, warrants careful scrutiny. Understanding the clinical significance of these outliers becomes paramount, as it directly influences the refinement of our predictive model.

Unbalanced Data Consideration

A notable aspect arising from our data examination is the potential imbalance in the dataset, primarily due to the distribution of patient survivability. The concentration of patients within a certain timeframe and the presence of outliers may contribute to this imbalance. Addressing data imbalance is critical for optimizing the performance of predictive algorithms. Techniques such as resampling or employing algorithms robust to imbalanced datasets may be considered to enhance the predictive accuracy. However we will need to try two different approaches with balanced and imbalanced data to check wether this influe on the results or not. In fact because we are only interested in the 1 year survival rate, this might not interfer with the result produced.

In conclusion, this comprehensive data understanding phase has provided valuable insights into the dataset's structure, key features, and potential challenges. The next steps involve refining the dataset based on defined criteria, delving deeper into outlier analysis, and implementing strategies to handle data

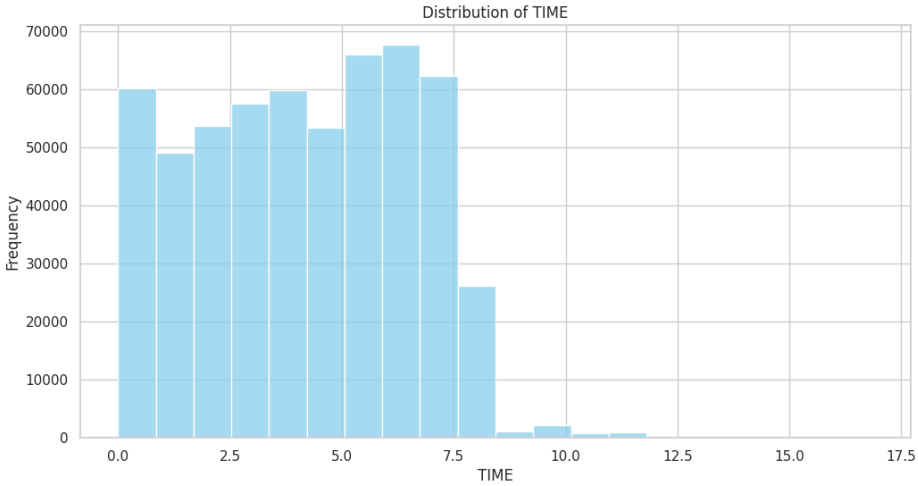

Figure 1: Distribution of TIME feature.

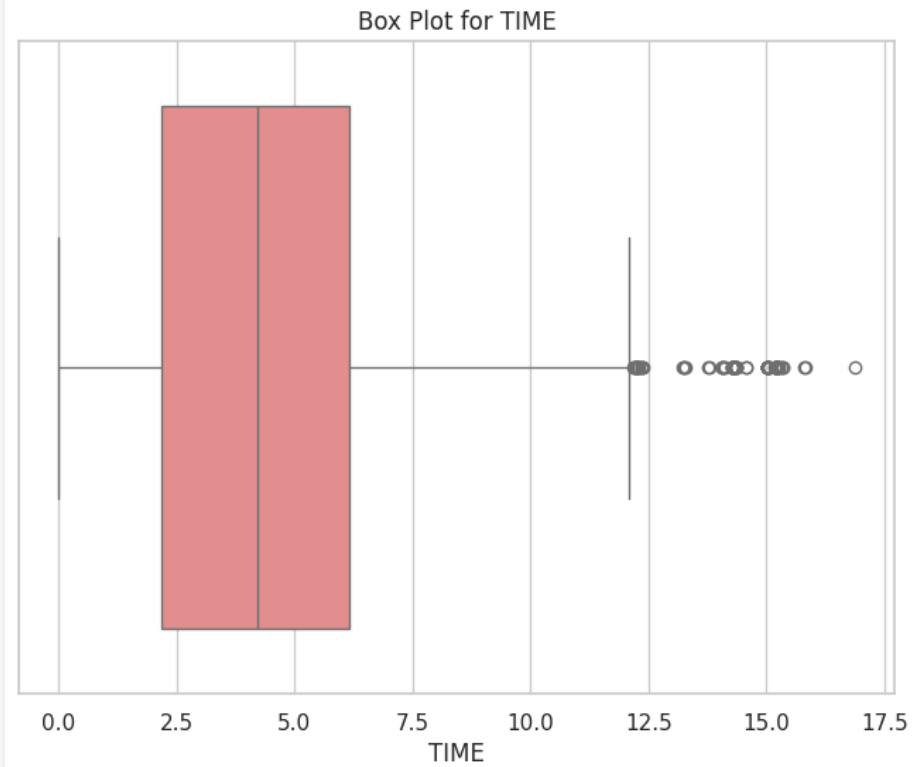

Figure 2: Box plot of TIME feature.

imbalance. These efforts are pivotal as we progress into subsequent stages of the CRISP-DM process, paving the way for effective data preparation and model development.

Project Plan

Week 1: December 1 - December 11

1. **Data Exploration and Understanding:**
   - **Task:** Load the dataset, examine its structure, and perform initial exploratory data analysis.
   - **Team:**
     - Henry Marie MONT: 4 hours
2. **Feature Engineering:**
   - **Task:** One-hot encode 'DEFINITION_ID' and handle time-related features.
   - **Team:**
     - Jean-Côme CARISSAN: 6 hours
3. **Data Cleaning:**
   - **Task:** Handle missing values, outliers, and ensure data integrity.
   - **Team:**
     - Henry Marie MONT: 5 hours
4. **Feature Selection:**
   - **Task:** Identify and remove less relevant or noisy features.
   - **Team:**
     - Jean-Côme CARISSAN: 6 hours
5. **Model Selection:**
   - **Task:** Start with logistic regression and random forest for baseline models.
   - **Team:**
     - Henry Marie MONT: 8 hours

6. **Model Evaluation:**

- **Task:** Split data, evaluate models using appropriate metrics.
- **Team:**
  - Jean-Côme CARISSAN: 6 hours

Week 2: December 12 - December 15 (Buffer for Iteration and Refinement)

7. **Addressing Class Imbalance:**

- **Task:** Implement techniques to handle class imbalance.
- **Team:**
  - Henry Marie MONT: 4 hours

8. **Cross-Validation:**

- **Task:** Implement cross-validation to ensure model consistency.
- **Team:**
  - Jean-Côme CARISSAN: 4 hours

9. **Interpretability:**

- **Task:** Explore model interpretability techniques.
- **Team:**
  - Henry Marie MONT: 6 hours

10. **Documentation and Finalization:**

- **Task:** Document the entire process, results, and recommendations.
- **Team:**
  - Jean-Côme CARISSAN: 8 hours

11. **Presentation Preparation:**

- **Task:** Prepare a presentation summarizing the project for stakeholders.
- **Team:**
  - Henry Marie MONT: 6 hours

Total Estimated Hours:

- Henry Marie MONT: 33 hours
- Jean-Côme CARISSAN: 30 hours

Notes:

- This plan provides some buffer time for unexpected challenges and allows for iterations and refinements in the second week.
- Regular check-ins between team members will be performed for coordination and problem-solving.
- Adjustments will be made based on progress and feedback received during the project.