Lung Cancer Survival Prediction

Group: B5
Project repository: https://github.com/malkmaria/IDS
Authors: Hugo Martin Teemus, Maria Malk, Robert Israel

Introduction

Lung cancer poses a significant health challenge globally, and survival rates hinge on various factors, especially the treatment path. This project focuses on predicting survival rates after the initial lung cancer diagnosis using synthetic medical data. Analyzing fictional patient treatment trajectories, our goal is to spot the key procedures, drugs, and different treatments that impact survival. Our aim is to develop a predictive tool to guide healthcare decisions for better patient care.

Our objective

Our objective is to enhance the precision of lung cancer survival prediction, measured through the ROC-AUC metric, by emphasizing pivotal features crucial for accurate forecasts and extracting key players that determine the results.

Methods used

Our flow for data processing used vectorization to encode features into one-hot vectors. Furthermore, we leveraged Principal Component Analysis (PCA) to refine the model's input space, presumably resulting in a more efficient representation of vital features and an overall enhancement in predictive accuracy.

Our approach

- Loading and preprocessing the patient data
- Applying PCA for key feature identification
- Employing ML models, based on top scorers from initial testing:
  - RandomForest
  - SVM
  - KNeighbors
  - Gradient Boosting
- Evaluating the effectiveness of models by ROC-AUC scores
- Visualizing feature distributions in PCA space
- Conducting correlation analysis for further insights
- Establishing a robust predictive modeling framework

Overview of the data

Three datasets in .csv format:

- Synthetic training dataset containing treatment trajectories of lung cancer patients.
- Synthetic validation dataset containing treatment trajectories of prostate cancer patients.
- Real dataset for testing the entire workflow. Dataset contains sensitive information, so the workflow testing was done by our instructor.

SUBJECT_ID	DEFINITION_ID	TIME
1	drug_217	0.00480733849108219
1	condition_1922	0.00864298398606479
1	condition_785	0.0277924723829692
1	drug_49	0.032514720763973
1	measurement_132	0.0567648616358154

SUBJECT_ID - Unique patient id.

DEFINITION_ID - Coded medical intervention that took place (drug, procedure, measurement, observation, condition).

TIME - time in years when the intervention took place.

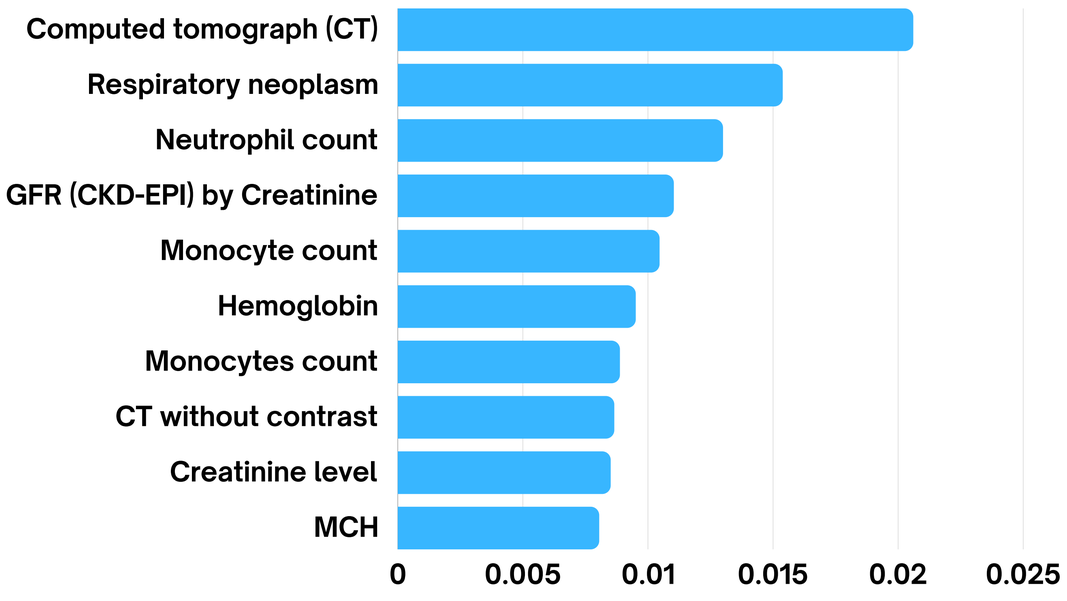
Model analysis

Best model (highest AUC-ROC score achieved) on the test data:

- Random Forest Classifier
- Exclusion of ‘drug_’ features
- No utilization of PCA
- ROC-AUC score: **0.874**

- Employing PCA appeared to raise ROC-AUC scores on synthetic data, while lowering scores when employed on real data.
- Highest ROC-AUC score achieved on synthetic data was **0.956**, using Gradient Boosting Classifier with PCA.

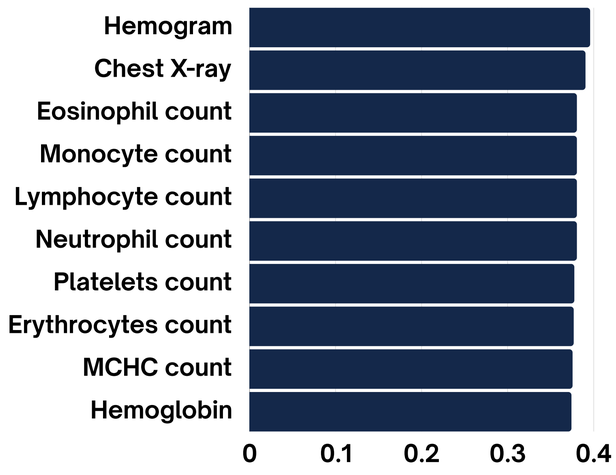
Top 10 features with the highest importance scores from the best model



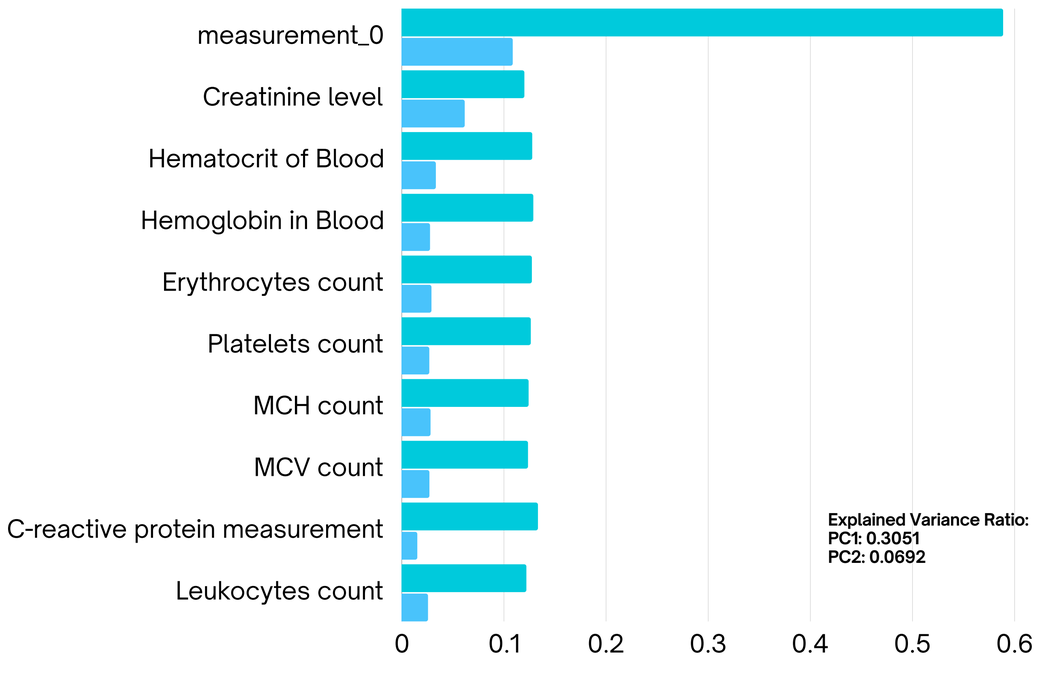
Data analysis

- Real data also contained categorized interventions with codes as ‘0’, which means the specific type is unknown. These data points seemed to be quite important, ranking high on model importance scores, having high correlation with dying and high variance values in PC components.

Top 10 correlations between death and medical interventions



Top 10 features with the highest variance scores from PC1 and PC2



Conclusion

The Random Forest Classifier stood out with an impressive AUC-ROC score of 0.874. Its performance was optimized by excluding 'drug_' features and avoiding PCA. Exploring features revealed key connections, providing valuable insights into how different elements correlate and influence patient outcomes. The comparison between synthetic and real data provided valuable lessons on navigating the complexities of healthcare datasets, guiding us toward more precise predictions.