Work Set 2 Report

Wojciech Lotko
Department of Physics and Astronomy and Computer Science,
Jagiellonian University, Cracow, Poland

(Dated: April 11, 2025)

Abstract

Below report describes medical data set analysis and linear model training. Features found in dataset were records of patients characteristics like age, gender, blood pressure as well as special parameters. Main point of this work was to find the best model that based on mentioned qualities could predict whether patient is eligible for treatment.

I. INTRODUCTION

This report details the development of a predictive model for medical treatment recommendations based on patient characteristics. The project aimed to identify which patient features most strongly influence treatment decisions and create an accurate classifier to predict appropriate medical interventions. The primary dataset contained both numerical variables (age, blood pressure, MeasureA, TestB) and categorical features (gender, blood_test, family_history, genetic markers), with treatment recommendations as the target variable.

The analysis began with comprehensive data exploration, identifying data quality issues including missing values and unrealistic blood pressure readings. Correlation analysis revealed significant relationships between numerical features and treatment outcomes, with age and blood pressure showing strong predictive power.

Multiple modeling approaches were implemented, starting with baseline logistic regression and progressing to more complex techniques. GridSearch was employed to optimize hyperparameters, while polynomial feature transformation captured non-linear relationships between variables. Recursive Feature Elimination with Cross-Validation (RFECV) identified the most impactful features, reducing model complexity while maintaining predictive performance. The final pipeline incorporated polynomial features (degree=3), standard scaling, and optimized logistic regression, achieving strong performance on validation and test datasets.

II. DATA EXPLORATION

Data exploration was based on plotting relationships between columns. Graphical representation of such data is the easiest solution for finding crucial points.

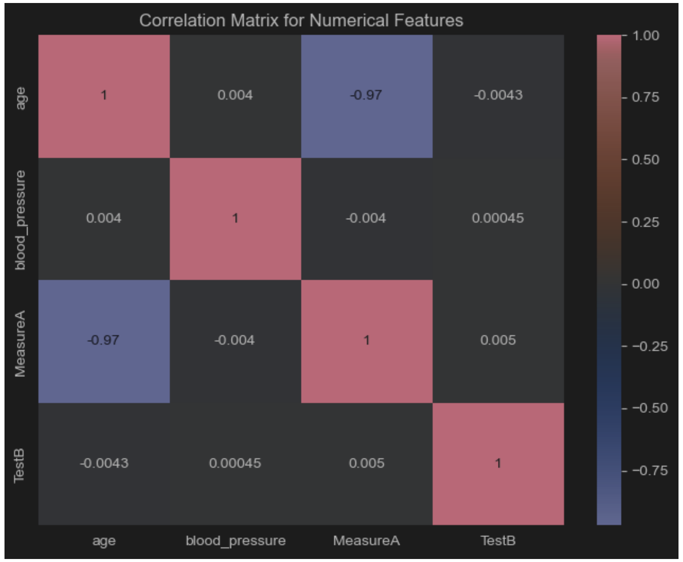

Figure 1 Correlation plot of numerical features

Correlation analysis highlighted no significant relationships between numerical features except negative one between blood_pressure and age. The feature importance analysis revealed that age and blood pressure had strong correlations with treatment recommendations.

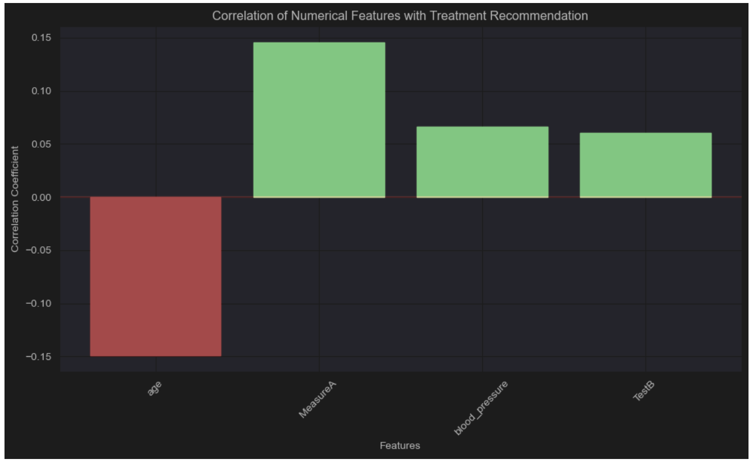

Figure 2 Correlation analysis between numerical features and treatment recommendation

Chi-square tests on categorical variables identified statistically significant associations between genetic markers (particularly GeneA, GeneB, and GeneC) and treatment decisions, suggesting these might be important diagnostic indicators.

Feature visualization by treatment groups showed distinct patterns, especially for blood_test results and family_history, which displayed clear separation between treatment groups. This exploration phase laid the groundwork for model building, highlighting which features most strongly influence medical treatment decisions in this dataset.

III. DATA PREPROCESSING

Data was pre-processed using simple pipeline encapsulated into one function. Missing values were identified in the dataset, and unrealistic blood pressure values (≤0) were filtered out during preprocessing. All categorical columns were one-hot encoded. Numerical values were well scaled, so there was no need to initially preprocess them.

IV. MODEL SELECTION AND FEATURE IMPORTANCE

Feature importance was measured using RFECV [1] on model whose hyperparameters were selected by GridSearchCV [2] method. This way it was determined what parameters are worth gathering and have the most impact during training.

Model was selected based on AUC score metric. After each training phase models were checked how well they predict on validation dataset. Based on given results AUC and ROC plot was drawn. This way it was easily interpretable how well it performs.

[1] Recursive feature elimination with cross-validation [https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html]

[2] GridSearchCV [https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html]

Starting from simple Logistic Regression next models were fine-tuned. Starting from GreadSearchCV to find best hyperparameters for given task. Additionally, PolynolmialFeatures [3] were introduced (with feature scaling afterwards) to fine-tune model performance and it was had the most significant impact.

V. MODEL EVALUATION

After training LogisticRegression [4]models each was evaluated on validation sets. Metric used to determine which fits the job the best was AUC score – the highest the better model performance. Below you can find plots presenting Confusion Matrixes (Figure 3,5,7) and ROC diagrams (Figure 4,6,8) calculated based on predictions on validation sets.

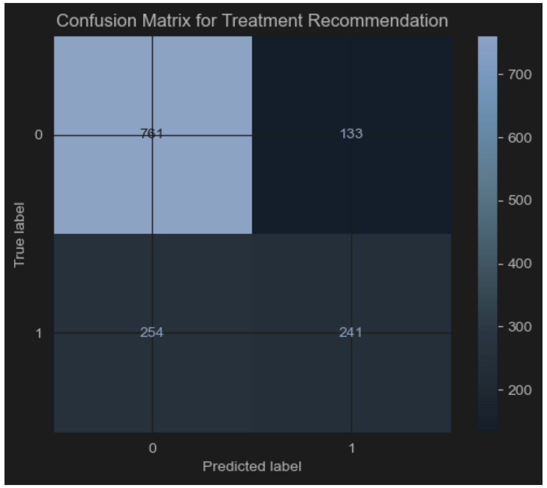

Figure 3 Base Logistic Regression Confusion Matrix

[3] PolynolmialFeatures [https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html]

[4] LogisticRegression [https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html]

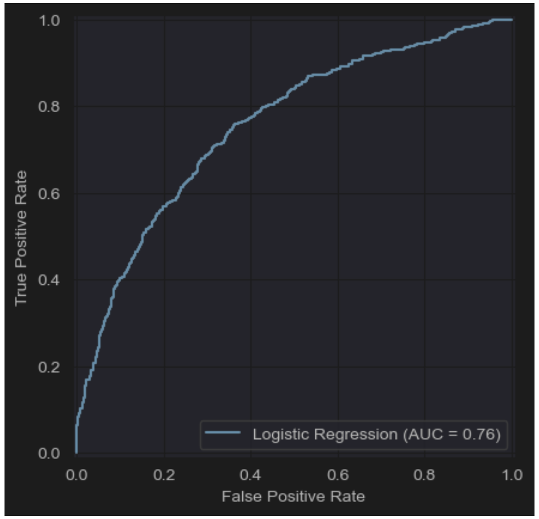

Figure 4 Base Logistic Regression ROC plot

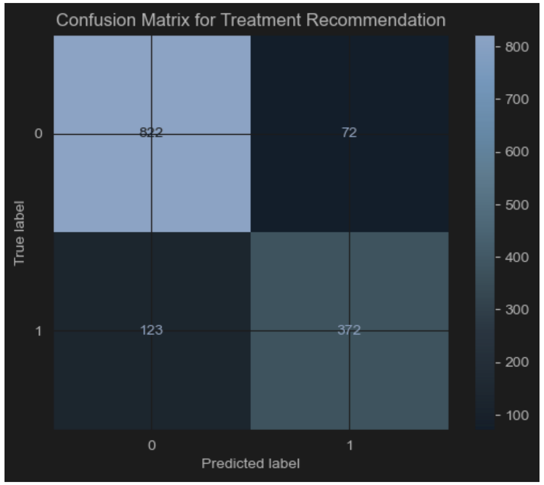

Figure 5 Logistic Regresion with PolynolmialFeatures(2) Confusion Matrix

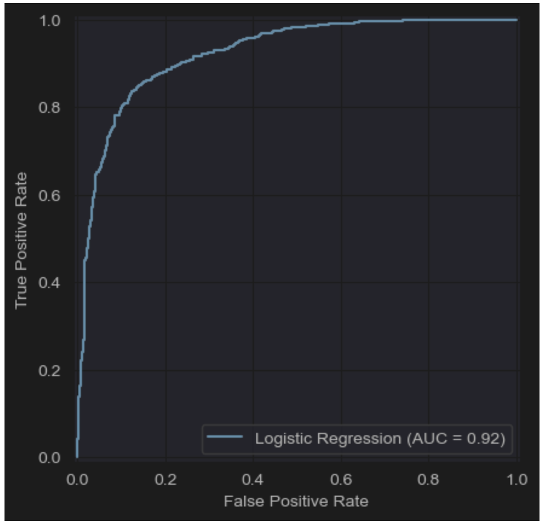

Figure 6 Logistic Regression with PolynolmialFeatures(2) ROC plot

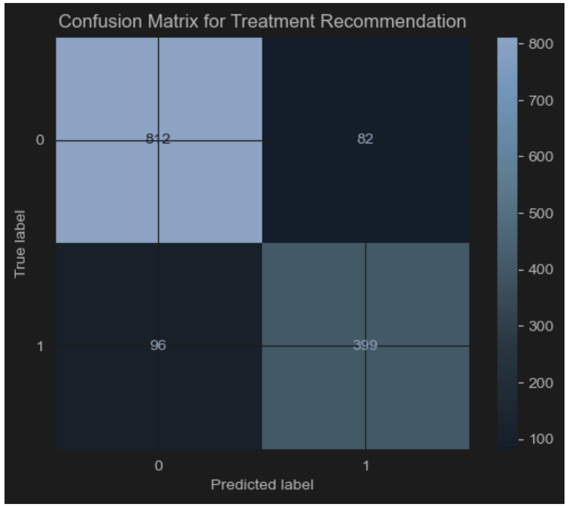

Figure 7 Logistic Regresion with PolynolmialFeatures(3) Confusion Matrix

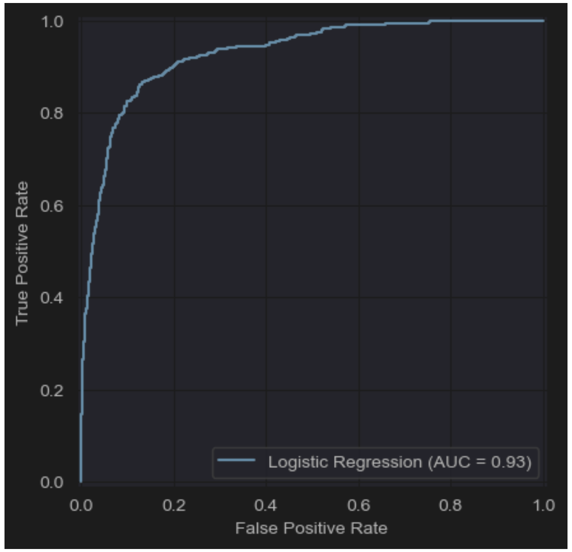

Figure 8 Logistic Regression with PolynolmialFeatures(3) ROC plot

You can see that polynomial features are significantly increasing model performance.

VI. INTERPRETATION

Data quality analysis identified several issues requiring preprocessing. We found missing values that needed handling and unrealistic blood pressure values (≤0) that were filtered out. The target variable showed an imbalanced distribution, which influenced our choice of evaluation metrics, prioritizing ROC-AUC over simple accuracy.

Correlation analysis highlighted significant relationships between numerical features and treatment outcomes. Age and blood pressure demonstrated strong correlations with treatment recommendations, providing initial indicators of clinical relevance.

Feature visualization by treatment groups revealed distinct separation patterns, especially for blood_test results and family_history variables.

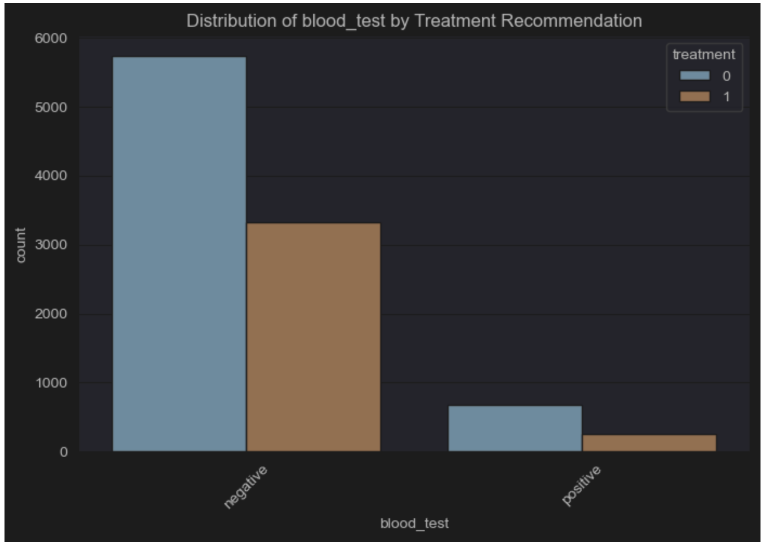

Figure 9 Plot of distribution of blood_test feature by treatment recommendation value

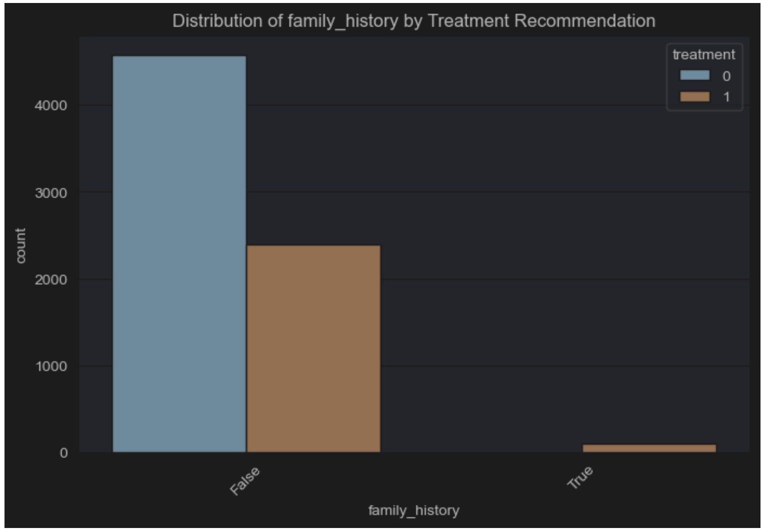

Figure 10 Plot of distribution of family_histor feature by treatment recommendation value

CONCLUSIONS

The RFECV feature selection process identified an optimal subset of features that maintained predictive power while reducing model complexity. Three topmost important features for model training are: “age”, “blood_pressure” and “TestB”, and the fourth “gender”. This may indicate that gathering expensive data “MeasureA”, GenA,B,C” may be unnecessary and has little importance for model predictions.