Projectreport for Machine Learning 2

Maluna Menke, Ari (Sara) Wahl, Pavlo Kravets

2023-12-05

Contents

1. Introduction

2. The Dataset

"The *Youth Risk Behavior Survey (YRBS)* measures health-related behaviors and experiences that can lead to death and disability among youth and adults.[...] Some of the health-related behaviors and experiences monitored are: - Student demographics: sex, sexual identity, race and ethnicity, and grade - Youth health behaviors and conditions: sexual, injury and violence, bullying, diet and physical activity, obesity, and mental health, including suicide - Substance use behaviors: electronic vapor product and tobacco product use, alcohol use, and other drug use - Student experiences: parental monitoring, school connectedness, unstable housing, and exposure to community violence [1]. It is a national survey conducted by CDC (Center for Disease Control and Prevention) and includes high school students from both private and public schools within the U.S. Data is collected from 1991 through 2021, we are only using the most recent data from 2021 though.

2.1 Preprocessing of the dataset

First, the variables were named instead of encoded with the questions number.

Then, the categorical variables are factorized. For ordninary variables, the factors are ordered as well.

To preprocess the dataset, we first ran a summary on our dataset. The number of NAs seems to depend very much on the question. The variable "orig_rec" only contained NAs and has therefore been removed, as well as the variable "site" which only contained "XX" entries. Variables q4 and q5 are already aggregated in "raceeth" and have also been deleted. The variable "record" seems to be an ID for the observations. This has to be considered later.

2.2 Simple Synopsis of the Dataset

number of observations: number of variables:

datatypes: - nominal variables: - ordinal variables: (numeric variables:) - discrete variables: - continuous variables:

2.3 Data Imputation

2.4 Reducing and balancing the dataset to 2000 observations

2.5 Target Variable

As a target variable, we decided to calculate a score from 5 questions that reflects the suicide risk of the person (observation) in question. This score is aggregated with a rule based approach according to the accompanying Data User Guide.[2].

3. Splitting the Dataset

According to the project requirements we split our data in 60% Training, 20% Validation and 20% Testing Data.

4. Machine Learning Models

4.1 Short Mathematical Overview on the used Methods

4.2 Fitting process

4.3 Hyperparameter Optimization

Mithilfe eines Likelihood-Ratio Tests wird überprüft welches Modell das bessere ist.

Das optimierte Modell ist signifikant besser als das erarbeitete zum festgelegten Signifikanzniveau, die Nullhypothese des LR-Tests, dass das optimierte Modell nicht besser ist als das komplexere, kann verworfen werden. Wir stellen außerdem fest: Beide Modelle weisen einen signifikanten Erklärungsgehalt auf, sie sind besser als das Nullmodell.

5. Comparison of the Models / Model's Performance on Test Data

5.1 Quantitative

5.1.1 Confusion Matrix

5.1.2 Accuracy

5.1.2 Precision, Recall and F1-Score

5.2 Qualitative

5.3 Overfitting Check

6. Visual Representation

7. Final Discussion

8. References

[1] https://www.cdc.gov/healthyyouth/data/yrbs/overview.htm

[2] https:
//www.cdc.gov/healthyyouth/data/yrbs/pdf/2021/2021_YRBS_Data_Users_Guide_508.pdf