

# Sick dataset analysis

Wojciech Bogucki

16/04/2020

## Contents

|                                  |           |
|----------------------------------|-----------|
| <b>Explanatory Data Analysis</b> | <b>2</b>  |
| First look at raw data . . . . . | 2         |
| Mistakes in data . . . . .       | 3         |
| Missing data . . . . .           | 4         |
| Imputation . . . . .             | 6         |
| <b>Creating models</b>           | <b>9</b>  |
| Training . . . . .               | 9         |
| Measures . . . . .               | 9         |
| <b>Conclusion</b>                | <b>11</b> |

# Explanatory Data Analysis

## First look at raw data

Firstly, let's have a look at selected training data from the original dataset *sick*:

Table 1: Features of dataset sick

|                      |       |
|----------------------|-------|
| rows                 | 3016  |
| columns              | 30    |
| discrete_columns     | 23    |
| continuous_columns   | 6     |
| all_missing_columns  | 1     |
| total_missing_values | 4878  |
| complete_rows        | 0     |
| total_observations   | 90480 |

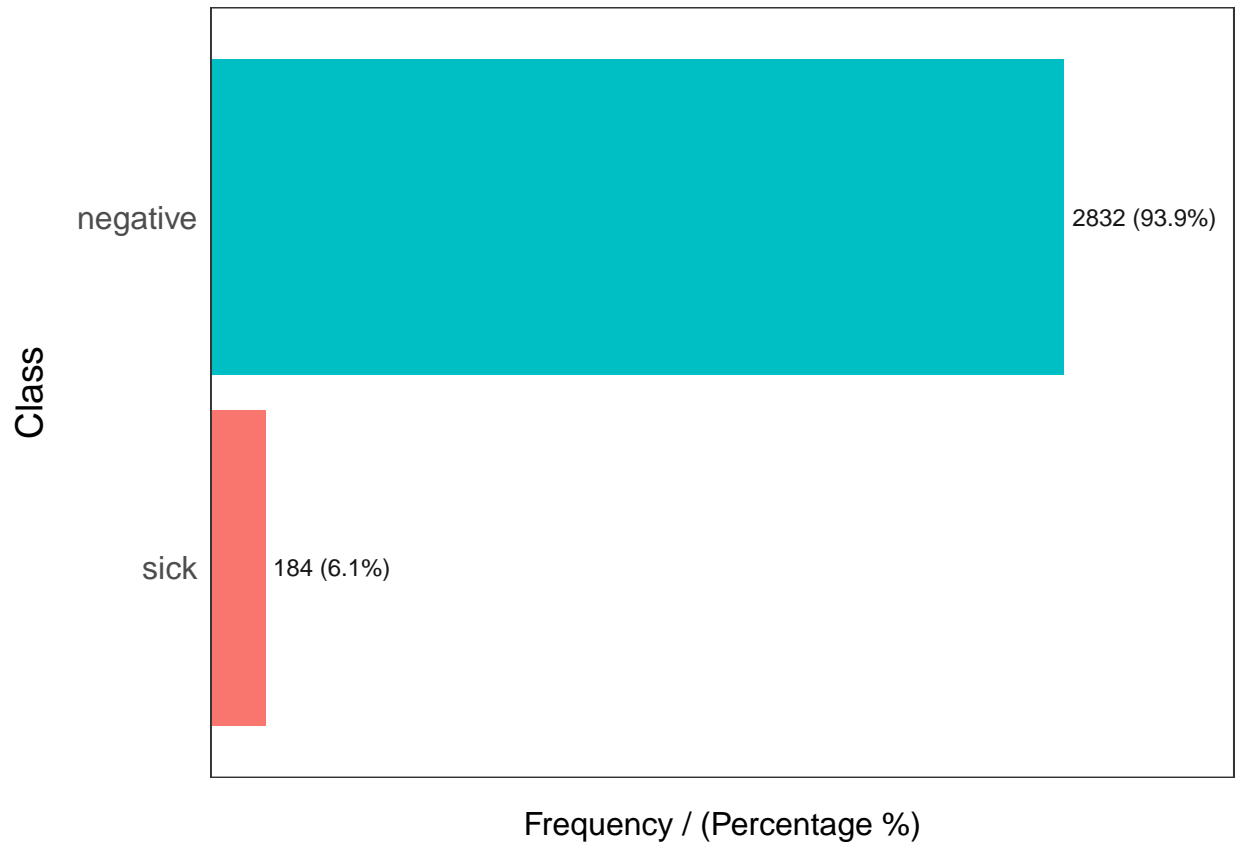
Table 2: Metrics for each variable of dataset sick

| variable                  | q_zeros | p_zeros | q_na | p_na   | q_inf | p_inf | type    | unique |
|---------------------------|---------|---------|------|--------|-------|-------|---------|--------|
| age                       | 0       | 0       | 0    | 0.00   | 0     | 0     | numeric | 88     |
| sex                       | 0       | 0       | 119  | 3.95   | 0     | 0     | factor  | 2      |
| on_thyroxine              | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| query_on_thyroxine        | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| on_antithyroid_medication | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| sick                      | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| pregnant                  | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| thyroid_surgery           | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| I131_treatment            | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| query_hypothyroid         | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| query_hyperthyroid        | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| lithium                   | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| goitre                    | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| tumor                     | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| hypopituitary             | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 1      |
| psych                     | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| TSH_measured              | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| TSH                       | 0       | 0       | 301  | 9.98   | 0     | 0     | numeric | 272    |
| T3_measured               | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| T3                        | 0       | 0       | 621  | 20.59  | 0     | 0     | numeric | 67     |
| TT4_measured              | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| TT4                       | 0       | 0       | 191  | 6.33   | 0     | 0     | numeric | 222    |
| T4U_measured              | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| T4U                       | 0       | 0       | 316  | 10.48  | 0     | 0     | numeric | 140    |
| FTI_measured              | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |
| FTI                       | 0       | 0       | 314  | 10.41  | 0     | 0     | numeric | 217    |
| TBG_measured              | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 1      |
| TBG                       | 0       | 0       | 3016 | 100.00 | 0     | 0     | numeric | 0      |
| referral_source           | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 5      |
| Class                     | 0       | 0       | 0    | 0.00   | 0     | 0     | factor  | 2      |

As we can see in the table, variable TBG has 100% missing values and variables TBG\_measured and hypopituitary have only one unique value. Therefore, I removed this three variables from training and test datasets.

Our target class 'Class' is very imbalanced.

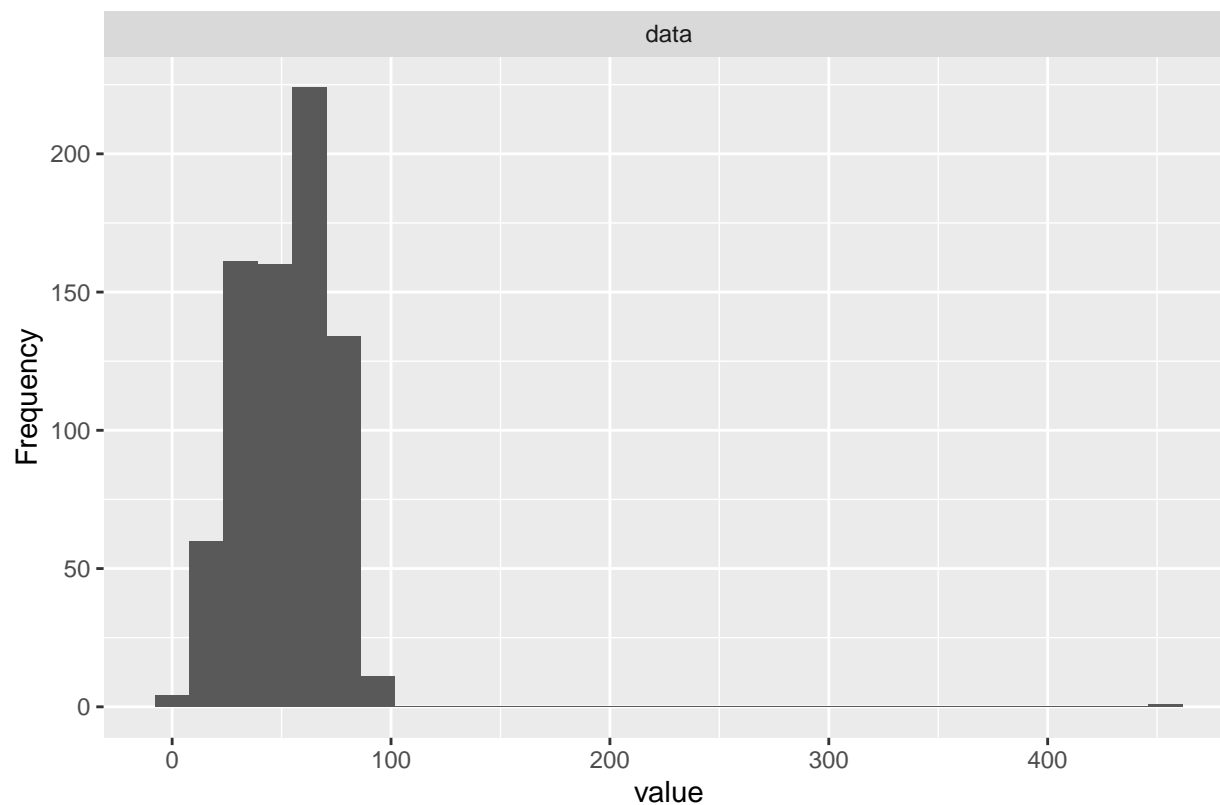
```
freq(sick_train, input=c('Class'))
```



```
##      Class frequency percentage cumulative_perc
## 1 negative      2832         93.9           93.9
## 2    sick       184          6.1          100.0
```

## Mistakes in data

When we look closely at the continuous variables in test dataset, we can observe that age variable has some big values.



```
## sick[, "age"]
##      n missing distinct      Info      Mean      Gmd      .05      .10
##  3771      1       93        1    51.74    21.98     20     25
##    .25    .50    .75    .90    .95
##    36    54    67    75    79
##
## lowest :   1   2   4   5   6, highest:  91  92  93  94 455
```

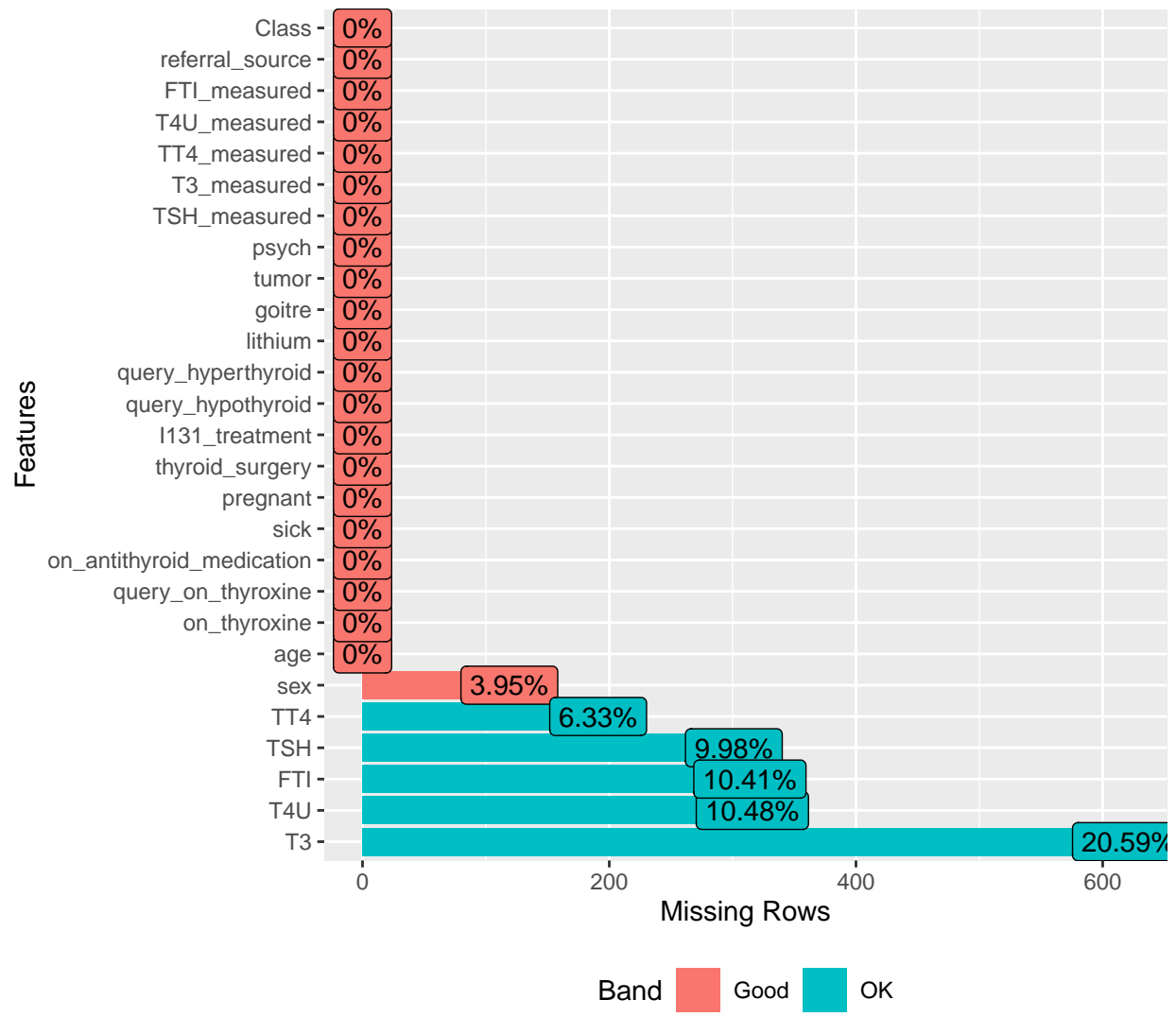
Age for one observation is 455, which is obviously a mistake. To avoid such mistakes data will be filtrated. Impossible age values will be changed to missing values.

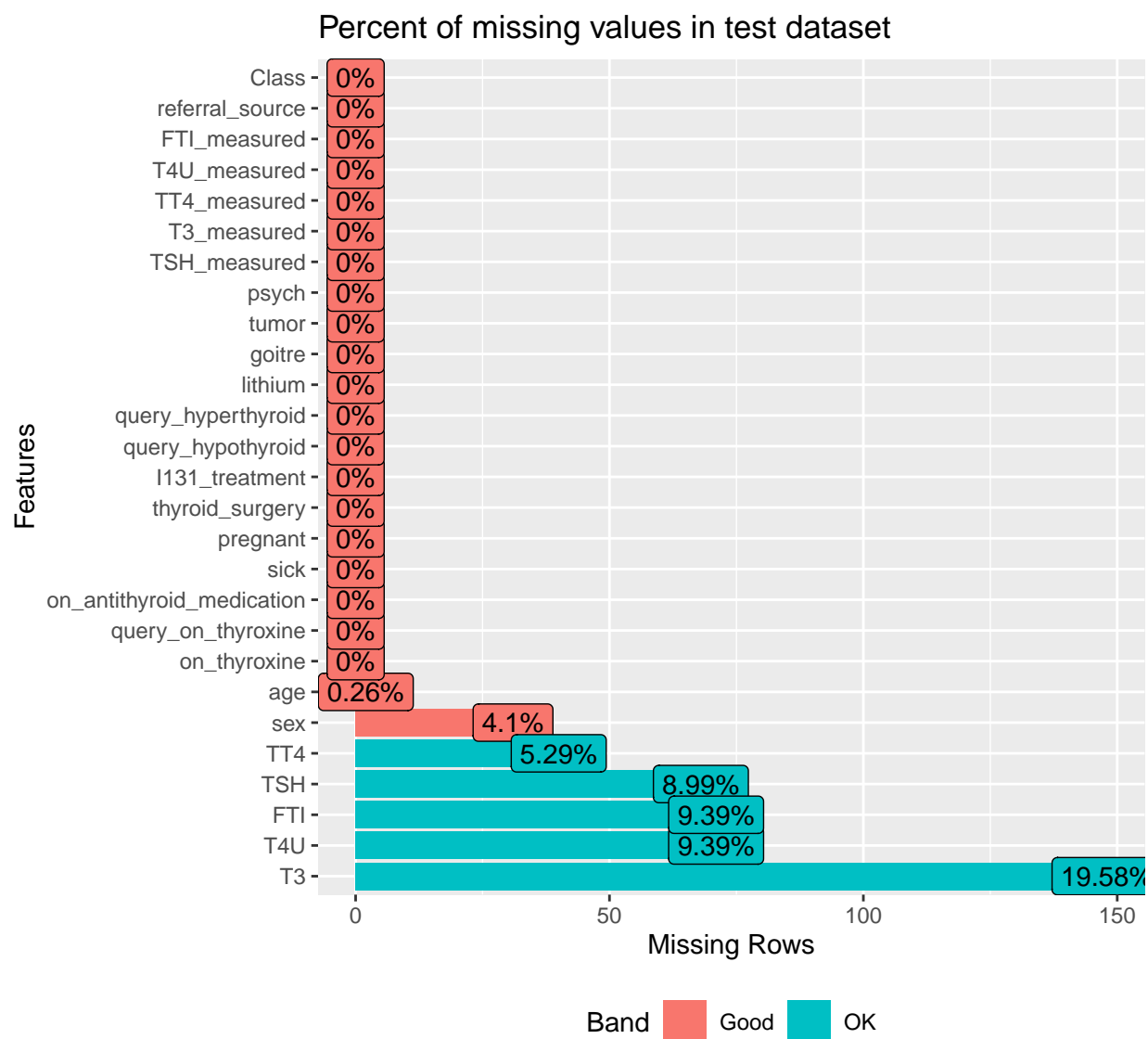
```
sick_train <- sick_train %>% mutate(age=replace(age, age>130 | age<0, NA))
sick_test  <- sick_test  %>% mutate(age=replace(age, age>130 | age<0, NA))
```

## Missing data

Dataset contains some missing values.

Percent of missing values in training dataset



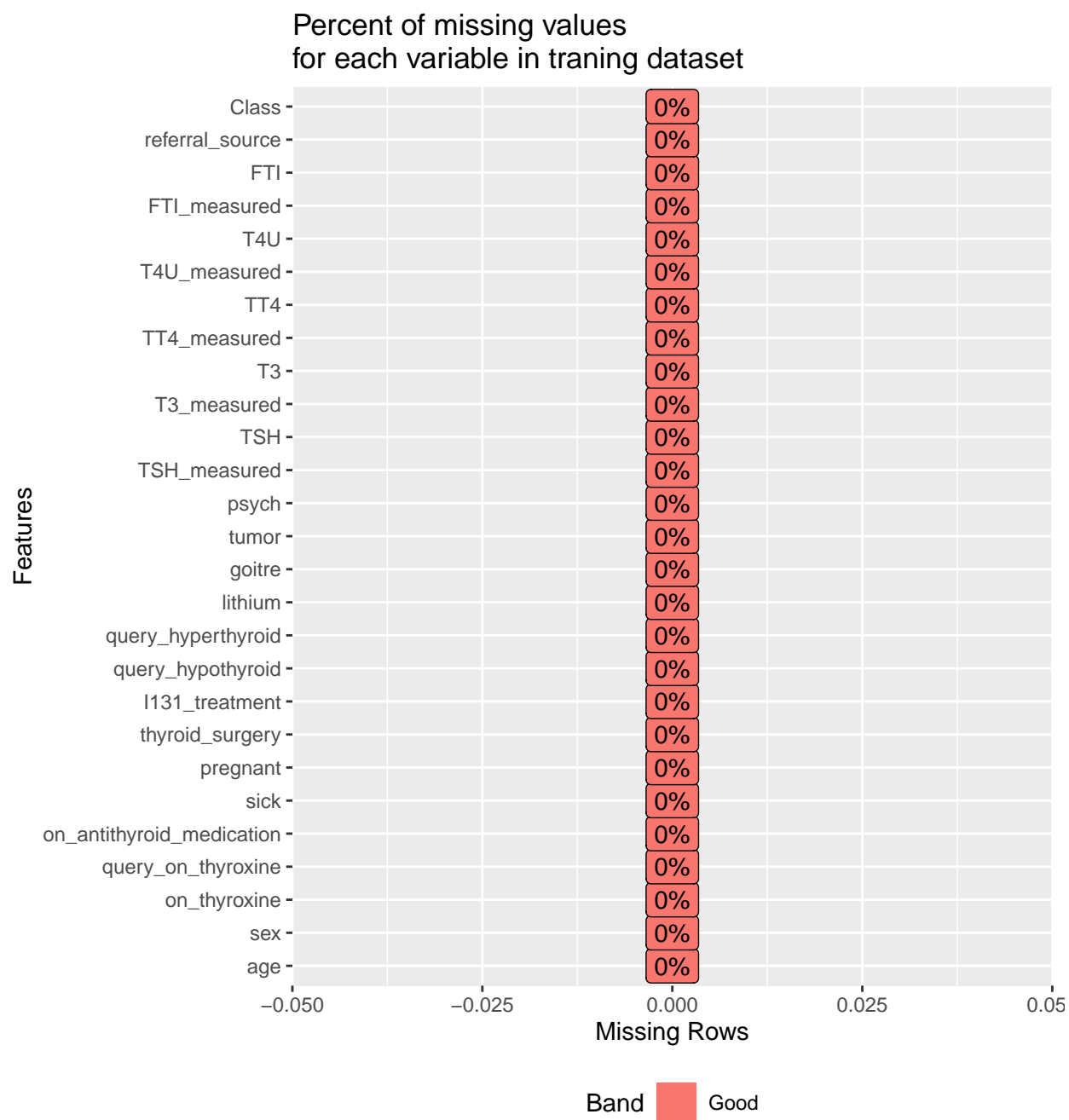


## Imputation

To eliminate missing data in training set I used package `mice`.

Table 3: Imputation method for each variable

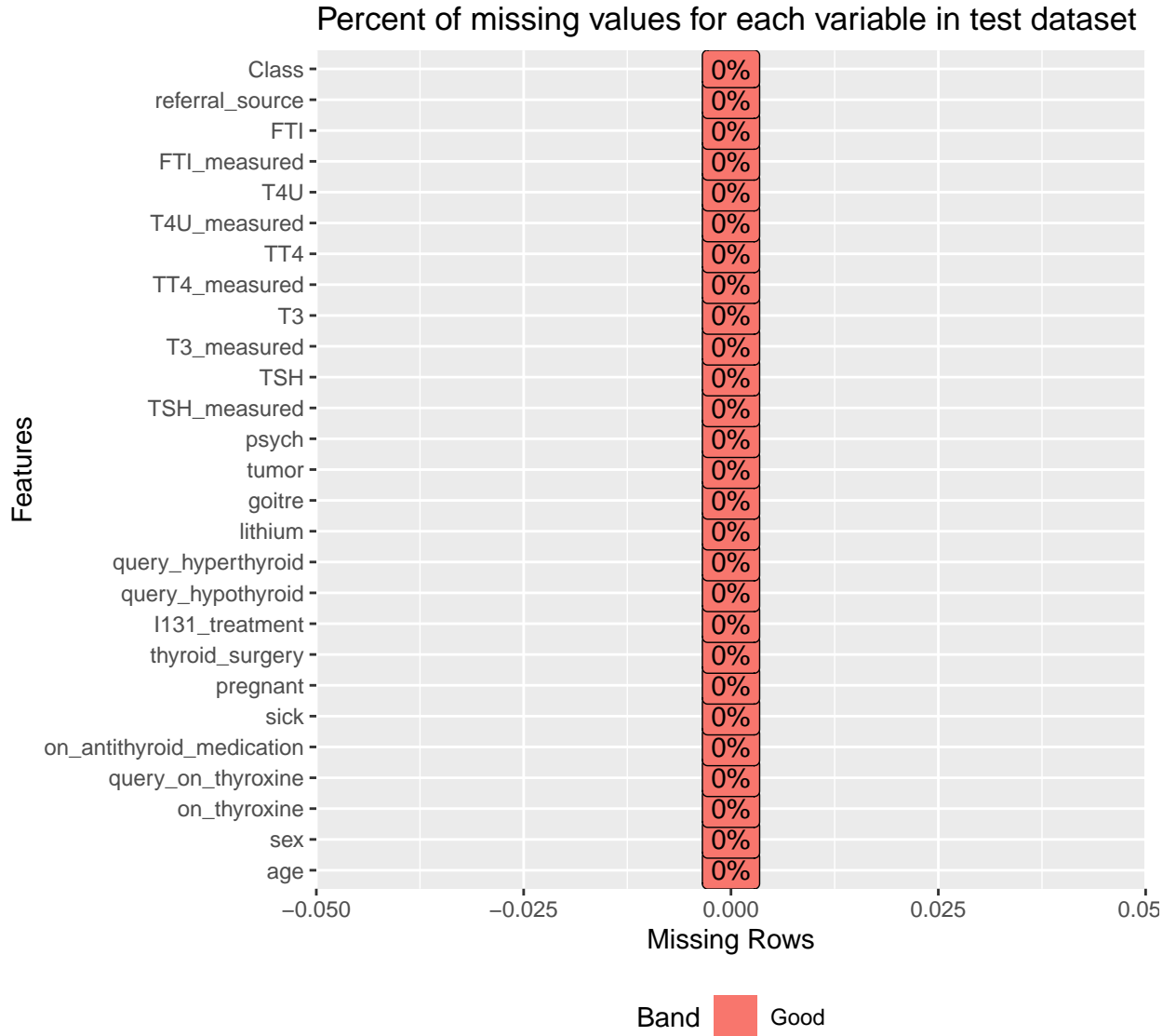
| variable | imputaton method         |
|----------|--------------------------|
| sex      | Logistic regression      |
| TSH      | Predictive mean matching |
| T3       | Predictive mean matching |
| TT4      | Predictive mean matching |
| T4U      | Predictive mean matching |
| FTI      | Predictive mean matching |



Then I performed imputation on dataset containing training data after imputation and test data.

Table 4: Imputation method for each variable

| variable | imputaton method         |
|----------|--------------------------|
| age      | Predictive mean matching |
| sex      | Logistic regression      |
| TSH      | Predictive mean matching |
| T3       | Predictive mean matching |
| TT4      | Predictive mean matching |
| T4U      | Predictive mean matching |
| FTI      | Predictive mean matching |





## Creating models

For prediction I chose three interpretable models:

- logistic regression
- decision tree
- naive Bayes

## Training

During training I performed hyperparameters tuning on decision tree. These hyperparameteres were used:

- Decision tree:

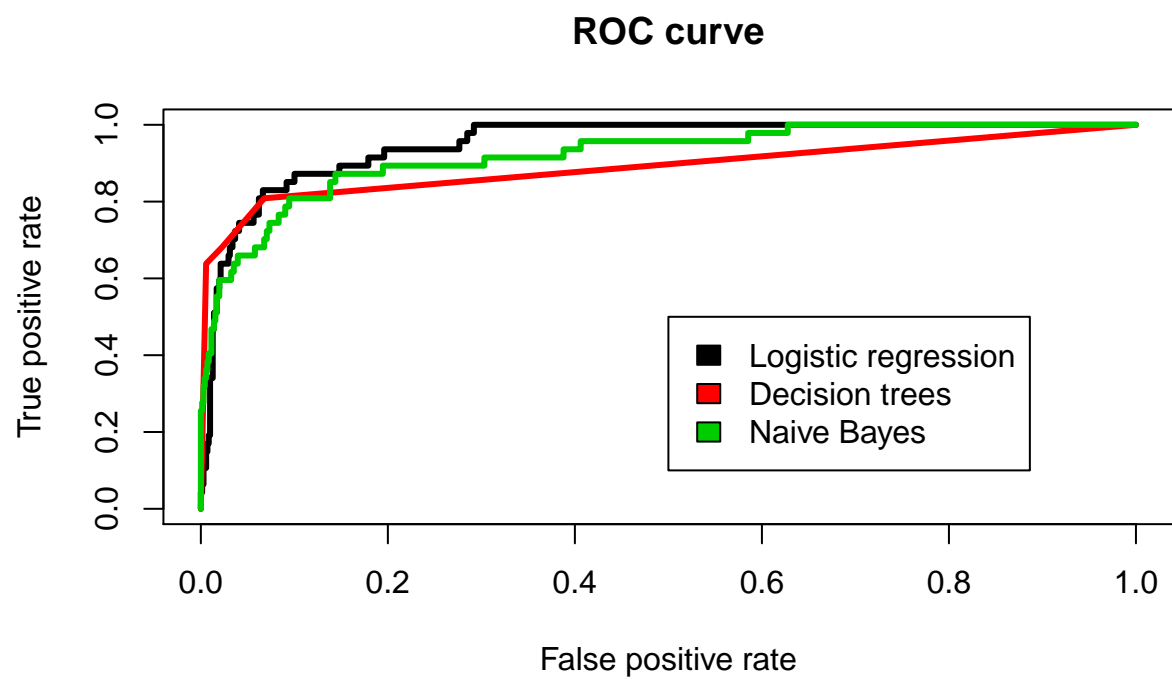
```
## $minsplit
## [1] 8
##
## $minbucket
## [1] 18
##
## $cp
## [1] 0.0335254
```

## Measures

Next, I calculated measures of goodness of prediction: AUC and AUPRC. I also created a comparison of ROC curve and precision-recall plot for each model.

Table 5: Measures of goodness of prediction for each model

| model               | auc       | auprc     |
|---------------------|-----------|-----------|
| Logistic regression | 0.9509048 | 0.5493493 |
| Decision trees      | 0.8896108 | 0.6598853 |
| Naive Bayes         | 0.9204453 | 0.6175951 |



## Conclusion

Even though decision tree has smaller AUC than logistic regression and naive Bayes, it has significantly higher AUPRC, which is better measure, when target variable is imbalanced like in this example. Decision tree from package `rpart` is an example of an interpretable model. We can see the decision tree.

