NATIONAL RESEARCH
UNIVERSITY

Faculty of Economic Sciences,
NRU HSE

Prediction of student's performance

Orange Team:
Gorlevich Daniil
Chertsov Pavel
Achikyan Eduard
Andreichiva Polina
Zubkov Ivan

Main task

- To **predict a student's final grade** in Portuguese language based on the data of student achievement in secondary education of two Portuguese schools.

  Note: Other approaches use marks for 1st and 2nd-periods grades which are highly correlated with the final grade. This makes prediction easier, but less applicable.To make the task more interesting we dropped previous grades.

General information about data set

What the collected data is used for?

Data approach to predict student achievement in secondary education of two Portuguese schools

What kind of data is it?

The data attributes include student grades, demographic, social and school related features

How data was collected?

It was collected by using school reports and questionnaires

Data description

Data set contains:

- 33 features:
  - 16 features - int64
  - 17 features - object
- 649 observations

More detailed info:

- 13 binary variables
- 5 categorical variables
- 11 ordinal variables
- 1 numeric variable

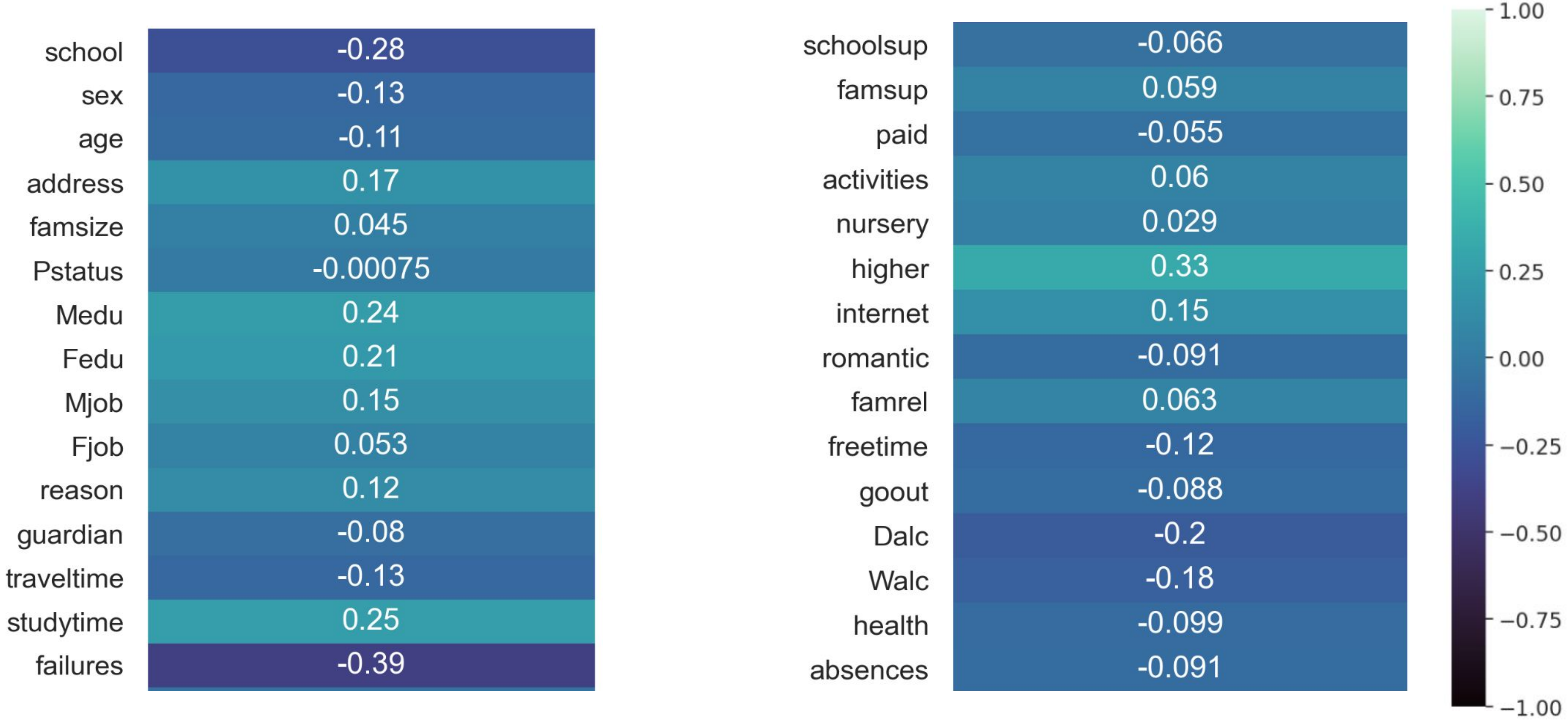
Data set preparations:

- no missing values
- 2 features were dropped because of multicollinearity

Exploratory Data Analysis

Target Histogram

Correlation of variables with target



Correlation of variables with target

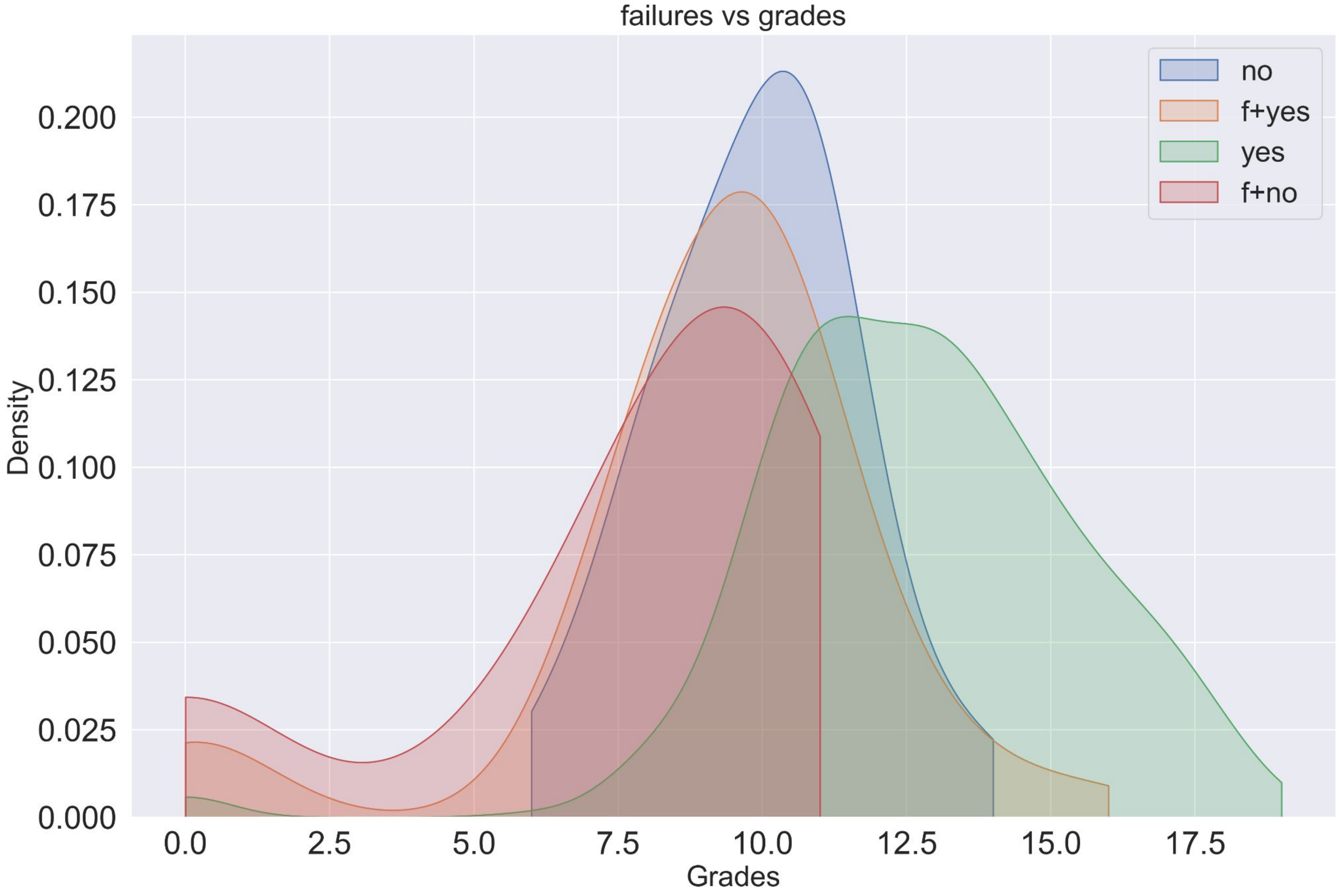
Final grade (G3) has a:

- Positive correlation with time of study a week (0.25)
- Positive correlation with desire to pursue higher education (0.33)
- Negative correlation with previous failures (-0.39)
- Negative correlation with school of education (-0.28) - MS school seems to be worse

Specific probability distribution

Probability distributions students' final grade with particular value of 'higher' and 'failures' variables':

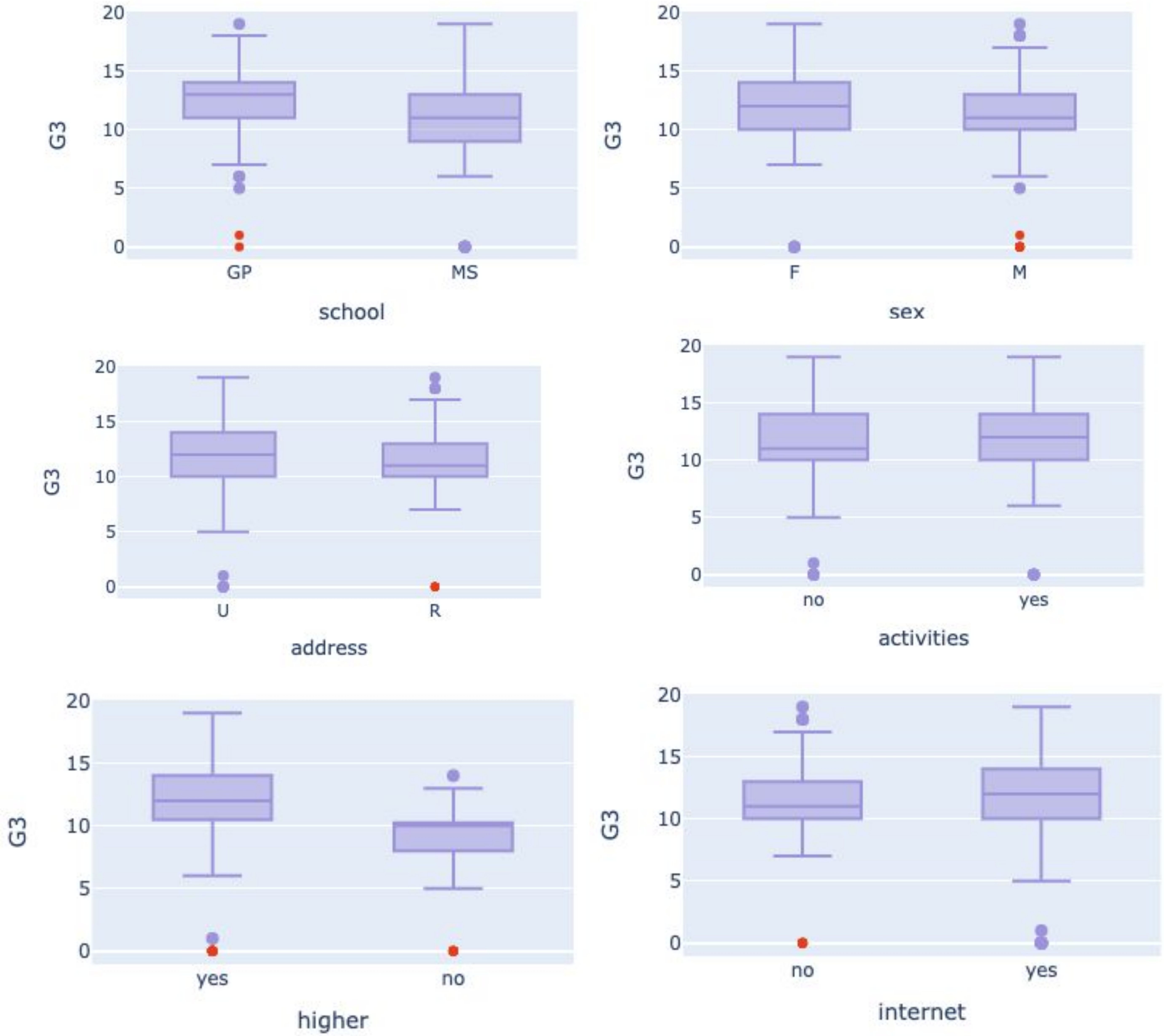
- no - doesn't want a higher education and has zero failures
- f+ no - doesn't want a higher education and has one or more failures
- yes - wants a higher education and has zero failures
- f+ yes - wants a higher education and has one or more failures



Data description. Box-plots. Categorical columns.

Key insights

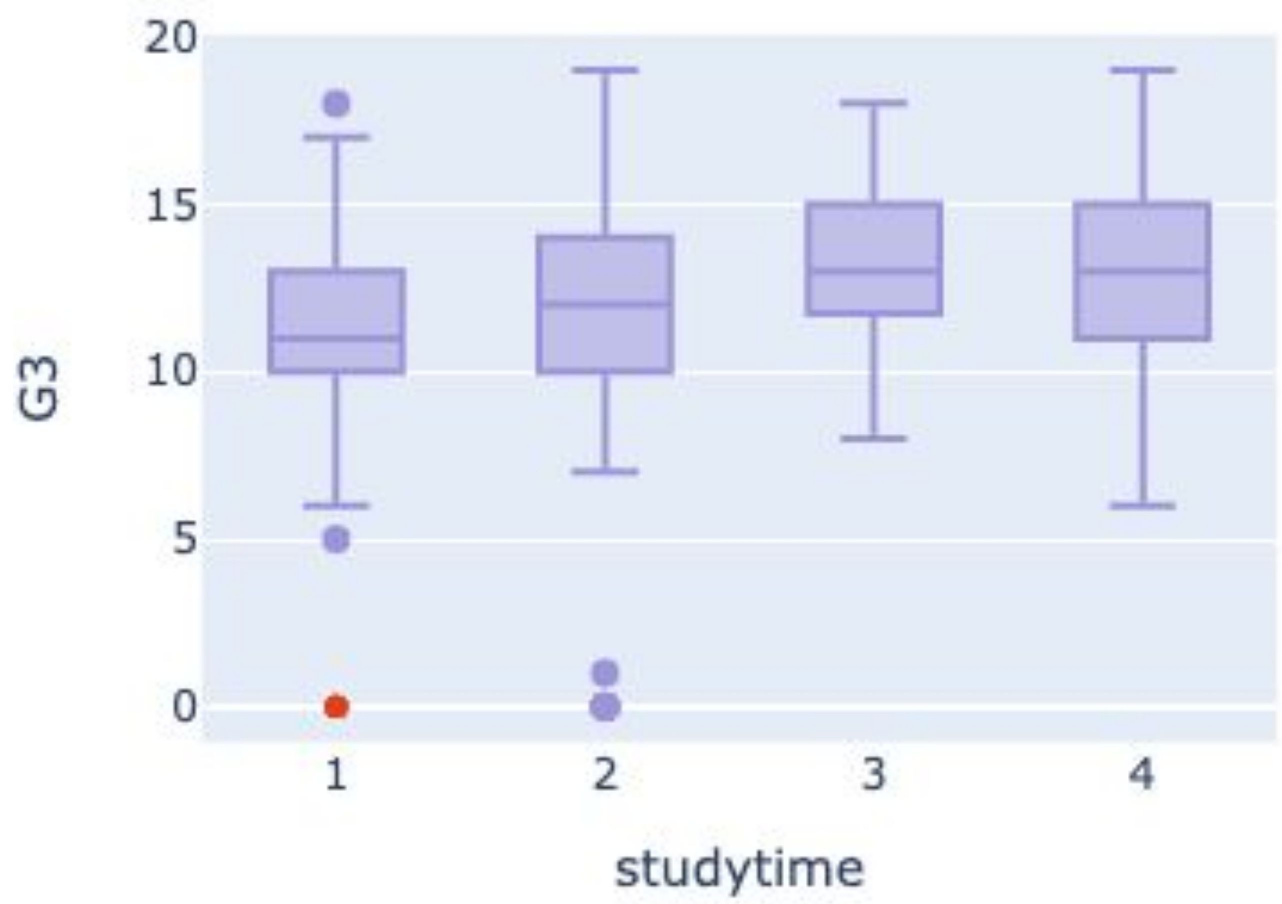
- Students from **Gabriel Pereira** school have more chances to get higher mark than their colleagues from Mousinho da Silveira.
- Median marks of **females** (12) are slightly higher than marks of males (11)
- Students from **urban area** tend to have moderately higher mark (12) than students from rural areas (11)
- The median mark of students, who have **extra-curricular activities**, is higher than median mark of those who do not.
- Students who want to pursue **higher education** in future have higher marks.

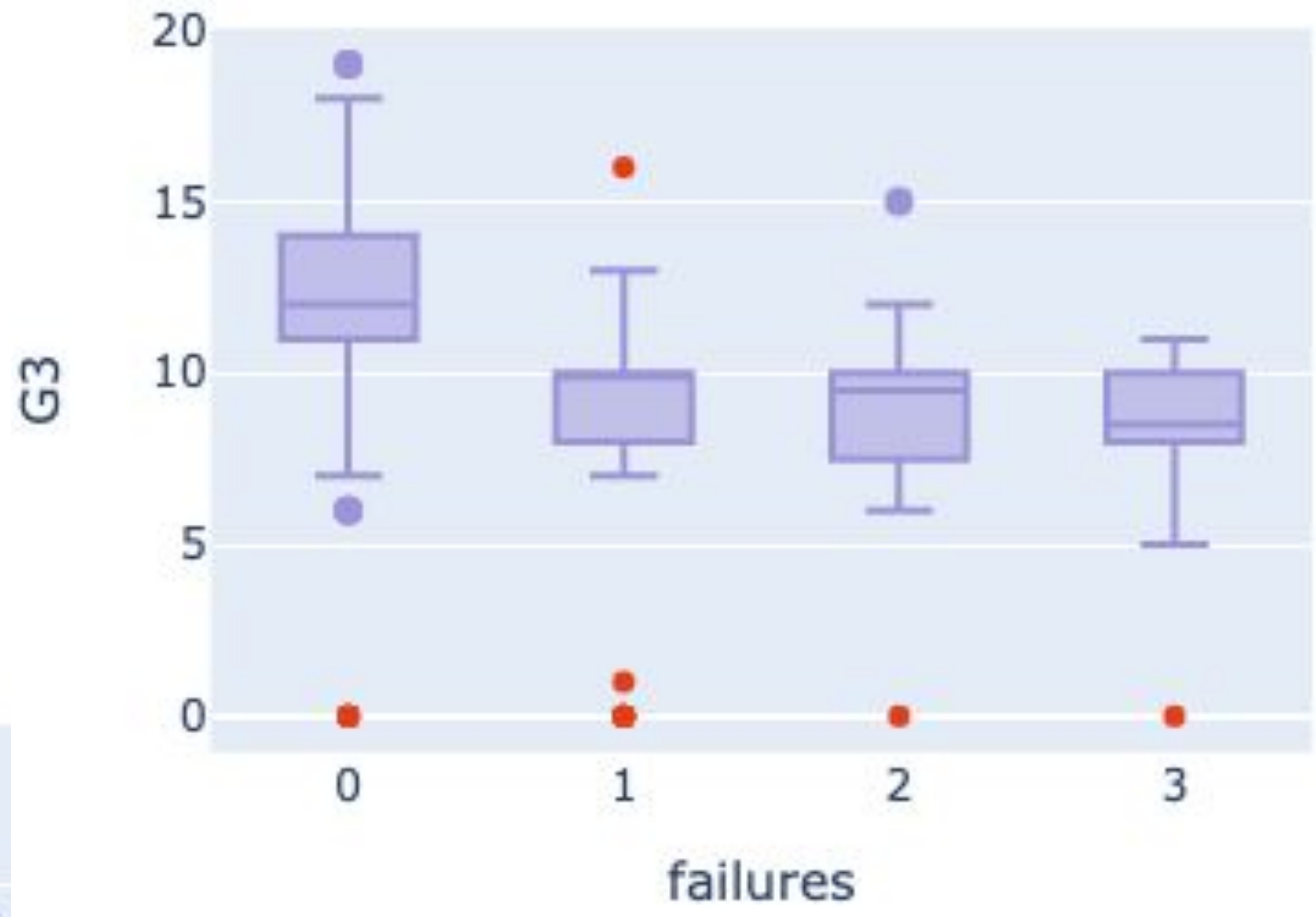


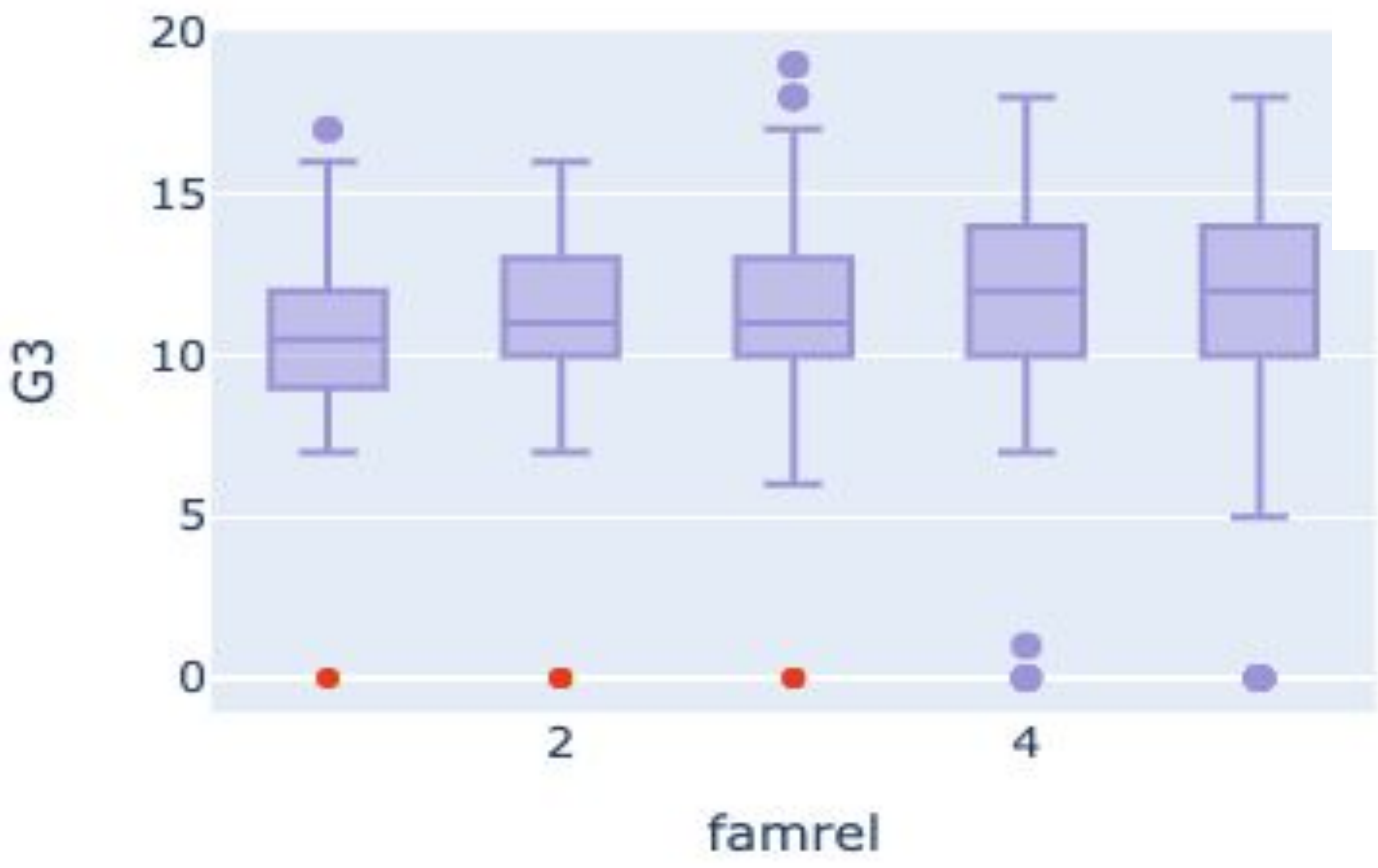
Data description. Box-plots. Numeric columns.

Key insights:

- Study time has a positive influence on grades, but the influence decrease, as the study time increases.
- Students, who have never failed, perform stronger, than other groups.
- Family relations have a positive linkage with grades.







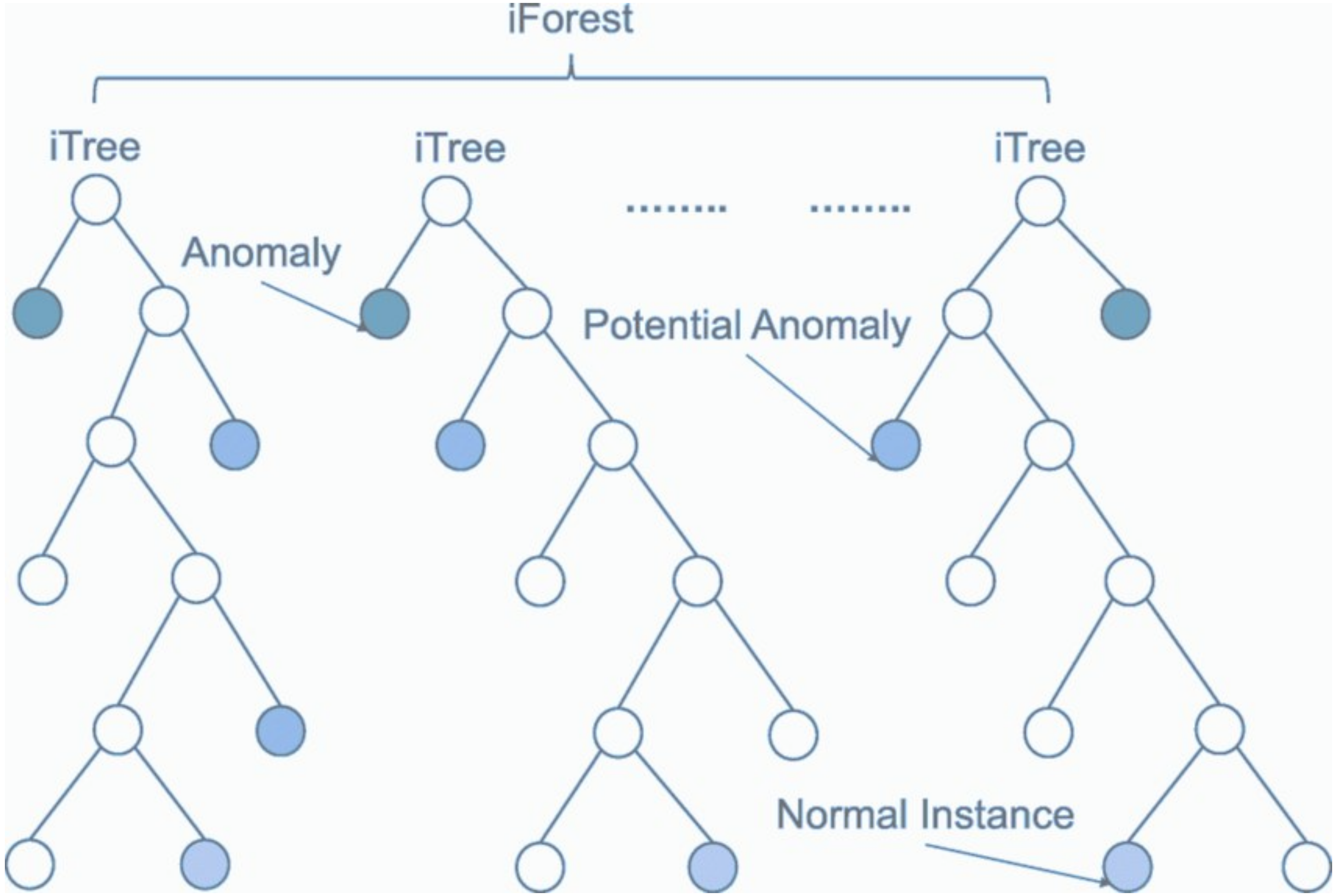
Outlier detection

We used **isolation forest** algorithm to remove the outliers

- It seemed to be the most adequate measure, since it marked 5% of the data as outlier, while classic methods marked 40%

- The method is based on the idea of building trees with random splits

- Points which are easily separated end up in the high leaves, while normal instances are found deep in the tree



Models

Broad model test

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	1.9390	6.8455	2.5853	0.2839	0.3366	0.1707	0.1430
catboost	CatBoost Regressor	1.9531	7.2192	2.6643	0.2341	0.3475	0.1723	0.9890
gbr	Gradient Boosting Regressor	2.0265	7.2018	2.6667	0.2265	0.3424	0.1795	0.0340
lar	Least Angle Regression	2.0122	7.4528	2.6972	0.2221	0.3520	0.1765	0.0150
br	Bayesian Ridge	1.9901	7.4329	2.6958	0.2209	0.3527	0.1755	0.0070
huber	Huber Regressor	2.0360	7.5929	2.7222	0.2061	0.3546	0.1769	0.0180
ridge	Ridge Regression	2.0434	7.5532	2.7185	0.2048	0.3528	0.1803	0.0060
lr	Linear Regression	2.0484	7.5766	2.7230	0.2020	0.3531	0.1808	0.1280
ada	AdaBoost Regressor	2.0732	7.6580	2.7321	0.1978	0.3509	0.1873	0.0700
omp	Orthogonal Matching Pursuit	2.0599	7.7553	2.7578	0.1840	0.3559	0.1837	0.0080
et	Extra Trees Regressor	2.2155	8.9082	2.9507	0.0598	0.3674	0.1964	0.1380
en	Elastic Net	2.2590	9.2543	3.0150	0.0331	0.3767	0.2012	0.0070
knn	K Neighbors Regressor	2.1608	9.1103	2.9937	0.0325	0.3753	0.1984	0.0100
lasso	Lasso Regression	2.3657	9.9326	3.1256	-0.0409	0.3833	0.2105	0.0070
llar	Lasso Least Angle Regression	2.3657	9.9329	3.1256	-0.0410	0.3833	0.2105	0.0080
par	Passive Aggressive Regressor	2.6173	12.0295	3.3514	-0.3183	0.3910	0.2410	0.0070
dt	Decision Tree Regressor	2.8968	15.2274	3.8830	-0.6664	0.5727	0.2597	0.0110

Decision trees

Algorithm

This algorithm builds a tree of data split on optimal conditions. The main idea:

- Greedy splitting data into nodes, which optimize the given criterion
- Previous nodes does not change on each step

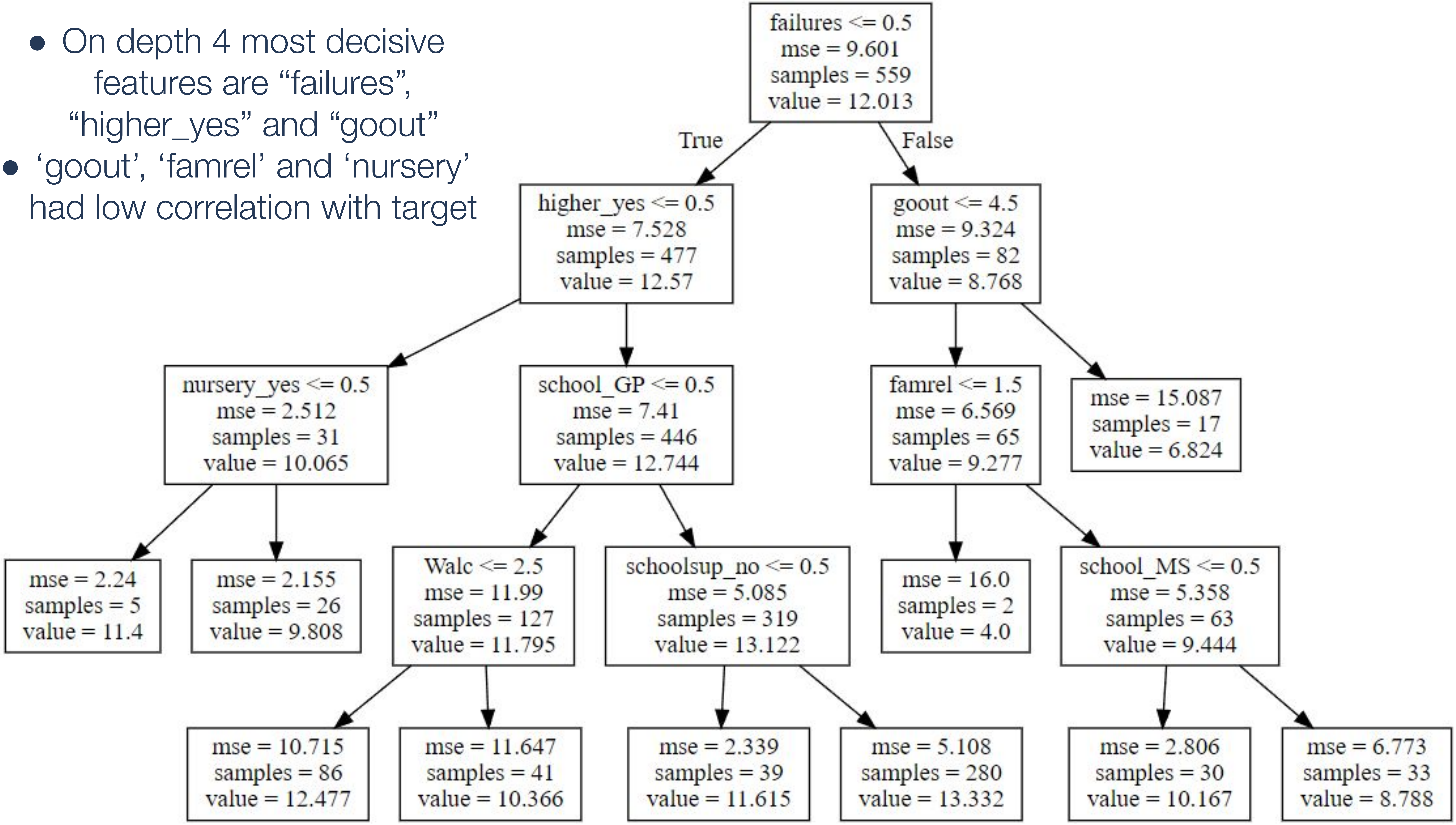
CV results:

- 'max_depth': 4, 'min_samples_split': 30
- depth of the tree is 4
- 16 leaf nodes = 'max_depth'^2
- minimal value for split is 30

Quality of estimator

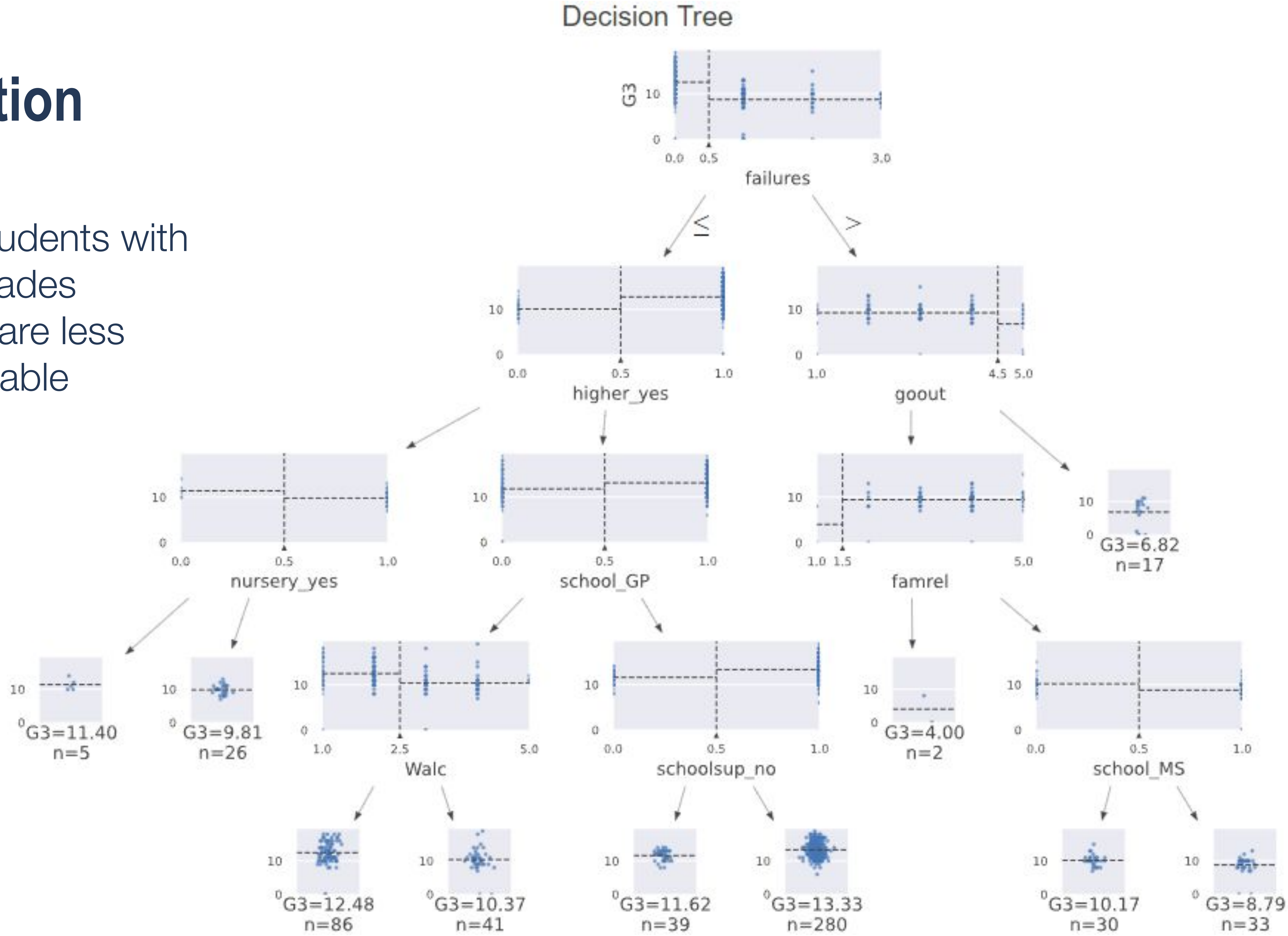
- **Baseline MAE: 2.2**
- **Tuned MAE: 2.07**

- On depth 4 most decisive features are “failures”, “higher_yes” and “goout”
- ‘goout’, ‘famrel’ and ‘nursery’ had low correlation with target



Visualization

- Model disserns students with very low grades
  - Other groups are less distinguishable



Linear Regression

Simple OLS model

MAE = 2.041 (CV score)

Cleaning data before OLS

- For linear regression drop collinear variables:
- From pairs ['Fedu', 'Medu'], ['Mjob', 'Medu'] and ['Dalc', 'Walc'] which have correlation with each other more 0.5, drop Fedu, Mjob and Walc as they have lower correlation with estimated variable G3

- Also completely insignificant features were dropped -- 'reason', 'guardian' and 'traveltime' (P-value >0.070, >0,123, 0.533 respectively)

- Results on test data is lower in comparison with train-- overfitting sign --> wil try regularization
  - MAE train = 1.965
  - MAE test = 1.852

The most important variables:
Failures, Higher, Fjob, Schoolsup, School

Regularization

Ridge

MAE = 2.033 (CV score)

Lasso

MAE = 2.017 (CV score)

Best alpha for Ridge regression -- 10 (Cross Validated)
For Lasso -- 0.011 (Cross Validated)

The most important variables
Ridge: Failures, Higher, Medu, Schoolsup, Health
Lasso: Failures, School, Dalc

Polynomial Model

Simple OLS model

MAE = 2.041 (CV score)

High overfitting on different from 1 degrees:
MAE test data (degree 2) = 59619068.7985
MAE test data (degree 3) = 5.0635

Lasso and Ridge Regularization Feature Importance

Random Forest

Algorithm

This algorithm build a composition of decision trees. The main idea is to:

- Build a number of trees, each on the random subset of features
- Aggregate the predictions of all the models
- This method helps to decrease the variance of the model

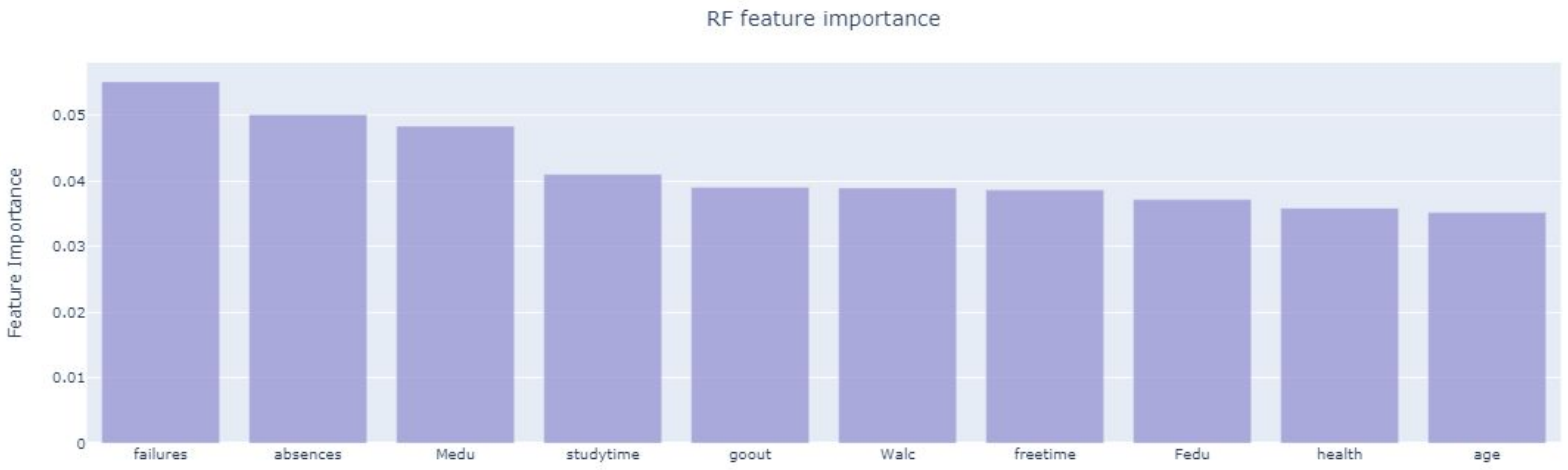
CV results:

- 'max_depth': None, 'max_features': 'sqrt', 'max_samples': 0.9, 'n_estimators': 150
- For optimal regression we need to use 150 trees
- 90% of X in each tree
- sqrt(number of features) features in each tree

Quality of estimator

- **Baseline MAE: 2.32**
- **Tuned MAE: 1.80**

Random Forest Feature Importance



Gradient Boosting Regressor

Algorithm

This algorithm build a composition of simple models. The main idea:

- Each model needs to be trained using previous model's prediction
- Subsequent models “fix” the errors of previous models

CV results:

- 'max_depth': None, 'max_features': 'sqrt', 'max_samples': 0.9, 'n_estimators': 150
- For optimal regression we need to use 150 trees
- 90% of X in each tree
- sqrt(number of features) features in each tree

Loss

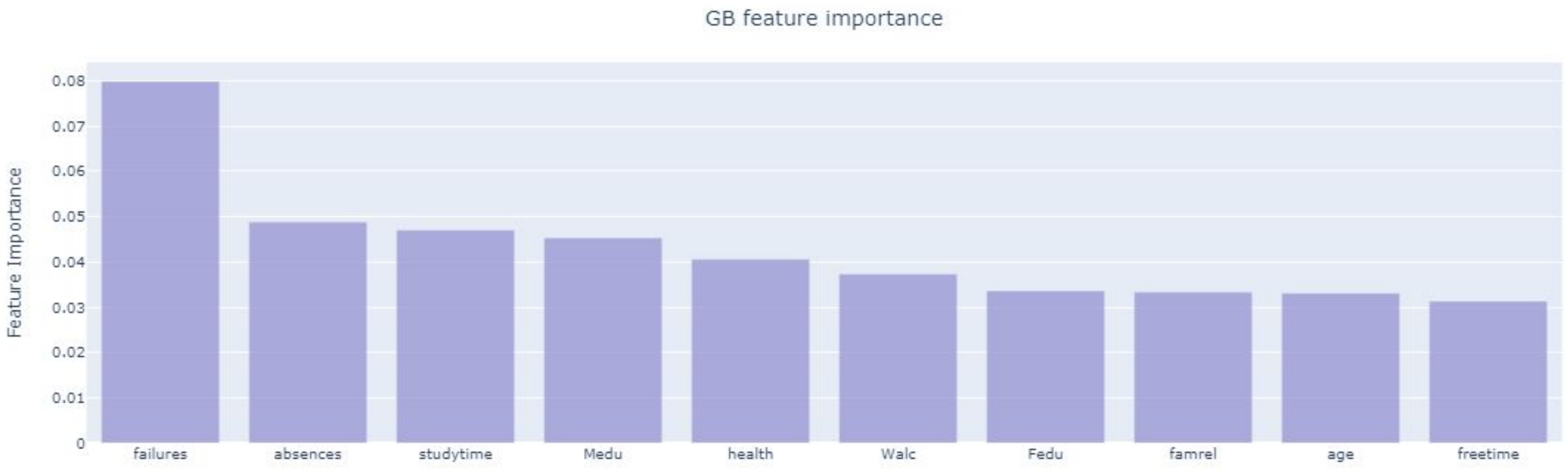
$$\frac{1}{l}\sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + b_N(x_i)) \rightarrow min_{b_N(x)}$$

$a_{N-1}(x_i) \rightarrow$ ответы предыдущих моделей, $const$

Quality of estimator

- **Baseline MAE: 2.28**
- **Tuned MAE: 1.78**

GB Feature Importance



KNN for regression

Algorithm

The main idea is to:

- Predict the necessary value by taking average results of **k nearest neighbors** (objects)
- KNN regression tries to predict the value of the output variable by using a local average

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

K?

- To define the optimal K nearest neighbors → GridSearchCV with MAE minimizing scoring
- Optimal K: `{'n_neighbors': 11}`
- Interpretation: it counts the average final grade of 11 closest points to the observation

Quality of algorithm

- **MAE (CV) : 2.18**
- **R^2 (CV) : 0.06**

poor results of method

Broad model test

Model	MAE
Decision Trees	2.07
Linear Regression	1.85
KNN	2.18
Random Forest	1.80
Gradient Boosting	1.78

Interpretation of the best model

- Previous failures may be seen as lack of motivation and negatively impact the grade
- Absences show lack of preparation and negatively impact the grade
- More studytime is good for a grade
- Parent’s education is impactful

Conclusion

- The most important variable across all models seems to be **failures**
- The models performance on the data is <u>mediocre</u> → Need more data or features
- The best model is Gradient Boosting Regressor