Breakdown Risk Analysis

Tamás Barczikay

December 1, 2024

Objectives of the Analysis

Objective 1: Understanding the sample

- What patterns can be seen from the data?
- What analytical and/or statistical statements can we make based on the data?

Objective 2: Estimating breakdown risk

- Which models would suit the task?
- Which attributes would be relevant for forecasting the risk of a breakdown?

Data Cleaning

- The raw data table consisted of 316 observations with 15 features on engine tests.
- Two observations didn't have any data in them except for the operating hours (`oph`).
- Entry error in `oph` deleted.
- No variation in `op_set_1` and `op_set_3` columns.
- `op_set_2` column is empty.

Feature Selection I.

Variable	Correlation	P-value
oph	-0.219	0.000
bmep	-0.144	0.011
ng_imp	-0.093	0.102
rpm_max	0.416	0.000
number_up	-0.420	0.000
number_tc	0.335	0.000

Table: Correlation of Numeric Variables

Feature Selection II.

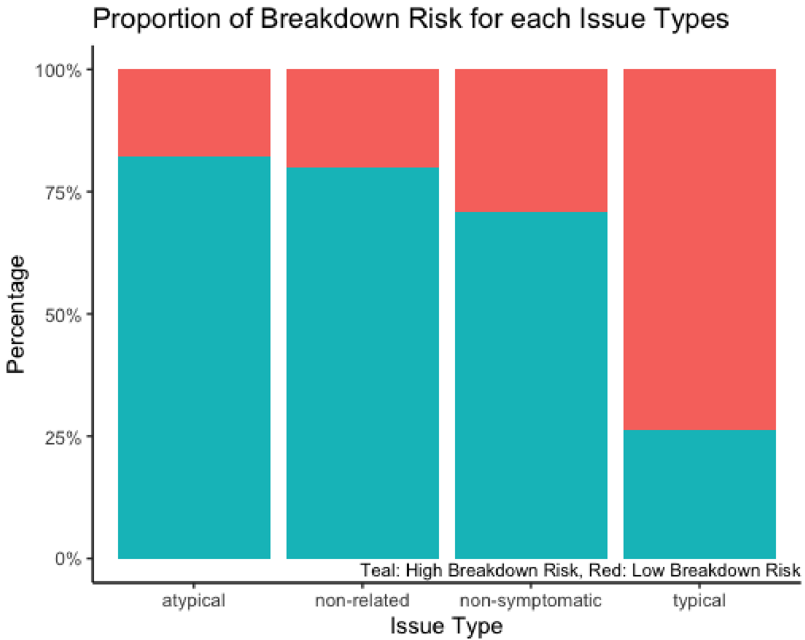
Phi Coefficients

Variable	**Phi Coefficient**	**P-value**
pist_m	0.293	0.000
past_dmg	0.036	0.634
full_load_issues	0.450	0.000

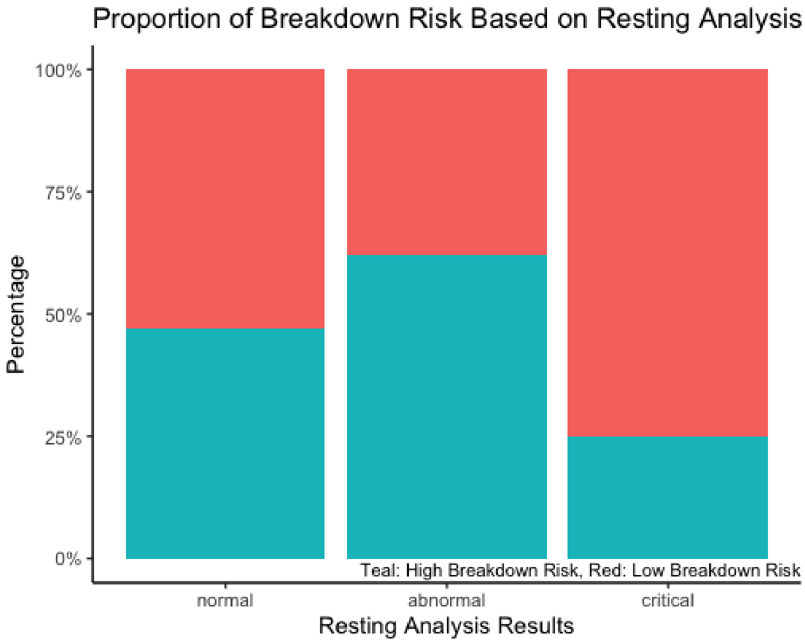
Cramer V

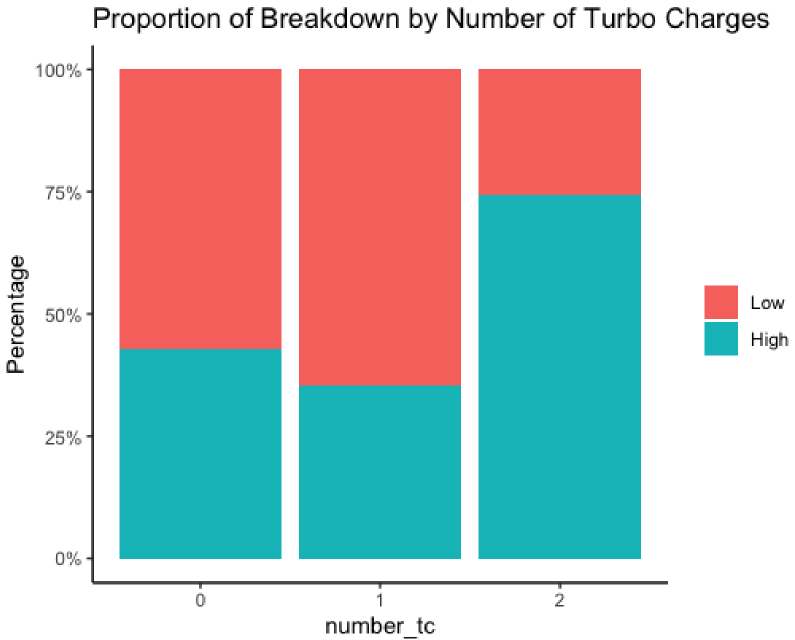
Variable	**Correlation**	**P-value**
issue_type	0.534	0.00
resting_analysis_results	0.164	0.01
number_tc	0.382	0.00

Table: Correlation metrics for categorical variables



Proportion of Breakdown Risk for each Issue Types
100%
75%
50%
25%
0%
Percentage
Teal: High Breakdown Risk, Red: Low Breakdown Risk
atypical
non-related
non-symptomatic
typical
Issue Type


Proportion of Breakdown Risk Based on Resting Analysis
Percentage
100%
75%
50%
25%
0%
Teal: High Breakdown Risk, Red: Low Breakdown Risk
normal
abnormal
critical
Resting Analysis Results


Proportion of Breakdown by Number of Turbo Charges
100%
75%
50%
25%
0%
Percentage
0
1
2
number_tc
Low
High

Dimensionality Reduction

Goal was to end up with a limited amount of features, that are statistically significant and easy to interpret.

- `issue_type` was recoded to the `typical_issue_type` variable
- `resting_analysis_results` was recoded to the `critical_rest_result` variable (Note: there were only 4 critical cases)
- `number_tc` was recoded to the `number_tc_2` variable (148 cases consist around the half of the overall)

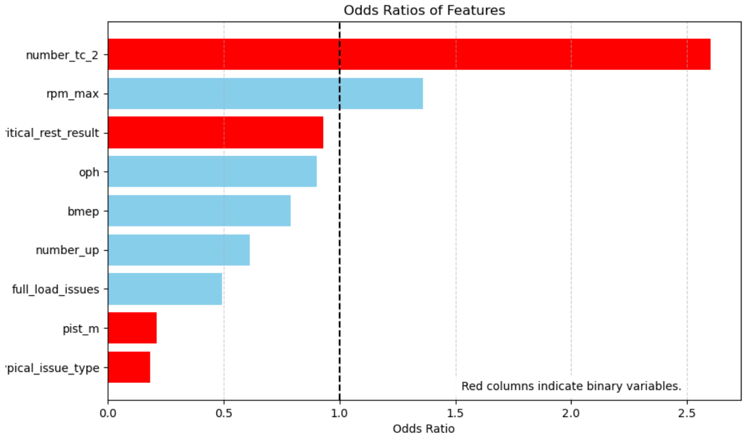
Modelling Breakdown Risk

- We used logistic regression and SVM models.
- Grid search with 10 fold cross-validation was performed.
- All numeric variables were standardized.

Metrics	**LR**	**SVC**
Accuracy	0.841	0.825
Recall	0.886	0.886
Precision	0.838	0.816
F1 score	0.861	0.849

Table: Metrics for best fitting models

Feature Importance of the Best Fitting LR Model



Conclusions

Objective 1: Understanding the sample (i.e. important patterns)

1. *47% of the engines had two turbo chargers.*
2. *68% of the engines had better piston material.*

Objective 2: Estimating breakdown risk

- Appropriate models: cross-validated LR (or similar classifiers)
- Attributes lead to breakdown:
    1. *Having two turbo chargers compared to having less is associated with a 2.6 times increase in the odds of experiencing a high breakdown risk.*
    2. *A one st. dev. increase in RPM is associated with a 1.4 times increase in the odds of experiencing a high breakdown risk.*
    3. *Having an engine piston made out of a better material, is associated with an odds ratio of 0.2, which corresponds to a decrease in the probability of a breakdown by 16.67%*
    4. *Typical issue type, though has a low value, might be endogenous.*