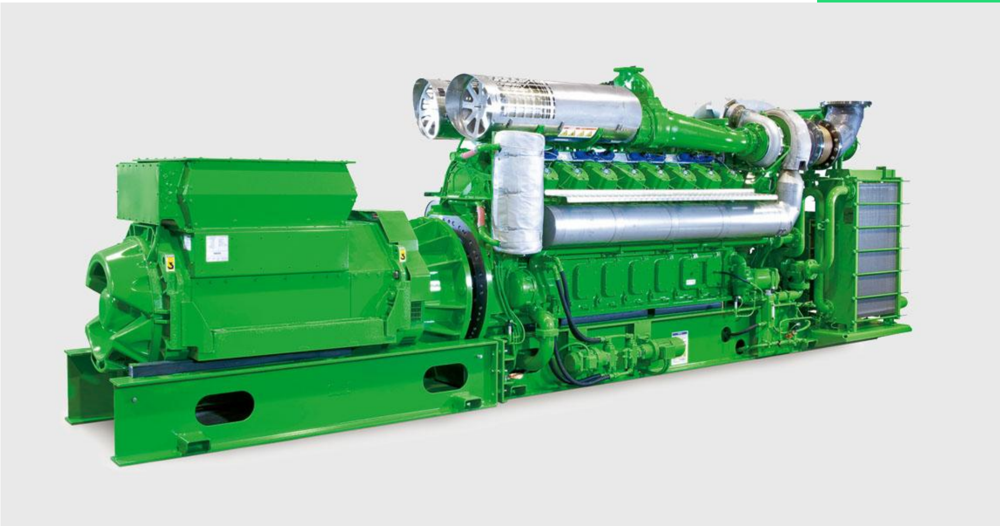
Engine project

Budapest, June 2024

Content



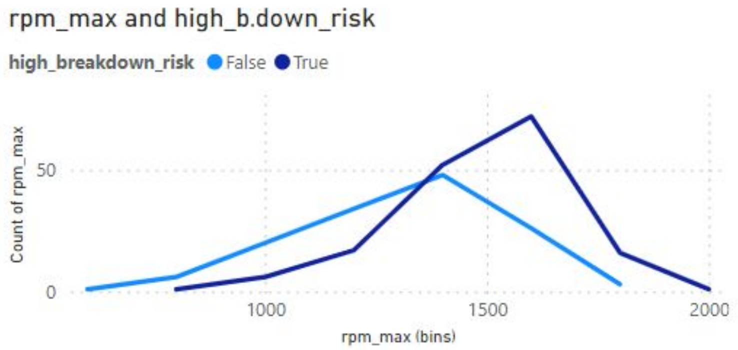
Summary

- In this analysis, I aim to identify key factors contributing to engine breakdowns and recommend predictive models to mitigate such risks. By examining various attributes of engine performance and maintenance, I seek to provide actionable insights for improving engine reliability and operational efficiency.

- In conclusion, my analysis identified key factors influencing engine breakdown risks, such as **RPM, turbochargers, and piston material**. By employing **Random Forest** and **Gradient Boosting Classifiers**, we can achieve accurate predictions, enabling proactive maintenance strategies to reduce downtime and improve engine reliability. Future work should focus on real-time data integration and continuous model improvement.

Trends and patterns

- **Higher RPM and Breakdown Risk:** Engines with higher maximum RPMs consistently show a trend toward higher breakdown risk, emphasizing the need for careful monitoring of high-performance engines. - page 24



- **Impact of Turbochargers:** More turbochargers correlate with increased breakdown risk, highlighting the complexity and potential vulnerabilities introduced by additional turbocharging systems. - page 27

High_breakdown_risk

number_tc	False	True	**Total**
0	12	9	**21**
1	91	50	**141**
2	35	106	**141**
Total	**138**	**165**	**303**

Trends and patterns

- **Resting Analysis Results:** Abnormal resting analysis results are moderately correlated with breakdown risk, underlining the importance of regular post-operation analysis for predictive maintenance. - page 23

High_breakdown_risk

resting_analysis_results	False	True	Total
0.0	79	68	**147**
1.0	56	96	**152**
2.0	3	1	**4**
Total	**138**	**165**	**303**

- **Unplanned Events and Breakdown Risk:** More unplanned events correlate with lower breakdown risk, suggesting effective corrective actions are being taken following each unplanned event. - page 26

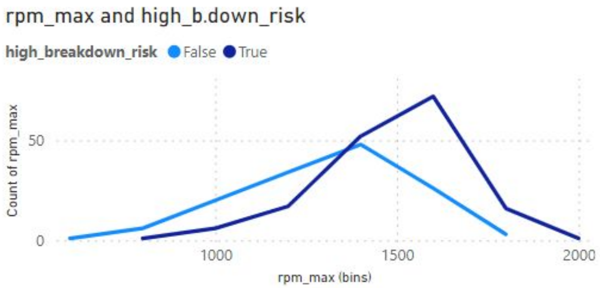
High_breakdown_risk

issue_type	False	True	Total
atypical	9	41	**50**
non-related	18	68	**86**
non-symptomatic	7	16	**23**
typical	104	40	**144**
Total	**138**	**165**	**303**

- **Combustion Issues:** Engines with non-related and atypical combustion issues show higher breakdown risks, emphasizing the need for timely and effective resolution of combustion problems. - page 19

Attributes analysis

RPM Max (rpm_max) and Breakdown Risk: page 24



- **Pattern:** Higher maximum RPM values are associated with an increased risk of breakdown.
- **Analytical Statement:** Engines that achieve higher maximum RPMs tend to have a higher breakdown risk. The correlation coefficient is approximately 0.42, indicating that it is relevant for forecasting the risk of a breakdown.

Number of Turbo Chargers (number_tc) and Breakdown Risk: page 27

High_breakdown_risk

number_tc	False	True	**Total**
0	12	9	**21**
1	91	50	**141**
2	35	106	**141**
Total	**138**	**165**	**303**

- **Pattern:** A greater number of turbochargers is associated with a higher breakdown risk.
- **Analytical Statement:** The presence of more turbochargers correlates with an increased likelihood of breakdowns (relevant for forecasting the risk of a breakdown).

Resting Analysis Results and Breakdown Risk: page 23

High_breakdown_risk

resting_analysis_results	False	True	**Total**
0.0	79	68	**147**
1.0	56	96	**152**
2.0	3	1	**4**
Total	**138**	**165**	**303**

- **Pattern:** Engines with abnormal or critical resting analysis results are more likely to break down.
- **Analytical Statement:** There is a moderate positive correlation (r ≈ 0.14) between resting analysis results and breakdown risk. Engines with abnormal or critical resting analysis results show a higher incidence of breakdowns compared to those with normal results. (relevant for forecasting the risk of a breakdown)

Attributes analysis

INNIO

Issue Types (issue_type) and Breakdown Risk: page 19

- **Pattern:** Different types of combustion issues have varying impacts on breakdown risk.
- **Analytical Statement:** Engines with non-related and atypical combustion issues show a higher risk of breakdowns compared to those with typical or non-symptomatic issues. The correlation between issue type and breakdown risk highlights the significance of addressing combustion issues promptly to prevent breakdowns. (relevant for forecasting the risk of a breakdown)

	High_breakdown_risk		
issue_type	False	True	**Total**
atypical	9	41	**50**
non-related	18	68	**86**
non-symptomatic	7	16	**23**
typical	104	40	**144**
Total	**138**	**165**	**303**

Full Load Operation Issues (full_load_issues) and Breakdown Risk: page 25

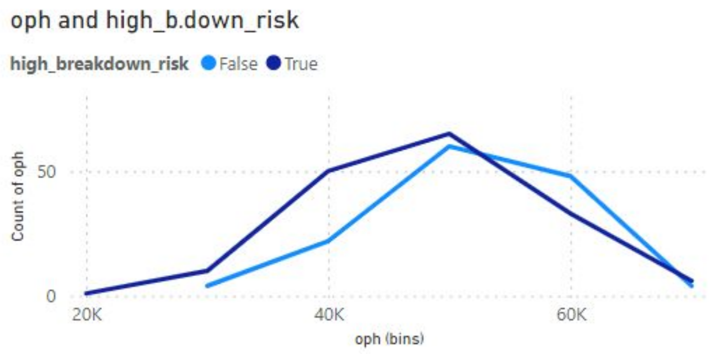
- **Pattern:** There is a strong negative correlation between full load operation issues and breakdown risk.
- **Analytical Statement:** Engines experiencing issues during full load operations tend to have a lower risk of breakdown (r ≈ -0.44), which could be due to more frequent maintenance and checks in response to these issues. (relevant for forecasting the risk of a breakdown)

	High_breakdown_risk		
full_load_issues	False	True	**Total**
False	62	142	**204**
True	76	23	**99**
Total	**138**	**165**	**303**

Attributes analysis

Operating Hours (oph) and Breakdown Risk: page 17

- **Pattern:** There is a negative correlation between operating hours and breakdown risk.
- **Analytical Statement:** Contrary to initial expectations, the data reveals a negative correlation (r ≈ -0.22) between operating hours and breakdown risk, indicating that engines with higher operating hours might have undergone more maintenance, reducing the risk breakdown. (relevant for forecasting the risk of a breakdown)



Piston Material (pist_m) and Breakdown Risk: page 18

- **Pattern:** The type of piston material shows a negative correlation with breakdown risk.
- **Analytical Statement:** Engines with certain piston materials are less likely to break down, with a correlation coefficient of approximately -0.28, indicating that specific materials contribute to improved engine reliability. (relevant for forecasting the risk of a breakdown)

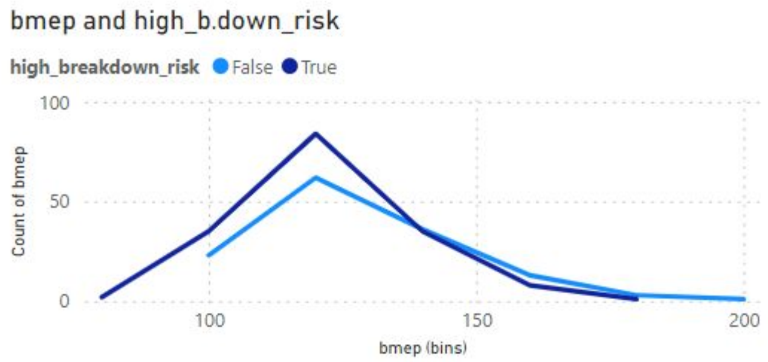
High_breakdown_risk

pist_m	False	True	Total
False	24	72	**96**
True	114	93	**207**
Total	**138**	**165**	**303**

Attributes analysis

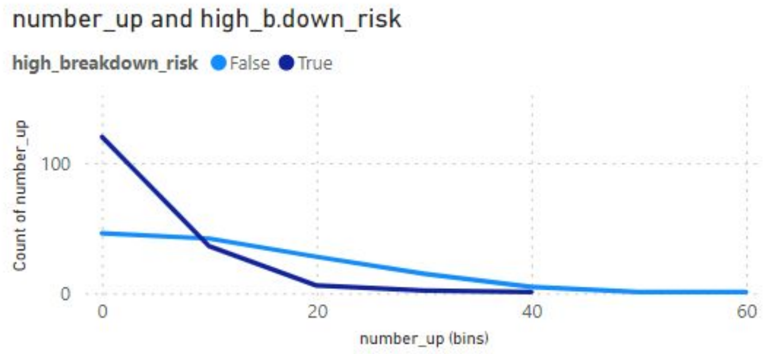
Break Mean Effective Pressure (bmep) and Breakdown Risk: page 20

- **Pattern:** There is a slight negative correlation between BMEP and breakdown risk.
- **Analytical Statement:** Higher BMEP values correlate with a reduced likelihood of breakdowns, suggesting that engines operating under higher pressure conditions might be less prone to failures.



Number of Unplanned Events (number_up) and Breakdown Risk: page 26

- **Pattern:** There is a strong negative correlation between the number of unplanned events and breakdown risk.
- **Analytical Statement:** Engines with more unplanned events show a lower risk of breakdown (r ≈ -0.43), possibly due to increased monitoring and preventive measures taken after each event.



Attributes analysis

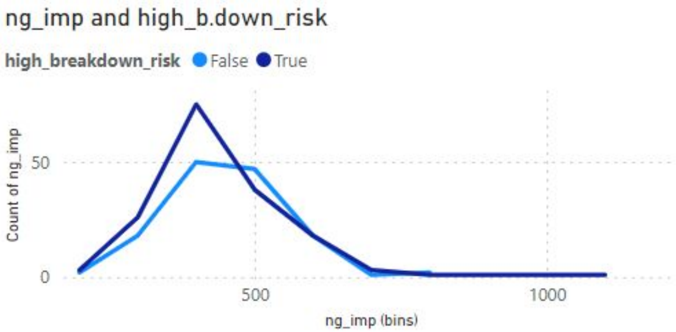
Past Damages (past_dmg) and Breakdown Risk: page 22

- **Pattern:** Past damages have a slightly negative correlation with breakdown risk.
- **Analytical Statement:** Surprisingly, engines with past damages have a weak negative correlation (r ≈ -0.03) with breakdown risk, suggesting that past damages alone are not a strong predictor of future breakdowns.

	High_breakdown_risk		
past_dmg	False	True	**Total**
False	116	142	**258**
True	22	23	**45**
Total	**138**	**165**	**303**

Natural Gas Impurities (ng_imp) and Breakdown Risk: page 21

- **Pattern:** Natural gas impurities show a slight negative correlation with breakdown risk.
- **Analytical Statement:** The correlation coefficient between natural gas impurities and breakdown risk is close to zero, indicating that impurities in the fuel have zero relationship with breakdown risk.



Correlations Between Attributes - page 29

- **Operating Hours and Resting Analysis Results:**
  - There is a moderate positive correlation ($r \approx 0.25$) between operating hours and resting analysis results, suggesting that engines with more operating hours tend to show more abnormalities during resting analysis.
- **RPM Max and Number of Turbo Chargers:**
  - A moderate positive correlation ($r \approx 0.3$) exists between maximum RPM and the number of turbochargers, indicating that engines capable of higher RPMs often have more turbochargers installed.
- **Natural Gas Impurities and BMEP:**
  - A moderate positive correlation ($r \approx 0.2$) is observed between natural gas impurities and BMEP, suggesting that engines with higher fuel impurities tend to operate under higher pressure conditions.

Model recommendation

Model	Accuracy	AUC	Precision	Recall	F1 Score
Random Forest	87.8%	0.9686	90.1%	85.6%	87.8%
Gradient Boosting	93.9%	0.9921	92.3%	95.6%	93.9%
Logistic Regression	75.0%	0.8487	77.0%	80.6%	78.5%

For forecasting the risk of engine breakdown using the provided dataset, the Random Forest and Gradient Boosting Classifier models are highly recommended due to their superior performance in handling complex interactions and providing accurate predictions. These models, along with the identified relevant attributes, will ensure robust predictive maintenance strategies, enabling proactive measures to mitigate breakdown risks effectively.

Random Forest Classifier:

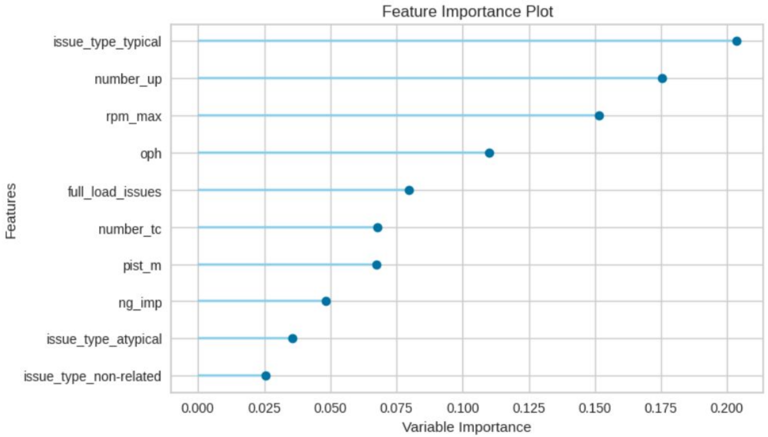
- **Suitability:** The model demonstrated an accuracy of 87.8% and an AUC of 0.9686 after tuning. This indicates that Random Forest effectively captures complex interactions between variables and provides reliable predictions. It also provides feature importance, which helps in understanding the impact of different attributes on breakdown risk. - page 13

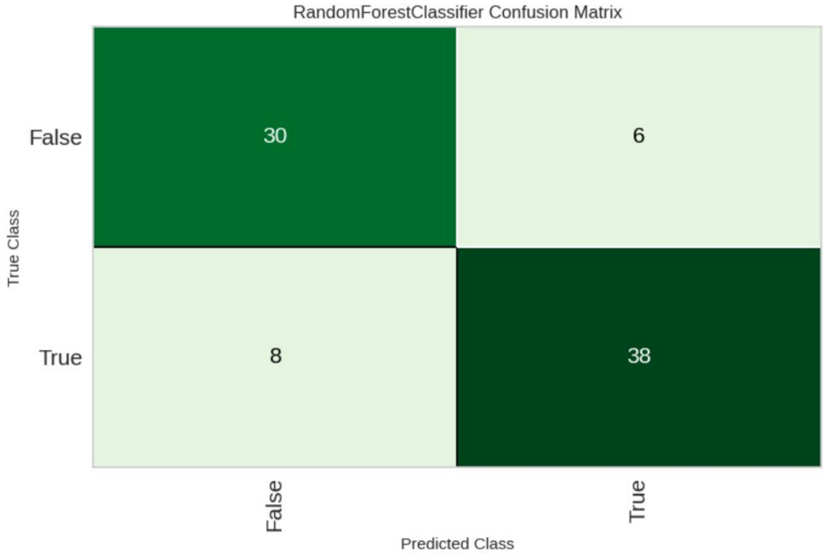
Gradient Boosting Classifier:

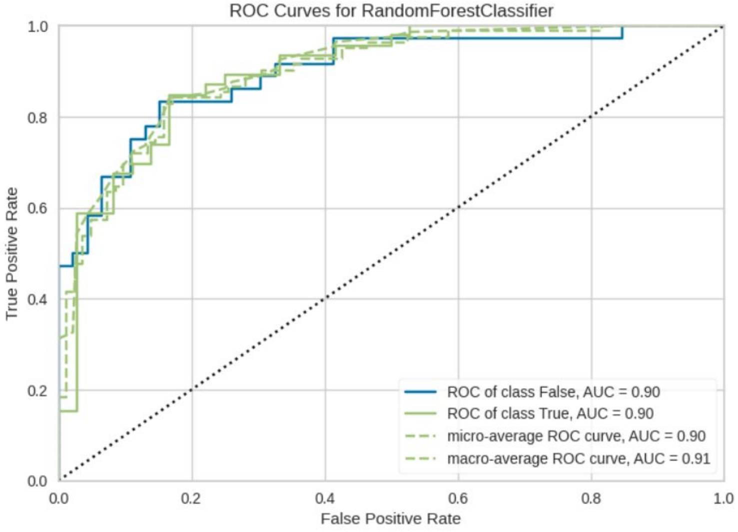
- **Suitability:** This model achieved the highest accuracy (93.9%) and AUC (0.9921) after tuning, indicating excellent performance in predicting breakdown risk. - page 14

Random Forest Classifier

Steps	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
create_model() -mean	0.7847	0.8515	0.8236	0.8164	0.8127	0.5618	0.5738
tune_model() -mean	0.7958	0.8562	0.8036	0.8352	0.8136	0.587	0.5967
predict_model(tuned)	0.8293	0.8979	0.8261	0.8636	0.8444	0.6555	0.6563
predict_model(final)	0.878	0.9686	0.8696	0.9091	0.8889	0.7539	0.7548
unseen_prediction	0.7	0.8688	0.6923	0.6429	0.6667	0.3946	0.3955

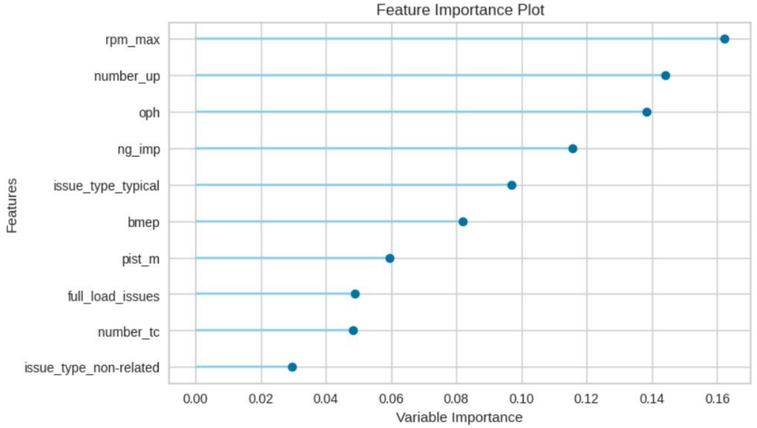


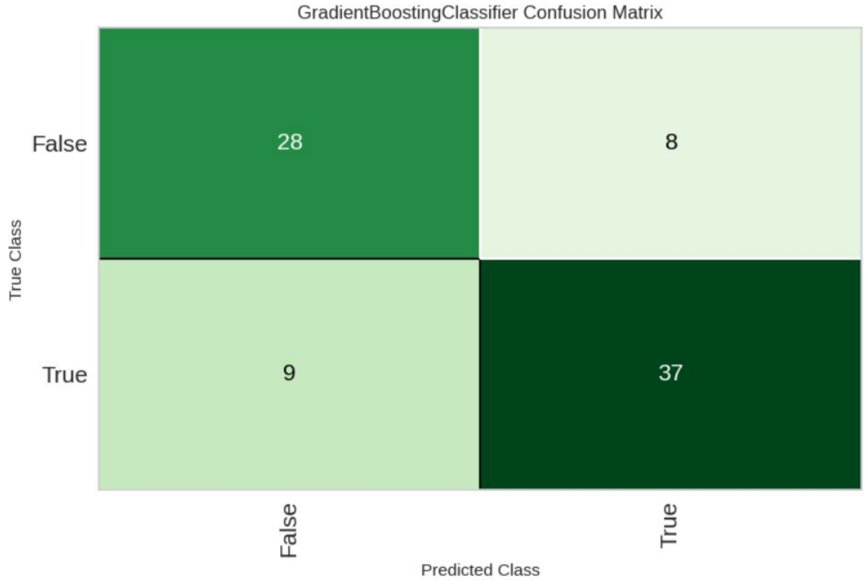


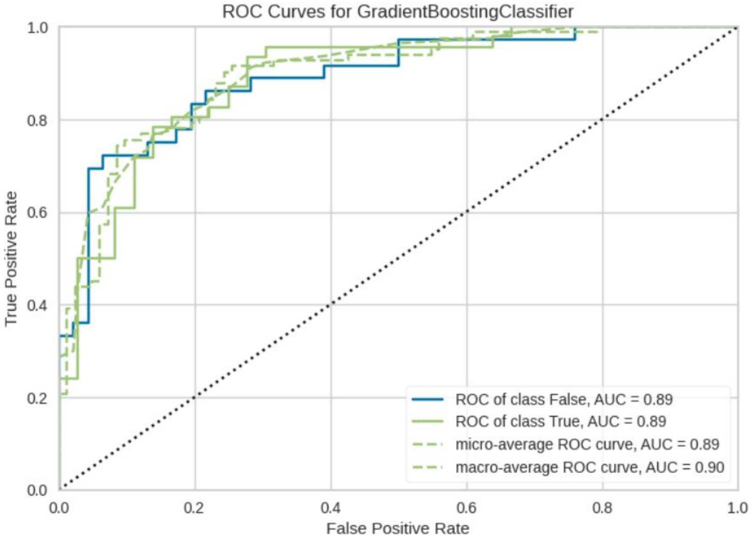


Gradient Boosting Classifier

Steps	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
create_model() -mean	0.7689	0.8473	0.8036	0.8003	0.7954	0.5297	0.5395
tune_model() -mean	0.7797	0.8696	0.8618	0.7814	0.8152	0.5453	0.5577
predict_model(tuned)	0.7927	0.8901	0.8043	0.8222	0.8132	0.5804	0.5806
predict_model(final)	0.939	0.9921	0.9565	0.9362	0.9462	0.8758	0.8761
unseen_prediction	0.7	0.8824	0.7692	0.625	0.6897	0.4053	0.4135







Annex

Data cleaning steps

Original dataset with 15 columns and 316 entries

- Exclude duplicates
- Drop 3 columns
- Drop outliers
- Handle missing values
- Convert data types

Cleaned dataset with 12 columns and 303 entries

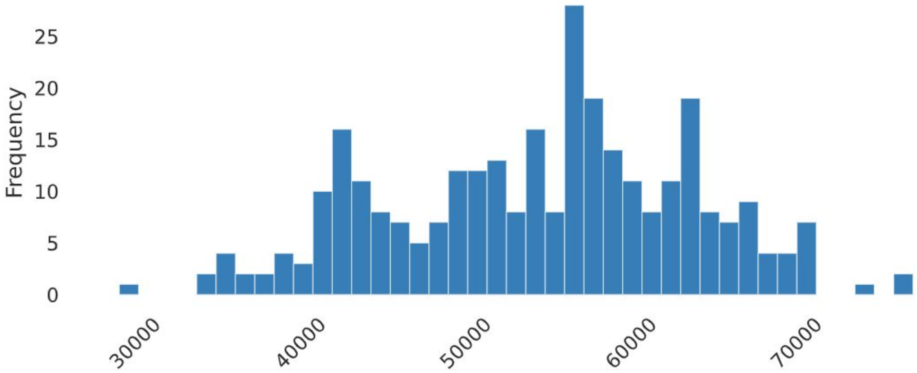
Attribute analysis with ydata-profiling

Attribute analysis - oph

oph

Real number (ℝ)

Distinct	41	**Minimum**	29000
Distinct (%)	13.5%	**Maximum**	77000
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	54468.647	**Memory size**	2.5 KiB



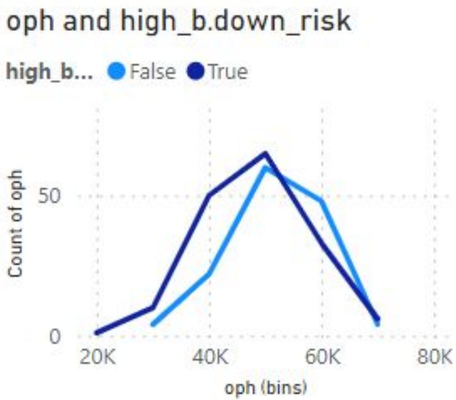
Histogram with fixed size bins (bins=41)

Quantile statistics

Minimum	29000
5-th percentile	40000
Q1	48000
median	56000
Q3	61000
95-th percentile	68000
Maximum	77000
Range	48000
Interquartile range (IQR)	13000

Descriptive statistics

Standard deviation	9071.7253
Coefficient of variation (CV)	0.16654949
Kurtosis	-0.53514629
Mean	54468.647
Median Absolute Deviation (MAD)	6000
Skewness	-0.20365472
Sum	16504000
Variance	82296199
Monotonicity	Not monotonic



Attribute analysis - pist_m

pist_m
Boolean

Distinct	2
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	431.0 B

High_breakdown_risk

pist_m	False	True	**Total**
False	24	72	**96**
True	114	93	**207**
Total	**138**	**165**	**303**

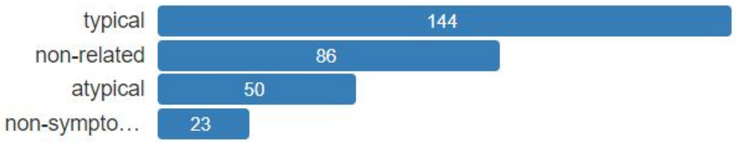
Attribute analysis - issue_type

issue_type

Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	1.3%
Missing	0
Missing (%)	0.0%
Memory size	635.0 B

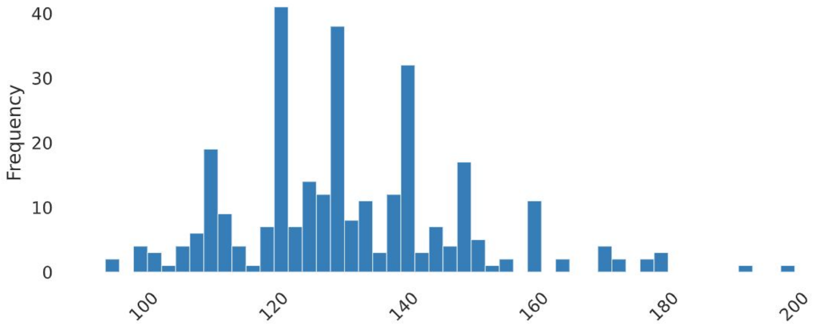


	High_breakdown_risk		
issue_type	False	True	**Total**
atypical	9	41	**50**
non-related	18	68	**86**
non-symptomatic	7	16	**23**
typical	104	40	**144**
Total	**138**	**165**	**303**

Attribute analysis - bmer

bmep
Real number (ℝ)

Distinct	49	Minimum	94
Distinct (%)	16.2%	Maximum	200
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	131.59736	Memory size	2.5 KiB



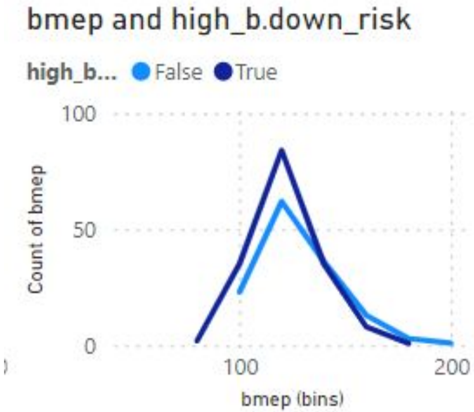
Histogram with fixed size bins (bins=49)

Quantile statistics

Minimum	94
5-th percentile	108
Q1	120
median	130
Q3	140
95-th percentile	160
Maximum	200
Range	106
Interquartile range (IQR)	20

Descriptive statistics

Standard deviation	17.534533
Coefficient of variation (CV)	0.13324381
Kurtosis	0.93662805
Mean	131.59736
Median Absolute Deviation (MAD)	10
Skewness	0.71859366
Sum	39874
Variance	307.45986
Monotonicity	Not monotonic

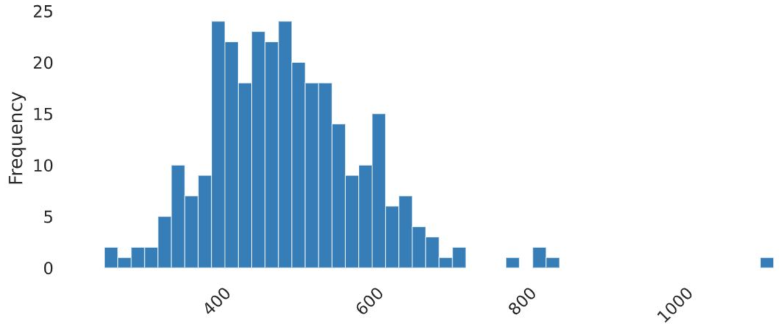


Attribute analysis - ng_imp

ng_imp
Real number (ℝ)

Distinct	153
Distinct (%)	50.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	492.9604

Minimum	252
Maximum	1128
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	2.5 KiB



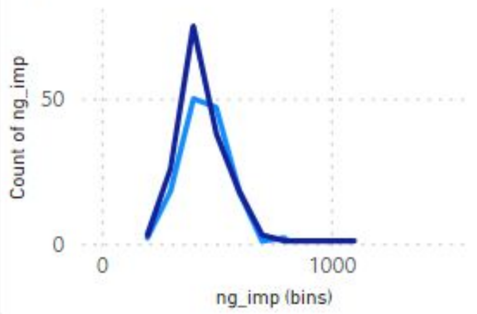
Histogram with fixed size bins (bins=50)

Quantile statistics

Minimum	252
5-th percentile	350.2
Q1	422
median	481
Q3	549
95-th percentile	653.8
Maximum	1128
Range	876
Interquartile range (IQR)	127

Descriptive statistics

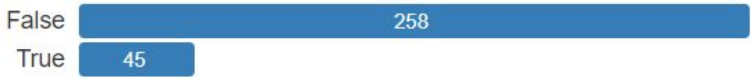
Standard deviation	103.33777
Coefficient of variation (CV)	0.20962691
Kurtosis	4.5684594
Mean	492.9604
Median Absolute Deviation (MAD)	63
Skewness	1.1502869
Sum	149367
Variance	10678.694
Monotonicity	Not monotonic



Attribute analysis - past_dmg

past_dmg
Boolean

Distinct	2
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	431.0 B



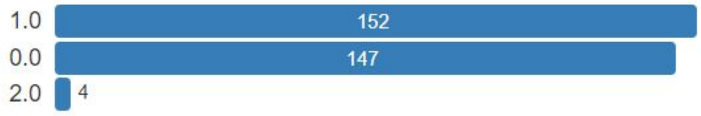
High_breakdown_risk

past_dmg	False	True	**Total**
False	116	142	**258**
True	22	23	**45**
Total	**138**	**165**	**303**

Attribute analysis - resting_analysis_result

resting_analysis_results
Categorical

Distinct	3
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Memory size	563.0 B



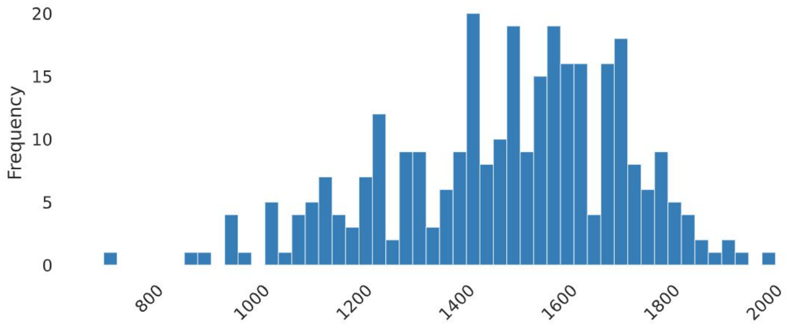
High_breakdown_risk

resting_analysis_results	False	True	**Total**
0.0	79	68	**147**
1.0	56	96	**152**
2.0	3	1	**4**
Total	**138**	**165**	**303**

Attribute analysis - rpm_max

rpm_max
Real number (ℝ)

Distinct	92	Minimum	710
Distinct (%)	30.4%	Maximum	2020
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1495.7921	Memory size	2.5 KiB



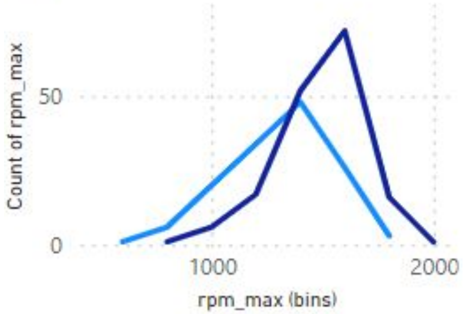
Histogram with fixed size bins (bins=50)

Quantile statistics

Minimum	710
5-th percentile	1081
Q1	1335
median	1525
Q3	1660
95-th percentile	1819
Maximum	2020
Range	1310
Interquartile range (IQR)	325

Descriptive statistics

Standard deviation	228.66196
Coefficient of variation (CV)	0.15287015
Kurtosis	-0.051828192
Mean	1495.7921
Median Absolute Deviation (MAD)	155
Skewness	-0.53477053
Sum	453225
Variance	52286.291
Monotonicity	Not monotonic

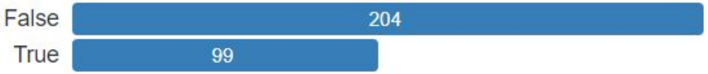


Attribute analysis - full_load_issues

full_load_issues

Boolean

Distinct	2
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	431.0 B



	High_breakdown_risk		
full_load_issues	False	True	**Total**
False	62	142	**204**
True	76	23	**99**
Total	**138**	**165**	**303**

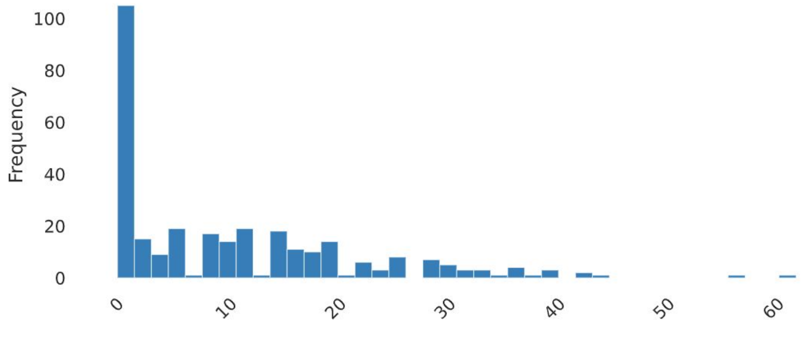
Attribute analysis - number_up

number_up

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	40	Minimum	0
Distinct (%)	13.2%	Maximum	62
Missing	0	Zeros	98
Missing (%)	0.0%	Zeros (%)	32.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	10.422442	Memory size	2.5 KiB



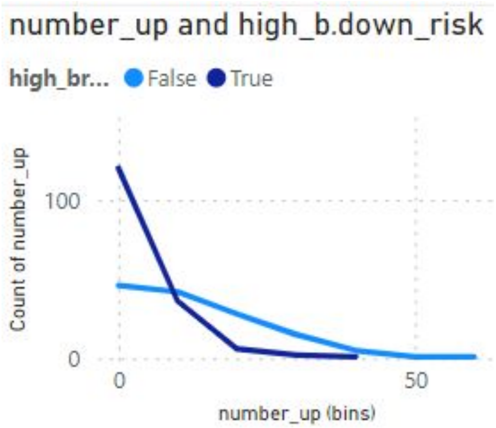
Histogram with fixed size bins (bins=40)

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	8
Q3	16
95-th percentile	34
Maximum	62
Range	62
Interquartile range (IQR)	16

Descriptive statistics

Standard deviation	11.596118
Coefficient of variation (CV)	1.1126105
Kurtosis	1.5851753
Mean	10.422442
Median Absolute Deviation (MAD)	8
Skewness	1.2700288
Sum	3158
Variance	134.46996
Monotonicity	Not monotonic



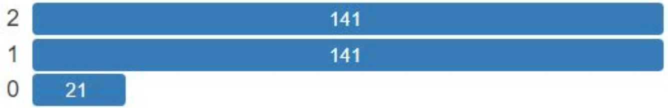
Attribute analysis - number_tc

number_tc

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Memory size	2.5 KiB



High_breakdown_risk

number_tc	False	True	**Total**
0	12	9	**21**
1	91	50	**141**
2	35	106	**141**
Total	**138**	**165**	**303**

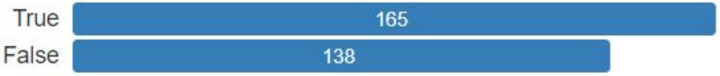
Attribute analysis - high_breakdown_risk

high_breakdown_risk

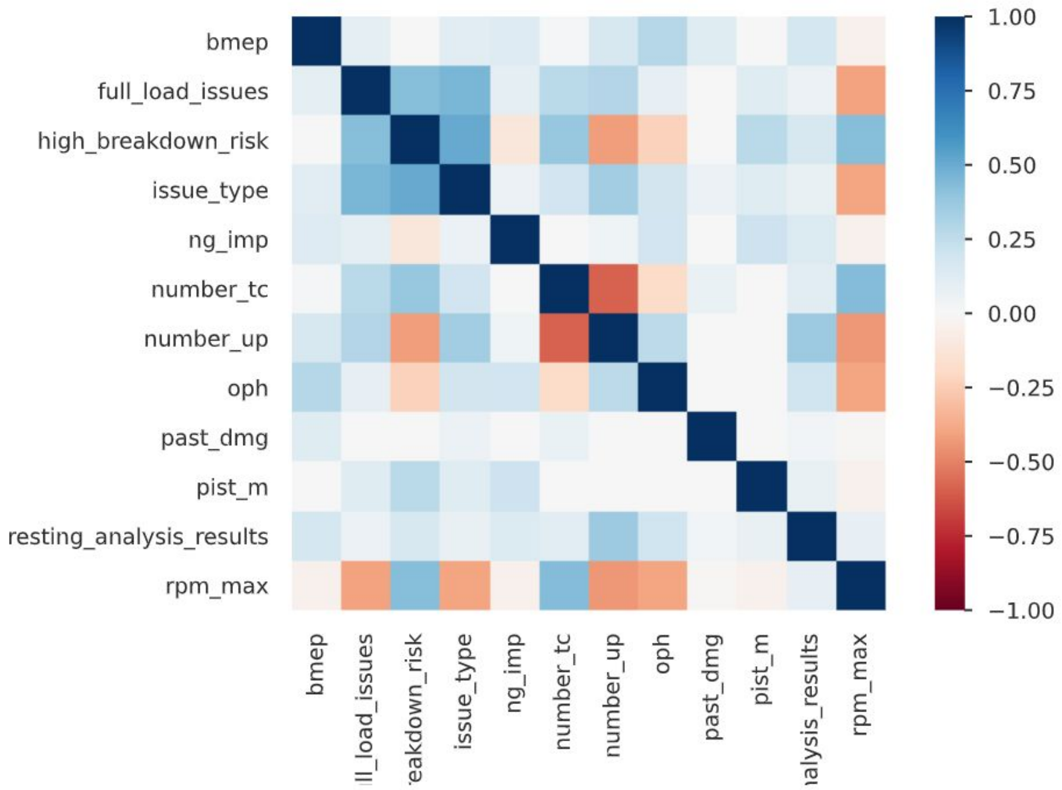
Boolean

HIGH CORRELATION

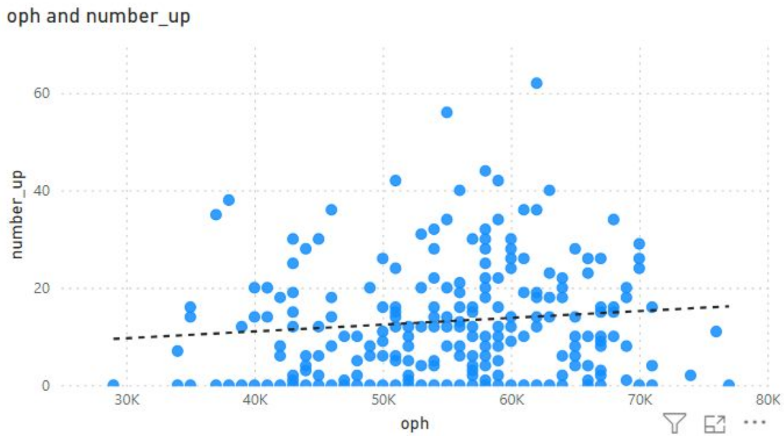
Distinct	2
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	431.0 B

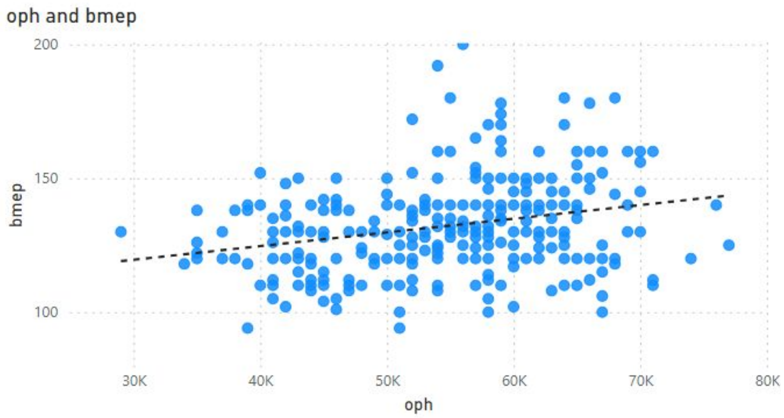


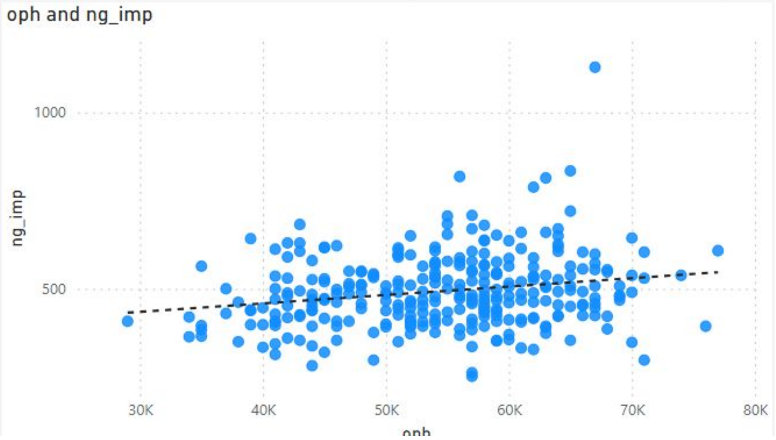
Correlations

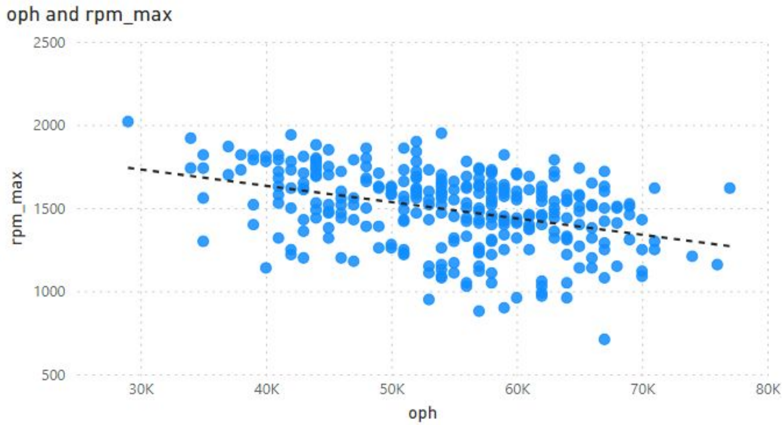


Interactions









Model and feature selection with PyCaret

(pycaret.classification Module)

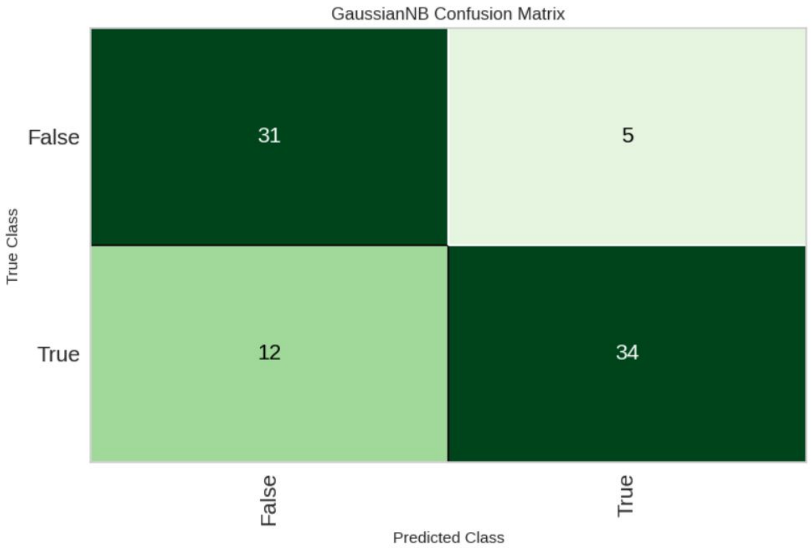
- **Getting Data:** Cleaned dataset with 12 columns and 303 entries (90% data, 10% as unseen data for the calculation)
- **Setting up Environment:** use 273 entities (70-30% training and test dataset), target is the ‘high_risk_breakdown’, using StratifiedKFold (fold number = 10)
- **Create Model:** create a model, perform stratified cross validation and evaluate classification metrics
- **Tune Model:** automatically tune the hyper-parameters of a classification model
- **Plot Model:** analyze model performance using various plots
- **Finalize Model:** finalize the best model at the end of the experiment
- **Predict Model:** make predictions on new / unseen data
- **Save / Load Model:** save / load a model for future use

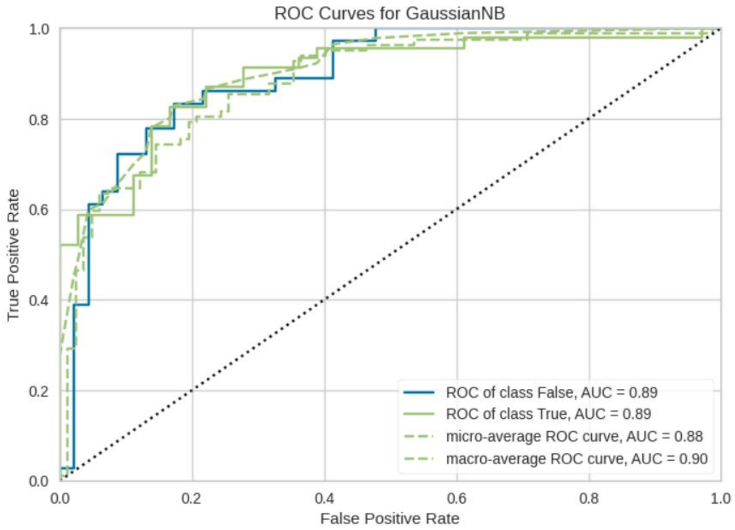
Compare_models()

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
nb	Naive Bayes	0.7850	0.8610	0.7964	0.8262	0.8043	0.5675	0.5794	0.0930
rf	Random Forest Classifier	0.7847	0.8515	0.8236	0.8164	0.8127	0.5618	0.5738	0.4080
ridge	Ridge Classifier	0.7742	0.0000	0.8327	0.7939	0.8060	0.5384	0.5545	0.1800
gbc	Gradient Boosting Classifier	0.7689	0.8473	0.8036	0.8003	0.7954	0.5297	0.5395	0.3070
lda	Linear Discriminant Analysis	0.7689	0.8494	0.8227	0.7915	0.8010	0.5279	0.5391	0.1800
et	Extra Trees Classifier	0.7589	0.8209	0.7773	0.8014	0.7822	0.5147	0.5244	0.3630
lightgbm	Light Gradient Boosting Machine	0.7587	0.8595	0.8055	0.7941	0.7873	0.5117	0.5316	0.2450
lr	Logistic Regression	0.7487	0.8487	0.8064	0.7703	0.7805	0.4875	0.5028	0.1150
ada	Ada Boost Classifier	0.7266	0.7808	0.7845	0.7454	0.7593	0.4440	0.4514	0.3120
dt	Decision Tree Classifier	0.6800	0.6761	0.7091	0.7253	0.7097	0.3500	0.3570	0.1690
knn	K Neighbors Classifier	0.6239	0.6410	0.6909	0.6560	0.6694	0.2342	0.2378	0.1930
qda	Quadratic Discriminant Analysis	0.5600	0.5865	0.5009	0.6404	0.5127	0.1048	0.1292	0.1030
dummy	Dummy Classifier	0.5550	0.5000	1.0000	0.5550	0.7135	0.0000	0.0000	0.1120
svm	SVM - Linear Kernel	0.5024	0.0000	0.3891	0.3322	0.3027	0.0437	0.0569	0.1080

Naive Bayes

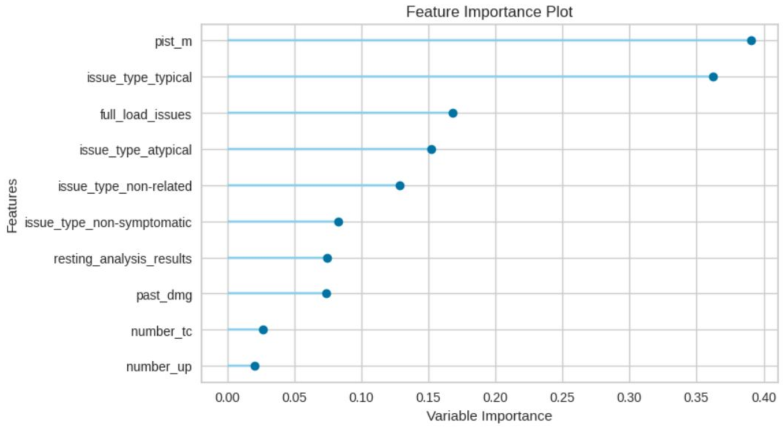
Steps	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
create_model() -mean	0.785	0.861	0.7964	0.8262	0.8043	0.5675	0.5794
tune_model() -mean	0.7905	0.8664	0.8336	0.8055	0.814	0.5759	0.5866
predict_model(tuned)	0.7927	0.8937	0.7391	0.8718	0.8	0.5878	0.5965
predict_model(final)	0.8293	0.904	0.8043	0.881	0.8409	0.6575	0.6607
unseen_prediction	0.7	0.8597	0.8462	0.6111	0.7097	0.4156	0.4394

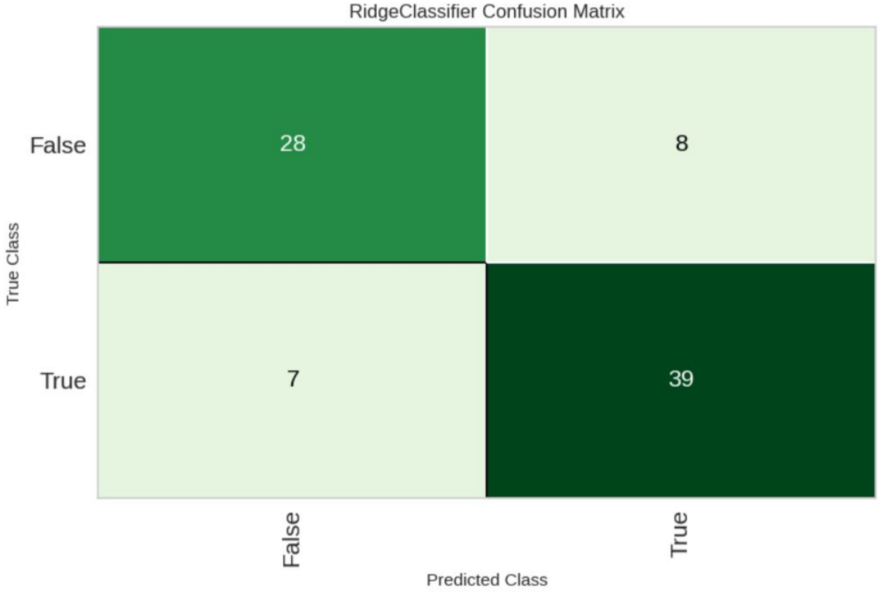




Ridge Classifier

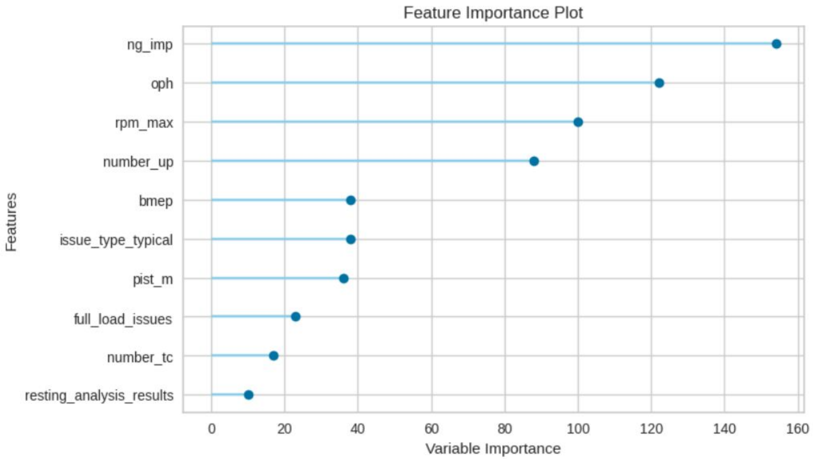
Steps	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
create_model() -mean	0.7742	0	0.8327	0.7939	0.806	0.5384	0.5545
tune_model() -mean	0.7795	0	0.8518	0.7894	0.8133	0.5473	0.5635
predict_model(tuned)	0.8171	0.8128	0.8478	0.8298	0.8387	0.6275	0.6277
predict_model(final)	0.8537	0.8454	0.913	0.84	0.875	0.6993	0.7028
unseen_prediction	0.7667	0.776	0.8462	0.6875	0.7586	0.5374	0.5483

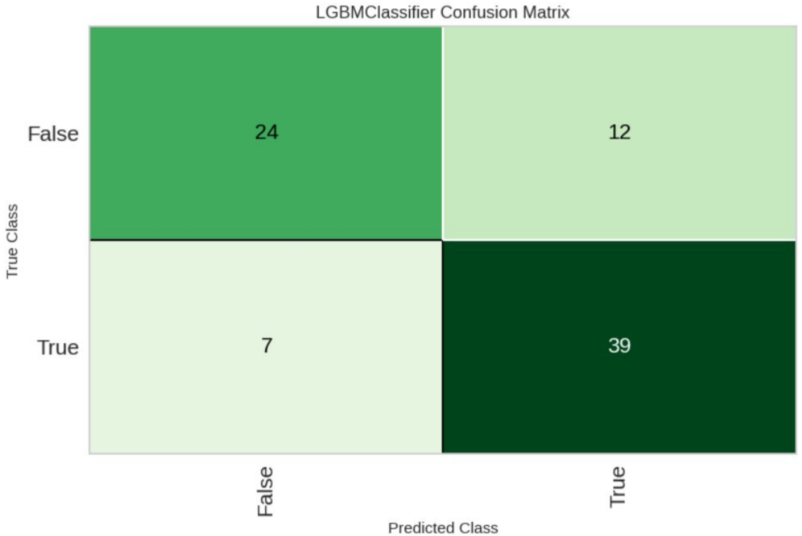


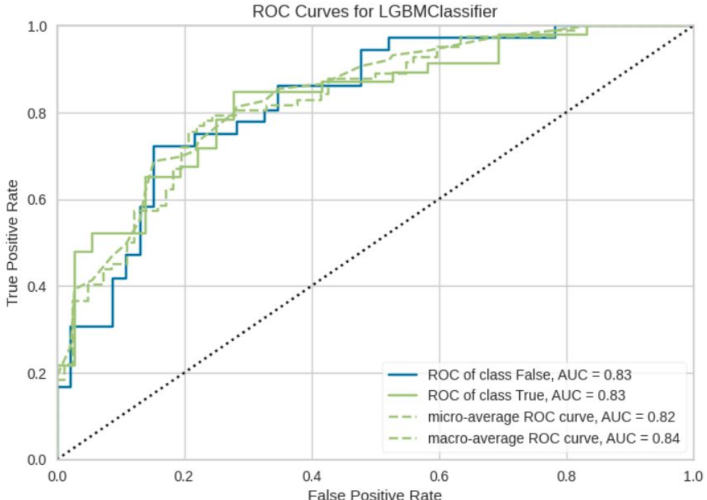


Light Gradient Boosting

Steps	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
create_model() -mean	0.7587	0.8595	0.8055	0.7941	0.7873	0.5117	0.5316
tune_model() -mean	0.7587	0.8643	0.8518	0.7656	0.8002	0.5005	0.5156
predict_model(tuned)	0.7683	0.8273	0.8478	0.7647	0.8041	0.5224	0.5266
predict_model(final)	1	1	1	1	1	1	1
unseen_prediction	0.7	0.7783	0.7692	0.625	0.6897	0.4053	0.4135







Logistic Regression

Steps	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
create_model() -mean	0.7487	0.8487	0.8064	0.7703	0.7805	0.4875	0.5028
tune_model() -mean	0.7795	0.8222	0.8409	0.7867	0.81	0.5482	0.5554
predict_model(tuned)	0.8415	0.9076	0.8043	0.9024	0.8506	0.6829	0.6881
predict_model(final)	0.6951	0.7591	0.7391	0.7234	0.7312	0.3792	0.3793
unseen_prediction	0.7	0.8869	0.9231	0.6	0.7273	0.4255	0.4757