MSG LIFE
TUM DATA INNOVATION LAB

ANALYZING THE GOODNESS OF FIT: NEURAL NETWORK REGRESSION MODELS

JULY 2021

CONTENTS

INTRODUCTION: OVERVIEW

	1 SITUATION	2 TASK	3 ACTION
GOAL	**To provide consumers with innovative & market-competitive tools confidently**	**Analysis of the fit of a Neural Network Regression Model**	**Creating reproducible and automatic procedures**
PROCEDURE	1. **Offer** new machine learning techniques 2. **Ensure** methods retain high standard of accuracy	1. **Given** the predicted values and future related values 2. **Data:** Large Sample size: 100,000 **Input values:** Key Attributes of individuals • Includes binary, continuous and nominal variables	1. **Goodness of fit:** predicted, observed and error values • Regression Assumptions • Distribution Fitting 2. **Anomaly Detection Methods** 3. **Risk Analysis**
OUTCOME	To guarantee "that msg life software solutions are up to the **highest standard of accuracy."**	Given the nature of neural networks, ensure stable findings in **post-evaluation stage**	Every procedure proposed can be implemented to analyze the goodness of fit and implications

CHATPER 1:
EVALUATION OF A REGRESSION MODEL

REGRESSION ASSUMPTIONS

Assess if regression model assumptions are fulfilled:
- Goodness of Fit tests, Diagnostic plots and statistical tests

DISTRIBUTION FITTING

General Workflow of Distribution Fitting
- Find the Error Distribution
- Find the Predicted Value Distribution
- Results and Discussions on given data set

ASSESS THE MODEL FIT

msg Life

PURPOSE

Post-Evaluation Stage: assess specification and statistical significance of model aspects

Evaluate: structural fit and prediction power via "what is left unmodelled"

STRATEGY

Residuals of Model: should behave like "white-noise" (random error)

Analyze: statistical properties of error terms tells us if there is evidence of "specification bias"

APPROACH

Perform Adequacy or Diagnostic tests

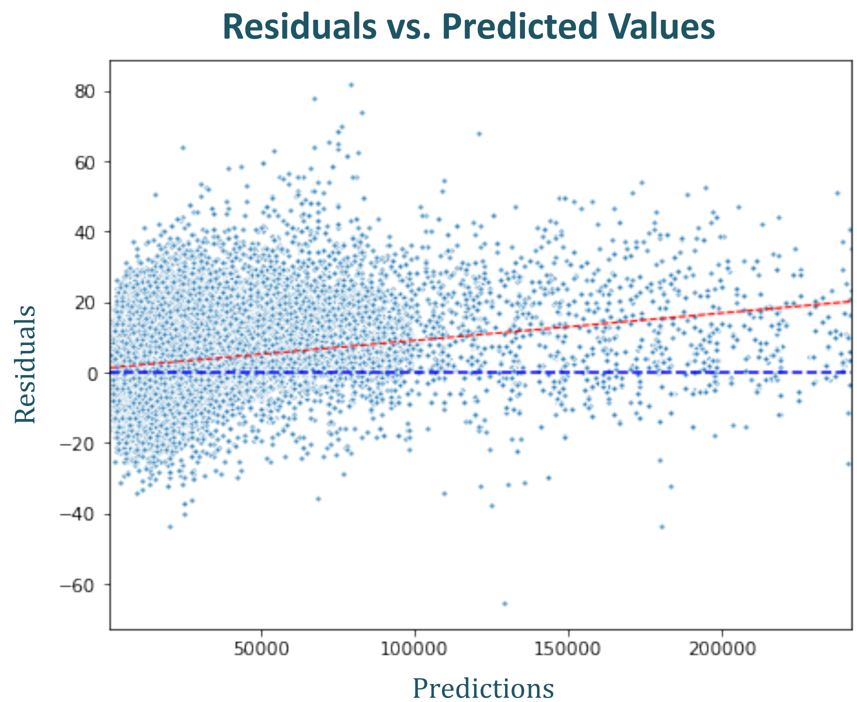
Part 1: **Assess:** identically independently distributed residuals with zero mean & constant variance

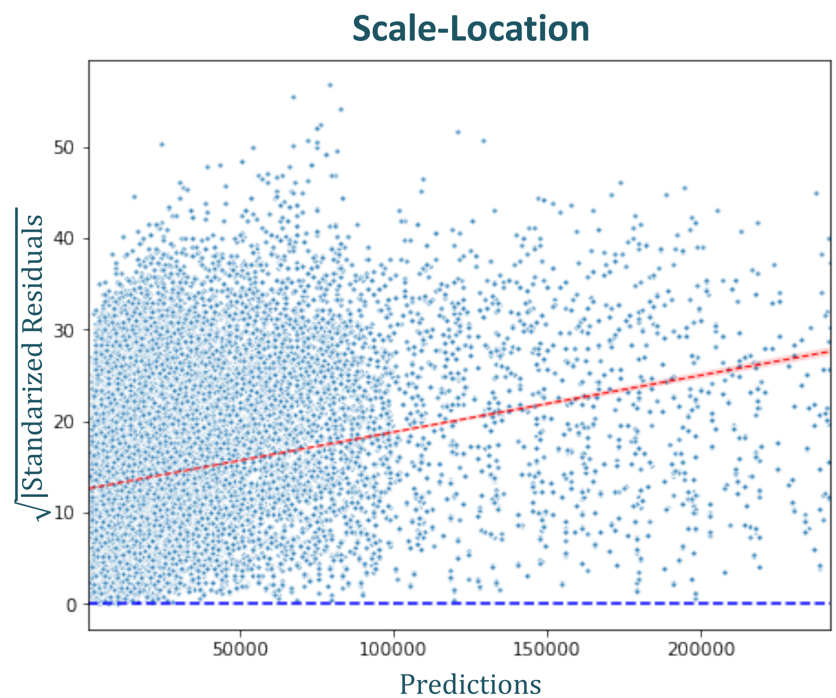
1) $E[\epsilon_i]$ = 0 **(Zero mean)**
2) ϵ_i are independent random variables **(Independence)**
3) Constant variance: $Var(\epsilon_i) = \sigma^2$ **(Variance homogeneity)**
4) Normally distributed (assumption **not required** though ideal)

→ Visualizations for screening & Statistical tests

Part 2: **Characterize:** the distribution of error to gain insight on prediction accuracy

PART 1: REGRESSION ASSUMPTIONS





- No “strong” signs of **Heteroskedasticity** (i.e. if “funnel-shape” pattern signs of non-constant variance)
- No signs of violations against **Independence assumptions** (random scattering above and below 0)
- Signs of clustering (lower values of predictions) – confirm via **statistical tests**

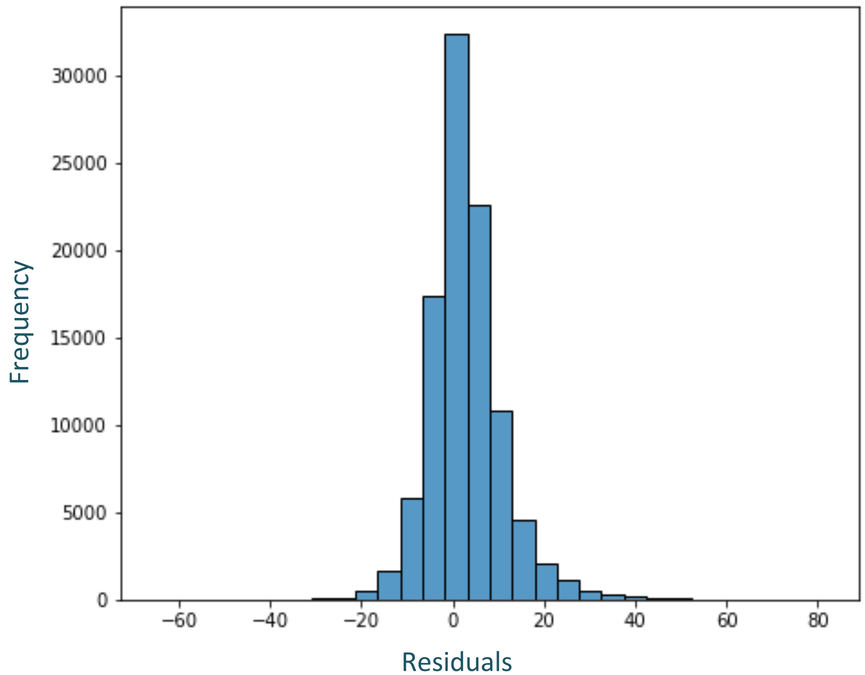
1) Leven test for Heteroskedasticity:

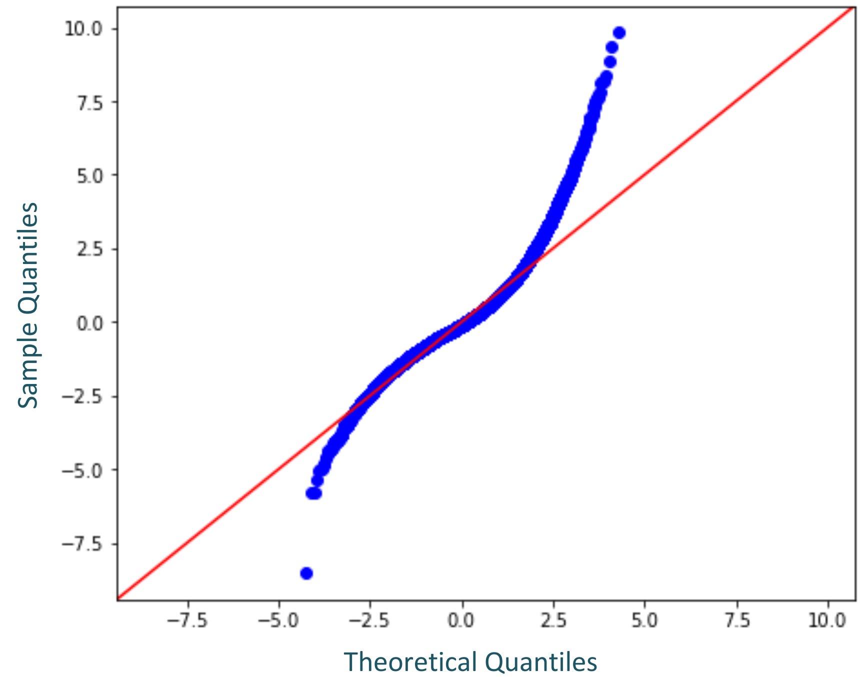
- **Fail to reject the null hypothesis** of variance homogeneity at a **5%** significance level (**p-value** = 0.1009)

2) Variance inflation factor (VIF): quantifies the correlations between the model variables:

- No strong evidence of multicollinearity

PART 1: REGRESSION ASSUMPTIONS

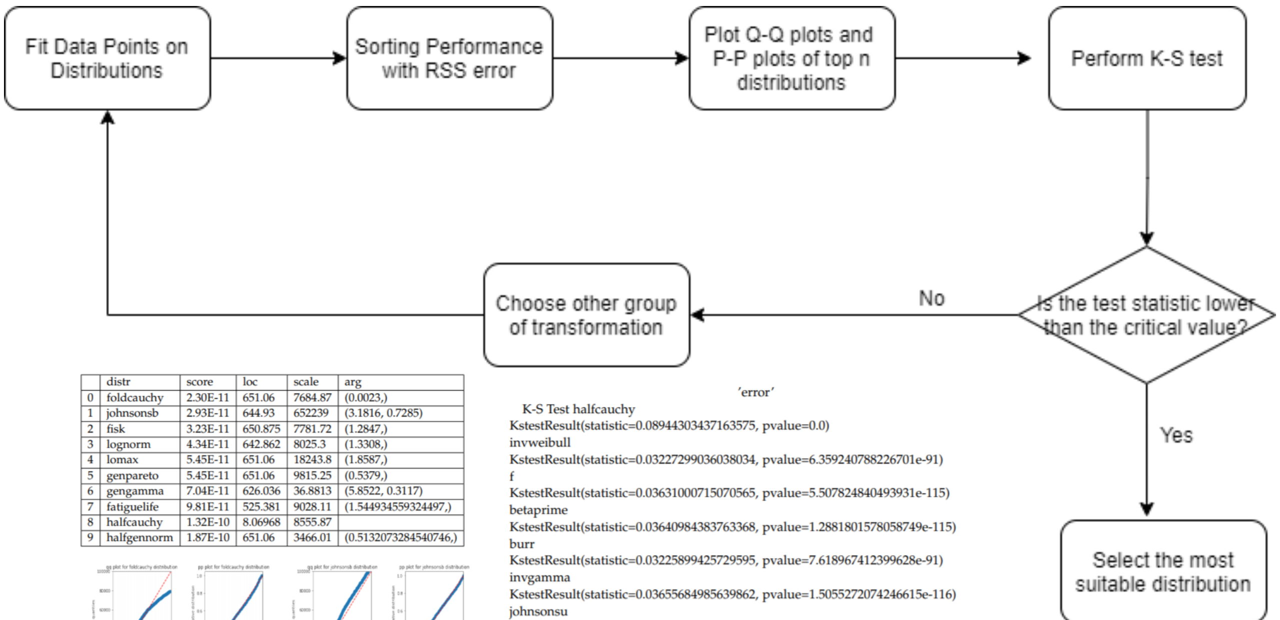




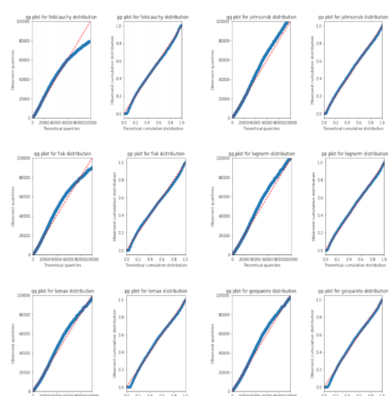
- **Normality assumption** is arbitrary but ideal
- If **residuals are normally distributed** - it makes interpretation and mathematical derivations more convenient
- Based on initial observations – the residuals are **not** normally distributed
- An accurate error distribution is **essential** – to compare predictive powers of the model or (fat tails) of target

PART 2: DISTRIBUTION FITTING

HOW DO WE FIT THE ERROR DISTRIBUTION?

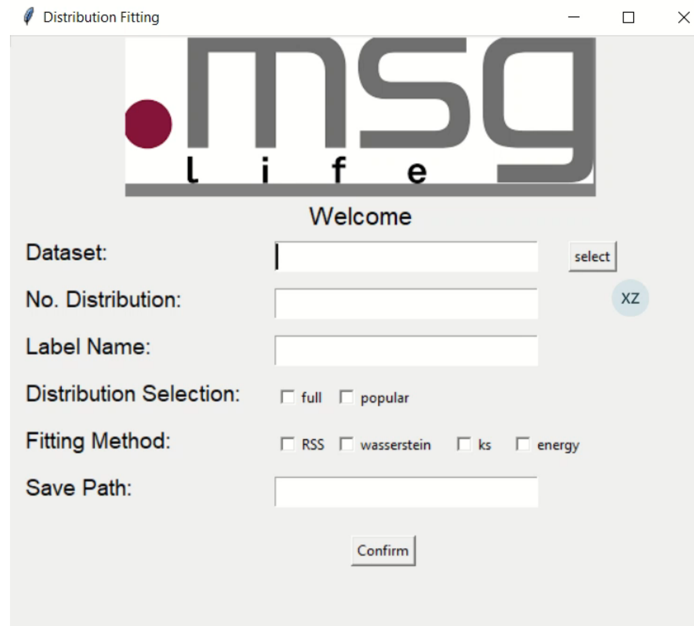


	distr	score	loc	scale	arg
0	foldcauchy	2.30E-11	651.06	7684.87	(0.0023,)
1	johnsonsb	2.93E-11	644.93	652239	(3.1816, 0.7285)
2	fisk	3.23E-11	650.875	7781.72	(1.2847,)
3	lognorm	4.34E-11	642.862	8025.3	(1.3308,)
4	lomax	5.45E-11	651.06	18243.8	(1.8587,)
5	genpareto	5.45E-11	651.06	9815.25	(0.5379,)
6	gengamma	7.04E-11	626.036	36.8813	(5.8522, 0.3117)
7	fatiguelife	9.81E-11	525.381	9028.11	(1.544934559324497,)
8	halfcauchy	1.32E-10	8.06968	8555.87	
9	halfgennorm	1.87E-10	651.06	3466.01	(0.5132073284540746,)



'error'

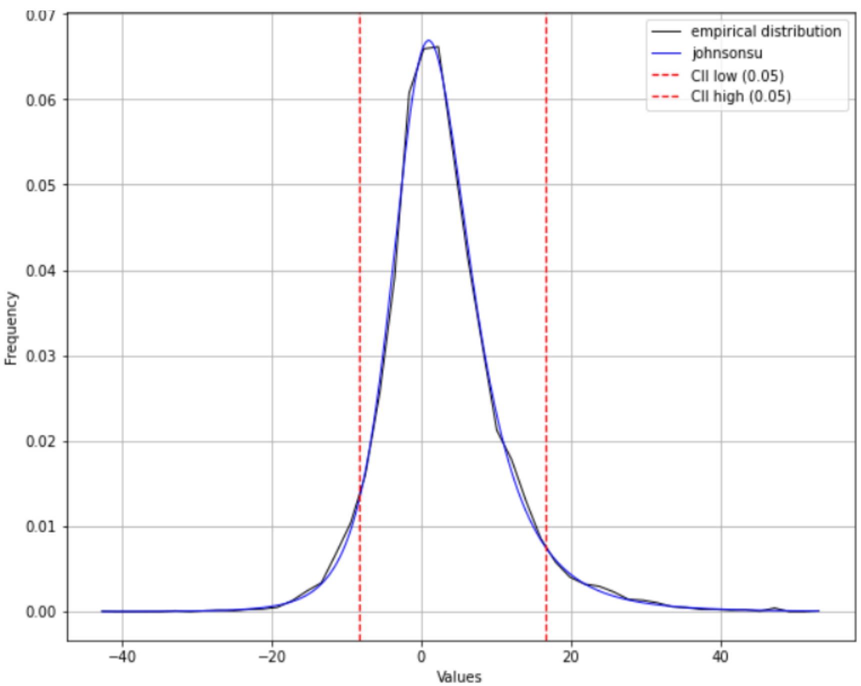
K-S Test halfcauchy
KstestResult(statistic=0.08944303437163575, pvalue=0.0)
invweibull
KstestResult(statistic=0.03227299036038034, pvalue=6.359240788226701e-91)
f
KstestResult(statistic=0.03631000715070565, pvalue=5.507824840493931e-115)
betaprime
KstestResult(statistic=0.03640984383763368, pvalue=1.2881801578058749e-115)
burr
KstestResult(statistic=0.03225899425729595, pvalue=7.618967412399628e-91)
invgamma
KstestResult(statistic=0.03655684985639862, pvalue=1.5055272074246615e-116)
johnsonsu
KstestResult(statistic=0.013754513174349059, pvalue=7.308881092159986e-17)
foldcauchy
KstestResult(statistic=0.04733704658896072, pvalue=3.611165977966616e-195)
powerlognorm
KstestResult(statistic=0.01806752912046075, pvalue=8.708984295765239e-29)
invgauss
KstestResult(statistic=0.016572096549770127, pvalue=2.7566963142091437e-24)
exponpow
KstestResult(statistic=0.21849460331167114, pvalue=0.0)
fisk
KstestResult(statistic=0.0334562279565314, pvalue=1.1079731661704413e-97)
loglaplace
KstestResult(statistic=0.062433335077018315, pvalue=0.0)



ERROR DISTRIBUTION

Kolmogorov–Smirnov (K-S) Test
Results of Top 3 distributions:

Distribution	Test statistic
Johnson SU	0.0083
t	0.02600
Double Gamma	0.0300



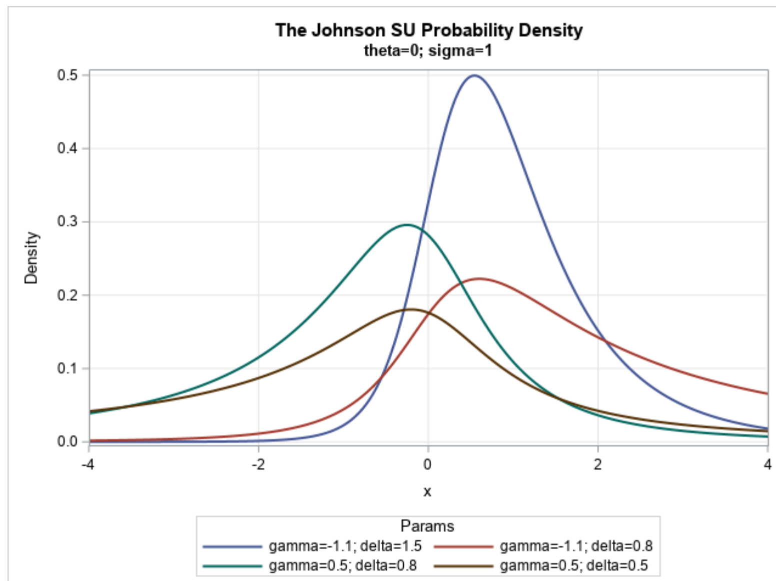
Probability Density Function of Johnson SU distribution:

$$f(y, a, b) = \frac{b}{\sqrt{y^2+1}} \varphi(a + b \log(y + \sqrt{y^2+1}))$$

Where y = $\frac{x-loc}{scale}$, $\boldsymbol{\varphi}$ is the pdf of a normal distribution and **a,b** are the shape parameters.

Johnson SU distribution: Flexible

It deals with different skewness and kurtosis



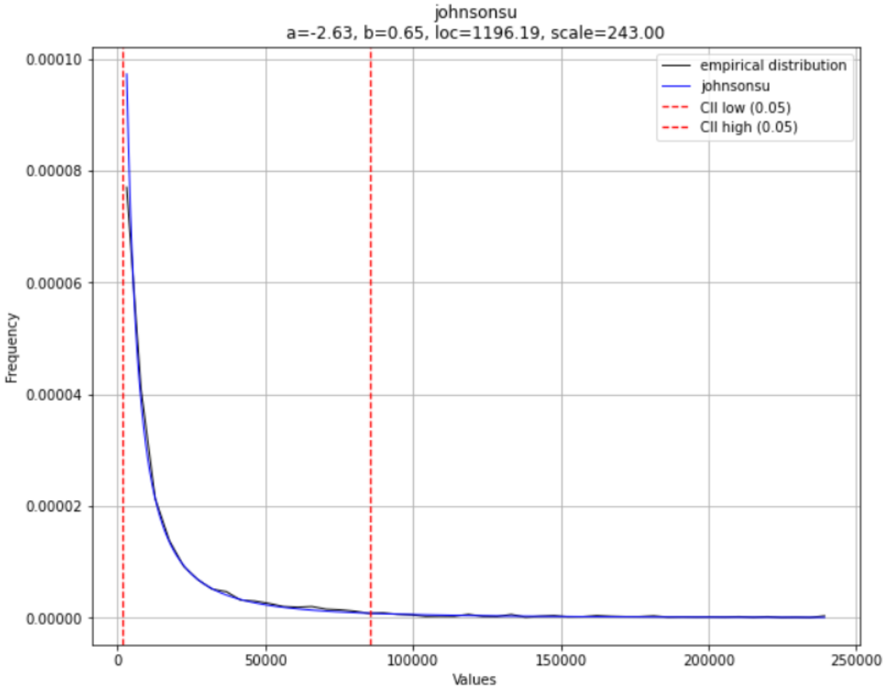
[1]

DISTRIBUTION FITTING

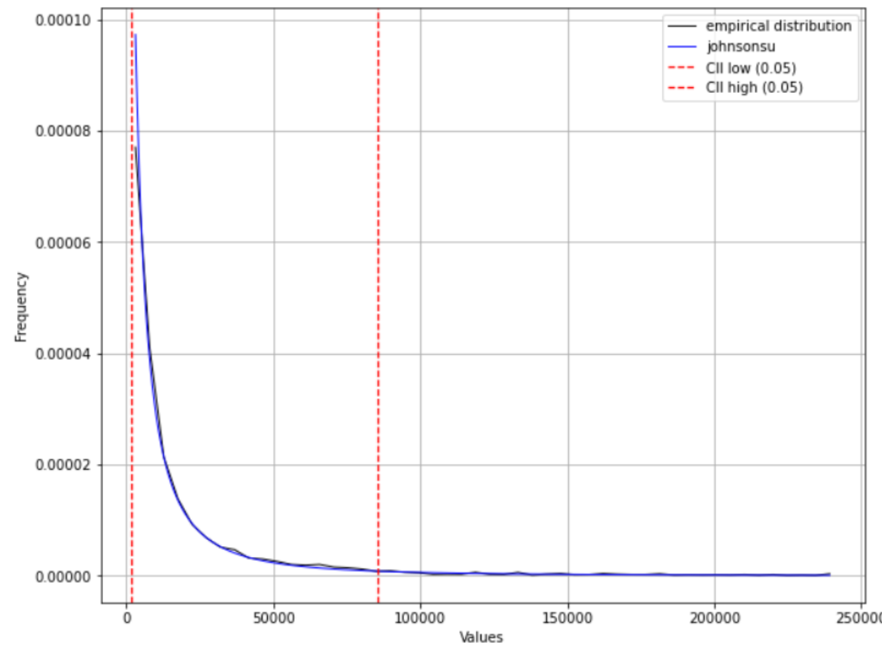
DISTRIBUTION OF THE PREDICTED VALUES

K-S test Results of Top 3 Distributions:

Distribution	Test statistic
Johnson SU	0.01635
Inverse Gaussian	0.01682
Power Lognormal	0.01881



Fitting the Johnson SU distribution on Target values:



Two-Sample K-S Test Statistic

(sample size = 10000):

0.0159

CHATPER 2: ANOMALY DETECTION

INTRODUCTION	**Overview:** Situation, Insights & Strategy, Questions & Goals and Tasks
APPROACH	**Outline**: Comparison of Methods (Previous Studies) and Challenges
METHODS	**Overview**: Implementations (**3 - step approach)**
RESULTS	**Summary:** all algorithms *(Abby & Xiaoyu)*

INTRODUCTION: OVERVIEW

STRATEGY

Anomaly Detection: identify **deviating patterns** (i.e. ‘outliers’)

- Do not fulfill **expectations**
- Significant **impact** on conclusions drawn
- Accounting for outliers ensures **stable** findings

Algorithms integrated in today’s applications:

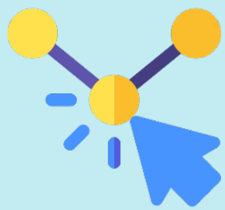
- Requires: **high accuracy, high detection performance, with fast execution.**
- For example: credit card fraud analytics, network intrusion detection, etc.

TASKS & GOALS

1. Identify ‘**infrequent**’ and evidently ‘**different**’ **instances**, given the distribution
2. Ensure that proposed methods can also **predict and identify all ‘new’ anomalies** – given a new dataset.

OVERVIEW: APPROACH

1. Definition of Anomalies



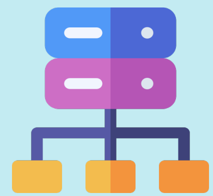
What is exactly is an 'Anomaly'?

Outlier Detection Methods

- Identify extreme events; via statistical and outlier detection methods

Challenges:
- Different operational definitions
- Parametric Methods
- Large Sample Size

2. Anomaly Detection Algorithms



Which algorithm to implement?

Novelty Detection Methods

- Unsupervised, Supervised or Semi-Supervised
- Clustering, Classification or Outlier Ensemble Methods

Challenges:
- Different operational definitions
- Parameter choice (bias)
- Dependency on 'ground truth'

3. Performance Measure

How to evaluate the methods?

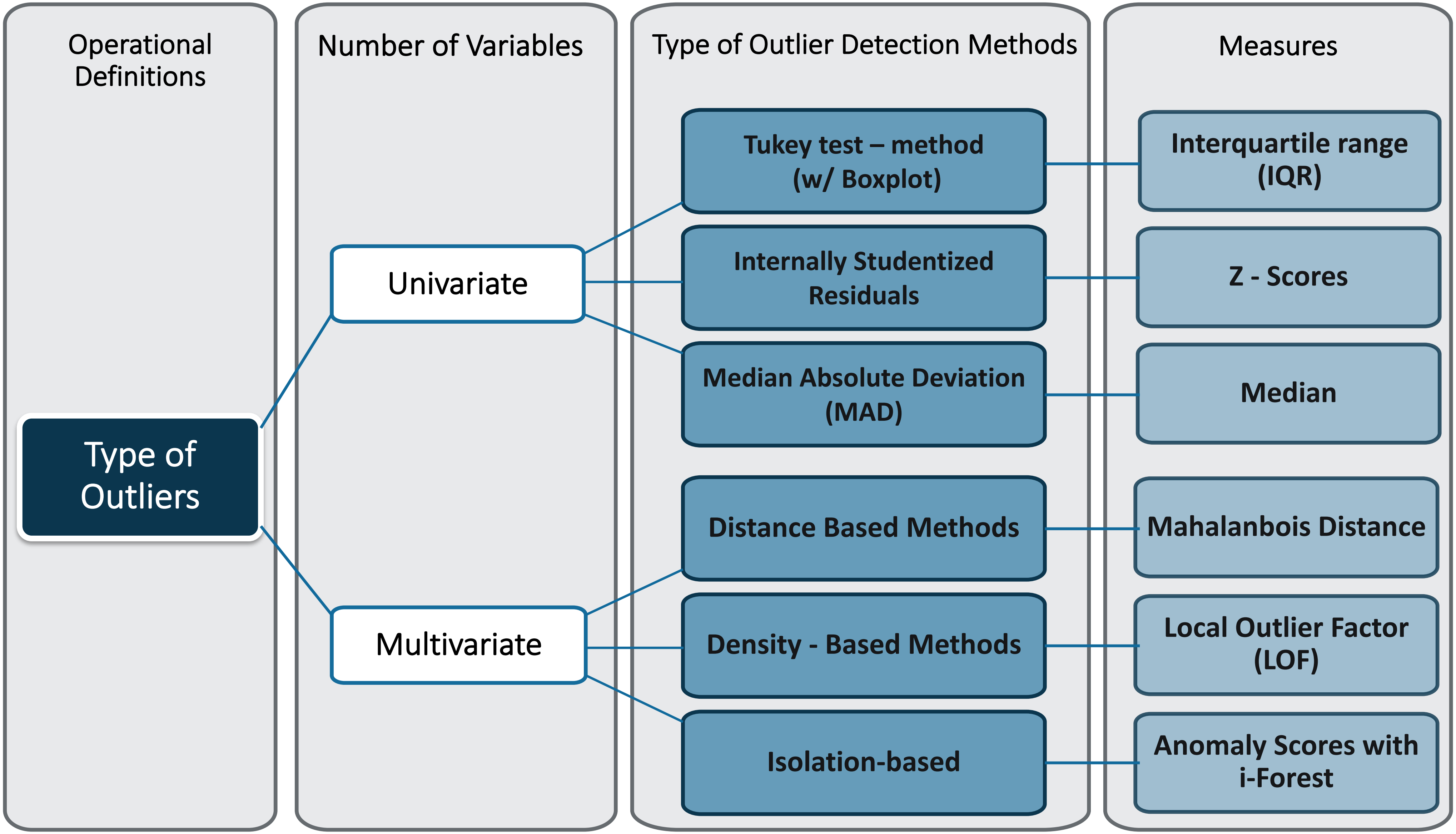
Performance Measures

- No universal "good" benchmark -but use a `standard' performance measure
- AUC, median-AUC, and Average Precision

Challenges:
- Influential Factors
- Dependency on benchmarks

GOAL: Anomaly detection algorithm with high accuracy & detection performance, and fast execution

OVERVIEW: METHODS INVESTIGATED



OVERVIEW: SELECTED METHODS

UNIVARIATE OUTLIER DETECTION

TUKEY'S RANGE TEST

- **Outliers**: residuals, predicted, & target values (individually)
- Detection via **Boxplots** (visually) & **Interquartile range** - IQR (quartiles)
- **Extreme values** lie outside of:

(1) **Inner fence**: [Q1−1.5∗IQR, Q3 + 1.5∗IQR] (2) **Outer fence**: [Q1−3∗IQR, Q3 + 3∗IQR]

MULTIVARIATE OUTLIER DETECTION

MAHALANBOIS DISTANCE

- **Training Data-set Only:** outliers considering in **errors, target, prediction** only
- Potential outlier if **large Mahalanobis** distance from the distribution

$$\boldsymbol{D}(\boldsymbol{X},\boldsymbol{\mu}) = \sqrt{(X - \mu)^T \Sigma^{-1} (X - \mu)}$$

ISOLATION FORREST (I-FOREST) DETECTION

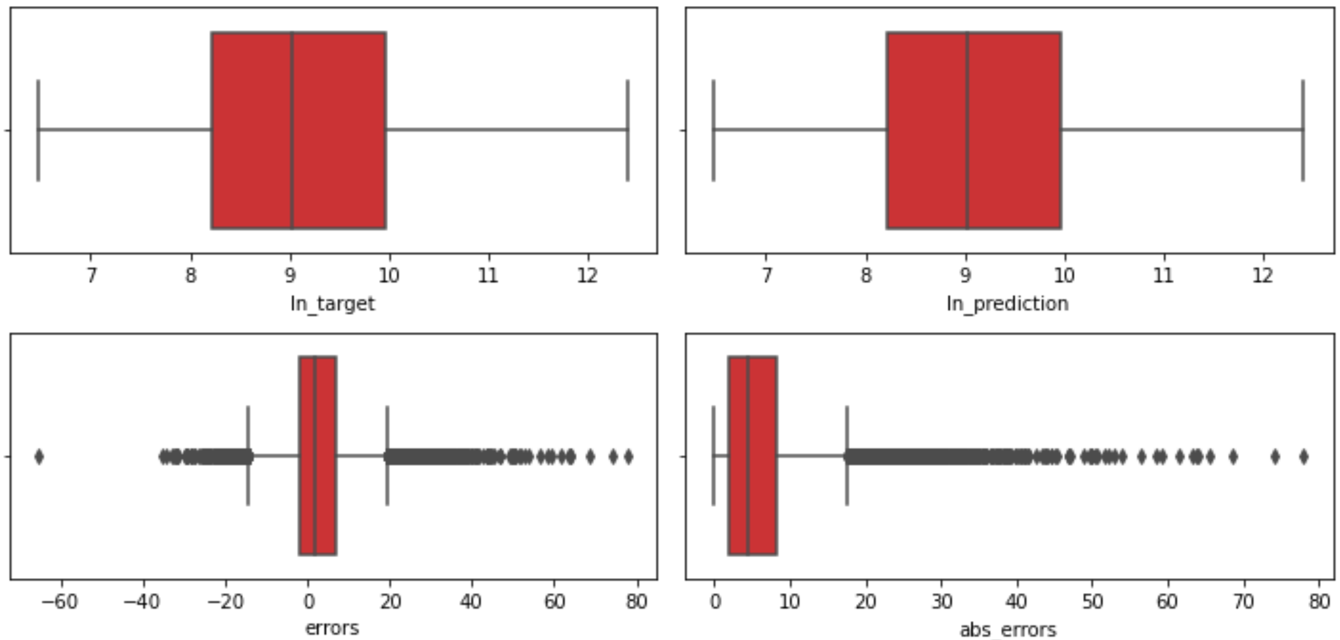
ANOMALY SCORES WITH I-FOREST

- **3 Datasets:** training, validation and test subsets
- **Main Idea:** **isolating anomalies** is an easier task compared to isolating the normal instances

RESULTS: TUKEY'S RANGE TEST

UNIVARIATE OUTLIER DETECTION

Type of Outliers	Sample Size	No. of Outliers: (log) Target	No. of Outliers: (log) Prediction	No. Of Outliers: Errors
Probable Outliers	25,000	0	0	578
Possible Outliers	25,000	0	0	1885

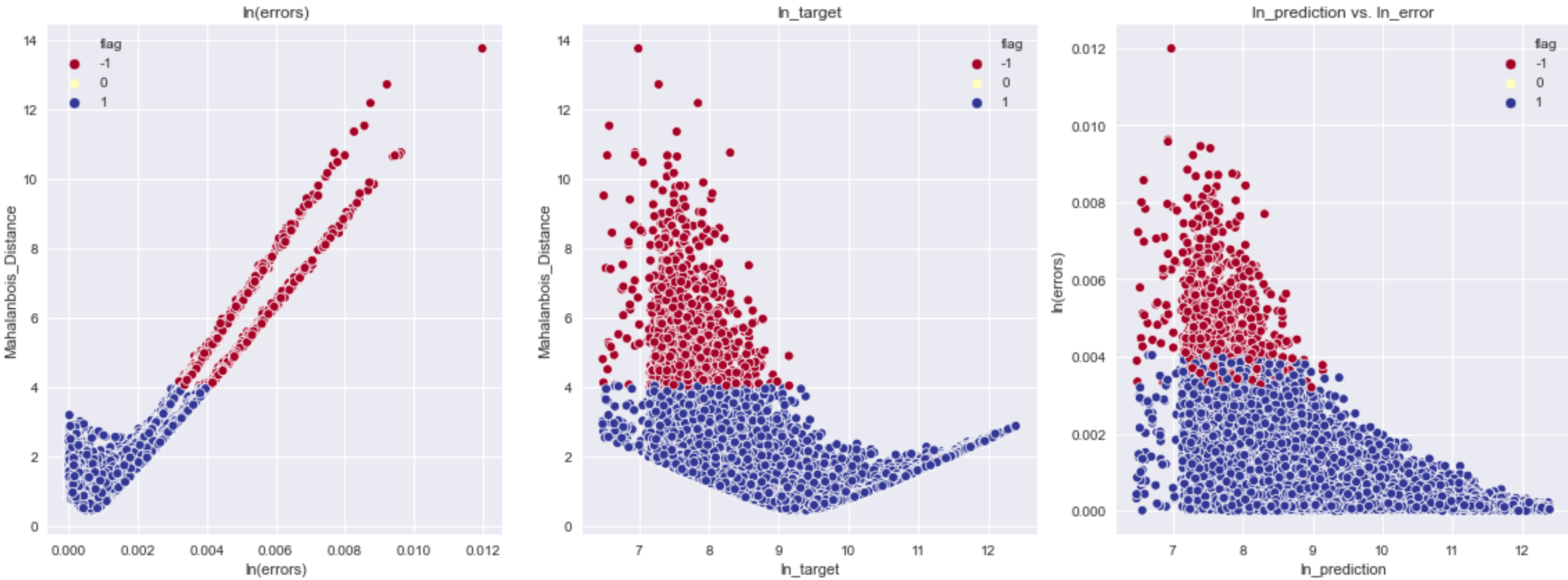


Tukey method extended to the log-IQ method

- **Zero outliers** : (log) target & prediction values
- **Errors & Absolute error :**
    - 7.54 % contamination rate for **possible outliers** (1885 outliers)
    - **578 Probable Outliers:** contamination rate, at 2.31%
    - Outliers detected had a mean for prediction (log) values at 7.78

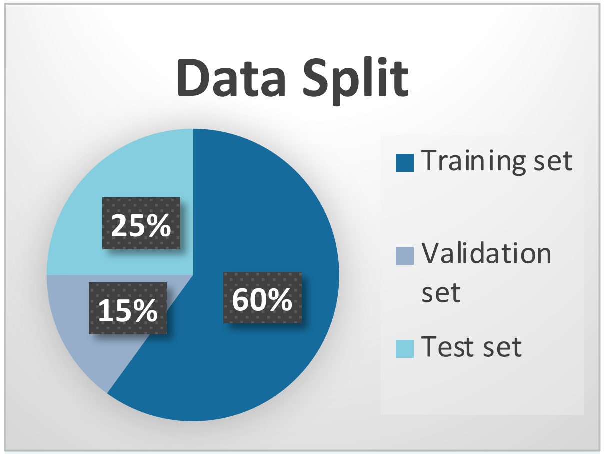
Indicates: we need to account for outliers in **lower predicted values** - located between the minimum and the first quartile

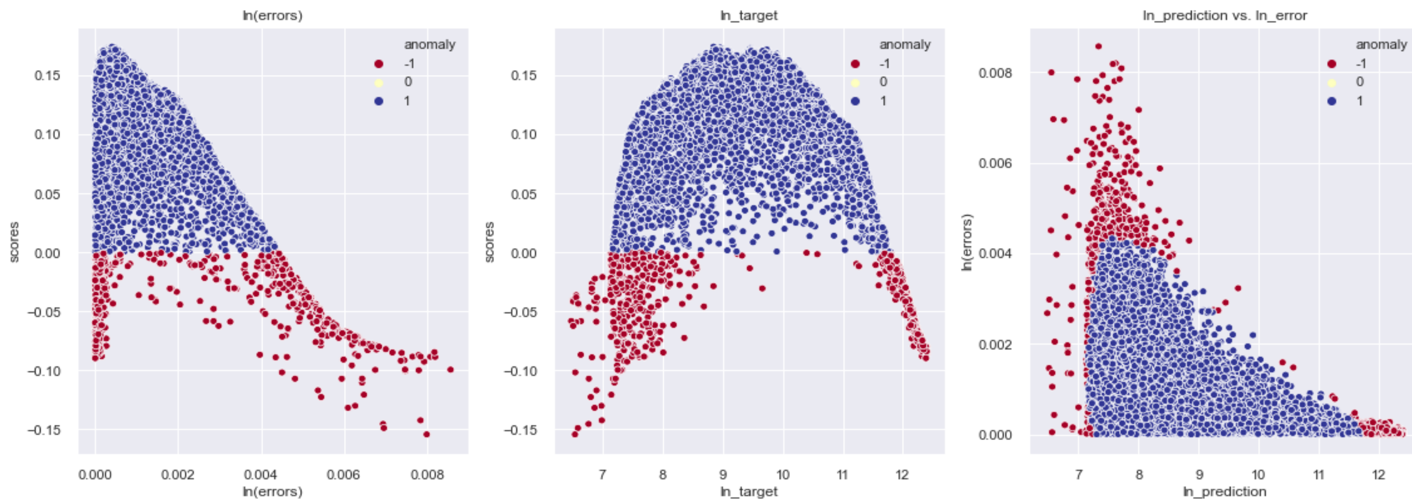
RESULTS: MAHALANOBIS DISTANCE
MULTIVARIATE OUTLIER DETECTION



- **60,000 instances,** total of **1584 multivariate outliers** observed
- Similar **contamination rate** to the univariate outliers detected at **2.64**%;
- Instances with over 4.03 **Mahalanobis distance** flagged as outliers
- **Low observed response** values are more prone to being underestimated or overestimated by the model

RESULTS:ISOLATION FORREST
ANOMALY DETECTION METHODS



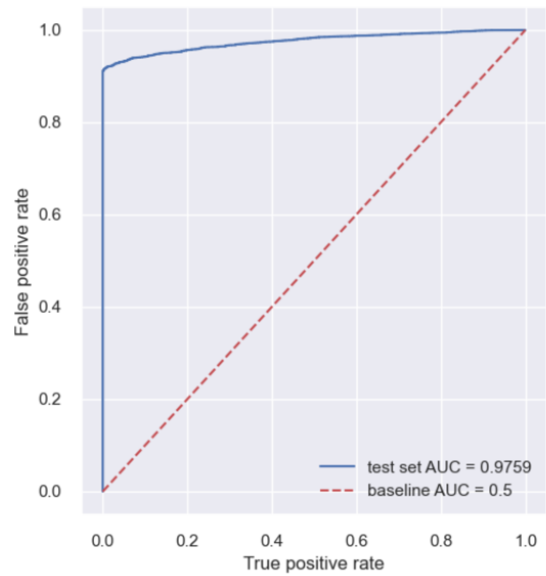


- **Training set**: Build a model
- **Validation set**: Verify the performance of the model
- **Test set**: Final evaluation of the model

- **60,000 instances,** total of **2820** multivariate outliers detected
- **Outliers** cover all range of error values
- Outliers **predicted** have either very high or low target values
- Our model could predict 66% outliers calculated by Mahalanobis distance
- **Precision** is relatively low: about 37%

PERFORMANCE & EVALUATION
ANOMALY DETECTION METHODS

IN THE TEST SET



- **25,000 instances,** total of **1212** multivariate outliers detected
- **Model** could predict 441 out of 644 outliers calculated by Mahalanobis distance (i.e. 68% Recall)
- **Precision** is about 53%
- **High AUC:** 0.9766
- **Relative high $\boldsymbol{F_1}$ score:** 0.6

CHATPER 3: RISK ANALYSIS

INTRODUCTION	How to find the trustful 95% VaR and CVaR?
LOSS & DATA	Loss Functions: absolute loss, percentage loss and logarithm loss.
	Two perspectives: whole dataset and breakdown according to 3 risk classes.
METHODS	Parametric Method, Historical Simulation, Bootstrap, and Extreme Value Theory.

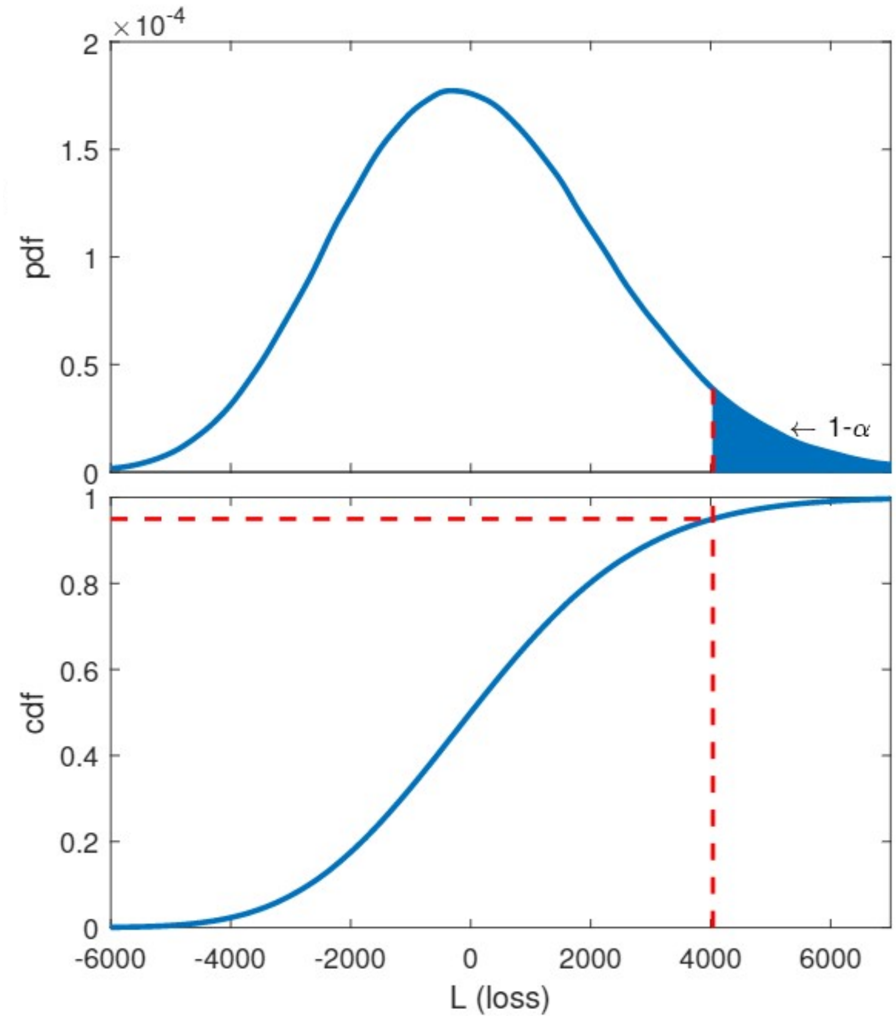
VALUE AT RISK - VaR

- Given the loss L and a confidence level $\alpha \in (0,1)$, VaR is given by the smallest number x such that the probability that the loss exceeds x is not larger than $1-\alpha$.

EXPECTED SHORTFALL- CVaR

- "If things do get bad, what is the expected loss?
- CVaR is the expected loss given that the loss is greater than the VaR.

- We use 95% VaR and CVaR as risk measurements.



LOSS FUNCTIONS AND DATA

LOSS FUNCTIONS

- $\boldsymbol{Absolute\ loss} = |target - prediction|$
- $\boldsymbol{Percentage\ Loss} = \dfrac{absolute\ loss}{target}$
- $\boldsymbol{Logarithm\ loss} = |\ln(target) - \ln(prediction)|$

DATA

- Whole dataset
- Splitting according to risk classes: high, middle, and low risk class
- To explore if the 95% VaR and CVaR of these 3 risk classes vary dramatically.

METHODS

PARAMETRIC METHOD

- Fit loss into different distributions and find the top 3 distributions.
- Take 95% percentile of the distribution as the 95% VaR.

HISTORICAL SIMULATION

- Find 95% worst loss of the historical loss as 95% VaR.

EXTREME VALUE THEORY

- The threshold is set as the **95%** percentile of the historical loss.
- **β** and **h** are the scale and shape of the best GPD fit. q is the confidence level (e.g. 95%).

$$VaR = u + \frac{\beta}{h} * \left[\left(\frac{N * (1-q)}{K} \right)^{\{-h\}} - 1 \right]$$

$$CVaR = \frac{VaR + \beta - h * u}{1 - h}$$

- The probability that the actual loss will be greater than a certain value **M** can be calculated by the equation:

$$Probability(Loss > M) = \frac{K}{N} * \left(1 + h * \frac{M - u}{\beta} \right)^{\left\{\frac{-1}{h}\right\}}$$

BOOTSTRAP

- **95% confidence interval of VaR** : We resample $\boldsymbol{T}$ times and get $\boldsymbol{T}$ VaR values, and then find the 95% confidence interval of VaR (2.5% quantile, 97.5% quantile).

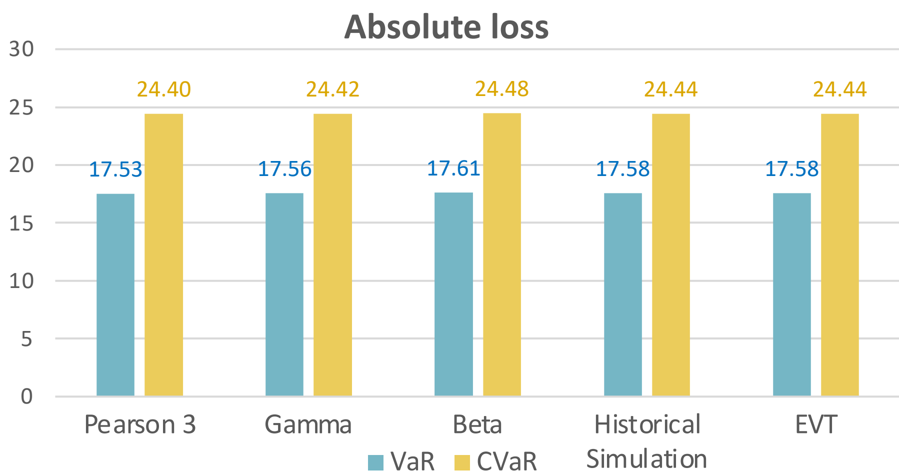
$$2Ln(LR) = 2 * \left[(T-N) * \ln\left(\frac{1-\frac{N}{T}}{1-p}\right) + N * \ln\left(\frac{N}{T * p}\right)\right]$$

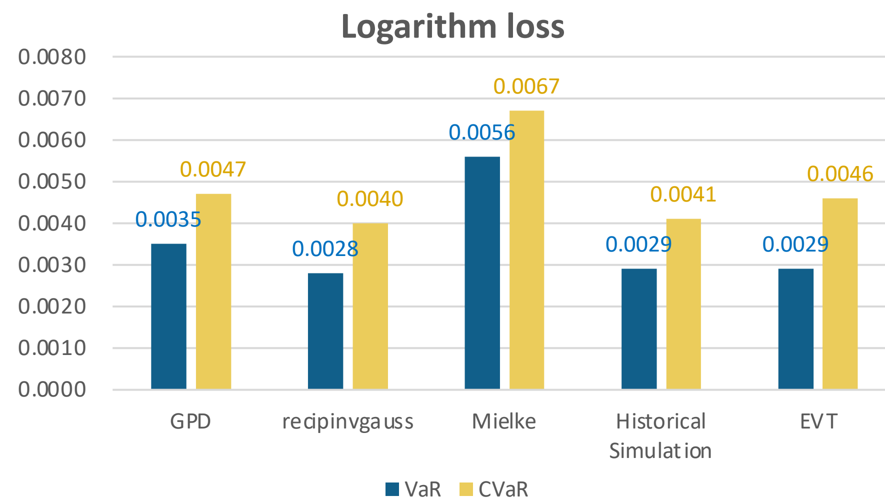
- **Accuracy Test of the mean of the interval:** $Kupiec - LR$ test
- LR is likelihood ratio. If $actual\ loss > VaR$, we denote this event by 0, Otherwise, we denote it by 1.
- N is the number of Event 0. $1 - P$ is the confidence level of VaR. $\boldsymbol{T}$ is total number of events.

- For $p = 0.05$, if $2\ln(LR) < 3.841$ → accurate
- For $p = 0.05$, if $2\ln(LR) > 3.841$ → not accurate

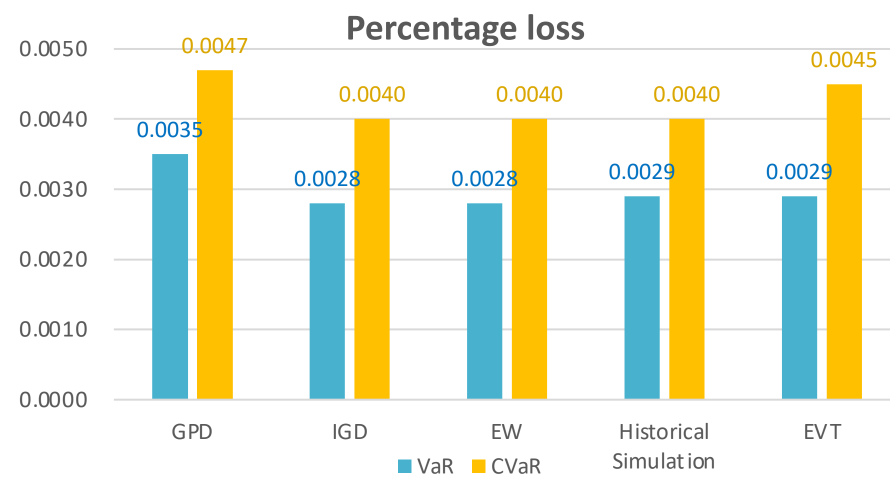
RESULTS

WHOLE DATASET



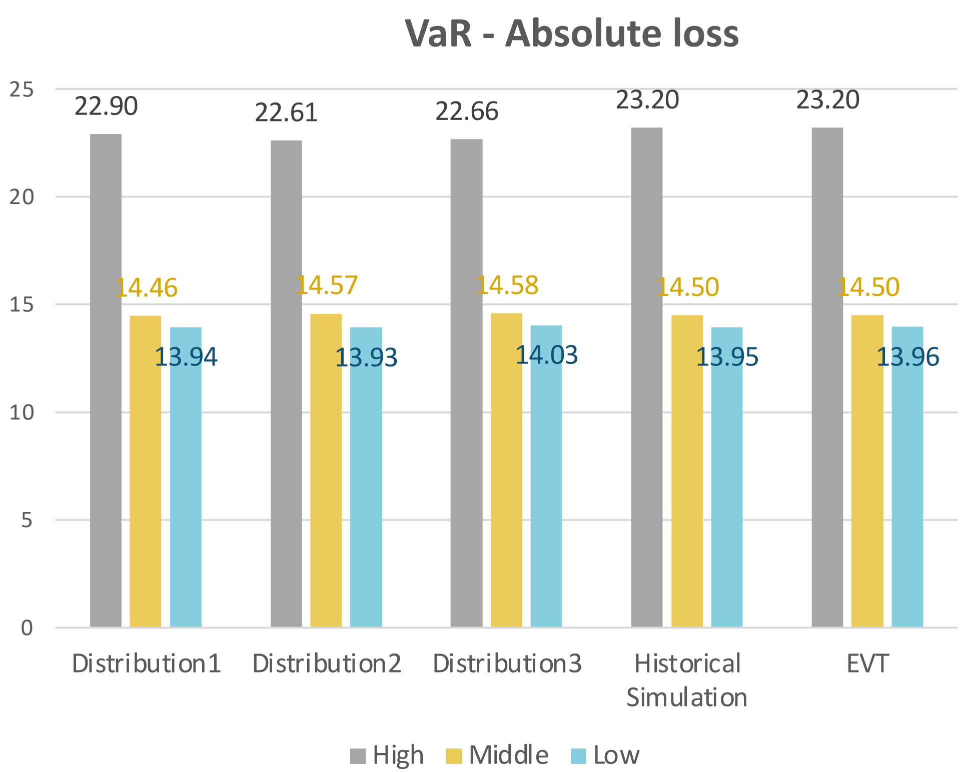


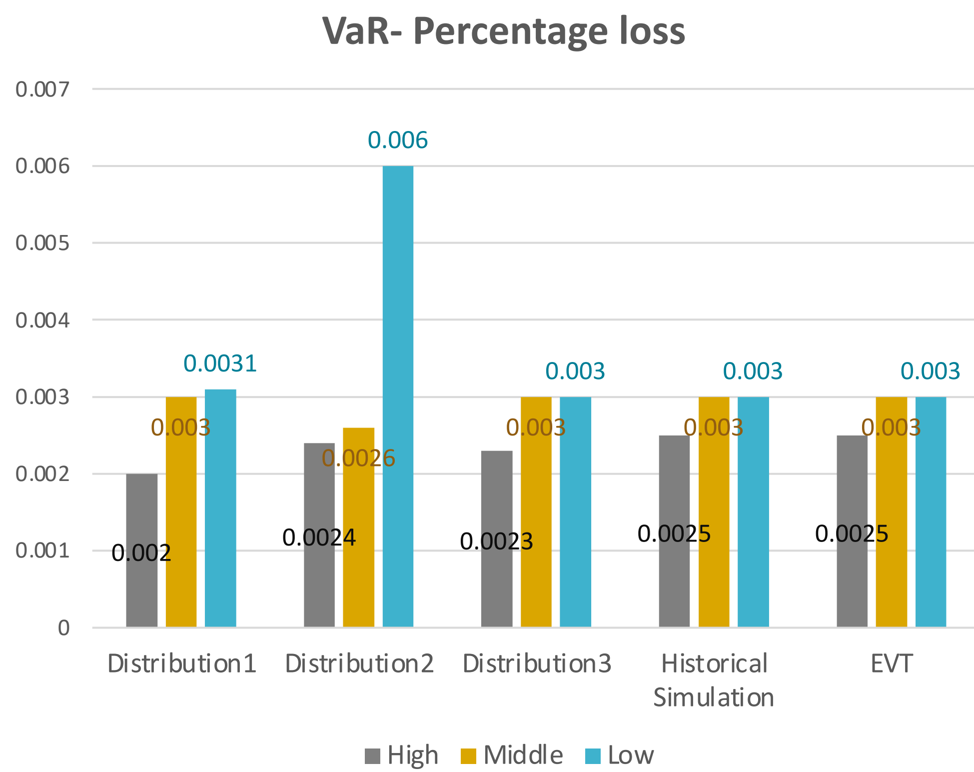
- We are 95% sure that the absolute loss of a new contract will not be greater than 17.6 EUR.
- We are 95% sure the loss of a new contract will not be greater than 0.3%.



RESULTS

RISK CLASS SPLIT





- High risk class tends to have higher absolute loss, but lower percentage loss

CONCLUSION & RECOMMENDATIONS

EVALUATION OF REGRESSION MODEL

- The **efficiency** of the process is low since we tried to find which distribution in our list may fit the data
- All distributions we fit can be **grouped** in different categories

ANOMALY DETECTION

- **Presence of outliers** has a significant impact on the conclusions drawn
- Further research other various outlier detection methods to detect **all types** of outliers (Unsupervised or Semi-supervised methods)

RISK ANALYSIS

- **Higher target** implies higher absolute loss, but the percentage loss could be lower.
- The **95% VaR** of a new contract disregarding of risk class is around 17.5 EUR or 0.3%.

RESOURCES & CITATIONS

[1] Rick Wicklin,The DO Loop, Statistical programming in SAS with an emphasis on SAS/IML programs

[2] Didit Budi Nugroho, Tundjung Mahatma, and Yulius Pratomo. Garch models underpower transformed returns: Empirical evidence from international stock indices.Aus-trian Journal of Statistics, 50(4):1–18, 2021.

[3] Christophe Leys, Marie Delacre, Youri L Mora, Daniël Lakens, and Christophe Ley.How to classify, detect, and manage univariate and multivariate outliers, with emphasison pre-registration.International Review of Social Psychology, 32(1), 2019.

[4] Hamid Ghorbani. Mahalanobis distance and its application for detecting multivariateoutliers.Facta Univ Ser Math Inform, 34(3):583–95, 2019.

[5] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In2008 eighth ieeeinternational conference on data mining, pages 413–422. IEEE, 2008

[6] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. Mathematical finance, 9(3):203–228, 1999.

[7] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-atrisk. Journal of risk, 2:21–42, 2000.

[8] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The annals of mathematical statistics, 9(1):60–62, 1938.

[9] Alexander J McNeil. Extreme value theory for risk managers. Departement Mathematik ETH Zentrum, 12(5):121–237, 1999.

THANK YOU!

Questions?