TECHNICAL UNIVERSITY OF MUNICH

Department of Mathematics

TUM Data Innovation Lab

msg life

"Analyzing the Goodness of Fit of Regression Models"

Authors	Abhilakha Das, Xian Jin, Xiaoyu Zhao
Mentor(s)	M.Sc. Marija Tepegjozova
Co-Mentor	Kilian Wohlleben
Project Lead	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Jul 2021

Abstract

In order to evaluate the fit of the regression model, firstly the basic regression model assumptions were checked. Since, if various assumptions cannot be fulfilled, most statistical methods and parametric methods based on normal distribution - often used to evaluate the goodness of fit - cannot be used in further analysis. To check these model assumptions, we utilize diagnostic plots (such as residual plots). Then, we try to determine the distribution of errors, targets, and predictions. We investigate various different distributions - to be further analyzed with the least residual sum of square error rates. By identifying the model's distribution, we are able to evaluate how well our model fits on data. Then we statistically validate the model distribution through the Kolmogorov-Smirnov test (KS test), visualize our findings via Exploratory Data Analysis (through the use of Quantile-Quantile plots, residual plots, etc.)- to confirm our findings on the best-fitted distribution of our regression model.

Secondly, we perform anomaly detection to analyze our extreme values and its' impact on our model conclusion. Our approach is to first determine the type of outlier we are interested in detecting based on the number of variables of interest (based on the number of variables of interest). Thus, we implement mixture of both univariate and multi-variate outlier detection methods - in order to find our threshold for possible outliers (required for predicting outliers). Then, for univariate outliers we implement the Tukey's method (using interquartile range), whereas. For multivariate analysis - we briefly discuss methods for the operational definitions, and implement the Mahalanobis distance to calculate initial thresholds for our data set. Then, based on our findings, we implement the Isolation Forrest to predict new outliers in our test sample. Lastly, we compare and assess our detection methods via the F_1 score, precision and recall.

Lastly, to find the possible loss caused by using our model, we analyze the absolute residual and percentage residual from two perspectives: Value at risk (VaR), which gives the maximal loss at a certain confidence level, and Expected Shortfall (CVaR), which focuses more on the tail risk. With these two risk metrics, a completed risk review can be generated. To obtain VaR and CVaR, four methods are used: Parametric method, Historical simulation, Bootstrap, and Extreme value theory. By comparing the VaR and CVaR values calculated in the four different methods, we are able to come up with a more accurate result. Additionally, we also tested the accuracy of the VaR value with Kupiec LR test.

Contents

List of Figures

List of Tables

1 Introduction

With over 40 years of experience, msg life's consumers depend on their exceptional industry-leading expertise and expect high standards in software solutions and products. Currently, msg life is implementing machine learning techniques-which in itself rises technical, and mathematical challenges. Thus, msg life's overall objective of this project is to ensure that the new machine-learning-based methods developed to retain their high standard of accuracy; enabling them to provide their consumers with innovative and market-competitive tools with confidence.

For this study, we focus on neural networks developed to primarily calculate insurance premiums - which can also in the future be used to calculate other key related values. The aim of this study is to 1) propose methods to measure the goodness of fit of this neural network, and essentially, on the entire sample space, 2) visualize the model fit through explanatory data analysis, and, 3) ensure every procedure proposed is reproducible and can be implemented for different data-sets and future models. Our biggest challenge here, firstly, arises from the nature of the neural networks ("the black-box"). Ensuring stable findings and accuracy in predicted values for new instances (not used during the testing and training phase) is extremely challenging to validate. Secondly, the growing sample size of the data set, and a large number of features to consider permits us from testing every different scenario, when predicting and validating premium rates. Thus, this also creates additional challenges in terms of the computational resources required.

Hence, to validate the performance and fit of the regression model - to essentially verify it's quality given the sample space - we break down our approach in three different steps (briefly describe in the next section). Firstly, we assess the overall adequacy of the model, proposing methods to find the best distribution fit (for residuals, and response variables), and, then utilizing statistical methods and visualizations to confirm our findings. Then, we analyze our extreme events or outliers by implementing different outlier detection models - to ensure stable findings, and assess the impact on our conclusions. Lastly, we investigate the financial risk and loss through various risk assessment measurements and tools - while analyzing the overall financial implications of our findings.

1.0.1 Data

Data was provided by msg life, with includes both input and output values with 100,000 observations. We consider five covariates (or independent variables) in addition to the observed response variable and its predicted values (obtained by the neural network). Thus, as the goal of the regression neural network model is to predict the premiums based on the input values; our observed dependent variable (referred to as "target") and the fitted (or predicted) dependent variable (referred to as "prediction") represents the number of premiums for the individual policyholder, i. Whereas, the covariates of interest (attributes of the policyholders, and their variable class) includes the following: 1) sex of the policyholder (binary), 2) risk class (nominal, with high, medium, low classes), 3) age of the policyholder (continuous), 4) policy years (continuous), and the 5) time the policyholder wants to pay premiums (in years, continuous). Table 6.1 provides a summary of the statistics of our whole data set, per risk class.

2 Evaluation Of A Neural Network Regression Model

To assess the goodness of fit of the neural network - through the use of statistical methods and diagnostic plots - is an essential step of a model's post-evaluation stage. As only investigating the models' predictive abilities does not provide us with full insights on the model's structural fit, estimation nor the statistical significance of various aspects of the derived model [1]. Hence, the purpose of this chapter is to analyze how the specified model structure describes the adequacy of the model and to ensure that interesting attributes of input variables (features) are not included in the 'unmodeled' portions of the model (the model's residuals). Specifically, if the model's error terms behave like "white noise" - in other words, has zero mean and constant variance and are "independent identically distributed"(i.i.d) - then we can conclude the regression model is of adequate structure. Since if there is "strong" evidence against these model properties (with respect to the error term) this indicates a type of specification bias in our model, resulting in sub-optimal accuracy, predictive powers, and questionable model fit.
Hence to test the adequacy of a neural networking regression model, we utilize: screening and detection tools (through visualizations such as normality plots), and statistical testing procedures appropriate for testing the adequacy of the model's fit.

2.1 General Setting of a Neural Network Regression Model

The goal of a neural network regression model is to formulate a function that maps a set of input variables to a quantitative variable with some unknown values (the estimate). Thus, similar to "standard" regression analysis, the goal is here to also to find functions of the given input variables (covariates) adequate for the defined task. [2]
Suppose we define the target variable as $y_i \in \mathbb{R}$, for i observations, the vector with k input variables as $x_i \in \mathbb{R}^k$ as $\bar{x}_i = (1, x_i')'$,and error terms of the model as ϵ_t. Then, the following neural network regression model can be (generally) formulated as:

$$y_i = f(x_i; \theta) + \epsilon_i \tag{1}$$

Here, generally, the function $f(x_t; \theta)$ one layer(with n hidden neurons) of a multi-layer feedforward neural network [1]. Thus, between the input and out puts (the regression estimates) - lays the hidden units of the network. However, due to the scope of this study we mainly focus on the post-evaluation stage of the model and it's regression estimates - specifically analyzing the error terms of the model [3]

2.2 Regression Assumptions

Moreover, ϵ_i is (generally) specified as $\epsilon_i = r_i$, with r_i residuals (such that the estimates of errors are calculated by the difference between the target and predicted values). Thus, we check our assumptions such that the ϵ_i terms are independent and identically distributed random variables. Additionally, we also check if these terms are normally distributed (even though it is not a necessary and an arbitrary assumption) as normality provides us with desirable properties, such as simpler interpretations and calculations for further analysis of the errors. Hence, if we find that the random variables ϵ_i are indeed normally distributed, then it satisfies the following assumptions:

1. $E[\epsilon_i] = 0$ (Zero mean)
2. ϵ_i are independent random variables
3. Has constant variance or Variance homogeneity: $Var(Y_i) = Var(\epsilon_i) = \sigma^2$

In this case, ϵ_i are independent identically distributed random variables with distribution: $\epsilon_i \sim N(0, \sigma^2)$. Which also implies that the errors contain no additional structure, and we can conclude there is no strong evidence of against the fit of the model.

Homoscedasticity. Homeoscedasticity is the assumption that the variance of the residuals or error terms is constant across all fitted values. Violations of this assumption (referred to as "heteroscedasticity") when analyzing the model not only imply that the error variance is changing (with the fitted values) but can also imply impaired efficiency and that the "standard" measurements used for determining the coefficient standard errors are also inaccurate. Additionally, heteroscedasticity makes estimating the actual standard deviation of forecast errors problematic, resulting in confidence ranges that are too large or too tight.

Confidence intervals for out-of-sample predictions will be likely to be unnecessarily tight if the variance of the errors increases with the expectation of the predictor. When estimating coefficients, heteroscedasticity may also have the consequence of giving too much weight to a small subset of the data. To detect any violations against homoscedasticity, we utilize residual diagnostic plots and analyze further dependencies among the residuals by interpreting a quantile-comparison plot [4]. Here, assumptions are violated if errors get systematically larger in one direction by a significant amount. Later, we perform Leven's test for heteroscedasticity to verify our observations.

Independence of errors and Zero Mean. If there is no evident relationship nor structural dependency between the residuals and the variables in the model - then both the assumption of error independence and zero mean hold. As violations of the independence assumption indicate the sample values may be correlated (known as "multicollinearity") with strong evidence of a dependency (or a structural pattern). Hence, even if the model is correctly specified, correlations among the residuals, with a non-zero mean may influence the model to under-predict or over-predict depending on the configuration of the covariates.

As, firstly, these patterns contain information that the regression model was unable to capture during its training on the training set, resulting in a model that is either inadequate or sub-optimal. [5]Secondly, a non-zero error mean indicates skewed model estimates. Such that, for instance, the model would underestimate with large positive error mean and overestimates large negative error means. To check these assumptions, we once again utilize residual diagnostic plots to visualize correlations between the predictors and residuals. Then we also briefly analyze the variance inflation factor (VIF) to quantify the correlations between the model variables.

Normality. As stated, though the assumption of normally distributed errors is arbitrary, leveraging quantile-comparison plots to analyze the distribution of the error estimates, the residuals, is very effective in assessing the model fit. Given quantile-comparison plots are often utilized to examine tail behavior of the model residuals, such as outliers, skewness, and light or heavy tails. This is crucial as highly skewed distributions compromise the interpretation of the model and the estimation of coefficients[6]. Additionally, interpreting quantile-comparison plots for normality also helps us interpret and possibly adjust for dependencies among the residuals[7].

2.2.1 Results & Discussion

To detect non-constant error variance, we first analyze the scatter plots of residuals and standardized residuals versus predicted value, shown in Figure 2.1. We observed the residuals are randomly disturbed almost uniformly across the y-axis with no evident "funnel-shaped" structures. Though there is a slight increase in residuals as the predicted values increase (often commonly expected) - there are no strong indications of heteroskedasticity as the relationship is relatively not extreme. We confirm these findings by performing the Levene's Test (grouping by the risk classes), where we fail to reject the null hypothesis of variance homogeneity at a 5% significance level (with p-value = 0.1009, and 5 degrees of freedom).

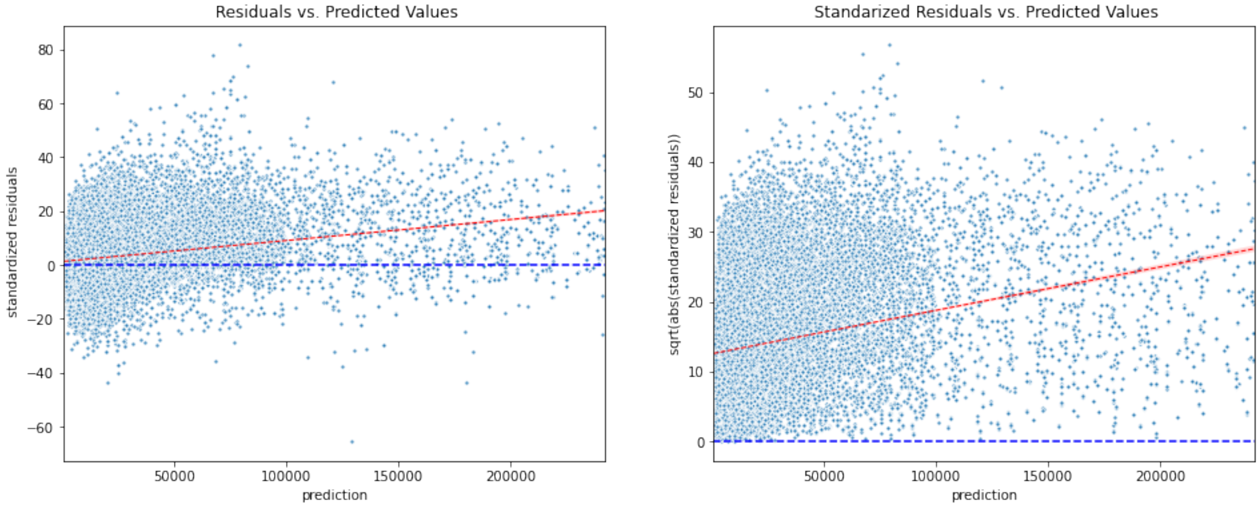


Figure 2.1: (1) Residual vs. Fitted Values Plots, (2) Squared Standardized Residuals (Scale-Location)

However, we also observe more positive values for the residuals as there is more residuals concentrated above 0, especially for small predicted values. This indicates that the residual mean may be above zero for low predicted values, implying that the model underestimates for large positive error mean. This is especially evident when we analyze the standardized residuals(2.1), and when computing the mean of the raw residuals for all observations($E(r_i) = 6.003$). Additionally, it is clear that there may a large amount of outliers which may also contribute to the skewed model estimates. To verify our results, we perform the Levine-s test of variance homogeneity (against the null hypothesis that the variances are equal across all instances). Based on our results, at 5% significance level, we concluded that we did not have enough evidence to reject the null hypothesis (with p-value 0.1054) of constant variance.

In addition to check the assumptions of the independence of errors between covariates (and to detect multicollinearity), we utilize scatter plots with residual and independent variables. Regarding age, duration of the contract, gender, and risk class, residuals randomly scattered around the horizontal-line of zero, with no apparent pattern. However, the variance of residual decreases while age is increasing, which we have to account for during further analysis. We further verify these findings through the VIF. Where we found age and the number of policy years to have VIF values between 8-10 (age with 8.42 VIF, and 9.44 the policy years).

This is expected as both variables are dependent on each other (as the number of policy years is determined by the age of the policyholder).Hence, we conclude, there is no strong evidence of multicollinearity. This occurs if our regression model differs from the actual model or we have outliers. As a result, the model will not predict optimally for many of the observations. It is recommended that either: the implications of the results should be repeated after the removal of outliers,redundant variables are removed, or re-conceptualizing the meaning of the predictor may help solve this issue. [8]

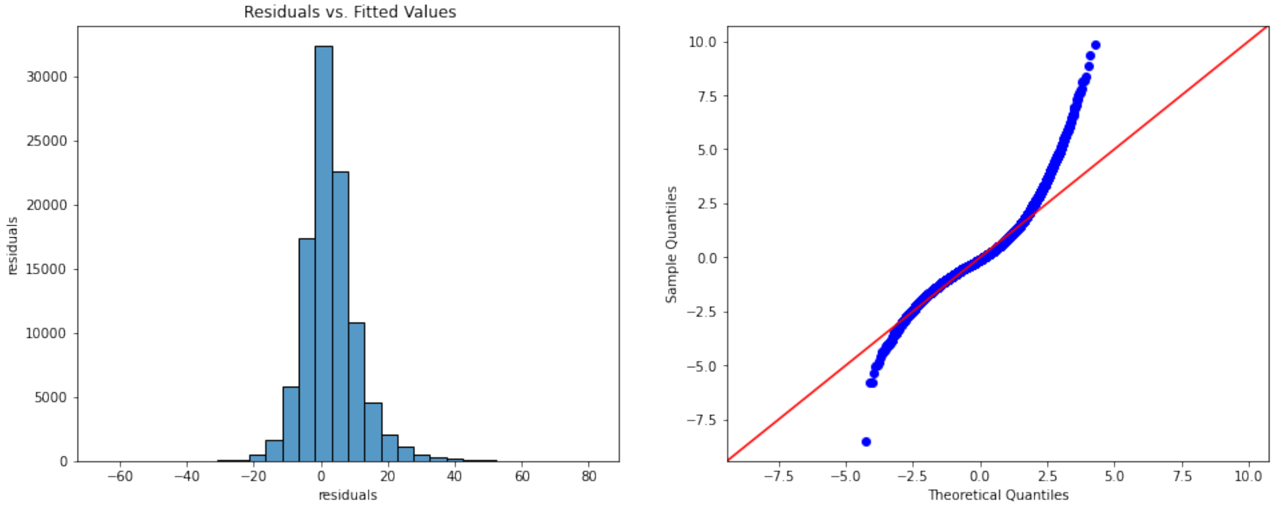


Figure 2.2: (1) Distribution of Residuals, (2) Normality Quantile-Quantile Plot

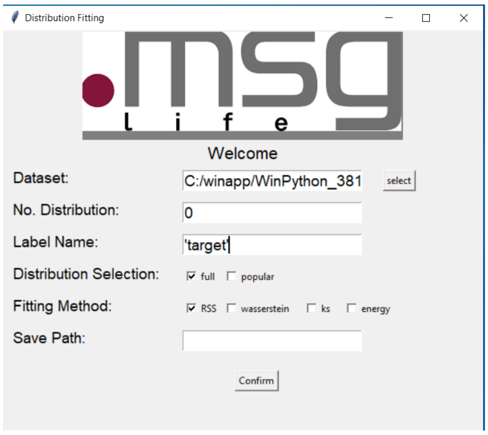
Q-Q plot and P-P plot were used to identify normality or residual. It is evident that normality does not meet since the distribution of residuals (in Figure 2.2) do not follow the normal distribution. To solve this issue, we tried applying various transformations on the predicted values (for example, logarithmic and polynomial transformations), and Box-cox transformations. However, there were still somewhat strong indications of the non-normally distributed errors. Therefore, we conclude the normality assumption cannot be fulfilled for model error terms.Thus, we have to further analyze, compare and fit different error distributions.

2.3 Distribution Fitting

Until now, we still have no clear indication regarding the underlying distribution of our variables of interest, i.e., the distribution of errors, targets, and predictions. If we know the distribution of these variables, we can assess how well the model fits. Thus, in the following section, we examine various distributions against our data and, utilize goodness-of-fit tests to determine which distribution and parameter estimates fit the given data the best.

2.3.1 Workflow of Distribution Fitting

To verify the distribution our data may follow, we hope to develop a tool to automate this process for any unseen dataset. This process is based on the python package ’distfit’. It compares the actual frequency(f) to the model’s anticipated frequency (f-hat), then computes the residual sum of squares (RSS). It returns the best-fitted distribution is returned with the corresponding parameters. To start with, we try to compare whether our data follows the eight commonly used distributions in 'distfit'[9]. These are normal distribution, generalized extreme value (GEV) distribution, exponential distribution, gamma distribution, pareto distribution, log normal distribution, weibull distribution, beta distribution, t distribution, and uniform distribution. If none of them has a good fit, we will expand to fit our data to a larger set of distributions derived from the 'scipy' library. Then, different goodness of fit methods was performed to evaluate how well the data follows each distribution. In addition, Q-Q plots and P-P plots help us visualize the goodness of fit intuitively. In order to make this process faster and to reuse the same procedure for any data set, we created a tool to automate the above process.



Welcome

Dataset: C:/winapp/WinPython_381

No. Distribution: 0

Label Name: 'target'

Distribution Selection: ☑ full ☐ popular

Fitting Method: ☑ RSS ☐ wasserstein ☐ ks ☐ energy

Save Path:

Figure 2.3: User Interface of the Tool developed

Figure 2.3 shows us the user interface of the tool we developed. Users are able to import the data set into the program and select the corresponding variable they want to fit. Then they can define the category of the distributions, the top n distributions they want to pick, and the sorting criteria of the result distributions. In the end, a PDF file will be generated including the RSS error values, Q-Q plots, and P-P plots of the top n distributions and the K-S test result of the corresponding distributions to the path user provided.

2.3.2 Goodness of Fit Measurement

In this section, different goodness of fit measurements are introduced to select the distribution our data follows. We used basic metrics, Q-Q plot, P-P plot, and Kolmogorov-Smirnov Test(K-S Test) to measure this project.

Metrics Regression metrics for evaluation are most commonly used in the goodness of fit analysis since it is easy to calculate and provides an easy way to compare different models. In this project, the residual sum of squares(RSS) was used, which is defined as in eq.2 where y_i is the actual value and $f(x_i)$ is the predicted value. It calculates the sum of the square of the difference between the true and predicted value. If RSS is low, we got a model which is close to our raw data.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 [10] \tag{2}$$

Kolmogorov-Smirnov Test (K-S Test) The Kolmogorov-Smirnov one-sample distribution test is often used to examine whether the samples come from a given distribution [11]. It is based on the empirical distribution function(ECDF). ECDF is defined as:

$$E_N = \frac{n(i)}{N}[12] \tag{3}$$

where n(i) is the number of data smaller than the value X_i and X_i is the ordered data point from smallest to largest. Then the KS-test calculated the maximum distances between two curves[13].
The test is defined by:
H_0: The data follow a specified distribution
H_a: The data do not follow the specified distribution

The Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1 \leq i \leq n}(F(X_i) - \frac{i-1}{N}, \frac{i}{N} - F(X_i))[12] \tag{4}$$

$F(X)$ is the theoretical cumulative distribution of the distribution tested. The hypothesis is rejected if the test statistic is greater than the critical value from the table, depending on the confidence interval and the sample size[12].
The advantage of using the K-S test is that the test is exact and non-parametric. Moreover, we can use the test statistic to compare the goodness of fit easily. Unlike other goodness of fit tests, the K-S Test can test whether samples follow any distribution without making some assumptions. In addition, the critical value of the K-S Test does not depend on reference distribution, which gives us a clear comparison of the goodness of fit performance.

2.3.3 Results and Discussion

In this section, methods in 2.3.2 were implemented on target and error values to investigate how they are distributed.

Fitting a distribution on error value As mentioned in 2.1, choosing a correct error distribution is critical for us to evaluate the performance of the model.
We started by finding out a suite of candidate probability distributions distributions with the least RSS values. Q-Q plots and P-P plots give an intuitive measurement of goodness of fit. To verify which distribution best fits the error estimates, the K-S Test was performed on a subset with a sample size of 10000. Table 2.1 shows the three lowest test statistic values of all distributions tested.

Table 2.1: Test statistic of K-S Test on error value

Distribution	Test statistic
Johnson SU	0.0083
t	0.02600
Double Gamma	0.0300

Moreover, the critical value of 0.05 significance level with a sample size of 10,000 can be calculated from [12] as

$$C = \sqrt{-\frac{1}{2}ln(\frac{\alpha}{2})}[14] \tag{5}$$

$$D_{crit} = \frac{C}{\sqrt{10000}} = 0.0136. \tag{6}$$

where α is the confidence interval. We fail to reject the null hypothesis of the K-S Test aganist the Johnson SU distribution fit. Therefore, we suggest that the Johnson SU distribution fits the error estimates the best. Figure 2.4 shows a comparison of empirical error distribution and the fitted error distribution.

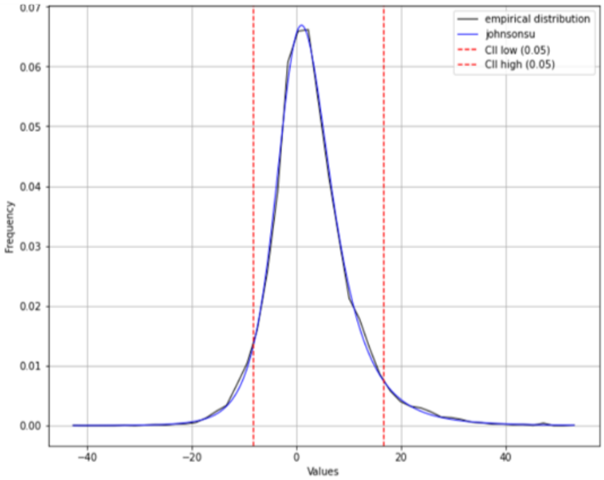


Figure 2.4: Empirical vs. Fitted error distribution

Johnson SU distribution is a transformation of the normal distribution, which can solve a skewed normal distribution with four parameters. The probability density function for johnson SU is defined as in eq.7.

$$f(y, a, b) = \frac{b}{\sqrt{y^2 + 1}}\phi(a + b * log(y + \sqrt{y^2 + 1}) \tag{7}$$

where $y = \frac{x-loc}{scale}$, ϕ is the probability distribution function(pdf) of a normal distribution and a,b are the shape parameters.

Location and scale parameters are used to transform the distribution into the standardized form. The shape parameter 'a' corresponds to the skewness of the distribution and 'b' tells the kurtosis of the data. A positive skewness corresponds to a skew-left distribution, and a negative skewness means we have a right-skewed distribution. Kurtosis is a statistical measure that defines how heavily the tails of distribution differ from the tails of a normal distribution which can only be greater than 0 by definition.

The corresponding shape parameters, location, and scale parameters were derived by maximizing the likelihood when fitting the distribution on error values with the sample size of 10000 using the python package 'stats'. The parameters estimated are (a=-0.4732,

b=1.3960,loc=-0.7811, scale=8.0959). The K-S Test indicates that the Johnson SU distribution fits our error value well. In our case, the error distribution is a light skewed-right normal distribution, which indicates that more errors are clustered around the left tail of the distribution, and the right tail is longer. In addition, we observed a positive kurtosis which indicates that the error distribution is peaker than the normal distribution, i.e., more error values are clustered around the mean value. However, it also tells us that the tail of the error distribution is heavier than the normal distribution, which indicates we have observed more error values with high absolute values, which can be a risk.
The Johnson SU distribution with four parameters provides us a highly flexible model, which can account for any conditional mean and variance in any degree of positive or negative skewness combined with positive levels of kurtosis[15]. With the help of flexible error distribution, we can capture more features of the error values to describe the data more accurately.

Fitting a distribution on predicted and target value Ten most likely distributions matched for the predicted value were sorted with the RSS value. K-S Test was performed on the top 10 distributions, and we selected three distributions with the lowest test statistic, and the results are shown in Table 2.2. The three distributions selected are the Johnson SU, inverse Gaussian, and power lognormal distribution.

Table 2.2: Test statistic of K-S Test on predicted value

Distribution	Test statistic
Johnson SU	0.01628
Inverse Gaussian	0.01682
Power Lognormal	0.01881

Figure 2.5 shows a comparison of empirical predicted value distribution and the fitted predicted Johnson SU distribution.

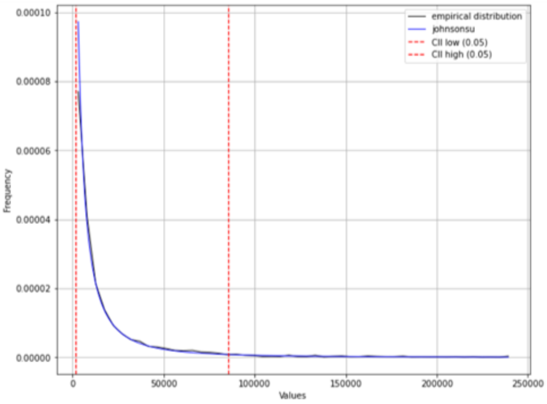


Figure 2.5: Empirical vs. Fitted predicted value distribution

The python package 'stats' estimated the corresponding shape parameters, location, and scale parameters when fitting the predicted values to a JohnsonSU distribution. The parameters are given as follows: $a = -2.6294, b = 0.6536, location = 1196.1891, scale = 243.0011$.

In this case, the distribution still has a negative skewness and a positive kurtosis parameter. It is noteworthy that the predicted value is a highly positive-skewed distribution, where most of the values are around 0. The kurtosis value tells that the distribution is not so peaky. However, we cannot accept the null hypothesis since the test statistic is greater than the critical value. On the other hand, the Johnson SU distribution is a two sided distribution between $-\infty$ and ∞. All predicted values are greater than 0. Although the Johnson SU distribution fits the predicted value well, we still have to reject that the predicted value follows the Johnson SU distribution. For further analysis, one should consider further transformations on the response variables, such as box-cox transformations etc.
To evaluate the goodness of fit of the regression model, people often try to compare whether or not the observed value distribution is consistent with the predicted value distribution by the goodness of fit tests[16]. The two-sample K-S Test was used to detect whether the target value and the predicted values are consistent. The test statistic for the two-sample K-S Test is 0.0159. Furthermore, the critical value of a two-sample K-S Test with a sample size of 10000 and a significance level of 0.05 is:

$$D_{crit} = 1.36\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.36\sqrt{\frac{10000 + 10000}{10000 \cdot 10000}} = 0.0192.[12] \tag{8}$$

Since the test result is lower than the critical value, we conclude that the regression model provides good predictions of the target values.

3 Anomaly Detection

3.1 Introduction

As the primary goal of this project is to ultimately investigate our methods to measure the goodness of fit over the entire sample space (in addition to visualizing the fit via explanatory data analysis); it is crucial we create procedures to analyze, and, more importantly, detect the presence of outliers (or 'anomalies') in our data set. As anomaly detection identifies certain deviating patterns given the distributions of the variables of interest - it allows us to further assess instances that do not fulfill our expectations and ensures stable findings.
In recent years, anomaly detection has played a significant role in a variety of fields and has been deemed crucial in several application domains [17]. The anomaly detection algorithms implemented in these types of applications (such as network intrusion detection, credit card fraud analytics, and fault detection) are not only essential - but they also require the corresponding algorithms to have the following key properties: high accuracy, and high detection performance with extremely fast execution.
Given these requirements, our primary goal in this chapter is to propose anomaly detection algorithms and methods for any given msg data set (including new 'unseen' and existing data). The key motivations of previous and existing comparative studies in anomaly detection can often be categorized into addressing the following topics: 1) explicitly defining anomalies 2) comparison of existing and new anomaly detection methods, and 3) addressing the lack of 'good' benchmarks and performance measures. [18] To achieve this - we must first ensure our proposed algorithm can identify the outliers in our dataset that are 'infrequent' and evidently 'different' from the assumed distribution. Then, secondly, ensure that our

proposed methods also can predict and identify all 'new' anomalies or outliers given a new dataset. [19]
Therefore, to design an algorithm that can complete the given tasks above with high accuracy, high detection functionality with fast execution - requires a three-step process with different methodologies (with the corresponding major challenges, comparable to previous studies, and the methodology):

1. **Operational Definitions.** There exists a variety of different operational definitions of anomalies or outliers - which then determines the different statistical and type of outlier detection methods required (dependent on the study design and data). Hence there is no 'universal' definition of an outlier nor a defined threshold for what is considered an extreme event. The question we aim to answer here is: What is an outlier or 'anomaly' in the data-set? To do this we implement methods to detect: univariate outliers (Tukey's Range Test) and multivariate outliers (Mahalanobis Distance).

2. **Anomaly Detection Methods.** Based on the operational definition and threshold pre-selected in the step before, we need to predict 'new' outliers on unseen data as well. This includes using a mixture of unsupervised, and supervised methods. Based on our research, we found the Isolation Forrest is the most suitable for this study.

3. **Performance Evaluation Measures.** Similarly to the definition of outliers, there is no universal 'good' benchmark to compare and assess the performances of the algorithms in the previous step. Yet, since we need to evaluate different types of algorithms to each other - we also require a 'standard' performance measure for comparability purposes. Thus, the main question we address here is: What are the appropriate or 'good' benchmarks and performance measures for our proposed detection methods?

3.2 Definition of Outliers

Before comparing different operational definitions and outlier detection methods - we must first distinguish the type of outlier we are interested in. Specifically, whether or not we aim to detect univariate or multivariate outliers. Since we are interested in the errors of the fitted response values and its' observed values, we also need to investigate univariate error outliers. Nevertheless, we must consider data sets containing our fitted or observed response variable (premiums) with errors - to gain more insight into the goodness of fit. Thus, we need to implement various outlier detection methods accordingly (as we cannot use the same algorithm to detect both univariate and multivariate outliers).

3.2.1 Univariate Outliers

To detect univariate outliers, we focus on statistical methods commonly used for analyzing goodness of fit. Implementing these methods provides us a threshold - by finding potential outliers and determine the initial range of extreme values we require for further investigation[20]. Since in the previous chapter, we found that we need to account for highly skewed distributions. This implies we may require robust and non-parametric methods (for new sets of data). Consequently, for these purposes, the Tukey's range test or the interquartile range (IQR) approach is the most effective for this scenario.

Tukey's Range Test. Essentially, the Tukey's range test (also known as the Tukey's method) through the use of boxplots (shown in 3.1 visualizes the dataset and graphically assess it by dividing it into five essential values (based on quartiles):

1. **First Quartile (Q1):** represents 25% quartile of the observations
2. **Median:** the center or middle value of the (sorted) data - representing 50% quartile of the observations
3. **Third quartile (Q3):** 75% quartile of the observation
4. **Interquartile range (IQR):** the range representing the distance between Q1 and Q3

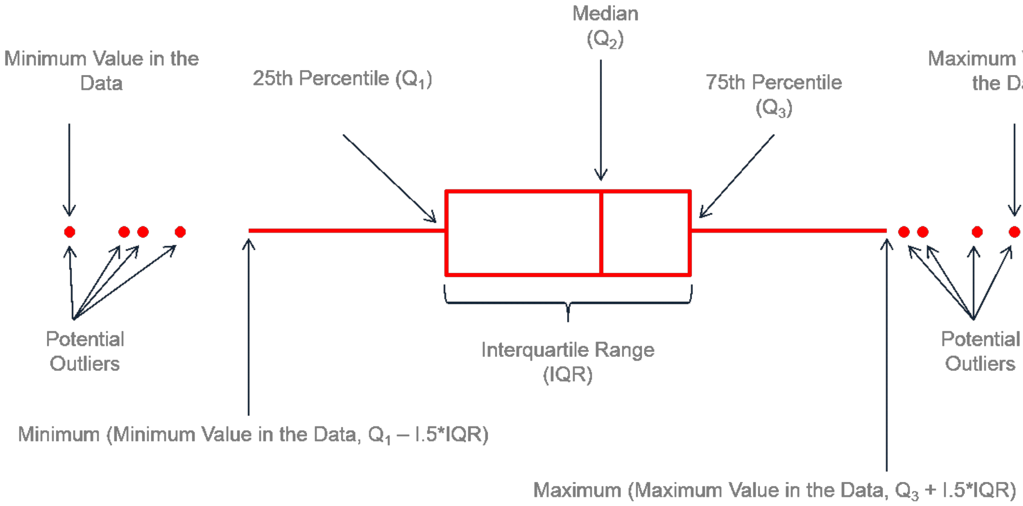

Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

Figure 3.1: Visualizing Uni-variate Outliers via Box-plots

Since these values are determined by the median, quartiles (indicating 25 % to 75% of the instances), and the minimal and maximal values of our observations. We can easily visualize and determine the extreme values through the use of IQR, and utilizing its fundamental feature referred to as 'whiskers' (see Figure). Assume we define the lower boundary and the upper boundary respectively as:

$$\begin{aligned} \textbf{Inner fence:}\ & [Q1 - 1.5 * IQR,\ \ Q3 + 1.5 * IQR] \\ \textbf{Outer fence:}\ & [Q1 - 3 * IQR,\ \ Q3 + 3 * IQR] \end{aligned} \tag{9}$$

Then the largest and smallest values of our dataset within the boundaries define our whiskers. Such that we distinguish between possible and probable outliers or extreme values - if any observations fall outside of boundaries or between the range. Specifically, extreme values located between the range of the inner fence and outer fence are labeled as 'possible outliers' whereas instances whose values exceed the boundaries of the outer fence are labeled as 'probable outliers'. Thus, the key advantage of the Tukey method is that - it does not require the data to be normally distributed - it can be extended to or can be adjusted for highly skewed distributions (found in this study) through the use of logarithmic transformations (i.e. the 'log-IG'method) [21].

3.2.2 Results & Discussion

Firstly, the data is split by training and testing subsets such that 25% of data is used for training (n_0 = 25000 samples), and the remaining 75% is used for testing (n_1 = 75000). Note, we pre-processed categorical and nominal covariates (such as gender, risk-class of policyholders) using dummy encoding. Our observed response variables are defined as 'target', where 'prediction' refers to the fitted values. For univariate outlier detection, only the response values and errors are of interest to us. Additionally, as we can: 1) extend the Tukey Range test to the logarithm scale (required for response values only), and 2) be interested mainly in the distance (error) between the fitted and response values. The logarithmic transformations of both target and prediction values (denoted as ln(target) and so forth) are first calculated and, then the absolute errors were derived (given by Equation (**??**)).

Table 3.1: IQR, 25% Quartile (Q1) and 75% Quartile (Q3) of Observations

Variables	IQR	Q1	Q3	mean	std	min	max
errors	8.43	-1.73	6.70	2.85	8.08	-65.40	-65.40
abs_errors	6.32	1.90	8.22	6.08	6.04	0.00	0.00
ln_target	1.75	8.21	9.97	9.15	1.19	6.48	6.48
ln_prediction	1.75	8.21	9.96	9.15	1.19	6.48	6.48

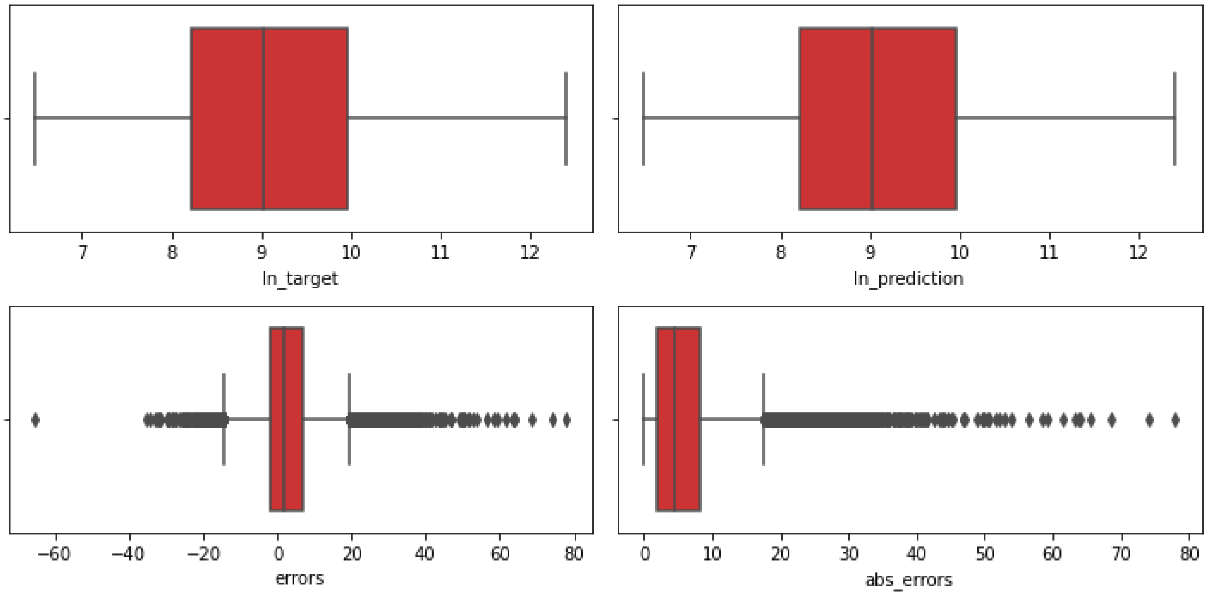


Figure 3.2: Visualizing Uni-variate Outliers via Box-plots

Whereas, based on the results for anomalies in absolute errors a high 'contamination rate' (the portion of detected outliers within the training samples) of 7.54% for possible outliers (1885 total possible outliers) was detected whereas a (somewhat expected) contamination rate, at 2.31%, was detected for extreme values located outside of the outer fence (578 total outliers, with error, mean of 2.7049). Comparing the summary statistics for probable outliers, the absolute error outliers detected had a mean log prediction value of 7.78 (or

2178.76) which indicates that we need to account for outliers for low predicted premium rates - located between the minimum and the 25th quartile (see Figure 3.3, (a), where 'flag = 1' indicated an outlier).

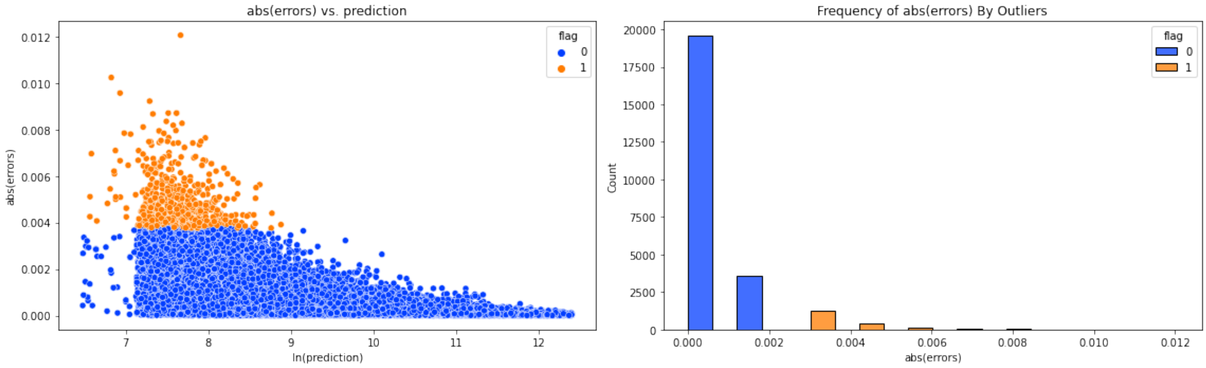


Figure 3.3: Visualizing Uni-variate Outliers via Box-plots

Furthermore, for the covariates, error anomalies on average were detected for policyholders with the following attributes: 28 years of age, with contracts covering 14.6 years, and the average time of 9.6 years which the policy wishes to pay the premiums. Hence, for policyholders with these key attributes, predicated premiums are more likely to deviate from the true values (on average, by 2.71). More specifically, the premiums predicted for policyholders (with attributes stated above) are more likely to be higher than the observed premiums.

3.3 Multi-variate Outlier Detection Methods

As previously stated, since different methods use different operational definitions - the goal here is to implement "flexible" algorithms suitable to detect outliers in high-dimensional data. Often (in most cases), we can group these definitions into the following categories: 1) Distance-based definitions (instances with fewer than p neighboring points) or (2 Density-based definitions (instances which are in regions of low density or low relative density.), and 3) Isolation-based anomalies (instances most susceptible to isolation.). However, for this study, identifying initial potential outliers by utilizing the Mahalanobis distance metric (defined in the next section) deems to be the most effective - in terms of computational simplicity, efficiency, flexibility, and accuracy [22]. Especially in multivariate analysis techniques, the Mahalanobis distance has a wide range of applications in various methods (such as deterministic classification and pattern analysis).

3.3.1 The Mahalanobis distance

Geometrically, we can identify multivariate outliers by calculating the shortest distance between the instances. The Euclidean distance is widely implemented for similar purposes to measure the shortest possible distance between two instances. However, if we have highly correlated variables, the Euclidean distance fails to account for the correlation between the variables and fails to account for the different scales.
An alternative to the Euclidean distance, the Mahalanobis distance (a scale-invariant metric) - utilizes the covariance or correlation between the variables, the variability of each variable,

and scales its contribution to the distance value [23]. It measures how much an instance deviates from the mean of distribution by the number of standard deviations. In other words, it determines the distance between the point $x \in R^p$ (distance point), sampled from a p-dimensional probability distribution given by $f_x(.)$, and the mean of the given distribution $\mu(x) = E(X)$. Thus, assuming there exists finite second-order moments of $f_x(.)$, and the covariance matrix is defined as: $\Sigma = E(X-\mu)^2$. Then the Mahalanobis distance is given by the following:

$$D(X,\mu) = \sqrt{(X-\mu)^T \Sigma^{-1} (X-\mu)} \tag{10}$$

Thus, an instance is classified as an potential outlier if an instance has large a Mahalanobis distance from the distribution. Note the Mahalanobis distance reduces to the Euclidean distance if we consider uncorrelated variables with unit variance (if the covariance matrix is equivalent to the identity matrix).

3.3.2 Results & Discussion

To find abnormal observations in at least one dimension, we need first to identify our groups or clusters of normal instances (also referred to as 'normal' or 'reference' patterns) within our training set. Only then we can identify the instances that do not behave like normal instances and label them as outliers. Hence, we first split our data-set once again (training data-set with 60% of the instances, 15% for validation, and 25% for testing) and implement the Mahalanobis distance algorithm. Then, we consider the data set containing only the response variables and the errors. It allows us to analyze further the impact of our error outliers concerning our premium values which are of high interest to us.
For the data-set with 60,000 instances, we observe a total of 1584 multivariate errors with a similar contamination rate to the univariate outliers detected at 2.64%; with an average Mahalanobis distance of 5.47 for instances flagged as outliers.

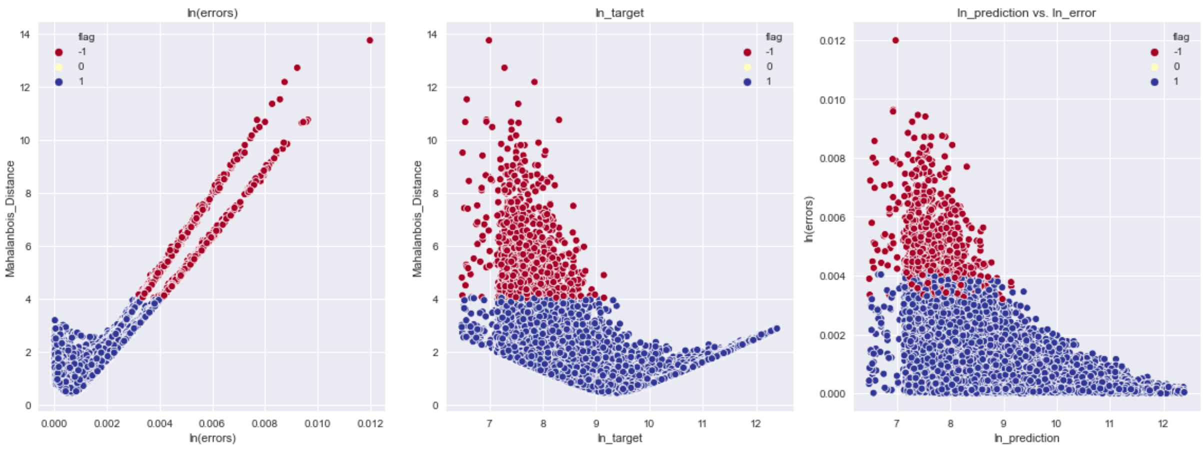


Figure 3.4: Multivariate Outliers Detection

Also, similar to our findings for univariate outliers, we observe that multivariate outliers (labeled as flag = -1 for outliers) in Figure 3.4 were again found for low premium values in target and predicted values in the training set. When analyzing our summary statistics of the only multivariate detected outliers (on the whole and original scale, given in Table 6.1), then we see that our multivariate outliers for the errors range from the range of [-34.15,35.39]. This indicates that premiums for an individual policyholder are underestimated by a maximum value of 34.15 while underestimating by a minimum value of 35.39. In general, for the covariates, outliers with an average of 28 years of age, with a duration of a contract of 13.7 years, and 7.8 years to pay for the premium were detected. Similar behavior also applies to the validation and test set.

However, as observed, since the Mahalanobis uses standard deviation, our results can be susceptible to the presence of outliers. Thus we must consider methods that utilize more robust estimators, such as the minimum covariance determinant (MCD), for calculating the Mahalanobis distance. For further studies, we must also account for the time-consuming computations required for measuring the Mahalanobis distance for large sample sizes.

3.4 Anomaly Detection Methods

Often in existing studies, anomaly detection is also referred to as 'outlier detection'. Even though there exists an overlap between the methods; there is an evident distinction between the two detection algorithms as well. Unlike the previous outlier detection methods discussed, here, we aim to determine if a new observation is an outlier. Thus, in this scenario, we refer to a "new" identified outlier as a 'novelty' anomaly. In contrast to outlier detection (the previous sections) the anomaly detection, implemented in this section, consists of both outlier detection and novelty detection methods. In this study, we implement the following detection method - for predicting outliers.

Forest Trees (or iForest): is the first isolation-based anomaly detection method proposed by Liu et.al (2008). The general idea of these methods is based on the notion that often isolating anomalies is an easier task compared to isolating than normal instances; because they are in 'few and different'. In other words, this model-based method explicitly isolates anomalies rather than describing normal instances by utilizing the following anomaly properties: 1) they are the minority group of instances, and 2) they contain attribute-values that are highly different from normal instances. [24]

The procedure consists of an ensemble of isolation trees (referred to as iTrees), such that each iTree is a special binary tree built from a subsample. Then, each iTree isolates all abnormal instances from the subsamples by identifying the shortest average pathlengths on the iTrees since the outliers can be split out easily bypassing only a few edges in the tree. Key advantages of this method include the following: 1) fast performance, and 2) can achieve high detection performance with high efficiency even with small subsamples.

3.4.1 Results & Discussion

In this section, we implement the isolation forest algorithm for only the training set to build our model and use the validation set to evaluate the performance and find the best model. The test set gives us the final performance measurement of our model at the end. Implementing the isolation forest provides us with comparable results about the previous Mahalanobis distance.

The isolation forest model we implemented was based on the python package 'scikit-learn.' The model contains three hyper-parameters: the number of trees estimated in the forest, the contamination rate, and the number of samples fed into each tree. Liu et al. already show that a subsampling size of 256 could achieve excellent model performance, not depending on the data size[25]. Therefore, we kept the subsampling size as a constant of 256. Grid search was performed to find the best hyper-parameter by a contamination rate between 0.001 and 0.06 with a step size of 0.001 and the number of trees between 50 to 500 with a step size of 50. We used the F_1 score to measure the model's performance, which describes the trade-off between precision and recall. A high F_1 value indicates the precision value is close to the recall value in a classification problem. The F_1 score is defined as: [26]:

$$F_1 = \frac{2}{R^{-1} + P^{-1}} \tag{11}$$

where R stands for recall and P stands for precision. We found our best model with a contamination rate of 0.047 and a number of trees of 300 with the highest F_1 score.

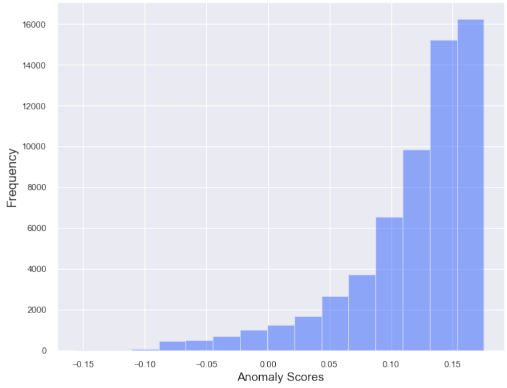


Figure 3.5: Distribution of scores of each data point

After obtaining this model, we applied it to the training set and validation set. Figure 3.5 shows the distribution of scores in the training set. All data points with negative scores were classified as outliers and the percentage of the outliers is exactly the contamination rate we set at the beginning.

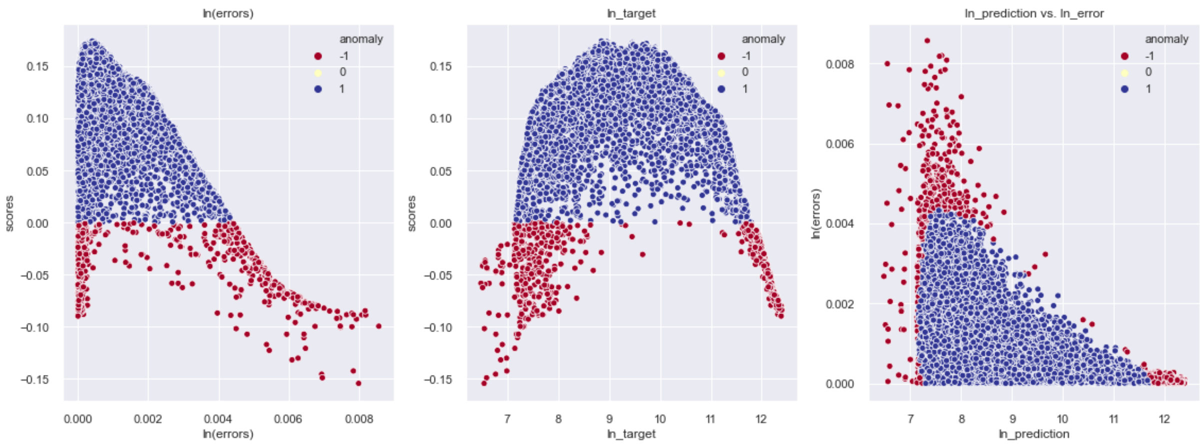


Figure 3.6: Multivariate Outliers Detection

Based on our results, we observed that the 2820 anomalies detected have an average score of -0.0394 in the training set. In the validation set, we found 772 outliers in 15000 data with an average score of -0.0389. The model can find 1046 outliers detected based on the Mahalanobis distance in the training set, i.e., we got 66% recall. For the validation set, we found 188 outliers among 398 calculated by Mahalanobis distance. However, only 1046 out of 2820, and 188 out of 772 outliers are detected correctly in the training and the validation set, which gives us a precision of around 37% and 24%, respectively.
The outliers observed concerning the target, prediction, and error are shown in Figure 3.6. The score of single data interprets how likely it could be an outlier(score '-1' indicates that data is more likely to be an outlier). From the outliers detected, we found that there are more outliers for target and prediction larger than 60000(logarithm of target and prediction larger than 11) comparing to 3.4. Regarding the error, outliers detected from the model have a value of error in all ranges. Our findings would seem to show that the model predicts the data with the logarithm of error smaller than 0.004 wrongly. We got a similar behavior for the validation set.

3.5 Performance Assessment & Evaluation

As previously stated, our goal here is to: without a 'universal' definition to provide a 'standard' for performance assessment - based on the type of variable of interest. Since we can not obtain a model with both precision and recall high, we use F_1 score to measure the trade-off between precision and recall. F_1 score has a range between 0 and 1, whereas 0 is the worst case, and 1 is the best case. The isolation forest model was applied for the test set to find the performance of the model. Additionally, since the dataset is unbalanced (as the sample size of anomalies and normal instances vary), we also use the Area Under the Curve (AUC) as a measurement to assess performance[27].

3.5.1 Results & Discussion

For the test set, 1212 outliers were detected from the isolation forest model. The model can find 441 out of 644 outliers calculated by Mahalanobis distance as the same for the training and validation set, which gives us 68% of recall. The precision of the model is around 53%. By applying eq.11, we got a F_1 score of 0.6. When we apply the model to our test set, we got relatively high precision, recall, and the F_1 score. Figure 3.7 shows us a Receiver Operating Characteristic curve(ROC). The ROC curve describes the performance of a binary classification problem at different thresholds. The Area under the ROC curve is called AUC, which provides an aggregate measure of performance across all possible classification thresholds ranging between 0 and 1[28]. A high AUC indicates that our model predicts class 1 as one and class -1 as -1 better. In the test set, the AUC score reached around 0.97, which shows us a good performance of the model.

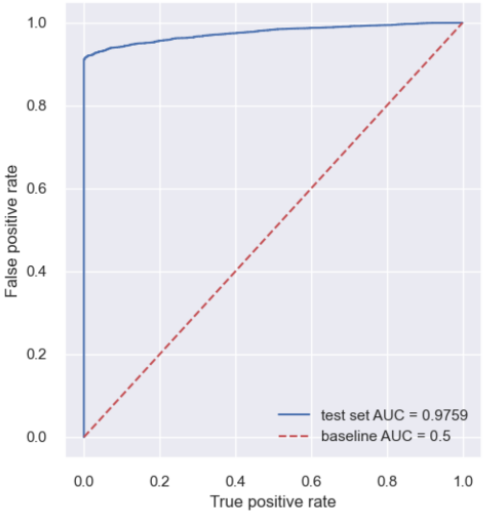


Figure 3.7: AUC of Isolation Forest model on test set

4 Risk Analysis

4.1 Introduction

To illustrate the risk of using our neural network model to do the data migration, we assess the potential loss of an individual contract that could be brought by the usage of our model. In this chapter, we consider two types of risk measurements: the standard deviation and Value at Risk (VaR) or Expected Shortfall (CVaR). These risk measurements allow us to analyze the potential loss at a high confidence level and thus, can be used to evaluate whether our model is trustful or not.

4.2 Theory

4.2.1 Standard Deviation

The standard deviation or Variance is the simplest method to show the risk level. The higher the standard deviation and variance, the higher the risk. Since Markowitz proposed the portfolio theory based on standard deviation as the risk indicator in 1952[29], standard deviation has become a common risk measurement widely used today. However, this traditional risk metric has a significant downside: only considering the average deviation is not enough to describe the large losses caused by the low probability events. Therefore, the standard deviation is not our optimal risk metric.

4.2.2 VaR and CVaR

Value at Risk (VaR), developed by J.P. Morgan Chase & Co., is a widely used risk metric by financial institutions and regulators worldwide - it shows the worst-case loss with a given probability. From the academic perspective, there are plenty of researches on different ways to calculate VaR. In terms of our model, some approaches are not applicable because our dataset is non-time-series. Although, it is still feasible to apply the VaR concept into our risk measurement framework.

Unlike the standard deviation, VaR is the maximal expected loss in a given time horizon under a certain confidence level. Given our data, since there is no time horizon and the loss of an individual contract does not follow the normal distribution. Therefore, we do not use the function derived by the so-called variance-covariance method, which has strong normal distribution assumptions. Instead, we simply take the original definition of VaR:

$$Probability(Loss > VaR) = 1 - p[30] \tag{12}$$

where 1-p is the confidence level.

However, we can not get a completed risk measurement framework with only VaR because it ignores the tail risks. Nevertheless, it is exactly the tail risk that decides the severity of the losses. Therefore, there are a lot of criticisms against VaR, among which the most important criticism is subadditivity. If the contract loss follows subadditivity, the risk of a contract portfolio will be lower than the sum of risks of its components. Subadditivity is one of the four conditions of coherent risk measurement, and the other three are positive homogeneity, monotonicity, and translation invariance [31]. Positive homogeneity means that the risk of an aggregation of X contracts is X times the risk of a single contract. Monotonicity means the greater the return, the higher the risk. Translation invariance states that if we add an amount k to the aggregation of contracts then the risk is reduced by the same amount. Therefore, VaR is not a coherent risk metric because it can not satisfy subadditivity.[30]

To get a coherent risk measurement, an additional risk metric, Expected Shortfall(CVaR), is introduced by Rockafellar and Uryasev (2000) [32]. CVaR measures the conditional expected value that the loss exceeds the VaR value at a given confidence level [32]:

$$CVaR = E(Loss|Loss > VaR) \tag{13}$$

where the confidence level of VaR and CVaR is 1-p.

To some extent, CVaR overcomes the shortcomings of VaR. As CVaR answers the question "if the loss occurs in a bad condition, how bad could the loss be?". If the VaR value is 150 Euro with a confidence level of 97%, the CVaR shows the average of extreme losses beyond 150 Euro with a confidence level of 97%. Thus, CVaR completes the processes of risk analysis in addition to VaR, and we will show the relevant results later.

We have investigated four methods to estimate 95% VaR and CVaR: Parametric method, Historical Simulation, Bootstrapping, and Extreme Value Theory.

Parametric Method: Our approach is to,firstly, find the best fit distribution of the losses, and then use the 95% percentile of the loss distribution as the 95% VaR.

Historical Simulation: The core of the Historical Simulation method is to use historical data to represent the future. Specifically, it takes the maximum historical loss value at a certain confidence level as the VaR. In our case, firstly comes sorting the individual loss (L$_i$) in ascending order, and then calculating the corresponding percentiles according to different confidence levels to obtain the VaR values.

Additionally, we also calculated the stressed VaR. As part of the regulatory requirements, the Basel Agreement mandates disclosure of stressed VaR. Similarly, we select the worst 10% scenarios and assume the objective function (loss) will follow one of the worst 10% scenarios.

Sorting the possible loss values from small to large, we can find the stressed VaR at a 95% confidence level.
Accuracy of VaR: There are many ways to test the VaR model, and the Proportion of Failures (POF) Test method proposed by Kupiec [26] is the most authoritative and practical one. The basic idea of the POF test method is if the actual loss is less than VaR, this event is denoted as 1. If the loss is greater than VaR, it is denoted as 0. If N is the number of events denoted by 0 (the number of events that the actual loss exceeds VaR), $1-p$ is the confidence level of VaR, T is the total number of events, Kupiec gives the maximum likelihood statistics [26]:

$$2Ln(LR) = 2 * [(T-N) * ln(\frac{1-\frac{N}{T}}{1-p}) + N * ln(\frac{N}{T * p})] \tag{14}$$

Since $2Ln(LR)$ is similar to the chi-square distribution with 1 degree of freedom. Therefore, when $p = 0.05$, the critical value is 3.841. That is to say, if the $2Ln(LR)$ value is smaller than 3.841, the VaR is accurate, otherwise, it is not accurate [33].
The historical simulation method has significant advantages. It does not need to make assumptions about the statistical distribution of losses. It completely relies on historical data to calculate the VaR. However, the biggest disadvantage of the historical simulation method is that it assumes the future risk patterns will be the same as past. This is not necessarily true in reality.

Bootstrap: In other VaR calculation methods, the focus is often on the point estimate of VaR. However, only a point estimate is not enough. We are more interested in the error between the point estimate obtained from the data and the true value of the statistic in the unknown population. Interval estimation, a confidence interval for the target statistic, can help us in this regard. Bootstrapping is a powerful approach to do interval estimation.
For bootstrapping, the most important part is to perform resampling with replacement, and the sample size should be the same as the amount of data we have. Then we do the re-sampling for T times. As for each sample, we sort the losses in ascending order and therefore get the 95% percentile as the 95% VaR. By re-sampling T times, we can get T VaR values. To get the 95% interval, we simply sort the T values and find the interval ranging from the 2.5% percentile to 97.5% percentile.

Extreme Value Theory (EVT): In financial VaR calculation, there are mainly two methods: Block Maxima Method (BMM) and Peak-Over-Threshold Method (POT) [34]. BMM is mainly used for time-series data. Therefore, we use the POT method to find the threshold. By using POT to find the threshold, the excess values follow General Pareto Distribution (GPD) [34].
Threshold: the most important setting in the POT method is to find a threshold to define the extreme values. Here, we investigated two methods to find a proper threshold. The first method is to use a mean excess plot, and another one is to use rules of thumbs in finance researches. Usually, in Financial Risk Management, the empirical tail estimator is $(N-K)/N$[34], where N is the total amount of data and K is the number of data that exceeds the threshold. If we set the confidence level as 95%: $0.95 = (N-K)/N$, K can be obtained easily and the threshold is thus the 95% percentile. By defining extreme values in such a way and fitting the excess values into GPD, the Q-Q plot can be easily obtained. By

comparing the Q-Q plots, we can choose the optimal method to set a threshold.
VaR and CVaR: After getting the best parameters of GPD with maximum likelihood estimation, we can obtain the estimations of VaR and CVaR with the following equations (u is the threshold) ([34]):

$$VaR = u + \frac{\beta}{h} * [(\frac{N * (1 - q)}{K})^{-h} - 1] \tag{15}$$

$$CVaR = \frac{VaR + \beta - h * u}{1 - h} \tag{16}$$

where $\beta isthescaleofthebestGPDfit, histheshapeofthebestGPDfit, andqistheconfidencelevel(e.g.95\%$
Besides, the probability that the actual loss will be greater than a certain value M can be calculated by the following equation [34]:

$$Probability(Loss > M) = \frac{K}{N} * (1 + h * \frac{M - u}{\beta})^{\frac{-1}{h}} \tag{17}$$

4.3 Empirical Result

Since both positive and negative residuals are not ideal, which brings losses to the clients, we take the absolute value of residuals, percentage loss, and logarithm loss into account. For $i = 1, 2, 3......100000$, we have the following three loss functions:

$$L_{i1} = |target_i - prediction_i| \tag{18}$$

To eliminate the impact of the magnitude of the target values on absolute losses, we also include the percentage loss (L$_{i2}$) to our dataset:

$$L_{i2} = \frac{|target_i - prediction_i|}{target_i} \tag{19}$$

Considering the outliers, we do the log-transformation. Note all the target and prediction values are greater than 10. Hence, the logarithm loss of each contract(L$_{i3}$) is:

$$L_{i3} = |ln(target_i) - ln(prediction_i)| \tag{20}$$

In addition to the difference between absolute loss, percentage loss, and logarithm loss, we also split the dataset into three parts according to risk classes: high, middle, and low-risk class. This way, we can compare the risk characteristics of different risk classes. Note that, in terms of the three risk classes analysis, the maximal i is no longer 100000, but the number of data in each risk class.

Table 4.1: VaR and CVaR for the whole dataset: L_{i1}

Method	Pearson 3	Gamma	Beta	Historical Simulation	Bootstrap	EVT
VaR	17.5319	17.5591	17.6113	17.5777	(17.4065,17.7462)	17.5787
CVaR	24.3953	24.4198	24.4762	24.4445	(24.1812,24.7129)	24.4442

Table 4.2: VaR and CVaR for the whole dataset: L_{i2}

Method	GPD	IGD	EW	Historical Simulation	Bootstrap	EVT
VaR	0.0035	0.0028	0.0028	0.0029	(0.0029, 0.0029)	0.0029
CVaR	0.0047	0.0040	0.0040	0.0040	(0.0040, 0.0041)	0.0045

4.3.1 VaR and CVaR for the whole dataset

The summary of empirical results of VaR and CVaR for the whole dataset are shown in Table 4.1, Table 4.2, and Table 4.3.

Note for the stressed VaR, we choose 5% of the largest losses from the historical losses. For the whole dataset, we would observe 5000 largest losses, and the potential loss of a new contract will have 5000 possible scenarios. Sort these 50000 losses and take the 95% percentile to get the 95% confidence stressed VaR. For the Extreme Value theory, we set the threshold as the 95% percentile of the loss. Thus, the CVaR values are given following the corresponding VaR.

From Table 4.1, we observe that the VaR values obtained are all in the range between 17.40 and 17.74. Additionally, the stressed VaR from Historical Simulation is 33.87. This states that if the new contract suffers from an extremely bad situation, the stressed VaR under that condition is 33.87. We also tested the accuracy of the mean of the 95% VaR confidence interval obtained from Bootstrapping, and the Kupiec-LR test result is 0.0033, which is smaller than 3.841. Therefore, there is no evidence of any inadequacy in the underlying VaR measure. The CVaR values are all lying between 24.18 and 24.71, and this range is slightly larger than VaR. Except for Bootstrapping, the results from all other methods are near 24.4. In terms of the result of CVaR, it states that if the losses are greater than the VaR, the severity of the losses is around 24.4.

The results shown in Table 4.2 indicate that a 95% confidence worst loss for a contract is about 0.29%, with accuracy test results $2Ln(LR)$ of 0.00336. Thus, the underlying VaR measure is accurate. The CVaR values are all lying between 0.40% and 0.47%, which means if the real percentage loss is larger than VaR, the expected loss will be about 0.4%. As for the logarithm loss, the VaR varies to some extent. Except the result from Mielke distribution, 95% VaR of L_{i3} is around 0.003, which means we are 95% sure that $|\frac{ln(target_i)}{ln(prediction_i)}|$(when target$_i$ > prediction$_i$) or $|\frac{ln(prediction_i)}{ln(target_i)}|$(when target$_i$ < prediction$_i$) is less than 1.00693. It means that we are 95% sure that the prediction and the target will not deviate more than 0.693% from each other. If the real L_{i3} is greater than the corresponding VaR, the expectation of this expression can be around 0.13%.

Table 4.3: VaR and CVaR for the whole dataset: L_{i3}

Method	GPD	recipinvgauss	Mielke	Historical Simulation	Bootstrap	EVT
VaR	0.0035	0.0028	0.0056	0.0029	(0.0029,0.0030)	0.0029
CVaR	0.0047	0.0040	0.0067	0.0041	(0.0040,0.0041)	0.0046

Table 4.4: VaR and CVaR for the high/middle/low risk class: L_{i1}

High	genexpon	mielke	burr	Historical Simulation	Bootstrap	EVT
VaR	22.9033	22.6052	22.6601	23.2005	(22.8186, 23.6032)	23.2031
CVaR	30.8127	30.5886	30.6330	31.1276	(30.5883, 31.6359)	31.1277
Middle	genexpon	beta	betaprime	Historical Simulation	Bootstrap	EVT
VaR	14.4604	14.5727	14.5777	14.5000	(14.3361, 14.6849)	14.5030
CVaR	19.1021	19.2201	19.2315	19.1523	(18.8523, 19.4499)	19.1537
Low	genexpon	weibull_min	beta	Historical Simulation	Bootstrap	EVT
VaR	13.9385	13.9284	14.0280	13.9536	(13.6793, 14.2277)	13.9574
CVaR	18.5291	18.5209	18.6095	18.5512	(18.2473, 18.8741)	18.5517

4.3.2 VaR and CVaR after splitting

To further analyze the impact of risk classes on losses, we once again split the data into three parts: high-risk class, middle-risk class, and low-risk class. Afterwards, we repeat the analysis (similar to Section 4.3.1), and the results of L_{i1}, L_{i2}, L_{i3} are shown in Table 4.4, Table 4.5, and Table 4.6.
Based on Table 4.4, the absolute loss of the high-risk class (around 23) is higher than the loss of the middle and low-risk classes (around 14). However, if we consider the percentage loss in Table 4.5, the percentage loss of the high-risk class is even slightly lower than that of the middle and low-risk classes. Table 4.6 can also confirm this characteristic. It shows that the greater the original premium, the more likely to have higher losses. As seen from Table 4.6, there is no obvious difference in the deviation of prediction from target relative to the smaller one between target and prediction regarding 3 risk classes.
To conclude, we are 95% sure the absolute loss of a contract will not be greater than 17.5 Euro. In addition, the 95% worst percentage loss is around 0.29%. In terms of risk class splitting, the high-risk class has greater absolute losses but lower percentage and logarithm losses.

5 Conclusion

Through the discussion above, we have successfully evaluated the provided neural network model in aspects of goodness of fit, anomaly detection and risk analysis. Firstly, we checked the adequacy of the model's structure by analyzing the residuals and their assumptions. We concluded there was no "strong evidence" of any violations of certain statistical properties,

Table 4.5: VaR and CVaR for the high/middle/low risk class: L_{i2}

High	GPD	powerlognorm	lomax	Historical Simulation	Bootstrap	EVT
VaR	0.0020	0.0024	0.0023	0.0025	(0.0024, 0.0026)	0.0025
CVaR	0.0033	0.0037	0.0036	0.0038	(0.0037, 0.0039)	0.0043
Middle	recipinvgauss	GPD	EW	Historical Simulation	Bootstrap	EVT
VaR	0.0030	0.0026	0.0030	0.0030	(0.0029, 0.0030)	0.0030
CVaR	0.0041	0.0037	0.0041	0.0041	(0.0041, 0.0042)	0.0046
Low	recipinvgauss	mielke	halfgennorm	Historical Simulation	Bootstrap	EVT
VaR	0.0031	0.0060	0.0030	0.0030	(0.0030, 0.0031)	0.0030
CVaR	0.0042	0.0067	0.0042	0.0042	(0.0041, 0.0042)	0.0047

Table 4.6: VaR and CVaR for the high/middle/low risk class: L_{i3}

High	GPD	powerlognorm	EW	Historical Simulation	Bootstrap	EVT
VaR	0.0028	0.0024	0.0024	0.0025	(0.0025, 0.0026)	0.0025
CVaR	0.0041	0.0037	0.0037	0.0038	(0.0029, 0.0031)	0.0043
Middle	recipinvgauss	GPD	EW	Historical Simulation	Bootstrap	EVT
VaR	0.0030	0.0038	0.0030	0.0030	(0.0029, 0.0031)	0.0030
CVaR	0.0041	0.0050	0.0041	0.0041	(0.0029, 0.0031)	0.0046
Low	recipinvgauss	GPD	mielke	Historical Simulation	Bootstrap	EVT
VaR	0.0031	0.0039	0.0059	0.0030	(0.0030, 0.0031)	0.0030
CVaR	0.0042	0.0050	0.0067	0.0042	(0.0030, 0.0031)	0.0047

and, hence, that the error terms are independent identically distributed (i.i.d.) process with zero mean and constant variance. However, we recommend further analysis on the functional fit of the model (beyond the scope of this study). For goodness of fit, we used the standard matrices, Q-Q plots, P-P plots and the K-S test to identify the distribution of the given data and developed a tool to automate this process. By applying the evaluation methods, we found that residual values follow a Johnson SU distribution. And the distribution of the target values corresponds to that of the predicted values.

Additionally, when assessing the goodness of fit of our model, we stated that accounting for outliers ensures stable findings; especially since the presence of outliers has a significant impact on the conclusions we draw from the model fit. Hence, we proposed various outlier detection methods to detect both univariate error outliers (through the Tukey's Range test), and multivariate outliers given the data-set(for prediction, target, and error values using the Mahalanobis distance). As for anomaly detection, the isolation forest method shows us a good performance for detecting new outliers - given the whole data set. Unfortunately, we were unable to determine a model with both high precision and recall values. Thus, we recommend for future studies, utilizing the area under the curve (AUC) as the performance

measure. Overall, we are able to detect existing univariate and multivariate outliers based on the given samples.
In terms of risk measurements, we investigated four methods to calculate VaR and CVaR, and these methods can verify each other pretty well. In such way, we have come up with the trustful approximate loss of a new contract from 3 perspective. Besides, we also found the risk class has influence on the absolute loss of a new contract, but not on the percentage loss or logarithm loss.
However, our work can be improved in the following aspects: for the goodness of fit, the tool we developed tries to fit some common distributions. In order to improve the efficiency of the process, firstly, all distributions may be grouped into various categories. Then, we can test and compare how well the values fit on each group of distributions. However, this requires further analysis before we can determine the candidate distributions which could fit the model. Furthermore, for anomaly detection, comparing and utilizing a mixture of other general approaches, with new sets of data (include statistical methods, classification-based methods, and clustering-based methods) to validate our findings, may also improve the accuracy of our outlier and anomaly detection methods. The risk analysis focuses mainly on the individual contract and we did not go to the portfolio level. The aggregated risk measurements can provide more information on a portfolio level. Besides, other coherent risk metrics like spectral risk measure is also worth investigating.

References

[1] NikosS. Thomaidis and GeorgiosD. Dounias. A comparison of statistical tests for the adequacy of a neural network regression model. *Quantitative Finance*, 12(3):437 – 449, 2012.

[2] Diane Duffy, Ben Yuhas, Arvind Jain, and Andreas Buja. *Empirical Comparisons of Neural Networks and Statistical Methods for Classification and Regression*, pages 325–348. Springer US, Boston, MA, 1994.

[3] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, 2015.

[4] Jim Frost. Heteroscedasticity in regression analysis-statistics by jim. *Statistics By Jim*, 15, 2019.

[5] Stéphane Lathuilière, Pablo Mesejo, Xavier Alameda-Pineda, and Radu Horaud. A comprehensive analysis of deep regression. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2065–2081, 2019.

[6] Richard A Groeneveld and Glen Meeden. Measuring skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 33(4):391–399, 1984.

[7] Nicole Augustin and Simon Wood. On quantile quantile plots for generalized linear models. *Computational Statistics Data Analysis*, 56, 08 2012.

[8] Huynh Huynh. A comparison of four approaches to robust regression. *Psychological Bulletin*, 92(2):505, 1982.

[9] distfit, https://pypi.org/project/distfit/.

[10] Eric Vittinghoff, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer, 2004.

[11] Pratnya P Oktaviana and Irhamah. Kolmogorov-smirnov goodness-of-fit test for identifying distribution of the number of earthquakes and the losses due to earthquakes in indonesia. *Journal of Physics: Conference Series*, 1821(1):012045, 2021.

[12] Carroll Croarkin and Paul Tobias. *NIST/SEMATECH e-Handbook of Statistical Methods*. 2012.

[13] I. M. Chakravarti, R. G. Laha, and J. Roy. Handbook of methods of applied statistics. vol. i: Techniques of computation, descriptive methods and statistical inference. *Handbook of Methods of Applied Statistics*, 1:392–394, 1968.

[14] Addison Wesley. *Knuth, D: Art of Computer Programming, Volume 2: Seminumerical Algorithms*. 1988.

[15] Catherine P. Lawas. Crop insurance premium rate impacts of flexible parametric yield distributions: An evaluation of the johnson family of distributions.

[16] Mauricio A Fernandez-Gonzalez. Metabolic allometry: A genomic approach to scaling. 2020.

[17] Zhangyu Cheng, Chengming Zou, and Jianwei Dong. Outlier detection using isolation forest and local outlier factor. *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, 2019.

[18] Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

[19] Niko Reunanen, Tomi Räty, Juho J Jokinen, Tyler Hoyt, and David Culler. Unsupervised online detection and prediction of outliers in streams of sensor data. *International Journal of Data Science and Analytics*, 9(3):285–314, 2020.

[20] Pete R Jones. A note on detecting statistical outliers in psychophysical data. *Attention, Perception, & Psychophysics*, 81(5):1189–1196, 2019.

[21] Didit Budi Nugroho, Tundjung Mahatma, and Yulius Pratomo. Garch models under power transformed returns: Empirical evidence from international stock indices. *Austrian Journal of Statistics*, 50(4):1–18, 2021.

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

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

[24] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[25] Fei Tony Liu and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, V(N), 2012.

[26] Paul Kupiec. Techniques for verifying the accuracy of risk measurement models. *The J. of Derivatives*, 3(2), 1995.

[27] Indra Waspada, Nurdin Buhtiar, Panji W. Wirawan, and Bagus D.A. Awan. *165Vol. 6 No. 2 | October 2020KHAZANAH INFORMATIKA | ISSN: 2621-038X, Online ISSN: 2477-698XPerformance Analysis of Isolation Forest Algorithm in Fraud Detection of Credit Card Transactions*, volume 6, page 165â“175. 2 edition, 2020.

[28] Jhon B. Valencia, Jeison G. Mesa, Juan J. Leon, Santiago undefined Madrinan, and Andres undefined Cortes. Climate vulnerability assessment of the espeletia complex on paramo sky islands in the northern andes. *Frontiers in Ecology and Evolution*, 8, 2020.

[29] Mark Rubinstein. Markowitz's" portfolio selection": A fifty-year retrospective. *The Journal of finance*, 57(3):1041–1045, 2002.

[30] John Hull. *Risk management and financial institutions,+ Web Site*, volume 733. John Wiley & Sons, 2012.

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

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

[33] 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.

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

Appendix

6 Appendix

Table 6.1: Appendix B.1 Summary Statistics of Error Outliers

	gender	risk	x	n	t	target	prediction	errors
count	1584.0	1584.0	1584.0	1584.0	1584.0	1584.0	1584.0	1584.0
mean	0.4451	0.8807	26.8535	13.7058	9.5764	2272.3323	2272.545	-0.2126
std	0.4971	0.7926	6.6395	9.9337	7.5996	1006.4824	1008.2294	11.3419
min	0.0	0.0	18.0	1.0	0.0	651.8595	654.4008	-34.1577
25%	0.0	0.0	21.0	5.0	4.0	1603.7722	1602.9882	-9.2877
50%	0.0	1.0	26.0	12.0	8.0	1975.1632	1976.2856	-5.3731
75%	1.0	2.0	31.0	21.0	14.0	2660.0725	2658.4964	9.4877
max	1.0	2.0	50.0	47.0	44.0	9390.0503	9390.478	35.3943

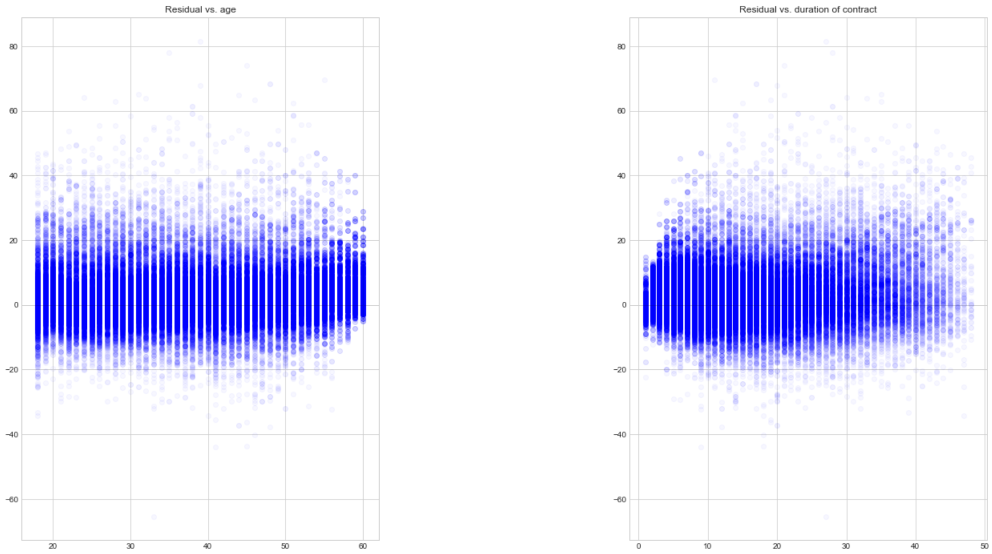


Figure 6.1: Residual vs. Age and Residual vs. Duration of contract Plot

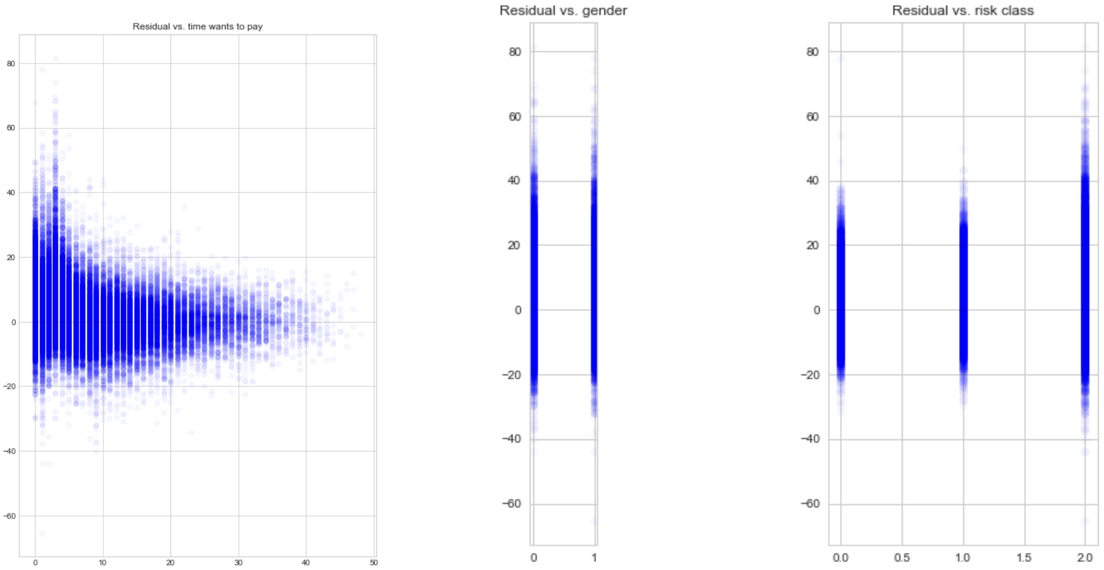


Figure 6.2: Residual vs. Time, Gender and Risk Class Plot

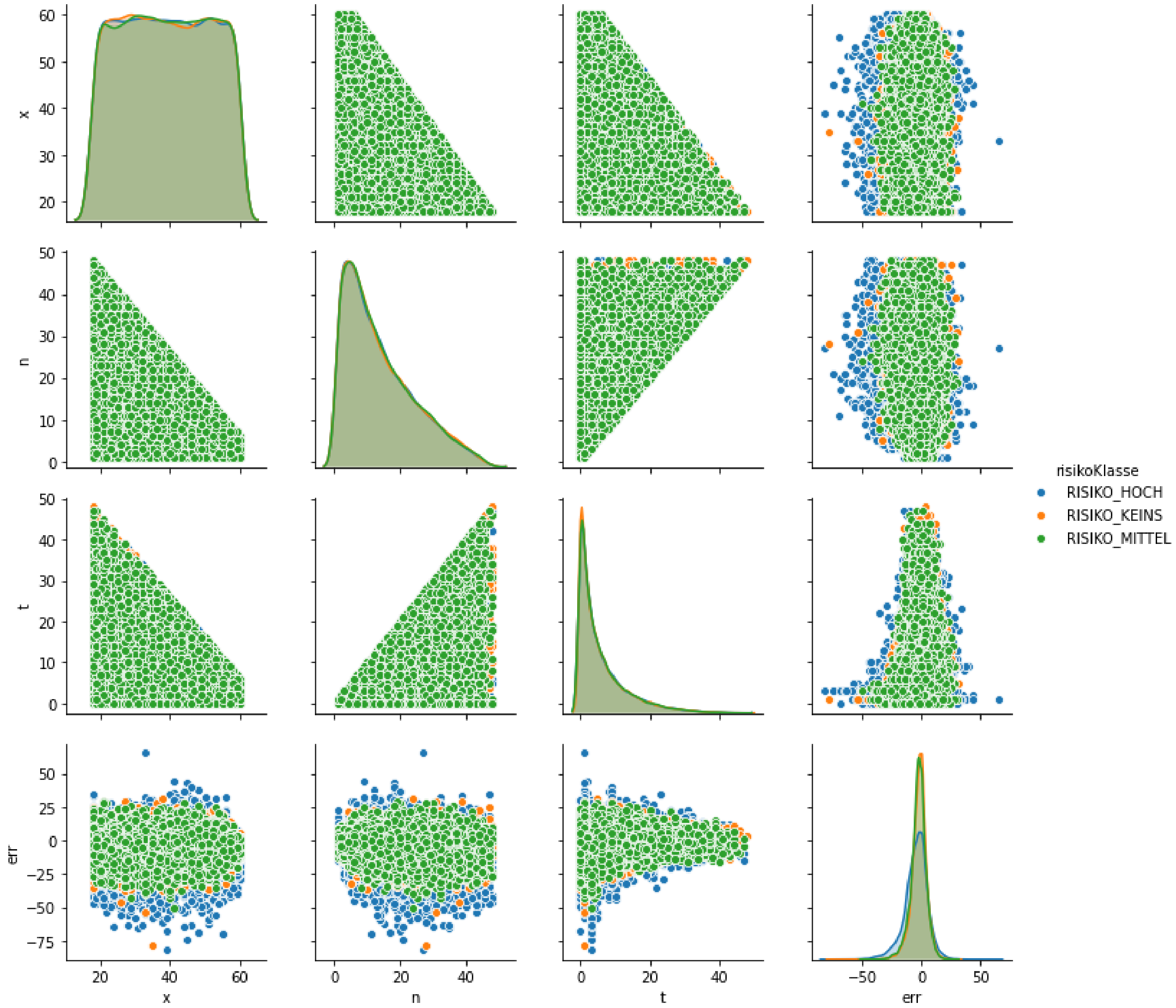


Figure 6.3: Visualizing Distributions of All variables