*buildings*

MDPI

*Article*

# Data Augmentation Approaches for Estimating Curtain Wall Construction Duration in High-Rise Buildings

**Sang-Jun Park [1], Jin-Bin Im [1], Hye-Soon Yoon [2] and Ju-Hyung Kim [1,*]**

[1] Department of Architectural Engineering, Hanyang University, Science and Technology Hall, 222 Wangsimni-ro, Seoul 04763, Republic of Korea
[2] Kunwon Engineering Co., Ltd., 167 Songpa-daero, Seoul 05840, Republic of Korea
* Correspondence: kcr97jhk@hanyang.ac.kr; Tel.: +82-02-2220-0304

**Abstract:** Reliable project management during planning stages of a building project is a meticulous process typically requiring sufficient precedencies. Typical construction duration estimation is based on previous cases of similar projects used to validate construction duration proposals from contractors, plan overall project duration, and set a standard for project success or failure. In cases of high-rise buildings exceeding 200 m, insufficient data commonly arise from the rarity of such projects, leading to a rough estimation of construction duration. Therefore, in this study, oversampling and data augmentation techniques derived from engineering principles, such as parametric optimization and data imbalance problems, are explored for curtain wall construction for high-rise buildings. The study was conducted in two phases. First, oversampling and data augmentation techniques, including Latin Hypercube, optimal Latin Hypercube, simple Monte Carlo, descriptive Monte Carlo, Sobol Monte Carlo, synthetic minority oversampling technique (SMOTE), and SMOTE–Tomek, were applied to 15 raw datasets collected from previous projects. The dataset was split into 8:2 for training and testing, where the mentioned techniques were applied to generate 500 virtual samples from the training data. Second, support vector regression was applied to forecast construction duration, where statistical performance criteria were applied for evaluation. The results showed that SMOTE and SMOTE–Tomek best represented the original dataset based on box plot analysis showcasing data distribution. Moreover, according to statistical performance criteria, it was found that the oversampling techniques improved the prediction performance, where Pearson correlation for linear, polynomial, and RBF increased by 0.611%, 4.232%, and 0.594%, respectively, for the best-performing sampling method. Finally, for the prediction models, probabilistic oversampling methods outperformed other methods according to the statistical performance criteria.

**Keywords:** data augmentation; insufficient data; high-rise buildings; curtain wall construction; duration estimation

check for updates

Academic Editor: Wenzhe Tang and Jianli Hao

Received: 9 December 2024
Revised: 7 February 2025
Accepted: 11 February 2025
Published: 13 February 2025

**Citation:** Park, S.-J.; Im, J.-B.; Yoon, H.-S.; Kim, J.-H. Data Augmentation Approaches for Estimating Curtain Wall Construction Duration in High-Rise Buildings. *Buildings* **2025**, *15*, 583. https://doi.org/10.3390/buildings15040583

**Correction Statement:** This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Estimating accurate and reliable construction duration is crucial for a building project from the project-management perspective. Construction duration estimation methods are constantly evolving, resulting in cost and duration predictions by the integration of multiple linear regression (MLR) analysis by incorporating various influencing factors [1,2]. Moreover, some studies have suggested superior accuracy by applying machine learning (ML) techniques depending on the availability of sufficient precedencies [3,4]. For example, ML models have been effectively applied to forecast construction duration [5]. However,

a drawback of forecasting application arises from insufficient data to learn patterns from existing samples, particularly from rare projects such as high-rise buildings. Despite these difficulties, project managers attempt to estimate construction duration during the planning stage prior to liaising with contractors as a standard practice. This estimation process aims to validate the construction duration proposals offered by the contractors and verify their feasibility. As mentioned previously, practicing solution to construction duration estimation requires reliable data acquisition and analysis-enhancing factors within construction, including efficiency, duration, and cost estimation [6], that determines project success or failure.

Various researchers have demonstrated that estimating the duration of construction for a building project is an essential process to overcome high level of risks and uncertainty, consequently setting a standard for project success or failure [7,8]. Previous studies has shown significant improvement in estimating a building project schedule for accuracy and a reduction in estimation error from 9 to 33% and 4.83%, respectively, by adopting a case-based reasoning method [9]. The typical method requires highly experienced experts who are costly and difficult to find. In particular, in cases for skyscraper projects, acquiring reliable data is challenging due to the limited number of skyscraper construction projects [7–9].

Skyscrapers are described as buildings exceeding 200 m or 50 floors. They are rising in popularity due to their architectural magnificence and socio-urban economic values [10]. According to skyscraper-related statistics provided by the Council on Tall Buildings and Urban Habitat (CTBUH), there has been an increase of approximately 100 buildings over 200 m per annum [11]. Specifically, focusing on notable countries, the number of skyscrapers is 1148 in China, 243 in the USA, 156 in the UAE, and 79 in South Korea. Aside from China, the number of skyscrapers increased exponentially by about 50 to 60 units annually since 2015. It is worth noting that typical skyscraper projects are vulnerable to many risks and challenges faced by insufficiently experienced managers and a lack of comprehensive data [12].

Recent studies regarding skyscrapers and high-rise buildings are scarce due to a sheer lack of relative projects within South Korea. Despite this, international journals contain studies mainly focusing on the technical and physical aspect, such as vibration prediction [13], wind flows and load simulations [14], and equipment layout decision frameworks [15]. From the perspective of construction management within the planning stage, studies on framework development for material selection to comply with codes or specific construction procedures were conducted [16,17]. Moreover, recent trends have shown a more modern method of predicting construction costs [18–20], where the researchers have analyzed the data collected from the general overall construction scenario. However, it is well known that duration is directly related to costs. Although previous studies have rightfully focused their studies on cost, it is also necessary to investigate specific stages of the overall construction that have an impact on the overall construction project outcome.

Curtain walls are an essential architectural feature typically used in skyscraper construction. They cover the exterior of a building, offering a beautiful façade made from materials such as glass. Similar to a concrete wall, they provide a weather-resistant barrier but allow natural light to penetrate deeper into the building. Moreover, curtain wall systems offer thermal efficiency, providing cost-effective heating and cooling for indoor environments [21]. The construction of a curtain wall system assumes a critical path role directly influencing the duration of project in whole. It also serves as a precedent process that sets the benchmark for subsequent finishing tasks [22]. This part is significant for construction requiring a substantial portion (approximately 10–15%) of the total project cost [23–25]. The challenge arises for curtain wall construction, as such walls are con-

sidered a non-structural member of a building, despite their self-supporting capability. Therefore, construction requires meticulous planning, as many aspects and tools need to be installed, such as mounting angles, mullions, and transoms, presenting a particularly hazardous task for construction participants [25].

The global scarcity of skyscraper buildings presents a challenge for artificial intelligence (AI)-based forecasting analysis due to limited available datasets. Furthermore, environmental factors, unique to the geographical location of skyscraper construction projects, such as weather, worker skill, and site layout, crucially affect construction productivity [5,26,27], contributing to the issue of data scarcity. Therefore, despite the introduction of AI, major construction management domains can only be predicted using data from specific regions, resulting in a steep challenge for collecting sufficient and reliable data based on other regions. As a potential solution to this issue, this study investigates a collection of existing oversampling methods to provide a sufficient number of data and overcome issues occurring from insufficient data necessary for executing AI learning methods of regression. This is made possible by taking advantage of data augmentation and oversampling techniques used in the field of data science, engineering, and optimization.

Data augmentation was adopted to generate new samples from existing datasets. This enhances the limitations of scant data, which are usually implemented for time-series analysis [28], image data [29,30], or text data [31,32]. Image augmentation techniques are frequently used when a dataset is visually accessible, such as a time series, images, or text [28,33]. Typical construction management domains, including construction duration and cost, consist of numerical data types necessitating alternative data augmentation methods. Statistical simulations and oversampling techniques that are widely employed for numerical datasets are a potential solution to overcome the mentioned issues [7,34–37]. Bootstrap and Monte Carlo (MC) methods are particularly prevalent in simulating the construction duration [38–41] due to their probabilistic characteristics. The strength of statistical simulation is that it does not require any assumptions regarding the probability distribution of the model error [42]. More recently developed oversampling methods, such as the synthetic minority oversampling technique (SMOTE), were implemented to solve problems from imbalanced datasets. Specifically, when the number of measurements of one variable in a dataset is smaller than that of another, oversampling increases the number of samples by generating the number of lower variables to match the number of higher variables. Therefore, oversampling is widely used for high-dimensional datasets to avoid losing sample properties [43].

This exploratory research predicts construction duration for curtain walls in skyscrapers by applying various oversampling techniques to overcome the limited availability of data from previous case studies. In this study, data on curtain wall construction were collected from 15 skyscrapers in South Korea. Owing to insufficient data to achieve reliable results, data augmentation techniques were adopted to oversample real data. A suitable oversampling technique was identified from initial data distribution analysis of the augmented dataset by comparing analysis between the original and augmented datasets. The result from the most appropriate oversampling technique was used to estimate the construction duration, using ML-based regression to identify the most appropriate combination to predict HRB curtain wall construction. This confirms the possibility of introducing data augmentation methods for managing HRB curtain wall construction during the planning stage.

The rest of the paper is organized as follows. Section 2 provides an overview of previous studies on general methods for construction period prediction. Section 3 presents

the data collection and methodology used in this study, and Section 4 presents the results. Section 5 discusses the results of the study and presents a conclusion.

## 2. Materials and Methods

### *2.1. Curtain Wall Construction Duration*

Curtain wall construction is a phase within a critical path that influences the outcome of a project, highlighting the importance of reducing the manufacturing space, construction time, and cost [44,45]. Curtain wall construction is generally divided into four stages: mock-up, production, construction, and post-construction [45–48]. The mock-up phase is a crucial process that verifies several vital elements, including the structural performance, insulation, and watertightness of the assembly members. According to international standards, tests are performed during the mock-up phase to evaluate airtightness, watertightness, structural performance, residual displacement, interlayer displacement, vertical displacement, and strength. Following the mock-up test is the production of curtain wall members, where a glass-processing company fabricates the double- or triple-glazed windows and subsequently assembles each member within the fabricating factory. The finished units are then transported to the construction site; inspected on-site; and placed on the corresponding floors, using lifts, before construction. The final stage involves finishing the roof and proceeding to the final lift section. These processes are performed simultaneously as construction work continues. Therefore, overlapping construction processes are considered when estimating the construction duration during planning. The general workflow of the HRB curtain wall construction process is shown in Figure 1.




**Figure 1.** Tasks in predicting curtain wall construction for skyscrapers.

Curtain wall construction plays a critical role in the overall skyscraper construction requiring a complicated and intricate decision-making process that stems from potential clashes between structural framing and finishing works. Delays in this construction phase substantially affect subsequent finishing tasks, which can amplify the fundamental risks associated with process delays. For larger and more complex projects, these delay factors increase uncertainty, leading to extensions and increased construction costs. Therefore, mitigating potential adverse consequences is necessary by implementing comprehensive, systematic, and organized risk-management strategies prior to construction. Deriving from identified processes shown in Figure 1, the formula for estimating the construction duration of an HRB curtain wall is as follows:

$$D_{total} = M + P + (F \times C_F) + P_{post} \quad (1)$$

where all of the units within the formula correlate to the raw data used in this study that were given in days; $D_{total}$ denotes the total construction duration; $F$ denotes the number of skyscraper floors; $C_F$ denotes the construction duration per floor; $M$ denotes the duration of the mock-up phase; and $P$ denotes the curtain wall unit production phase. Note that this formula has a caveat, as unworkable days, such as severe weather conditions, are not considered because they do not correspond to specific dates. However, this formula is suitable for the initial planning stage of a construction project and the data collected, as it already considers the duration overlap for the simultaneous mock-up, production, and construction phases.

It should be noted that this study focuses on the overall construction duration as the data input for the methodologies, despite the factors mentioned and the existence of other underlying factors, such as labor force, construction machinery, and others. Alhammadi et al. [49] distinguished between health and well-being in the context of construction workers engaged in infrastructure projects. Their study identified key determinants influencing workers' well-being and conducted an analysis of physiological effects, concluding that well-being does indeed impact productivity. However, their findings indicated that this effect is temporary and does not significantly affect the overall productivity of the construction project. Conversely, Tao et al. [50] mention that construction firms typically anticipate variables that cause construction delays, such as material-supply disruptions and regulatory delays. This, in turn, prevents the delays that would affect the construction duration. Prior studies that have focused on methodological approaches to specific solutions have further examined the fundamental factors influencing construction durations across different building typologies. Therefore, only total construction duration deriving from different stages of curtain wall construction was considered for machine learning-based estimation.

### *2.2. Construction Duration Estimation and Techniques*

Estimating construction duration has utilized various evolving techniques. Predicting and estimating construction duration is an essential aspect of any construction project. It is commonly understood that time is equal to money, a concept frequently discussed in the relative literature. Estimating the construction duration or cost is of particular interest to practitioners and researchers because it establishes a benchmark for the success or failure of a particular building project.

Czarnigowska and Sobotka [51] used various regression methods to predict the construction duration of public roads during the preplanning stage and highlighted the importance of accurately predicting the construction duration. In Ghana, the duration of a bridge construction project was estimated using 30 historical data points, and qualitative data analysis was conducted using regression methods and artificial neural network (ANN) techniques [3]. A case study based on MC simulation was conducted to renovate a multistory apartment building and estimate the construction duration [52].

Multiple linear regression (MLR) is used to consider various factors in construction projects. The MLR model utilizes various independent variables to estimate the construction duration, thereby enabling predictions using a range of variables. These include the cost, gross floor area, construction type, building units, number of floors, and size of parking lots. However, the MLR method has some limitations. It can only produce linear regression, which cannot simulate all diverse and complex construction properties, and is sensitive to outliers [1,2].

The development of ML has overcome the limitations of linear regression models, facilitating the examination of relationships between various variables. In particular, algorithms from the ANN family and SVM-based support vector regression (SVR) have

been widely utilized in estimation research. SVR utilizes kernel functions to perform high-dimensional mapping of input data, allowing for the inference of non-linear relationships without transforming the data into higher dimensions. SVR offers a robust framework for predicting construction duration by finding the optimal hyperplane in high-dimensional space to minimize errors [4,5]. This feature enables SVR to perform high-dimensional pattern analysis and prediction quickly, helping to estimate more accurately, even with complex elements of construction properties.

From the literature review, it was found that predicting the construction duration is essential to the project managerial process. However, most studies have mentioned that a large number of data are required to obtain reliable results. This poses the issue of cases where reliable data are difficult to obtain owing to limited data availability. Moreover, most related studies were case study-based and improved prediction models using the Bromilow time–cost (BTC) method, regression models, and ML and deep learning techniques. Additionally, most studies rely on available historical data, which are sufficient to produce reliable results. However, in special construction projects, such as curtain wall construction for skyscrapers, available data are scarce for conducting such studies.

### 2.3. Data Augmentation Techniques

When the quantity of data is too small or [53] imbalanced, data augmentation techniques serve as solutions to overcome these challenges [54]. Data augmentation is primarily employed when time series or image data are inadequate, using rotation, clipping, flipping, or adjustments in brightness or color to generate augmented data based on raw data [32,37]. For numerical data, random sampling and synthetic data generation were used to enhance the size and variability of the dataset [37]. For example, complaints about environmental disruptions from a construction project were predicted using various oversampling techniques to solve imbalance problems and fill missing data [55]. Similar to this study, construction equipment allocation was optimized using oversampling methods to create virtual simulated cases for construction noise emissions prior to construction [56]. In terms of appropriate sample size, a study was conducted to calculate an efficient method for determining sample size for imbalanced datasets specifically and warned against creating too many virtual sample sizes that could cause overfitting of a model [57]. It is commonly known that image-based AI learning methods require substantial numbers of training data in order to appropriately perform classification in the order of thousands [ref]. However, some studies suggest that the appropriate amount of sample size should be relative to the data size provided [58–61]. While there is no general rule of thumb for the appropriate sample sizes to be generated, this study generated 500 virtual samples, and this number was deemed appropriate, deriving from compensation between the small dataset used and the amount of training data used for typical image classifiers.

Data generation techniques, such as random oversampling and MC simulations, produce data with varying outcomes. MC, one of the most widely used techniques, combines probabilistic theory and fuzzy logic for uncertainty calculations. Its capabilities have been leveraged primarily for the uncertainty simulation of construction duration and time–cost simulation [7,38–41]. However, these methods may exhibit variations in augmentation accuracy and effectiveness depending on the number of repetitions. SMOTE was designed to synthesize new data samples from small datasets through interpolation, expansion of the decision space of minority classes, enhancement of model generalization, and prevention of data generalization through random oversampling [62]. SMOTE generates additional data by randomly sampling line segments between existing data points [63]. This technique is powerful for generating data while maintaining the properties of the dataset; however, caution is required against overfitting [38]. SMOTE–Tomek is a variation of

the SMOTE oversampling technique combined with an under-sampling technique called Tomek. This method has the advantage of avoiding the over-clustering of the data points generated by SMOTE.

The Latin Hypercube (LH) sampling method was adopted to explore the multidimensional distribution of data augmentation by uniformly sampling multivariate data across all dimensions. This method divides the entire range of each input variable into equal probability intervals and randomly selects values from each interval, allowing for uniform sampling across the range of variables. However, LH does not guarantee optimal sample placement and may result in data concentration when using non-uniform datasets. It employs a space-filling method to distribute the points evenly [64,65].

The optimal Latin Hypercube (OLH) enhances uniformity and representation by maximizing the minimum distance between generated data points. This feature renders it valuable for high-dimensional analysis or complex simulations. However, the OLH consumes substantial time and computational resources to generate data because it involves complex calculations that consider high-dimensional spaces and distances between data points. An illustration of the oversampling strategies is shown in Figure 2 to assist with visualization conveying the pattern of the generated data. It shows the MC methods in the first row, SMOTE method with 2 versions random sampling in the second and LH and OLH methods in the third.

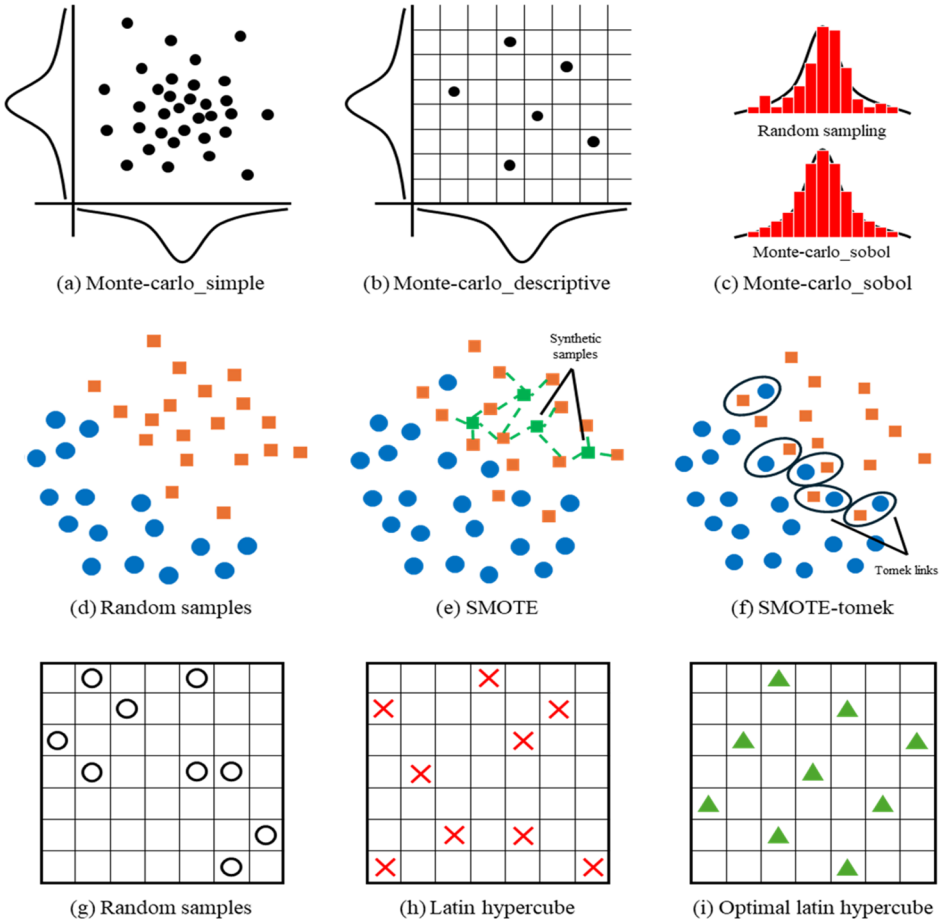



**Figure 2.** Illustration of various oversampling strategies.

This study focuses on widely used data augmentation techniques owing to their accessibility and potential applications. MC simulations are available in Oracle Crystal Ball version 11.1.2.4.600 software, LH, and OLH as part of the design experiment function in

Dassault Systems' iSight R2019X software, whereas SMOTE and SMOTE–Tomek can be applied through the Python package scikit-learn (sklearn).

#### 2.3.1. Drawbacks of Virtual Data Point Generation

As suggested by the previous studies, prediction capability of known regression methods improves with an increase in data provided. However, this suggestion comes with a caveat when using oversampling methods for generating sufficient virtual data for analysis. He and Garcia [66] comprehensively reviewed the state of development of solving imbalanced problems at the time. Discussions on various sampling methods have mentioned powerful potential in solving the issue of an imbalanced dataset while recognizing a definite risk in generating unrealistic data when thorough analysis of raw data is insufficient based on domain knowledge and experience. This is highlighted again in a study by Ye et al. [67] where the issue of the data oversampling technique, SMOTE, is mentioned explicitly where unnecessary generation of noise can occur when the initial data are not processed appropriately prior to oversampling. Ye et al. [68] predicted wind power generation within an extreme environment where sample data are scarce due to difficulty in collecting such data in dangerous situations. This issue was addressed with the use of generative network algorithm to create virtual data points to establish accurate prediction models. However, the study mentions the potential creation of an unrealistic dataset, causing the model to overfit to unrealistic data. The vast number of generated data may reflect the original raw data but also would increase computation time due to the sheer number of data to be computed.

Out of the several studies that applied oversampling methods across various disciplines, many have noted some of the drawbacks of using such an approach. Some of these drawbacks include generation of unrealistic data, requirement of specialized domain knowledge, generation of unnecessary noise, overfitting, increased computation time and resource requirement, and complexity in validation. Despite this, all of the studies applied this type of technique specifically for scenarios where data availability was restricted due to various reasons respective to their field. The potential benefits far outweigh the drawbacks when the oversampling methods are applied appropriately. Moreover, an investigation on the appropriate oversampling methods is necessary to suit particular scenarios, such as façade construction for a skyscraper. While past studies have investigated one specific method of oversampling, such as time-series-based generative network [68], SMOTE [67], and LH variants [64,65], there is a lack of studies where several methods are compared to analyze the prediction performance. It is also necessary to acknowledge previously known and validated methods, such as probability-based MC methods, while also investigating more modern methods derived from different disciplines prone to data deficiency.

#### 2.3.2. Probabilistic-Based Over-Sampling Techniques

To perform the MC simulation, the uncertainty in the variables was identified and defined as the probability distribution. Random values were extracted by generating random numbers from the defined probability distributions, and the resultant values were calculated based on the generated scenarios. MC simple (MCS) is the simplest method, randomly generating numbers between 0 and 1 and creating corresponding random variables.

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N}(x_i) \tag{2}$$

where $f$ denotes the function being integrated, $x_i$ are the random samples drawn from the distribution, and $N$ is the total number of samples.

The MC descriptive (MCD) is similar to the LH technique in that the probability distribution of each random variable is segmented into subsets of equal probability. However, the analysis uniquely incorporates each subset that corresponds to a random variable exactly once [69]. The MCD method follows the following formulas:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{3}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{4}$$

where $\mu$ is the population mean, $\sigma^2$ represents the population variance, and $X_i$ represents the population values.

The Sobol sequence generates quasi-random, low-discrepancy sequences that are used in MC simulations [70]. This method produced more evenly distributed sequences than pure random sampling. The Monte Carlo Sobol (MCsobol) method follows Formula (5):

$$I = \int_{[0,1]^d} f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{5}$$

where $N$ denotes the number of points in the Sobol sequence, and $x_i$ represents the points generated by the Sobol sequence. Function, $f$, is the average of the functions evaluated at a set of points, and $d$ represents the level of the unit cube dimensions.

#### 2.3.3. SMOTE and Variant

SMOTE was originally designed to mitigate class imbalance issues by leveraging the concept of nearest neighbors [63]. Specifically, when dealing with a data point from a minority class, SMOTE identifies n nearest neighbors belonging to the same class, which are considered to have equal weights and are instrumental in constructing a straight line. Subsequently, random synthetic data are generated along this line until a balanced class distribution is achieved. The choice of employing SMOTE over alternative methods in unstructured big data scenarios stems from its ability to generate synthetic data derived from existing instances, thereby minimizing the potential introduction of erroneous information.

This study applies SMOTE oversampling to a single dataset for expansion, as shown in Formula (6).

$$(V_1, V_2) = \sum_{i=1}^{n}\left|\frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2}\right|^k \tag{6}$$

where $V_1$ and $V_2$ represent the respective feature values, and $C_1$ denotes their sum. Additionally, $C_{1i}$ signifies the characteristic value, $V_1$, specific to class $i$. $C_{2i}$ and $C_1$ follow similar rules: The constant, $k$, is typically assigned the value of 1.

The SMOTE–Tomek (SMOTEt) technique is an under-sampling approach designed to clarify the ambiguous boundaries in a dataset by partially removing most data [71]. For each minority data point, $x_j$, located near the boundary between the majority and minority data, if there is no $x_k$, such that the distance $d(x_j, x_k)$ is smaller than the distance $d(x_i, x_j)$ to a majority data point $x_i$, then the relationship between $x_j$ and $x_i$ is defined as a Tomek link. A majority of data $x_i$ were subsequently removed from the Tomek link data. This method enhances the distinction between the dataset boundaries by selectively reducing the number of nearby majority samples. SMOTE–Tomek improves a classifier's ability to recognize minority data.

2.3.4. LH Over-Sampling Technique

LHS ensures uniform consideration of the dataset range over the probability distribution. This approach involved dividing each axis parameter equally, thereby creating hypercubes of identical sizes within the range of each axis [72]. Sampling was then conducted randomly within each cube, as shown in Formula (7):

$$xh_i = F^{-1}\left[\frac{(i-1+R_i)}{n}\right], i = 1, \ldots, n \tag{7}$$

where $R_I$ denotes an independent random uniform in $[0,1]$; i $= 1, \ldots,$ n; and $\mathrm{F}^{-1}(\mathrm{R})$, $\mathrm{R} \in (0,1)$ is the inverse transform for the modelled particular input distribution.

The random nature of LHS sampling may only partially reflect the dataset characteristics. This could induce correlations within the samples generated within each hypercube that were not present in the raw dataset. The OLHS was proposed to overcome these limitations by optimizing the spacing between the sampled data [73]. The optimization technique primarily focuses on maximizing the distance between data points, as given by Formula (8):

$$Min\ U = Min \sum_{p=1}^{P} \sum_{q=p+1}^{P} \frac{1}{L_{pq}^2} \tag{8}$$

where $U$ is the potential energy, $p$ and $q$ represent the points from the data, and $L_{pq}$ represents the distance between points $p$ and $q$.

Maximizing the entropy between raw and generated data is a crucial concept of the OLH. There are several methods to achieve this goal, such as the integrated mean squared error (MSE) or maximization of the closest distance between the raw and generated samples [74]. In the context of experiments (DOE) within a $d$-dimensional parameter dataset, a set of $n$ samples can be formalized as $S = \{x_{ij}\}$, where $x_i$ is a vector representing the $i$-th sample. Each component, $x_{ij}$, of vector $x_i$ pertains to the $j$-th dimension and is constrained to set $\{1, 2, 3, \ldots, n\}$. The $x_{ij}$ represents the permutation of the integers. The pairwise correlation coefficient, $r_{ij}$, is expressed as follows:

$$r_{ij} = \frac{\sum (x_{ik} - \overline{x_i})\left(x_{jk} - \overline{x_j}\right)}{\sqrt{\sum (x_{ik} - \overline{x_i})^2 \sum \left(x_{jk} - \overline{x_j}\right)^2}} \tag{9}$$

The OLH calculates the optimal $r_{ij}$ by switching the pairs. It enhances data quality by selecting the best samples from the data generated by the LH method [75].

2.3.5. Statistical Performance Criteria

$\mathrm{R}^2$, also known as the coefficient of determination, is the squared of the Pearson correlation, denoted by Formula (10). This signifies the average proportion of variation in the dependent variable that can be anticipated from the independent variable.

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n} [\![(x_i]\!] - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \times \sum_{i=1}^{m} (y_i - \overline{y})^2}} \tag{10}$$

where $x_i$ represents a specific value of the independent variable, and $\overline{x}$ denotes the average of the independent variable.

The mean squared error (MSE) is a prevalent statistical and ML metric that quantifies the average squared disparity between predicted and actual values. A smaller MSE indi-

cates proximity to the actual values, whereas a larger MSE indicates a greater deviation in the model's predictions. The MSE form is expressed as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

where $n$ is the total number of data points, $y_t$ is the actual value, and $\hat{y}_i$ is the predicted value.

Mean absolute error (MAE), a widely used metric in statistics and ML, aims to quantify the average absolute difference between predicted and actual values. It assesses errors in paired observations representing the same phenomenon and finds typical applications in time-series analyses [76].

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

The mean absolute percentage error (MAPE) is a metric used to assess the accuracy of forecasting or predictive models. It calculates the average absolute percentage difference between the predicted and actual values and is expressed as a percentage of the actual values. Additionally, it can be used for data assessment [77]. The MAPE is expressed as follows:

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_t - y_i}{y_i}\right| \tag{13}$$

This study aims to investigate the efficacy of various data augmentation techniques for improving the accuracy of predicting skyscraper curtain wall construction duration. The augmented dataset is validated in two phases: validating the augmented dataset through data distribution using the raw dataset and investigating which data augmentation method is suitable for the regression model performed. Based on the properties of each data augmentation technique, the specific research questions were as follows: (1) Which augmented dataset performs best in estimating the skyscraper curtain wall construction duration? (2) Which SVR kernel yields the best prediction performance with an augmented dataset?

## 3. Methodology

### *3.1. Research Framework*

Data from 15 skyscraper-building curtain wall construction projects were collected to predict or estimate the duration of the curtain wall construction. This study comprised two phases. The first phase analyzes various oversampling techniques employed in other engineering fields. For example, oversampling techniques are commonly employed in the design of experiments and sensitivity analyses, typically for the parametric studies of mechanical scenarios. Moreover, probabilistic methods are frequently employed to estimate managerial factors such as cost contingency and building-construction duration. Recently, oversampling techniques have been developed for the ML and deep learning sectors to address imbalanced datasets. Consequently, it is imperative to investigate multidisciplinary oversampling methods further, particularly when data availability is limited.

The dataset used in this study was split into an 8:2 ratio, producing 3 and 12 data points for testing and training, respectively. This step is particularly significant because the test data were strictly isolated from the regression processes to eliminate the risk of influence. Owing to a lack of data, the training data were oversampled using various techniques frequently employed in other fields of study. The oversampling techniques adopted in this study included LH, OLH, MCs, MCd, MCsobol, SMOTE, and SMOTEt.

A box plot was generated using the OriginPro 2016 statistical function to analyze the virtual data. The box plot visually represents the median, interquartile range, minimum

and maximum values, and outliers. This provides a visual summary of the data representation, quickly showing the mean value and dispersion of the generated data.

The virtual data were then used in novel regression techniques. This study applies SVR linear, polynomial, and radial basis functions (RBFs) to the oversampled data. The regression models were evaluated using four model performance metrics: $R^2$ score, MSE, MAE, and MAPE. The overall research flow is shown in Figure 3.

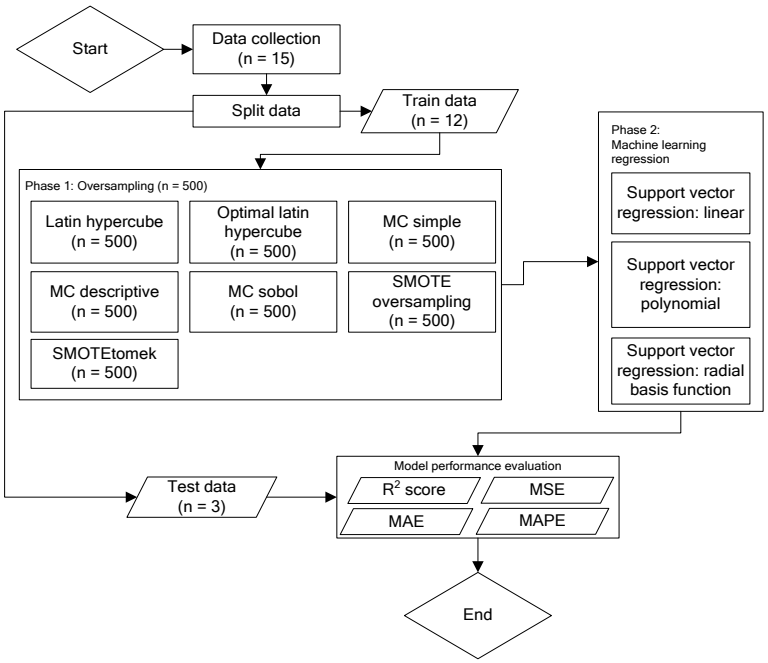



**Figure 3.** Research flowchart.

### *3.2. Data Collection for Curtain Wall Construction in High-Rise Buildings*

In this study, raw data were collected from 15 completed skyscrapers in South Korea to estimate the construction period of skyscraper curtain walls. These projects pertain to skyscraper curtain wall construction spanning more than 50 floors. The collected data included floors, mock-ups, production, construction per floor, and post-construction data, as listed in Table 1.

**Table 1.** Collected raw data for skyscraper curtain wall construction.

| Case | Floors | Mock-Up | Production | Construction per Floor | Post-Construction |
|---|---|---|---|---|---|
| 1 | 51 | 124 | 118 | 608/51 | 60 |
| 2 | 53 | 51 | 111 | 314/53 | 60 |
| 3 | 51 | 75 | 46 | 361/51 | 60 |
| 4 | 53 | 125 | 140 | 446/53 | 113 |
| 5 | 63 | 122 | 232 | 529/63 | 137 |
| 6 | 68 | 87 | 127 | 365/68 | 56 |
| 7 | 59 | 87 | 161 | 394/59 | 70 |
| 8 | 69 | 184 | 92 | 604/69 | 70 |
| 9 | 69 | 184 | 51 | 613/69 | 70 |
| 10 | 53 | 70 | 53 | 605/53 | 70 |
| 11 | 69 | 128 | 242 | 521/69 | 20 |
| 12 | 72 | 145 | 32 | 535/72 | 76 |
| 13 | 85 | 137 | 75 | 365/85 | 31 |
| 14 | 85 | 93 | 80 | 381/85 | 136 |
| 15 | 101 | 161 | 51 | 507/101 | 72 |

Of the 15 collected datasets, 12 were used as training data, and 3 were designated for testing. The training data were subjected to various oversampling and data augmentation techniques, expanding the dataset to 500 instances, representing the 12 original datasets, as shown in Figure 4. The black squares shown in Figure 4 represent the training data, and the red circles denote the test data.

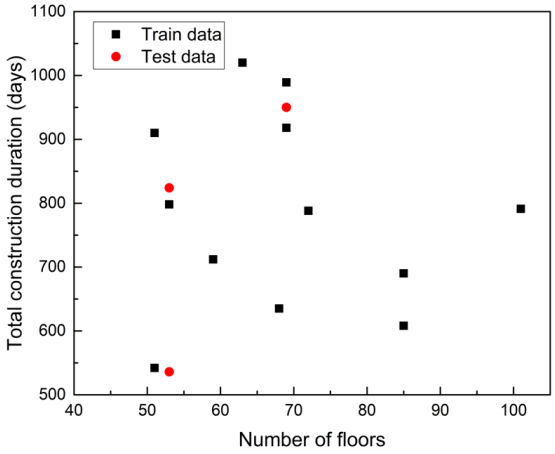



**Figure 4.** Scatterplot of collected raw data.

### *3.3. Data Oversampling and Augmentation*

In this study, seven oversampling strategies were executed on the limited curtain wall construction data for a skyscraper to generate 500 virtual data points. For Latin Hypercube strategies, the required inputs are the minimum and maximum values for each task, such as mock-up, production, construction per floor, and post-construction, which provide the appropriate boundaries for the oversampling process. For probabilistic oversampling methods, such as MC, a standard normal distribution is applied using the mean, standard deviation, and coefficient of variance as input values for the respective tasks. Conversely, for the SMOTE-based strategies, the total construction duration was calculated based on the previously mentioned formula before oversampling. This is necessary because of the nature of the SMOTE strategies. It requires dependent and independent variables to establish connections between datasets. In this study, LH and SMOTE variants of oversampling techniques were implemented using Python 3.10, with Scipy 1.12.0 used for LH and OLH and Scikit-learn 0.12.0 employed for SMOTE-based strategies. As mentioned previously, three variants of MC sampling were performed using Oracle, a popular project management software. Note that all oversampling methods employed in this study are proven methods that are openly and commercially available.

## 4. Results

This section presents the results of this two-phase study. The first phase displays the outcomes of the data augmentation techniques, where a scatterplot provides a concise visual observation. In addition, a box plot is included to provide an overview of the dataset. The graphs in this section were generated using OriginPro 2016.

### *4.1. Phase 1: Data Oversampling and Augmentation*

The data generated from the various oversampling strategies are shown in Figure 5, which shows a scatterplot of the 500 virtual data points, along with the 13 original training data points. Overall, all the virtual datasets were closely aligned with the training dataset, with no discernible boundaries between the virtual and original datasets. Close examination revealed that the LH, OLH, and SMOTE variants consistently produced virtual data

points within the boundary of the original dataset. Conversely, probabilistic methods tend to exhibit a wider spread owing to the nature of probabilistic sampling methods.

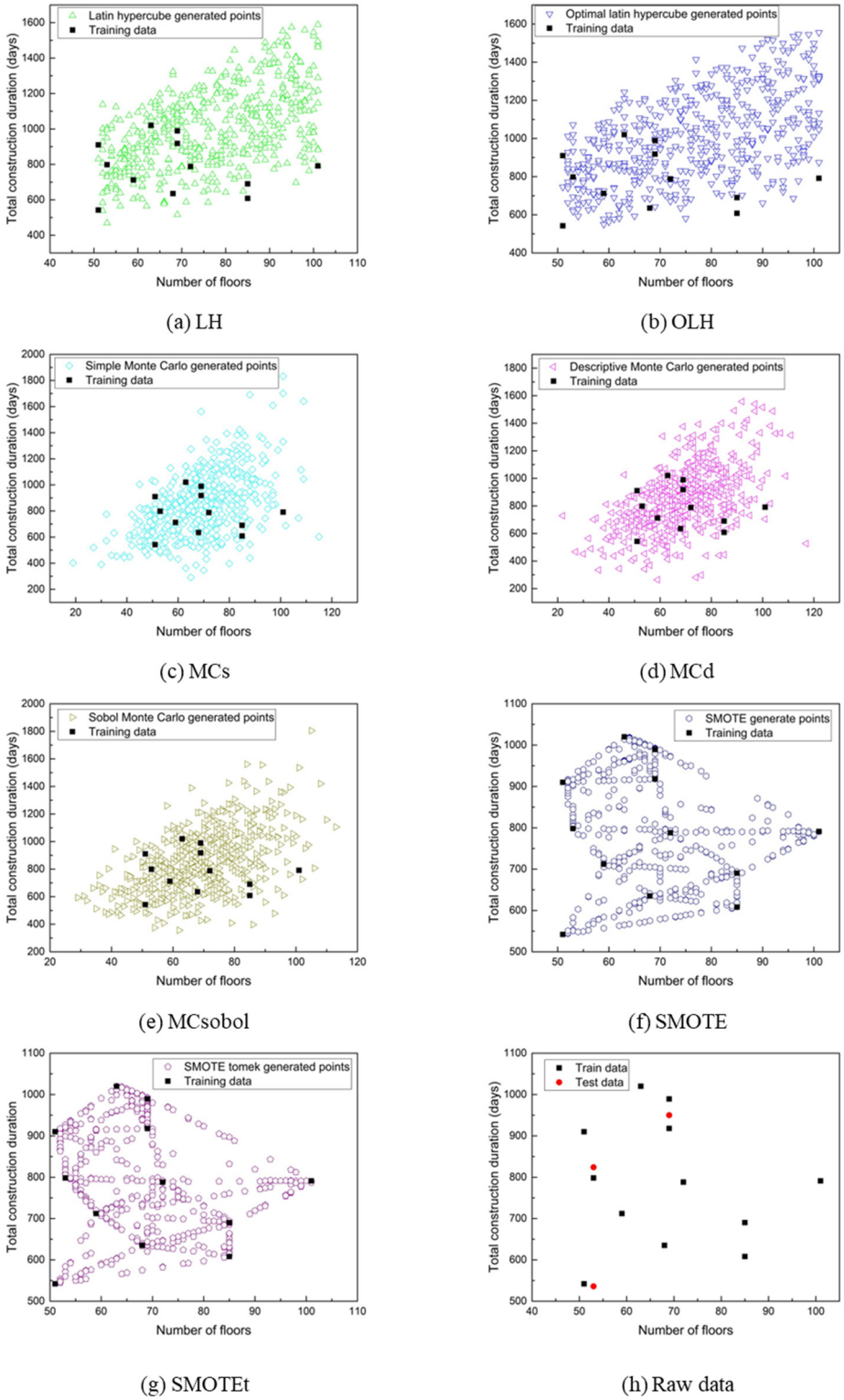



**Figure 5.** Scatterplot of virtually generated data against original data.

Figure 6 shows a box plot illustrating the data distribution by comparing the original training dataset with the generated virtual dataset. Overall, all the oversampling techniques effectively represent the original data. Note that the MC strategies exhibit the best representation of the original data, accompanied by an increase in the interquartile range,

which is consistent with the scatterplot shown in Figure 5. The LH and OLH oversampling methods demonstrated a general increase in the data distribution, with a higher maximum than minimum range. In contrast, SMOTE-based oversampling methods exhibit smaller interquartile ranges and median values.

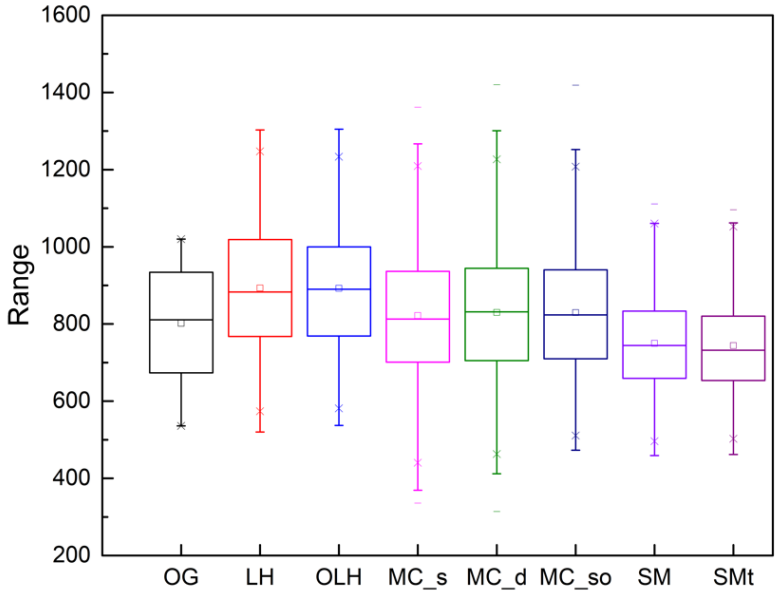



**Figure 6.** Box plot of the original data and the oversampling techniques.

### *4.2. Phase 2: ML-Based Regression*

In this phase, an ML-based regression called SVR was conducted. Three types of SVR kernels were used, which are commercially available in SciKit-Learn 1.4.1. The SVR kernels used in this study were linear, polynomial (n = 3), and RBF. The graphs include kernels with different lines of different styles representing the regression models produced by SVR. Prior to regression, the data were normalized using a standard scalar function to facilitate computing power and produce more accurate results.

Figure 7 shows the three SVR kernels, linear, polynomial, and radial basis function, plotted using OriginPro 2016. The *x*-axis represents the normalized number of floors, whereas the *y*-axis depicts the normalized total construction duration. This figure illustrates that the MC oversampling techniques generally result in a broad range of regression models. In addition, the figure includes the testing data randomly extracted from the 15 collected raw datasets. Despite not influencing virtual data generation, the test data align closely with the generated regression models. All regression models exhibited a similar skew, indicating the proximity of the generated virtual data.

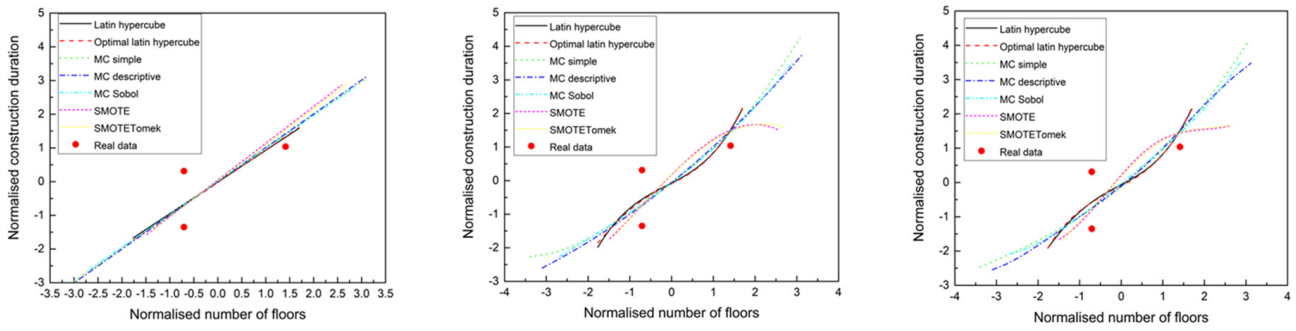



**Figure 7.** SVR regression models respective to linear (**left**), polynomial (**center**), and RBF (**right**) kernels.

In general, all three graphs show that the probabilistic method of Monte Carlo provides the widest band of predictions. Based on the visual analysis, RBF kernels best represented the diminutive real data compared to the other kernels where the model fitted had the least amount of variation. The linear kernel showcased the most appropriate trend model, as both the polynomial and radial basis function showed a decrease in construction duration past the midpoint of the model, which is clearly not realistic. Moreover, all of the oversampling methods showed overlap, representing the similarity of the virtual datapoints despite the different methods used.

## 5. Conclusions

### *5.1. Research Findings*

The findings indicate that MC probabilistic sampling methods have emerged as the optimal choice among the various oversampling techniques, primarily due to their ability to generate data points that extend beyond the boundaries of the original training dataset. Among the three sampling methods, simple, descriptive, and Sobol sampling, each was paired with its respective SVR techniques of linear, polynomial, and RBF kernels, and notable performance improvements were observed. The Pearson correlation coefficients exhibited consistently high values across the different sampling methodologies, confirming the viability of oversampling methodologies for enhancing the predictive capacity of small datasets. These results have substantial implications for the project management sector, particularly in facilitating accurate estimations of unique construction projects.

For the linear SVR kernel, the descriptive sampling method of the MC technique exhibited Pearson correlation, MSE, MAPE, and MAE values of 0.99997, 0.0001, 0.0094, and 0.0234, respectively. These results, evaluated against static-based performance criteria, demonstrated superior performance compared with other oversampling methods. Conversely, the least effective oversampling method was the SMOTE technique, with Pearson correlation, MSE, MAPE, and MAE values of 0.9896, 0.0104, 0.0563, and 0.0690, respectively. It displayed the highest MSE and MAE values and the lowest Pearson correlation values, indicating a comparatively lower reliability. However, considering the minimal difference of approximately 0.01 between the best-performing and worst-performing oversampling techniques, it remains a viable strategy for dealing with limited amounts of data.

For the polynomial SVR kernel with an order of three, the Sobol sampling method of the MC technique demonstrated the most favorable outcomes, achieving statistical performance criteria of 0.9999, 0.0001, 0.0094, and 0.0234 for the Pearson correlation, MSE, MAPE, and MAE, respectively. Conversely, the poorest-performing oversampling method mirrored that of the linear SVR kernel, with the corresponding Pearson correlation, MSE, MAPE, and MAE values of 0.9962, 0.0038, 0.0551, and 0.0690, respectively. The SVR RBF kernel analysis revealed that the simple MC sampling method yielded the most favorable results, whereas the LH sampling method performed the least effectively. The respective performance criteria values for the Pearson correlation, MSE, MAPE, and MAE were 0.9993, 0.0007, 0.0293, and 0.0260, whereas the worst-performing values were 0.9844, 0.0156, 0.1515, and 0.1234, respectively.

The original testing dataset was incorporated into the regression methods to establish a benchmark. Overall, the performance criteria of all oversampling methods closely paralleled those of raw data regression, thus confirming the feasibility of using oversampling techniques to generate virtual data based on real original data. Further analysis confirmed that the SMOTE and LH techniques were less suitable for predicting the skyscraper curtain wall construction duration. In particular, SMOTE in the SVR linear and RBF kernels and the LH and OLH methods in the SVR polynomial kernel worsened the estimation performance. Although SMOTE aims to address classification challenges by correcting data

imbalances, regression problems persist [78]. However, this method cannot be ignored because it produces moderately reliable results. As listed in Table 2, the SMOTE and SMOTEt methods yield more efficient results when combined with the SVR polynomial kernel. Consequently, a data augmentation method that considers the probability distribution is the most suitable for estimating skyscraper curtain wall construction duration.

**Table 2.** Statistical performance criteria for SVR by each kernel.

| Kernel | Case | $R^2$ | MSE | MAPE | MAE |
|---|---|---|---|---|---|
| Linear | Raw | 0.9939 | 0.0061 | 0.0470 | 0.0554 |
| | LH | 0.9962 | 0.0038 | 0.0538 | 0.0540 |
| | OLH | 0.9980 | 0.0020 | 0.0199 | 0.0269 |
| | MC s | 0.9998 | 0.0002 | 0.0149 | 0.0136 |
| | **MC d** | **0.99997** | **0.00003** | **0.00523** | **0.00521** |
| | MC Sobol | 0.9994 | 0.0006 | 0.0281 | 0.0234 |
| | SMOTE | 0.9896 | 0.0104 | 0.0563 | 0.0690 |
| | SMOTE–Tomek | 0.9954 | 0.0046 | 0.0402 | 0.0477 |
| Polynomial (n = 3) | Raw | 0.9593 | 0.0407 | 0.1876 | 0.0554 |
| | LH | 0.9812 | 0.0188 | 0.1345 | 0.0540 |
| | OLH | 0.9853 | 0.0147 | 0.1208 | 0.0269 |
| | MC s | 0.9965 | 0.0035 | 0.0592 | 0.0136 |
| | MC d | 0.99709 | 0.00291 | 0.04866 | 0.00521 |
| | **MC Sobol** | **0.9999** | **0.0001** | **0.0094** | **0.0234** |
| | SMOTE | 0.9962 | 0.0038 | 0.0551 | 0.0690 |
| | SMOTE–Tomek | 0.9990 | 0.0010 | 0.0299 | 0.0477 |
| Rbf | Raw | 0.9934 | 0.0066 | 0.0660 | 0.0685 |
| | LH | 0.9844 | 0.0156 | 0.1515 | 0.1234 |
| | OLH | 0.9864 | 0.0136 | 0.1380 | 0.1167 |
| | **MC s** | **0.9993** | **0.0007** | **0.0293** | **0.0260** |
| | MC d | 0.99529 | 0.00471 | 0.06244 | 0.06156 |
| | MC Sobol | 0.9996 | 0.0004 | 0.0224 | 0.0194 |
| | SMOTE | 0.9926 | 0.0074 | 0.1047 | 0.0797 |
| | SMOTE–Tomek | 0.9985 | 0.0015 | 0.0439 | 0.0382 |

Most appropriate oversampling method according to statistical performance criteria marked as bold.

### *5.2. Discussion*

This study explored research questions regarding the application of data augmentation techniques in predicting the skyscraper curtain wall construction duration. This research is necessary to overcome limitations for building projects where data availability is scarce. The approach used in this research is to implement well-known oversampling or data augmentation techniques to create virtual data frequently used in performing simulation, optimization, or AI within the field of engineering and machine learning-based computer vision.

The results indicate that the MC method is the most appropriate estimation technique. Out of the numerous methods of oversampling, MC-based sampling is the only probabilistic method explored. The reason for the most appropriate fitting compared to the other methods derives from this, where the boundaries also seen to have increased. One of the biggest advantages of the probabilistic method is that it reaches beyond the boundaries restricted by the raw dataset. Furthermore, it is also able to cover realistic and unrealistic scenarios that expand toward likely scenarios, unlike the Latin Hypercube and SMOTE methods, which only oversample systematically, despite being popular methods for optimization in engineering and solving data imbalance issues in AI learning fields. Focusing on the best performing oversampling method for each kernel, there was an increase in performance criteria respective to the kernels. The results show that the most appropriate

oversampling method differed for each kernel. MC descriptive, MC Sobol, and MC simple produced the best Pearson correlation value for linear, polynomial, and RBF kernels, respectively. They increased the performance criteria by 0.611%, 4.232%, and 0.594% when compared to the raw data, showing that oversampling methods generally improved the modelling capability.

In addition, the SVR linear model was the best regression model for predicting the skyscraper curtain wall construction duration as found from comparing the average of the statistical performance criteria across the sampling methods. Linear, polynomial, and RBF kernels each had 0.997, 0.989, and 0.994 Pearson correlation values, respectively. Other performance criteria followed the Pearson correlation.

These insights highlight the suitability of each data augmentation method based on the analysis results and properties. It offers the feasibility of applying data augmentation to overcome scant samples in construction duration estimations, such as skyscrapers or other specialized construction projects.

### *5.3. Limitations and Future Works*

This study collected data from the curtain wall constructions of 15 skyscrapers in South Korea to conduct experimental research and identified the most suitable oversampling techniques by comparing methods from engineering disciplines. It was conducted in two phases. The first phase involved the application of oversampling techniques and data distribution analysis. The second phase applied ML-based regression techniques to analyze the feasibility of oversampled virtual datasets.

The relationship between data augmentation techniques and regression models requires further exploration. Each regression model has unique properties addressing specific challenges. Understanding the properties of each regression model is crucial for identifying the most suitable data augmentation method for estimating construction duration. Further studies are required to investigate this type of research more extensively using larger datasets. Issues from the previous studies exist in this study where potential unnecessary data generation could cause an issue of overfitting the regression models. There is a need for future study that could potentially filter useful data from unnecessary ones that would, consequently, potentially save computation related resources. In addition, to produce more accurate estimates, unworkable days, labor force, construction machinery, material supply chain, and other resource factors that can cause short- and long-term construction delays should be included in the original data that can be further researched to find a potential solution.

**Author Contributions:** Conceptualization, S.-J.P. and J.-H.K.; methodology, S.-J.P.; software, S.-J.P.; validation, H.-S.Y. and J.-H.K.; formal analysis, S.-J.P.; investigation, J.-B.I.; resources, H.-S.Y.; data curation, J.-B.I. and H.-S.Y.; writing—original draft preparation, S.-J.P.; writing—review and editing, J.-B.I. and J.-H.K.; visualization, S.-J.P. and J.-B.I.; supervision, J.-H.K.; project administration, J.-H.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available upon request from the corresponding author due to restrictions on sensitive data.

**Acknowledgments:** We thank the participants in the research.

**Conflicts of Interest:** Author Hye-Soon Yoon was employed by the company Kunwon Engineering Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References


1. Jin, R.; Han, S.; Hyun, C.; Cha, Y. Application of Case-Based Reasoning for Estimating Preliminary Duration of Building Projects. *J. Constr. Eng. Manag.* **2015**, *142*, 04015082. [CrossRef]
2. Li, M.; Wang, G.; Cao, D.; Sun, J.; Jia, G. Construction Duration Estimation for Airport Terminals in China: Case-Based Reasoning Model Considering Case Retrieval Methods and Case Retention Strategies. *J. Constr. Eng. Manag.* **2023**, *149*, 05023012. [CrossRef]
3. Nani, G.; Mensah, I.; Adjei-Kumi, T. Duration Estimation Model for Bridge Construction Projects in Ghana. *J. Eng. Des. Technol.* **2017**, *15*, 754–777. [CrossRef]
4. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; The MIT Press: Cambridge, MA, USA, 2018; ISBN 9780262256933.
5. Zhang, S.; Li, X. A Comparative Study of Machine Learning Regression Models for Predicting Construction Duration. *J. Asian Archit. Build. Eng.* **2023**, *23*, 1980–1996. [CrossRef]
6. Han, S.; Ko, Y.H.; Hong, T.; Koo, C.; Lee, S. Framework for the Validation of Simulation-Based Productivity Analysis: Focused on Curtain Wall Construction Process. *J. Civ. Eng. Manag.* **2017**, *23*, 163–172. [CrossRef]
7. Nguyen, L.D.; Phan, D.H.; Tang, L.C.M. Simulating Construction Duration for Multistory Buildings with Controlling Activities. *J. Constr. Eng. Manag.* **2013**, *139*, 951–959. [CrossRef]
8. Lee, H.-S.; Shin, J.-W.; Park, M.; Ryu, H.-G. Probabilistic Duration Estimation Model for High-Rise Structural Work. *J. Constr. Eng. Manag.* **2009**, *135*, 1289–1298. [CrossRef]
9. Li, Y.; Lu, K.; Lu, Y. Project Schedule Forecasting for Skyscrapers. *J. Manag. Eng.* **2016**, *33*, 05016023. [CrossRef]
10. CTBUH an All-Time Record 97 Buildings of 200 Meters or Higher Completed in 2014. Available online: https://www.skyscrapercenter.com/research/CTBUH_ResearchReport_2014YearInReview.pdf (accessed on 4 June 2024).
11. CTBUH CTBUH Year in Review: Tall Trends of 2021. Available online: https://www.skyscrapercenter.com/year-in-review/2021/data (accessed on 4 June 2024).
12. Szolomicki, J.; Golasz-Szolomicka, H. Technological Advances and Trends in Modern High-Rise Buildings. *Buildings* **2019**, *9*, 193. [CrossRef]
13. He, Y.; Liu, Z.; Fu, J.; Hui, Y.; He, Y. Prediction Models for Modal Parameters of High-Rise Buildings Based on the Latest Statistics of Field Measurements. *J. Build. Eng.* **2024**, *98*, 111288. [CrossRef]
14. Tang, L.; Liu, H.; Zheng, C. Investigation of the Wind Loads and Flow Patterns of a High-Rise Building under Twisted Wind Flows Based on LES. *J. Build. Eng.* **2024**, *96*, 110478. [CrossRef]
15. Zhou, K.; Luo, X.-W.; Li, Q.-S. Decision Framework for Optimal Installation of Outriggers in Tall Buildings. *Autom. Constr.* **2018**, *93*, 200–213. [CrossRef]
16. Pham, V.H.S.; Dau, T.D.; Tran, L.A. Application of Multi-Criteria Analysis in the Selection of Formwork Material for High-Rise Building Construction Projects. *Cogent Eng.* **2024**, *11*, 2367121. [CrossRef]
17. Sachchithananthan, M.; Thayaparan, M. A Framework to Select Suitable Sustainable Materials for High-Rise Buildings in Sri Lanka. *Int. J. Constr. Manag.* **2024**, 1–15. [CrossRef]
18. Huang, J.; Ji, A.; Xiao, Z.; Zhang, L. Schedule-Cost Optimization in High-Rise Buildings Considering Uncertainty. *Eng. Constr. Archit. Manag.* **2024**. [CrossRef]
19. Liu, Z.; Lin, J. Combinatorial Machine Learning Approaches for High-Rise Building Cost Prediction and Their Interpretability Analysis. *J. Asian Archit. Build. Eng.* **2024**, 1–12. [CrossRef]
20. Mostofi, F.; Tokdemir, O.B.; Toğan, V.; Arditi, D. Predicting the Cost of Rework in High-Rise Buildings Using Graph Convolutional Networks. *J. Constr. Eng. Manag.* **2024**, *150*, 04024085. [CrossRef]
21. Qu, J.; Song, J.; Qin, J.; Song, Z.; Zhang, W.; Shi, Y.; Zhang, T.; Zhang, H.; Zhang, R.; He, Z.; et al. Transparent Thermal Insulation Coatings for Energy Efficient Glass Windows and Curtain Walls. *Energy Build.* **2014**, *77*, 1–10. [CrossRef]
22. Kassem, M.; Dawood, N.; Mitchell, D. A Decision Support System for the Selection of Curtain Wall Systems at the Design Development Stage. *Constr. Manag. Econ.* **2012**, *30*, 1039–1053. [CrossRef]
23. Dal Borgo, E.; Meneghetti, A. Production and Shipment Planning for Project Based Enterprises: Exploiting Learning-Forgetting Phenomena for Sustainable Assembly of Curtain Walls. *Comput. Ind. Eng.* **2019**, *131*, 488–501. [CrossRef]
24. Kim, T.; Kim, Y.W.; Cho, H. A Simulation-Based Dynamic Scheduling Model for Curtain Wall Production Considering Construction Planning Reliability. *J. Clean. Prod.* **2021**, *286*, 124922. [CrossRef]
25. Eom, J.; Kang, Y. Curtain Wall Construction: Issues and Different Perspectives among Project Stakeholders. *J. Manag. Eng.* **2022**, *38*, 04022054. [CrossRef]
26. Jung, M.; Park, M.; Lee, H.-S.; Kim, H. Weather-Delay Simulation Model Based on Vertical Weather Profile for High-Rise Building Construction. *J. Constr. Eng. Manag.* **2016**, *142*, 04016007. [CrossRef]
27. Hasan, A.; Baroudi, B.; Elmualim, A.; Rameezdeen, R. Factors Affecting Construction Productivity: A 30 Year Systematic Review. *Eng. Constr. Archit. Manag.* **2018**, *25*, 916–937. [CrossRef]

28. Rashid, K.M.; Louis, J. Times-Series Data Augmentation and Deep Learning for Construction Equipment Activity Recognition. *Adv. Eng. Inform.* **2019**, *42*, 100944. [CrossRef]
29. Xiao, B.; Zhang, Y.; Chen, Y.; Yin, X. A Semi-Supervised Learning Detection Method for Vision-Based Monitoring of Construction Sites by Integrating Teacher-Student Networks and Data Augmentation. *Adv. Eng. Inform.* **2021**, *50*, 101372. [CrossRef]
30. Dong, J.; Wang, N.; Fang, H.; Lu, H.; Ma, D.; Hu, H. Automatic Augmentation and Segmentation System for Three-Dimensional Point Cloud of Pavement Potholes by Fusion Convolution and Transformer. *Adv. Eng. Inform.* **2024**, *60*, 102378. [CrossRef]
31. Feng, D.; Chen, H. A Small Samples Training Framework for Deep Learning-Based Automatic Information Extraction: Case Study of Construction Accident News Reports Analysis. *Adv. Eng. Inform.* **2021**, *47*, 101256. [CrossRef]
32. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Text Data Augmentation for Deep Learning. *J. Big Data* **2021**, *8*, 1–34. [CrossRef]
33. Parente, A.P.; de Souza, M.B.; Valdman, A.; Mattos Folly, R.O. Data Augmentation Applied to Machine Learning-Based Monitoring of a Pulp and Paper Process. *Processes* **2019**, *7*, 958. [CrossRef]
34. Okmen, O.; Oztas, A. Uncertainty Evaluation with Fuzzy Schedule Risk Analysis Model in Activity Networks of Construction Projects: Technical Paper. *J. Van Die Suid-Afr. Inst. Van Siviele Ingenieurswese* **2014**, *56*, 10–20.
35. Moret, Y.; Einstein, H.H. Construction Cost and Duration Uncertainty Model: Application to High-Speed Rail Line Project. *J Constr. Eng. Manag.* **2016**, *142*, 05016010. [CrossRef]
36. Ahmadu, H.A.; Ibrahim, A.D.; Ibrahim, Y.M.; Adogbo, K.J. Incorporating Aleatory and Epistemic Uncertainties in the Modelling of Construction Duration. *Eng. Constr. Archit. Manag.* **2020**, *27*, 2199–2219. [CrossRef]
37. Hassan, A.; El-Rayes, K.; Attalla, M. Stochastic Scheduling Optimization of Repetitive Construction Projects to Minimize Project Duration and Cost. *Int. J. Constr. Manag.* **2023**, *23*, 1447–1456. [CrossRef]
38. Chen, L.; Lu, Q.; Li, S.; He, W.; Yang, J. Bayesian Monte Carlo Simulation–Driven Approach for Construction Schedule Risk Inference. *J. Manag. Eng.* **2020**, *37*, 04020115. [CrossRef]
39. Tokdemir, O.B.; Erol, H.; Dikmen, I. Delay Risk Assessment of Repetitive Construction Projects Using Line-of-Balance Scheduling and Monte Carlo Simulation. *J. Constr. Eng. Manag.* **2018**, *145*, 04018132. [CrossRef]
40. Hashemi, H.; Mousavi, S.M.; Mojtahedi, S.M.H. Bootstrap Technique for Risk Analysis with Interval Numbers in Bridge Construction Projects. *J. Constr. Eng. Manag.* **2011**, *137*, 600–608. [CrossRef]
41. Sonmez, R. Parametric Range Estimating of Building Costs Using Regression Models and Bootstrap. *J. Constr. Eng. Manag.* **2008**, *134*, 1011–1016. [CrossRef]
42. Morris, T.P.; White, I.R.; Crowther, M.J. Using Simulation Studies to Evaluate Statistical Methods. *Stat. Med.* **2019**, *38*, 2074–2102. [CrossRef]
43. Mohammadi, P.; Rashidi, A.; Asgari, S. Privacy-Preserving Culvert Predictive Models: A Federated Learning Approach. *Adv. Eng. Inform.* **2024**, *61*, 102483. [CrossRef]
44. Yu, S.N.; Lee, S.Y.; Han, C.S.; Lee, K.Y.; Lee, S.H. Development of the Curtain Wall Installation Robot: Performance and Efficiency Tests at a Construction Site. *Auton Robots* **2007**, *22*, 281–291. [CrossRef]
45. Han, S.; Lee, T.; Ko, Y. Implementation of Construction Performance Database Prototype for Curtain Wall Operation in High-Rise Building Construction. *J. Asian Archit. Build. Eng.* **2014**, *13*, 149–156. [CrossRef]
46. Jung, T.-S.; Kim, C.-D. The Research and Correspondence of Schedule Risk Management at the Pre-Construction Phase of Curtain-Wall. *Korean J. Constr. Eng. Manag.* **2005**, *6*, 71–79.
47. Kwon, W.; Chun, J.-Y. Collaborative Design Management System for Curtain Wall. *J. Archit. Inst. Korea* **2007**, *23*, 143–151.
48. Hwang, D.-W.; Yoo, S.-H.; Park, S.-Y.; Choi, Y.-K. Process Improvement Method of Unitized Curtain Wall Installation for High-Rise Building—Through Transform of CYCLONE Modeling Method. *J. Archit. Inst. Korea Struct. Constr.* **2017**, *33*, 29–40. [CrossRef]
49. Alhammadi, Y.; Radzi, A.R.; Alias, A.R.; Rahman, R.A. Modeling Workplace Well-Being Factors in Infrastructure Construction Projects: PLS-SEM Approach. *Buildings* **2024**, *14*, 2289. [CrossRef]
50. Tao, J.; Jin, X.; Cheng, H.; Wang, Q. Spatial Correlation Network Structure of Operational Efficiency and Its Driving Factors: A Case Study of Star-Rated Hotels in China. *PLoS ONE* **2024**, *19*, e0313500. [CrossRef]
51. Czarnigowska, A.; Sobotka, A. Estimating Construction Duration for Public Roads During the Preplanning Phase. *J. Eng. Proj. Prod. Manag.* **2014**, *4*, 26–35. [CrossRef]
52. Yoon, J.; Yu, I. Estimating Normal Duration of Renovation for Multistory Apartment Building Considering Extension-Type Renovation Projects. *J. Civ. Eng. Manag.* **2019**, *25*, 156–167. [CrossRef]
53. Iwana, B.K.; Uchida, S. An Empirical Survey of Data Augmentation for Time Series Classification with Neural Networks. *PLoS ONE* **2021**, *16*, e0254841. [CrossRef]
54. Kovács, G. An Empirical Comparison and Evaluation of Minority Oversampling Techniques on a Large Number of Imbalanced Datasets. *Appl. Soft Comput.* **2019**, *83*, 105662. [CrossRef]
55. Hong, J.; Kang, H.; Hong, T. Oversampling-Based Prediction of Environmental Complaints Related to Construction Projects with Imbalanced Empirical-Data Learning. *Renew. Sustain. Energy Rev.* **2020**, *134*, 110402. [CrossRef]

56. Baek, H.; Jung, S.; Hong, J.; Hong, T. Simulation-Based Framework for Optimal Construction Equipment Allocation Considering Construction Noise Emissions. *J. Asian Archit. Build. Eng.* **2023**, *22*, 2959–2976. [CrossRef]
57. Lee, D.; Kim, K. An Efficient Method to Determine Sample Size in Oversampling Based on Classification Complexity for Imbalanced Data. *Expert Syst. Appl.* **2021**, *184*, 115442. [CrossRef]
58. Karimzadeh, S.; Mohammadi, A.; Hussaini, S.M.S.; Caicedo, D.; Askan, A.; Lourenço, P.B. ANN-Based Ground Motion Model for Turkey Using Stochastic Simulation of Earthquakes. *Geophys. J. Int.* **2023**, *236*, 413–429. [CrossRef]
59. Calvo-Echenique, A.; Sánchez, M.; Duvivier, E.; Valero, C.; Chiminelli, A. Assessing Feasibility and Performance of Ultrasonic Guided Wave–Based Numerical–Experimental Methodology for Debonding Monitoring of Adhesive Joints: Application to an Internal Beam of a Battery Box. *Struct. Control Health Monit.* **2025**, *2025*, 1711913. [CrossRef]
60. Maitra, V.; Arrasmith, C.; Shi, J. Introducing Explainable Artificial Intelligence to Property Prediction in Metal Additive Manufacturing. *Manuf. Lett.* **2024**, *41*, 1125–1135. [CrossRef]
61. Sun, X.; Wang, Y. Conformal Prediction with Censored Data Using Kaplan-Meier Method. *J. Phys. Conf. Ser.* **2024**, *2898*, 012030. [CrossRef]
62. Ul Sabha, S.; Assad, A.; Din, N.M.U.; Bhat, M.R. Comparative Analysis of Oversampling Techniques on Small and Imbalanced Datasets Using Deep Learning. In Proceedings of the 2023 3rd International Conference on Artificial Intelligence and Signal Processing, AISP 2023, Vijayawada, India, 18–20 March 2023. [CrossRef]
63. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2011**, *16*, 321–357. [CrossRef]
64. Lin, C.D.; Kang, L. A General Construction for Space-Filling Latin Hypercubes. *Stat. Sin.* **2016**, *26*, 675–690. [CrossRef]
65. Gu, L.; Yang, R.J.; Tho, C.H.; Makowski, M.; Faruque, O.; Li, Y. Optimization and Robustness for Crashworthiness of Side Impact. *Int. J. Veh. Des.* **2001**, *26*, 348–360. [CrossRef]
66. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
67. Ye, X.; Li, H.; Imakura, A.; Sakurai, T. An Oversampling Framework for Imbalanced Classification Based on Laplacian Eigenmaps. *Neurocomputing* **2020**, *399*, 107–116. [CrossRef]
68. Ye, W.; Yang, D.; Tang, C.; Wang, W.; Liu, G. Combined Prediction of Wind Power in Extreme Weather Based on Time Series Adversarial Generation Networks. *IEEE Access* **2024**, *12*, 102660–102669. [CrossRef]
69. Koch, P.N.; Yang, R.J.; Gu, L. Design for Six Sigma through Robust Optimization. *Struct. Multidiscip. Optim.* **2004**, *26*, 235–248. [CrossRef]
70. Dimov, I.; Georgieva, R. Monte Carlo Algorithms for Evaluating Sobol' Sensitivity Indices. *Math. Comput. Simul.* **2010**, *81*, 506–514. [CrossRef]
71. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
72. Saliby, E.; Pacheco, F. An Empirical Evaluation of Sampling Methods in Risk Analysis Simulation: Quasi-Monte Carlo, Descriptive Sampling, and Latin Hypercube Sampling. *Winter Simul. Conf. Proc.* **2002**, *2*, 1606–1610. [CrossRef]
73. Bates, S.J.; Sienz, J.; Langley, D.S. Formulation of the Audze–Eglais Uniform Latin Hypercube Design of Experiments. *Adv. Eng. Softw.* **2003**, *34*, 493–506. [CrossRef]
74. Bates, S.J.; Sienz, J.; Toropov, V.V. Formulation of the Optimal Latin Hypercube Design of Experiments Using a Permutation Genetic Algorithm. In Proceedings of the Collection of Technical Papers—AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Palm Springs, CA, USA, 19–22 April 2004; Volume 7, pp. 5217–5223. [CrossRef]
75. Park, J.S. Optimal Latin-Hypercube Designs for Computer Experiments. *J. Stat. Plan Inference* **1994**, *39*, 95–111. [CrossRef]
76. Hyndman, R.J.; Koehler, A.B. Another Look at Measures of Forecast Accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
77. Kim, H.-J.; Kim, J.-H.; Im, J.-B. Forecasting Offline Retail Sales in the COVID-19 Pandemic Period: A Case Study of a Complex Shopping Mall in South Korea. *Buildings* **2023**, *13*, 627. [CrossRef]
78. Takefuji, Y. Converting Detailed Estimates to Primary Estimates with Data Augmentation. *Adv. Eng. Inform.* **2021**, *49*, 101354. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.