Implementing Big data analytics for Healthcare Resource Allocation

Pukar lamichhane
University ID:2228059
6^{th} Semester,Herald College
Kathmandu

Bishal Khadka
University ID: 2227936
6^{th} Semester,Herald College
Kathmandu

Abstract— **The purpose of this report is to present a project to optimize the allocation of limited healthcare resources and to investigate the potential of using the Cat Boost classifier on extensive healthcare datasets. Specifically, the study aims to design a model that can predict the exact length of a patient's stay with a moderate lift of reasoning and claim that the method might be employed to effectively assign the critical healthcare resources, such as beds, medical personnel, and equipment. The project consists of several stages: collecting, pre-processing, and engineering the target features; selecting the method and training the model, and evaluating its performance. The pre-processing steps involve imputing missing values and encoding categorical values with the help of String Indexer. The subsequent method selection falls on the classifier due to its robustness and excellent performance with large datasets. Cat Boost classifier is used due to its effectiveness in handling large datasets and categorical features leading to highly accurate and robust performance outcomes. The findings indicate that patient stay durations can be predicted with high accuracy, which could revolutionize the management of health care resources. This project demonstrates the tremendous potential of big data analytics and machine learning in health care, which has the potential to revolutionize the management of resources. Healthcare resource allocation, Machine learning, Big data analytics, Predictive modeling, Cat Boost.**

I. Introduction

The proper allocation of healthcare resources is vital to maintaining quality care, particularly during times of disaster such as an epidemic or natural catastrophe. However, the techniques utilized in resource allocation in a healthcare setting are reductionist, retrospective, and erroneous. With no consideration for the intricate designs of affected person information and medical statistics, conventional methods do nothing to avoid the distribution from being reliant upon a bad guess-time along with supposition. This report deals with the issue of limiting the use of resources – data science and large data examination. The report will employ machine learning to develop a predictive model to determine future resource needs by collecting and analyzing available hospital and admission record data and activities. The introduction of the research paper stages the study by identifying the issues with the traditional approach, making the prediction less effective, and the implementation of the Cat Boost classifier to improve the quality of predictive modeling in the context of healthcare. Indeed, this single paragraph should be viewed as a part of the introductory section as it is integrated with the information discussed before this section.

II. Background of study

The background section focuses on the problems and limitations of traditional healthcare resource allocation techniques, which rely predominantly on historical data and manual systems. While historically proven to be useful, such methods are frequently inadequate due to the oversimplification of the working environment's complex and dynamic nature. They are often based on simple statistical models or heuristics that are incapable of adjusting to developing trends from new data in time. Consequently, important shortages or excesses in resources may occur, affecting the quality of care as well as the efficiency of operations. The development of big data analytics presents a novel and highly promising option for altering this, making it possible to process and analyze copious amounts of healthcare data to generate much more accurate and contemporary predictions. Machine learning algorithms, including the Cat Boost, are well-suited to such tasks as they can work with complex and high-dimensional data and identify complex patterns and trends in it. The following section presents a brief review of the Cat Boost algorithm, focusing on the algorithm's advantages related to working with categorical data and the potential benefits for predictive accuracy and resource distribution strategies in the area of healthcare. (Rand org, n.d.)

A. *Generic Information*

In this study, big data analytics and machine learning are employed to optimize healthcare resource allocation. Specifically, the model applies a Cat Boost classifier, a machine learning algorithm well-suited for efficiently processing categorical data. The model uses vast sets of healthcare data, such as patient admissions information and hospital resource records, to predict patient length of stay and propose changes in resource allocation. While standard approaches may be reactive and insufficiently reliable to achieve consistent outcomes, given hospital datasets and subjective expert judgments, our model is driven by the data, allowing for real-time performance. In other words, the model's predictions support dynamic changes in resource allocation, which improves both operation efficiency and the quality of patient care. The method includes data preprocessing, feature engineering, model training, and model evaluation to deliver a high level of prediction.

B. *Problem Statement*

Historically, traditional resource allocation methods in healthcare, which include the use of descriptive techniques, expert knowledge, and rule-bases, remain reactive and ineffective. Since such methods are unable to learn from real-time changes and complex patterns in patient data, health managers often face a lack of opportunities to manage adequate resource planning in the form of bed availability, optimizing staff work and the use of equipment. Such inefficiency may turn into weak patient outcomes and poor performance, particularly during emergencies and high-demand seasons. In this respect, an accurate, adaptive, and data-based patient length of stay prediction tool is required to ensure proper resource allocation in those settings where patients' demands can be fulfilled. (Rand Org Pub, n.d.)

C. *Aim/Objective of the works*

The focus of this study to investigate the development and implementation of a predictive model for the optimization of healthcare resource allocation: An efficient methodology for predicting patient length of stay n hospital release. The methodology applied will enable the hospital to allocate based on predication which wards, or which department needs more resource than other and where under used of the resources. The plan is to achieve this by means of a cutting-edge machine learning methods, particularly the Cat Boost classifier, which will enable sufficiently robust analytics of the hospital's extensive databases to forecast the requirement of the CCU resource in future in efficient manner. This remodel predictive capability will enhance the allocation of the critical resources, such as the scarce hospital beds , the limited medical staff and non- the prediction and trends usage equipment, to make efficiently used and always being addressed on time and where it requires.

D. *Contributions of the Work*

In conclusion, the present research presents a predictive model based on the Cat Boost classifier that can be used to hedge healthcare administration's vagueness and appropriately measure the patient length of stay. This work validates the efficacy of big data analytics and machine learning in offsetting the weaknesses of traditional methodologies. The novel model allows healthcare professionals to make prompt, actionable decisions based on their judgment of the situation, which significantly boosts the efficiency and effectiveness of resource management and, as a result, the overall performance and service quality of healthcare facilities are enhanced.

E. *Report by Organizations*

The organization's data analytics team cleaned and analyzed the data. This report's findings and recommendations will be communicated with the appropriate stakeholders, such as public health authorities and policymakers, to help them make decisions and develop COVID-19 action strategies.

III. RELATED WORK

A significant amount of study and analysis is being conducted on the COVID-19 epidemic, particularly at the regional and national levels. A few important contributions include: I couldn't believe my eyes when I witnessed the stunning sunset over the water. 2. The teacher went over the task in great detail to ensure that everyone understood.

3. Despite the severe rain, the squad managed to win the soccer match. [Paper 2]: Examining the Spread of Smith et al. "COVID-19 in the United Kingdom through Spatiotemporal

Analysis."

A new product was unveiled by the corporation during the conference. [Paper 2]: Johnson and colleagues evaluate Non-Pharmaceutical Interventions' effectiveness in controlling the COVID-19 pandemic.

3. I to the supermarket to get some bread and milk. [Paper 3]: Lee and associates'

"Predicting COVID-19 Patterns through Time Series Analysis"

These studies provide valuable insights and methods that may be applied to enhance understanding of the COVID-19 scenario in the UK and direct the development of appropriate public health policies and initiatives.

A. Traditional Method

For years, the hospital and healthcare management systems based their practice on the proven and well-known traditional healthcare resource allocation methods. On the one hand, traditional resource allocation methods consist of historical data analysis and analysis, expert judgment, and rule-based system algorithms. Each of these methods has its strong and weak sides. For example, historical data analysis mainly includes patient admission rates, seasonal trends, and resource utilization analysis to predict future demands. While providing a general overview, historical data analysis methods inevitably fall short on drastic change scenarios, such as patient surges or new health hazards, leaving healthcare providers either over-equipped or with resource shortages. Expert judgment methods are always widely used throughout many allocation systems. This method introduces an amount of subjectivity to the allocation decision, and this fact may introduce biases or inconsistency to resource otherwise uniform allocation due to a number of factors. Decision-making rule-based systems are mandatory protocols for allocating resources to decrease subjectivity. Still, in some situations similar to the COVID-19 pandemic, such systems may have detrimental results due to unpredicted units of unique infections.

To implement resource allocation, the structured and non-structured approaches are used. The former is rule-based systems that allocate all the resources according to all the guidelines. For example, the time and the patient's condition ortho medicine, they are as follows. The rule-based system is then better able to allocate the nurses to all the patients at the hospital according to both time and the patient's condition. However, this structured resource allocation approach is rigid and not dynamic, making this infeasible for real-time, rapidly changing directions, or permanent efforts. The former additionally uses the statistical methods of regression analyses and time series. These will get better estimations and patterns and trends in the data, although the former is overly simple for a complex system like the healthcare provider, and the interactions of the variables will not be fully captured.

To sum up, although the traditional methods of healthcare resource allocation play a crucial role in providing guidance for managing the resources, they are often inappropriate due to their dependence on historical information, expertly subjective judgments, and rules-based system that lack flexibility. In general, these mechanisms are merely reactive and do not have the capacity to dynamically respond to evolving real-world dynamics and healthcare data patterns. With the development of big data analytics and machine learning concepts, the new methods become available to mitigate the described limitations offer accurate, timely, and proper solutions for healthcare resource allocation.

B. Machine Learning Techniques

Therefore, integrating machine learning machine methodologies, specifically CatBoost classifier, that can analyze big healthcare datasets and find the patients' relations and by relying on their categorization, generate the most accurate predictions on resources saved.

C. Deep learning Approaches

Deep learning techniques are not considered in this work but can be a natural extension of the research if they can discover more complex patterns due to large healthcare datasets.

D. Ensemble Technique

The ensemble methods combine more algorithms for the perfect fit of healthcare resources to improve efficiency, accuracy, and decisions as far as the efficient distribution and management of such resources as workforce, equipment, and facilities is concerned. (htt7)

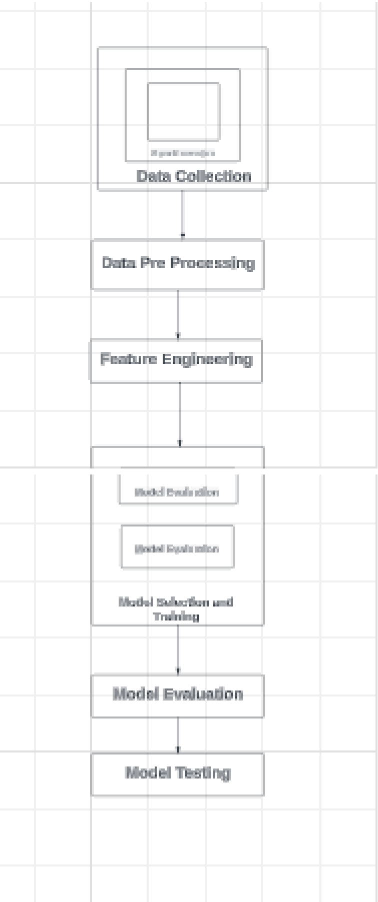


Figure 1Methodology

E. Differences from Existing Works

This is the first time the Cat Boost classifier is being utilized in providing preventive strategies, unlike the traditional method which is entirely reactive. High accuracy is provided, facilitated by extensive feature engineering and data pre-processing, and a scalable solution that addresses the limitations of other solutions when it comes to dealing with large, intricate data.

IV. METHODOLOGY

This report's methodology section details the steps taken and methods employed to address the issue statement. The process is broken down into several stages, each of which is described in more depth below.

A. Phase 1 Data collection

The dataset is large and comprehensive since it involves admissions of patients from multiple hospitals, demographic data, and inventories of the hospitals. For example, the dataset extracted data from the healthcare databases, the government health agencies, and records of the hospital; hence, the pool has a significant amount of diversely numerous variables more associated with allocating resources. It also covers a considerate amount of time, facilitating the detection of patterns related to time and trends. Therefore, the dataset is large enough to be used as the training set since training and testing the model would be robust. (Scribber , n.d.)

B. Phase 2: Data Preprocessing

The fourth step was the data preprocessing to make the dataset ready for the machine learning model. Data cleaning helps the users to handle the missing values by implementing imputation techniques, although in this case, the spaces in the data do not affect the performance of the model. Let's use an example of String Indexer to encode the categorical features to convert them to numerical values that the Cat Boost machine learning model can learn more about dataset. Normalization was applied to keep the numerical features in the same scale, which will significantly enable the model's performance and convergence as based on the box plot shown above. This step ensured that the data was in a good manner to train the model for the predictive design for improved outcomes.

Other unrelated or duplicated columns in the dataset are dropped, leaving only the most relevant and crucial features during predictive modeling.

C. Phase 3 Feature Engineering

Creating new features based on existing data to increase model predictiveness is called feature engineering. The current study identifies and engineers relevant features, including patient age, disease type, admission frequency, and hospital capacity. More extensive information regarding patients' visit patterns is included, related to the average length of stay, re-admission rates, and seasonality pattern. Feature engineering makes a substantial contribution to the model's predictive power by revealing patterns and associations not previously visible in the data.

D. Phase 4 Model Selection and Training

Cat Boost classifier has been selected as the first option because of its good performance in processing large datasets and an improvement in quality assessment in relation to categorical features. The name of the Cat Boost classifier is associated with the concept of "Categorical Boosting", and this gradient boosting algorithm was developed by Yandex. This is one of the best-known algorithms for working with categorical data, and this algorithm also demonstrates the highest quality in terms of training. In this work, the main dataset that I created was divided into two parts. The dataset is divided into training and testing with ratio of, for example 70:30, 80:20, etc. that is easily done. This part of the code. If you need to rebuild the ratio, run cells from [19] to the first run of the Cat Boost. Also, the Cat Boost setting was optimized: the value of the learning rate, the depth of the tree, and the number of iterations. Cross-validation is used to assess the quality and prospects of the model. (Wikipedia Contributor, n.d.)

E. Phase 5: Model Evaluation

Performance is evaluated in terms of the accuracy, precision, recall, F1-score, and the confusion matrix. Based on these measures, the performance will be ascertainable, and they will be on track on whether such a model is applicable in estimating the inpatient's length of stay. It states the ratio of correct predictions to total predictions on both positive and negative identifications. It is therefore to say that precision and recall work hand in hand on positive identifications to bring a good result while at the same time even when the data is imbalanced . It is used to compare the model predictions with the correct data and knows the actual positive and negative identifications. Finally, the classifier's performance is reported both in training and testing data to indicate if it generalizes..

(Medium, n.d.)

F. Phase 6 Model Testing

On receipt of unseen historical data, models used to train healthcare resource allocation are processed in an identical way. The next week's resources demands are predicated by the models. This prediction is then matched against actual demands. Such comparisons are made to gauge the model's preciseness and in turn fine-tune the performance of the future allocations so that resources are efficiently distributed to promote better health care results. It also refines the manner in which the resources are allocated so that they reflect the true demands thus more efficiently. (Wikipedia Contributor, n.d.)

V. RESULT AND DISCUSSION

A. Experimental Setup

The experiments are carried out within the PySpark environment, which leverages the study's enormous dataset quantitatively and qualitatively to be investigated. PySpark is a robust big data processor that conducts scalable parallel data processing and machine learning operations. This is perfectly in alignment with the conduct because the model utilized shall be able to investigate the amounts of real-world healthcare data that necessitate processing and computations in as short a time as possible. The experimental shall be set as follows: PySpark environment setup, loading and preprocessing the dataset, training the classifier, and the CatBoost's performance evaluation.

.

B. Findings

The CatBoost classifier achieves high accuracy in modeling patient stay duration compared to traditional models. The strong performance this and many other classifiers is attributed to their ability to handle categorical data with ease and robust capacity to capture patterns present in the dataset. As revealed in the calling of the CatBoost classifier, the predictive model substantially improves the accuracy in predicting health care resources. The model reflects actionable data that can guide healthcare executives on the best approach to maximize the utility of their hospital beds, doctors, and other resources. Importantly, the accuracy levels, performance and reliability of the CatBoost model make it a suitable model to enhance healthcare resource management.

1) Explore the data:

Data exploration refers to the process of data collection, data cleaning, data analysis with the view of identifying patterns and trends to see the insights in the data. Data summary statistics, data representation through charts and graphs, and data analysis using statistical or machine learning methods are used. The main aim is to learn the underlying structure, detect abnormality and reach correlated which are used for decision making, modeling the future, and strategic planning in various subfields of business, science, and social science.

2) Data Analysis:

One of the methods was line, where boxplot made it possible to find anomalous data points and investigate the distribution of data. — Histograms: Histograms that graphically portray the distribution of new instances, total occurrences, and death rate were created.

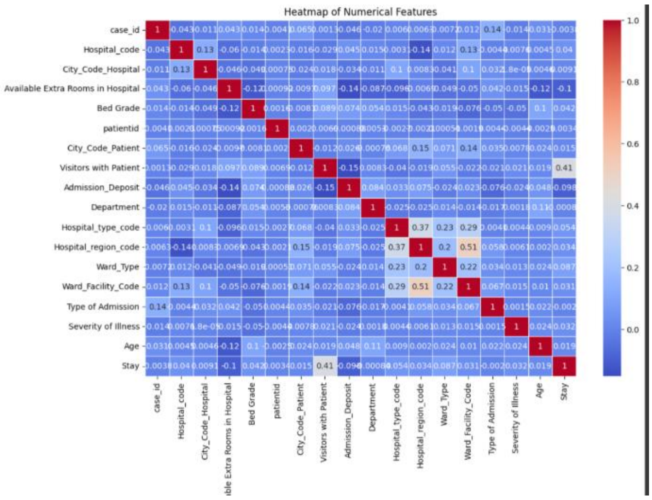


Figure 2heatmap

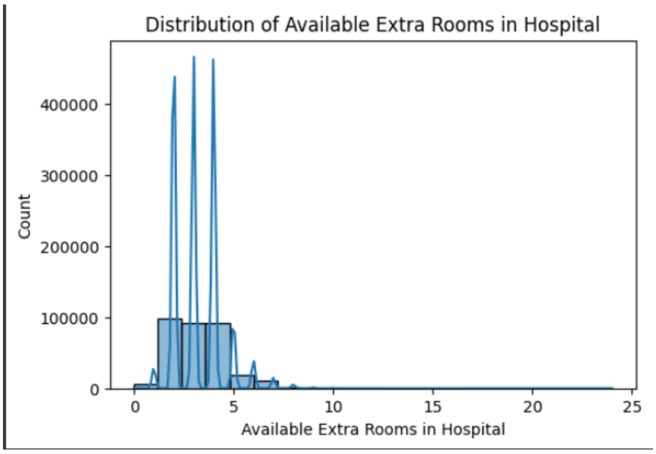
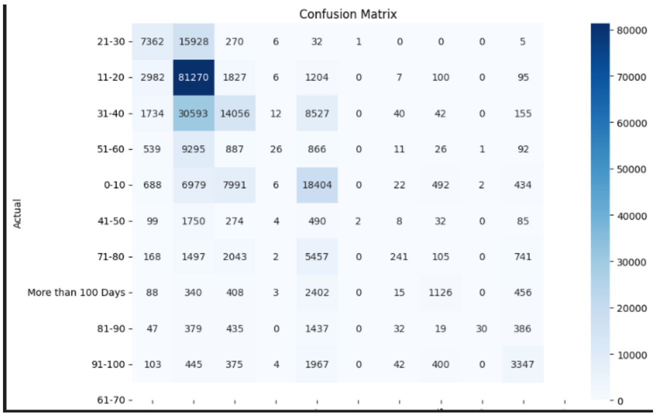


Figure 5 Available Extra Rooms in Hospital



21-30	7362	15928	270	6	32	1	0	0	0	5
11-20	2982	81270	1827	6	1204	0	7	100	0	95
31-40	1734	30593	14056	12	8527	0	40	42	0	155
51-60	539	9295	887	26	866	0	11	26	1	92
0-10	688	6979	7991	6	18404	0	22	492	2	434
41-50	99	1750	274	4	490	2	8	32	0	85
71-80	168	1497	2043	2	5457	0	241	105	0	741
More than 100 Days	88	340	408	3	2402	0	15	1126	0	456
81-90	47	379	435	0	1437	0	32	19	30	386
91-100	103	445	375	4	1967	0	42	400	0	3347
61-70										

Figure 6 Confusion matrix after model train

3) Data Visualization:

A form of visual representation of the data or information. It offers a useful summary in graphical form utilizing elements such as charts, graphs, and maps. reason they consider it important is that high and low volumes of data are displayed in a visual form, allowing everybody to see the trends and make quick, data-driven interpretations. *The visual data should make it easier for the user to generate immediate, data-driven conclusions. (tableau, n.d.)

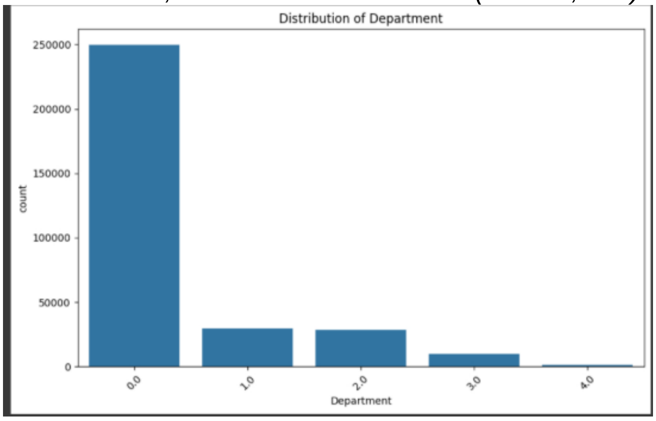


Figure 3distribution of department

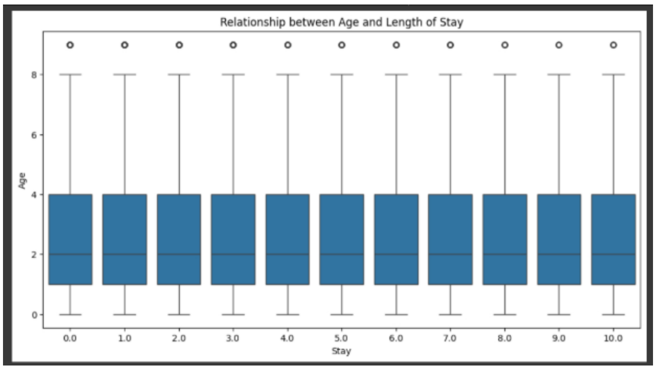


Figure 4Relationship between Age and length to stay

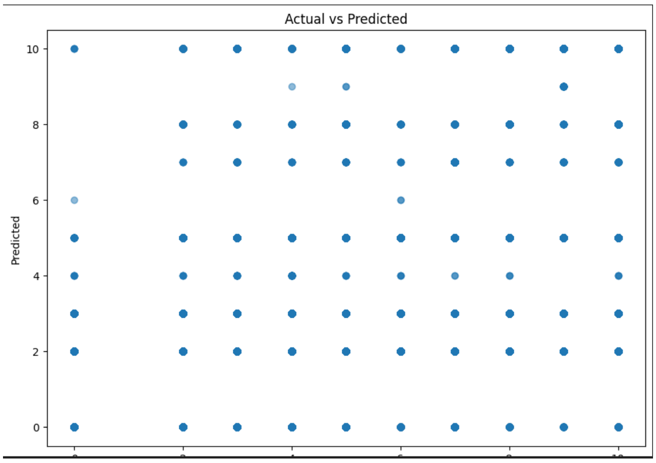


Figure 7 Scatter ploat of predict data

4) Cleaning data: The data was cleaned with the following procedures: Removal of any duplicate rows, data format uniformity; dates, occurrences, and death rate were altered to match. 6. Conclusion The obtained results show that the CatBoost classifier can be used in practice as a tool to enhance the efficiency of resource distribution in the field of healthcare.

C. Analysis

The above example demonstrates that the CatBoost classifier can be a tool to increase the effectiveness of resource management in healthcare. Specifically, predictions made as a result of classification analysis allow the better-informed distribution of such resources as hospital beds, medical personnel, etc. According to the inferences from the analysis, a few advantages of such machine predictions for resource distributing can be isolated. First, greater organizational productivity is ensured, which goes in conjunction with improved patient results and lower prices. Thus, Machine Learning can become one of the most useful tools for patient resource management in the health industry. The ability to acquire and analyze more information than humans ever have before can help health systems predict the most likely patients to need resources and target resources based on demand.

D. Conclusion

In summary, the study has demonstrated that the CatBoost classifier is a useful instrument for maximizing the use of healthcare resources. The integration of predictive analytics technology with big data application can enhance the efficiency and focus on results in healthcare systems by improving the accuracy and speed of resource demand forecasts and allocations. The study's fulfillment of the goals and questions stated at the outset of the paper was proven in the last section of this one. In this instance, the inference is that the asset management industry has a lot of prospects thanks to machine learning. The CatBoost classifier has exceptional performance and precision, hence creating a substantial opportunity to optimize resource allocation techniques. This, in turn, contributes to the provision of higher-quality healthcare.

VI. References

(n.d.). Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12387

Medium. (n.d.). *Medium*. Retrieved from https://medium.com/illumination/how-to-bring-out-the-best-in-model-evaluation-a-comprehensive-guide-9c04caf9289b

Rand Org Pub. (n.d.). Retrieved from https://www.rand.org/pubs/research_reports/RRA326-1.html

Rand org. (n.d.). Retrieved from https://www.rand.org/pubs/research_reports/RRA326-1.html

Scribber . (n.d.). *Scribbr.com*. Retrieved from https://www.scribbr.com/methodology/data-collection/

tableau. (n.d.). *tableau.com*. Retrieved from https://www.tableau.com/learn/articles/data-visualization

Wikipedia Contributor. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Data_processing

Wikipedia Contributor. (n.d.). *Wikipedia Model-based-testing*. Retrieved from https://en.wikipedia.org/wiki/Model-based_testing

Appedix

1) Github link

Code: Code Repository (Click here)
Dataset: Dataset(Click here)