Performing Analytics on Money Payments in Hospitals

Pranav Ramakrishnan
dept. MSc Data Analytics
National College of Ireland
Dublin, Ireland
x23107979@student.ncirl.ie

Asish Mathai Mathai
dept. MSc Data Analytics
National College of Ireland
Dublin, Ireland
x23173645@student.ncirl.ie

Jacob Saju
dept. MSc Data Analytics
National College of Ireland
Dublin, Ireland
x23166363@student.ncirl.ie

***Abstract*—This project aims to investigate how different factors, such as the number of patients who have to be readmitted, the number of surgery complications, or way too low care provided to the patient, impact the amount of money Medicare pays to the hospital in the United States. The project uses data from government sources to learn more about payment reduction programs and assess how well hospitals perform from a variety of perspectives. It involves the collection and preprocessing of various health-related data using MongoDB and PostgreSQL, exploratory data analysis, and trend visualization using Python libraries such as Matplotlib and Seaborn. The project also includes the training a logistic regression model for predicting payment reductions that include data balancing with SMOTE and feature selection with RFE. As a result, the project revealed that excess readmissions for heart attack and complications in hip/knee surgeries are high importance predictors. The findings of the study suggest that the quality of provided care, as demonstrated by healthcare outcomes, is crucial for determining the amount of Medicare payments. As a result, healthcare facilities should focus on improving patient outcomes to prevent financial losses associated with payment reductions. Thus, the study helps better understand the link between healthcare quality and financial incentives suboptimal.**

***Index Terms*—Python, Matplotlib, seaborn, RFE, MongoDB, PostgreSQL**

I. Introduction

The healthcare industry is always attempting to maximize patient care efforts while ensuring optimal resource allocation across the supply chain. One of the best ways to implement the goal is to drive data-derived insights in order to improve the performance of a hospital on different levels. The project discussed below entails the integration and analysis of several databases that can be combined to make comprehensive and actionable insights applicable to healthcare providers involved in different kinds of reimbursement programs. As a result, the current project covers the following three programs: the Hospital Readmissions Reduction Program , Hospital Acquired Conditions Reduction Program , and the Hospital Value-Based Purchasing program . The central idea behind the abovementioned reimbursement programs is to make sure the hospital is incentivized to develop a high-quality service evident by a low readmission rate and minimal number of acquired conditions .

Our project has involved heavy data retrieval and transformation efforts due to vast disparities in databases such as patient readmission rates, hospital-acquired condition scores, and value-based purchasing performance. The integration of datasets used one variable provider identifier allowing the merging of scores into a single dataset for millions of data points . The dataset produced from the integration of datasets enabled the use of exploratory data analysis methods to learn and visualize data distributions and patterns. Histograms, box plots, and maps were critical in understanding performance distributions and geography .

The main part of the project, which was based on machine learning techniques, focused on logistic regression classifying whether a hospital will have its payment reduced. Initially, the model produced poor results due to inadequate testing on untrained data and imbalanced classes involved with class variables. The model was refined through Synthetic Minority Over-sampling Technique and Recursive Feature Elimination feature selection to obtain the best model outcome. The model created a robust prediction model identifying valuable hospitals targeted for interventions to ensure better patient care results. The project has articulated the value of data-driven approach to decisions and the implication of combined databases in determining valuable actions for hospital performances and government policy.

II. Literature Review

Healthcare performance measurement and analysis have been a central research and clinical focus in the last decade, facilitated by increasing attention to improving patient outcomes, quality of service delivery, and resource utilization across health systems. Various studies, papers, journals, and reports have been published on the subject, expounding the various performance-based metrics used, the methodology for analysis, and performance-based strategies to drive healthcare delivery improvement. One of the most frequently examined aspects of healthcare performance is hospital readmission, especially for heart failure, pneumonia, acute myocardial infarction , chronic obstructive pulmonary disease , and elective hip/knee replacements, among others. Different studies have explored readmission-related factors like patient demographics, comorbidities, social economic status, quality of care, and

Identify applicable funding agency here. If none, delete this.

character characteristics like the ownership and specialty of healthcare providers. One such study is the research conducted by Jencks et al. , which discussed the significant financial burden readmissions posed to American taxpayers and the possible interventions to cease preventable readmissions. Moreover, other studies like those of Kansagara et al. and Dharmarajan et al. have elaborated on the essence of identifying high-risk patients and applying appropriate interventions to reduce readmissions. Similarly, healthcare performance has also been assessed through the value-based purchasing programs and reimbursement models. For example, the Hospital Value-Based Purchasing initiatives provide incentives for improving quality of service delivery by tying payments to sellers' performance on various high-quality metrics. On the other hand, Hospital-Acquired Condition Reduction Program penalize solidify high prevailing HAC rates to further incentivize quality improvement. Machine learning algorithms, including logistic regression, have also been used to define healthcare performance and to predict various outcomes like payment reductions, readmissions, and mortalities. For instance, various studies like those of Rajkomar et al. and Obermeyer et al. have demonstrated the effective use of machine learning to predict poor patient outcomes and identify at-risk patients. However, the literature agrees on the critical nature of healthcare analytics as a feasible way to drive the cost-cutting and eventual improvement of patient care delivery.

III. METHODOLOGY

A. Description of Data

a) Dataset 1: Hospital-Acquired Condition Reduction Program

The dataset methodology is a composition of data bits on the statistics of the medical facility about the performance of the fiscal year 2024 in the prevention of healthcare-associated infections . Besides, IHC consists of the Patient Safety Indicator PS 90 Composite Value, Central Line-Associated Bloodstream Infection CLABSI SIR, Catheter-Associated Urinary Tract Infection CAUTI SIR, Surgical Site Infection SSI SIR, Clostridioides difficile Infection CDI SIR, and many more. Furthermore, the dataset is added to the Hospital-Acquired Condition Reduction Program HACRP JSON data, which consists of the information about hospitals, overall HAC rating, and payments reduction factor . The HAC total score is calculated based on the scores of various measures and all hospitals with the score above 75 percent have their payments reduced.

b) Dataset 2: Hospital Readmissions Reduction Program

The dataset from the HRRP JSON comprises 19,300 observations, and each of them provides measure scores for 12 variables. The wasteful re-admissions are identified by the ratio. The ratio is computed as the predicted 30-day readmission rate for specific conditions, such as heart attack, heart failure, pneumonia, COPD, hip/knee replacement, and CABG, over what would be expected for any hospital responsible for a particular patient source.

c) Dataset 3: Hospital Value-Based Purchasing

The JSON dataset's focus is the file's Clinical Outcomes Domain Scores. It covers all 2,731 Medicare hospitals and provides varying metrics stated through 36 columns. Based on the file, It seems that the paper concentrates mainly on the hospitals under the VBP hospital scheme. This is because it contains performance rates and scores based on clinical outcome indicator . I presume that the indicators contained in the VBP are patients' outcome rates, the successfulness of treatments, and quality of care among others.

B. Data Pre-Processing

Data preprocessing is an essential step in any data analysis pipeline that ensures that the data is adequately treated for analysis and modeling. This section details all the preprocessing steps that are applied to the Hospital-Acquired Condition Reduction Program , Hospital Readmissions Reduction Program , and Hospital Value-Based Purchasing healthcare datasets. The first task in any typical preprocessing pipeline is handling missing values. Missing values are attributed to various reasons, such as data collection errors and incomplete records. We identified missing values and used appropriate imputation techniques to handle them. The mean or median imputation is a commonly used approach where the missing values are replaced . This approach maintains the dataset's integrity without affecting the subsequent analyses. The second preprocessing step is encoding the categorical variables. These are attributes such as payment reduction and state that must be converted to numerical values to be learned by machine learning algorithms . Two common encoding techniques used include one-hot encoding and label encoding. The choice between the two depends on the nature of the categorical variables and the requirements of the analysis. Data consistency is another integral preprocessing step. It involves conducting checks on duplicate entries, validating data ranges, and standardizing numerical variables. Removing duplicates eliminates redundancy, ensuring that the set has no duplicate entries. Validating data ranges ensures that numerical attributes are within expected bounds and do not act as outliers. Standardizing numerical variables scales the numerical attributes to a common scale or distribution. It helps improve the performance of machine learning algorithms by making the data points more homogeneous.Data preprocessing is pivotal as it prepares the datasets to be analyzed and modeled. By handling missing values, encoding the categorical variables, and ensuring data consistency, we can clean and format the data, ensuring that it is set for meaningful analysis and predictive accuracy.

C. Data Loading and Database

The current project was designed to use a comprehensive and systematic approach to data loading and data management inside the databases. From the very beginning, we initialized our MongoDB, and, thus, addressed the problem of uploading the data to this or that instance . Before we needed to start inserting the data into the instance used or getting it retrieved, we needed to solve the issue of populating the data related to the three major aspects of healthcare into the existing MongoDB collections on the Hospital Readmissions Reduction Program , the Hospital-Acquired Condition Reduction Program , and the Hospital Value-Based Purchasing . To load the existing data into our MongoDB server, we ensured that the collections with similar kinds of data were indeed either non-existent or already empty ; this allowed us to avoid duplication and logical inconsistency.

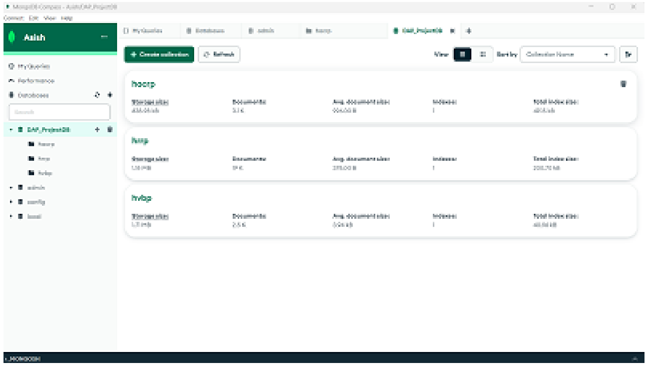


Fig. 1.

Nowadays, the data related to the three distinguished areas of healthcare could already be considered uploaded into our MongoDB; hence, it is high time to move to the data management servers with ending dap medicare . In the given data management system with a predefined name, each kind of structured data can be stored; since, previously, one could have dealt with the misfiles in the system; therefore, we needed to drop every database that had a name used for that aim . Thus, we managed to create the tables in our custom-defined data management server. Each file that we used, therefore, was personalized to contain the attributes and metrics concluded; meanwhile, the Python libraries like Pandas and SQLAlchemy enabled us to transfer the data from our struct and temp data management servers to all the shared postgre table without a major loss associated. All the major processing, like the necessary pre-processing, was already included as a function to the methods used. Thus, every step of our work was clear and easy to handle.

IV. Exploratory Analysis

a) Hospital-Acquired Condition Reduction Program (HACRP): Hence, the first step in handling the Hospital-Acquired Condition Reduction Program data is conducting descriptive statistical analysis. This will involve identifying the characteristics and distribution of the observations and variables in the dataset. To address, one has to examine the types of data, summary statistics, and the interconnections amid the variables. This will help the practitioner identify the data structure and pattern, leading to the identification of potential data problems such as anomalies, outliers, and data skewedness. At this level, the analyst could also aggregate the data by state to investigate the relationship between the total HAC scores and the payment reductions across all the states. This analysis can indicate geographical differences in health outcomes and financial incentives. For instance, examining the mean or median total HAC scores and the mean payment reductions per state can help the practitioner to understand the tendencies or disparities regarding the facility's performance and the implement of the reimbursement-oriented policy .

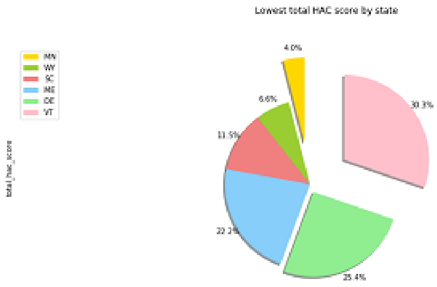


Fig. 2. .

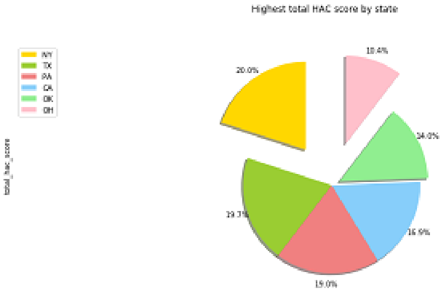


Fig. 3. E

Besides, the data can assist to compare one state with another to help one state to benchmark with another to share the best practices in the industry. In addition, visualizing the data applying such graphic tools, as heatmap, box plot and pie chart can help to examine the variable spread and the distribution. The heatmap illustrate the correlated variables and cluster the data. The box plot demonstrates the numeric variables spread and the central tendency to help identify the outliers, extreme values, and other deviations . The pie chart helps show the categorical variable proportions and frequency count also known as count by category. Together, these tools enable the practitioner to communicate finding in the data effectively for evidence-based decision-making and policy development, particularly in the facility management context.

b) Hospital Readmissions Reduction Program (HRRP): Descriptive Statistical Analysis In this step, I explored the distributions of expected, predicted, and excess readmission rates based on various medical conditions including heart attack , CABG , COPD , HF , THA/TKA , and pneumonia . By comparing the distributions, I had the opportunity to explore the extent of variability and central tendencies in readmission rates attributed to different conditions. The above

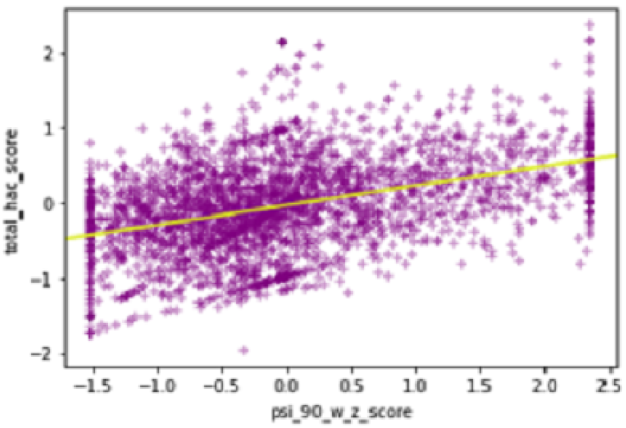


Fig. 4.

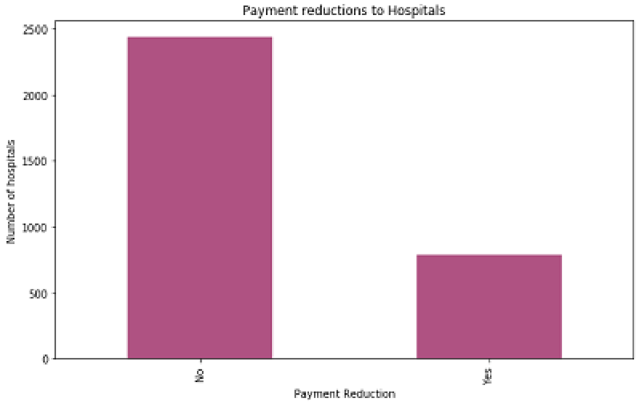


Fig. 5.

statistics are vital in determining the expected rate of health facilities management in responding to patient readmission. 2. Grouping by Measure Name and State I explored the data by grouping measure name and state. This was to derive the average readmission rates and excess ratios in each measure name and state group. Measuring data in this level is useful in exploring the relationships and patterns of the readmission measures necessary in guiding healthcare policies in reducing the level of readmissions. 3. Data Visualization of Target Variable Visualizing the distribution of the target variable, that is, the payment above a threshold value, was essential in informing the model. The visualization was through the count plot metrics that allowed me to present the proportions of hospitals that received payment reductions and those that did not. It was important in identifying the likelihood of model prediction class imbalance, and hence applied the resampling technique. The descriptive analysis of the HRRP dataset was instrumental in providing readmission rates and excess ratios and making baseline predictions of payment reductions, which are essential to health quality.

c) Hospital Value-Based Purchasing (HVBP): One of the critical exploratory analysis components is to understand the data, which allows investigating the structure, content, and implications of the dataset . In regard to the Hospital Value-Based Purchasing Table, the first step of the analysis is to take an overview of the dataset. Firstly, we obtain the dimensions of the given dataset, which is the first reflection of how many rows and columns the dataset has. This type of analysis allows me to predict the dataset's difficulty and volume, guiding the further analysis process. Besides, the next analysis is more detailed and delves deeper into certain aspects of the data. More specifically, my analysis involves metrics associated with hospitals' performance in various medical treatment areas . Metrics included achievement points, improvement points, and measure scores for certain conditions such as heart attack and pneumonia . It is important to understand I am looking to obtain the ways of how hospitals perform in separate areas of care based on the chosen metrics. Along with this, the exploration also attempts to find patterns and connections in the data, which implies hospitals performing well in one area perform in others as well. Such revelations help in finding the factors which drive the overall hospital performance. In the end, it is easier to formulate this information and make it clear for others with the help of visuals like graphs and charts. They can help see certain trends, outliers, and connections in an easier way and are more accessible for sharing with others for reporting. Exploratory analysis involves a systematic overview of the dataset, detailed focus on metrics associated with performance and their connection, and the effort to visualize other observations.

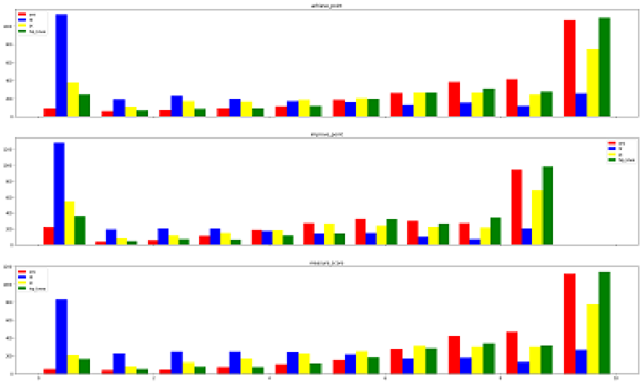


Fig. 6. .

d) Hospital Readmissions Reduction Program (HRRP):

V. Merging Databases for Comprehensive Analysis

Merging data from two types of databases: The databases called MongoDB and PostgreSQL. The databases provide lots of details about various health care programs. For example, we gathered data about reducing hospital readmission, preventing hospital-acquired conditions, and enhancing hospital value-based purchasing . Initially, each of these databases allows us to look at one side of the health coin: It's similar to watching part of a puzzle. When we assemble all parts of the puzzle, it’s like looking at a complete image . Despite that, each of the current databases has its information. To this extent, they give us a holistic picture of how health goes. Our previous databases showed little parts of the whole; for instance, MongoDB showed how well the hospital reduced the likelihood of a patient coming back for a reoperation surgery after seeing him. On the other side, PostgreSQL gave us information about how well the hospital is preventing patients from getting sick while in the hospital . By merging the databases named above, we collect everything we might need to obtain a variety of health indicators. It’s like having all the pieces of your puzzle in one place to see the whole picture and understand what’s happening. In this regard, merging our databases helps us conduct a complete analysis of health. Our study about health can include a variety of aspects: hospital experience management and how hospitals perform endlessly of our greatest concern and what other factors might be impacting health. Thus, this knowledge provided can help personnel in the health care system contribute to informed decisions and perceptions to enhance patient health and health-related metrics.

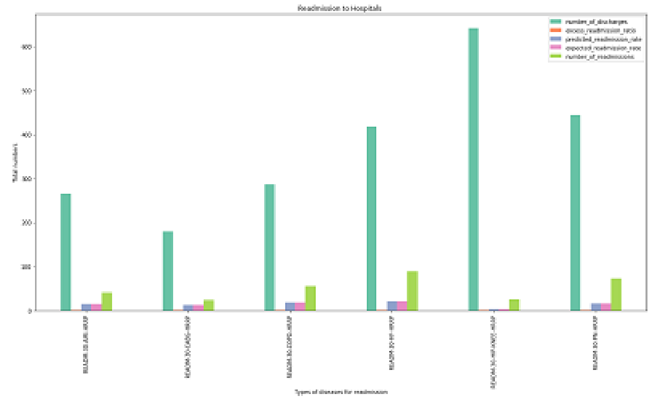


Fig. 7.

VI. Visualizing Target Variables

After merging data from the different healthcare programs, including the Hospital Readmissions Reduction Program , HACRP , and HVBP , the next step involves visualization of the target variables. This involves the creation of numerous plots using various plot types using the imported Matplotlib and Seaborn libraries in Python. The purpose of generating different types of plots is to draw more insights from the data. As provided for in this section through the code provided, the creation of the box plot, histogram, and scatter plot indicates information on how the measures are distributed and interact. One of the insights drawn from the box plot is the identification of measures distributed in various hospitals or states. As such, it identifies the issues and the areas of interest.

Additionally, the histogram gives the frequency distribution of the measure, wanting to identify whether these distribution differs. Furthermore, the scatter plot shows the relationship between various variables or whether they are correlated. Specifically, it helps find out whether a hospital with more achievement points in one program also has more high performance in the other program. Therefore, through these three types of visualizations, we can identify the patterns, trends, and disparities in healthcare performance. Target variable visualization is thus beneficial for our analysis since.

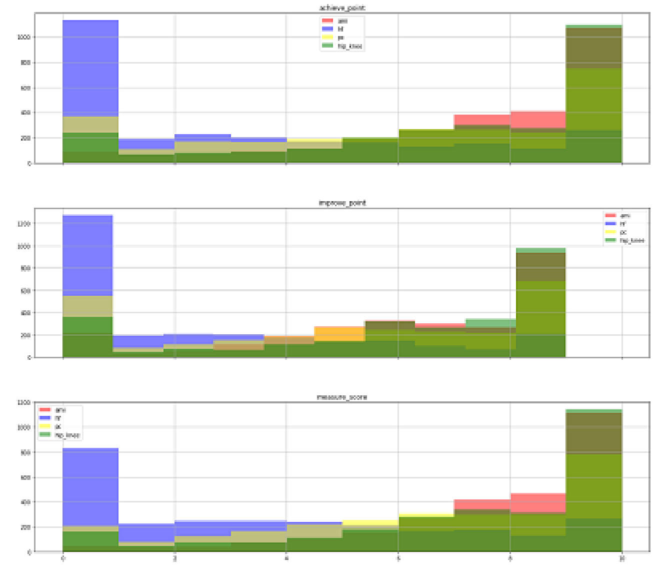


Fig. 8. .

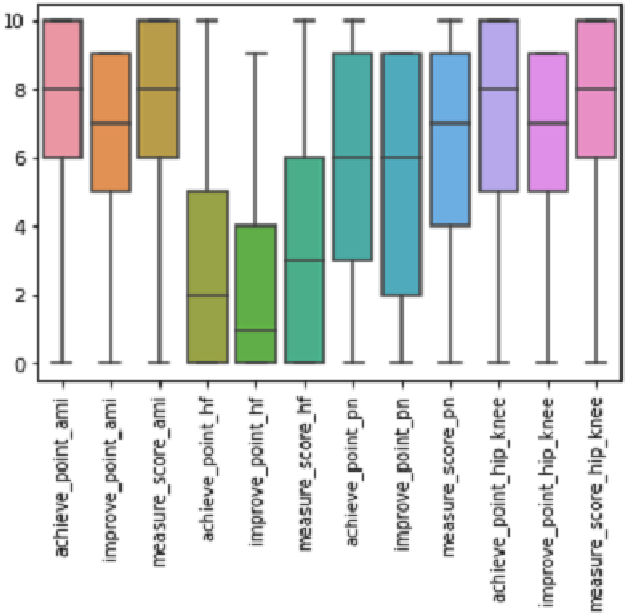


Fig. 9.

Firstly, it helps the stakeholders as they find it easy to interpret the data. It further solves the exploratory analysis helps identify areas that need focus and emphasis while interpreting the data . Finally, it is vital to our analysis when presenting the findings since it makes the results easy to interpret.

VII. Training Logistic Regression Model

Before we start explaining how the model is trained, I would like to mention that getting the data completely reviewed, examining all of the data and understanding the nuances is the hardest and most time-consuming part of our project. That being said, in order to proceed to the next step, we should train our logistic regression model. In other words, the logistic regression model implies that what we are training, guiding us in predicting an outcome in healthcare when we haveare uncertain about, the environment, and many more.

However, we need to select the ready-to-use logistic regression model, choose items to predict or the columns' pro. I have selected the logistic regression model since it is almost always a good choice for problems such as separating things into one of two categories such as a probability that the patient will be readmitted, or the patient is at risk for diabetes acquired during a hospital stay. Here, let us show what stands behind our choice. The binomial hypothesis testing is the most efficient in our case, as it is optimal to use when: when the dispersals are known two outcomes, or pro being readmitted or not , 5 percentage confidence, hospital acquired pneumonia 57.28 percentage chance, reinfarction 0.17 percentage chance of patients being readmitted, and 70.72 percentage patients not readmitted, Conclusion: Here is a 68 percentage pro readmitted rate between the patients data set on whether they have Diabetes and are within Medicare and private insurance, 1636756 case files. from which we developed a corollary. In order to develop the model, we use a merged dataset which the data is an outcome of many hours used on combining separate data provided in the different sources. Specifically, what this feature is expected to be delivered in the dataset is the features in the training set which are the features to be implements in the training process. Thus, because some types of features may not be promising as of related to the outcome, the omitted . After the model has prepared, the model can be used to predicts; data is inserted, and then data accessed, from where we determine the final output.

VIII. Evaluating Model Performance

During the Evaluating Model Performance phase, we take our logistic regression model for a ride. The different metrics we have discussed are accuracy, precision, recall, and F1 score. They provide us with numbers that how good predictor of healthcare outcomes a model is. While accuracy measures how frequently over the occurrences the model is correct, precision measures the proportion of positive occurrence among the predicted positive cases. Recall measures the ratio of positive occurrence that is accurately documented by the model. The final metric F1 score is a combined metric of precision and recall, giving us an overall score of the model. By considering these metrics, we can easily see where the model is shining with glory and where we need to put more focus to achieve the best. If a model has high accuracy but low recall, the model is not correctly representing all positive cases. It is not good because we don't want to lose any major relevant occurrences of data. And if the precision measures are low, we obtain a large number of false-positive cases. It is also unacceptable as it may turn into unnecessary actions which sometimes work counter to the actual case. Hence, it is emphasized that measuring model performance helps identify an excellent rank in a model and where the model stumbles. If there are flaws in it, then based on this we can work on a model by tweaking or by adding extra data, or we can try another algorithm. It is aimed at having a good-edged model providing healthcare professionals an accurate prediction of healthcare so that they can give their patients the best care.

IX. Improving Model Performance

Improving the performance of a logistic regression model is essential to ensure accurate predictions in the healthcare setting. We can utilize several methods for that purpose. One method utilized is feature engineering, where we create new features from the original features in the dataset in the hopes that the new features will contain more information relevant to the predictions. For example, we can form new f features by combining existing ones or creating interaction terms that capture possible relationships between variables. In doing so, we expect to enhance the model's ability to learn the data's underlying patterns. Another method we can employ is hyperparameter tuning, where used to refer to how the model learns from the data. The hyperparameters can be set to regulate how the parameters are optimized during the training of the model . By changing the hyperparameters, the model can be set to its optimal configuration for its best performance. This process may entail trying many formulations of hyperparameters, such as through grid search or random search, assuming various hyperparameter configurations and taking the best-performing one. Furthermore, ensemble methods may also be employed. Ensemble models combine the prediction of several models to produce a more accurate prediction collectively. Ensemble models can improve the predictions of the individual models by training on slightly different subsets of the training data. Two categories of ensemble models are bagging and boosting. Bagging models average out the prediction of "weak learners," a classifier that performs slightly better than random prediction. Boosting involves training models on new "weak learners" utilizing the prediction errors from the trained learners . In addition to creating a more robust model design, ensemble models also lessen the likelihood of overfitting by averaging the prediction errors of their members. In summary, various methods may be employed to improve the performance of a logistic regression model. These methods include changes to the model parameters and adding more data through feature engineering and creating more complex models in line with that data. Through these methods, the performance of the logistic model can be optimized, and predictions can be made with a higher level of confidence.

X. Addressing Class Imbalance

Class imbalance is a critical problem while working with health datasets due to its implications for the accurate functioning of the predictive model leveraged to identify potential patients at risk of readmission and other adverse outcomes. In the present study, class imbalance refers to a situation where one class, such as patients with readmission outcomes, is significantly less common compared to the other, such as those who do not experience this outcome . In particular, the class imbalance occurrence challenges the notion that any predictive model is biased if it creates scenarios where most instances are members of the majority class, therefore mistakes being at risk

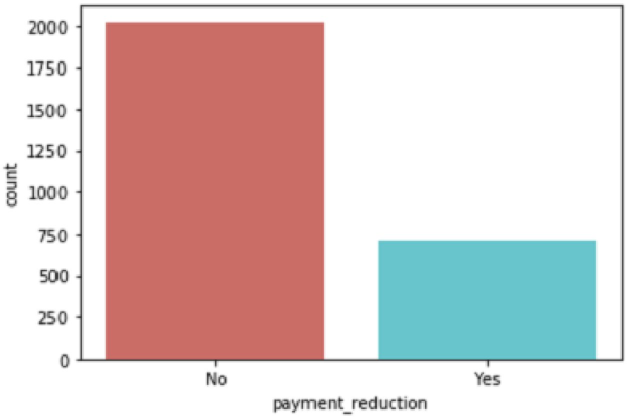


Fig. 10. .

. Oversampling is an approach that addresses the problem of class imbalance by either doubling the artificially increased number of instances in the minority class or creating their number, similar to the distribution of the majority class. Oversampling allows one's predictive model to overestimate the instances of the minority class, thus allowing it to reduce overrepresentation bias to make more accurate predictions. Undersampling is an approach that reduces the instances of the majority class to balance the two classes. While undersampling enables a model to be free from class imbalance when correctly implemented, it may disadvantage the organization by causing information loss and high bias when done incorrectly. Synthetic data generation enables an organization to create artificial instances of the minority class within the texture of the existing data. Specifically, the synthetic data generation technique produces samples that are similar in detail to the existing minority ones and introduces various data variations that greatly increase the data's diversity to make the model generalizable across most scenarios. Thus, health organizations can address the issue of class imbalance and ensure that their predictive models make accurate estimations. The best fixes ensure that a model remains generalizable and predictable, offering valuable insight into who is likely to face negative outcomes. Therefore, it is critical for an organization to check in on their model after completion, considering that this fix might not be entirely safe since data input features change once in a while.

XI. Recursive Feature Elimination

One of the key parts is the selection of features used to train the model while creating predictive models, particularly in complicated datasets, such as the ones in the healthcare field. For this, the popular technique is Recursive Feature Elimination . In general, RFE eliminates the unimportant features systematically, determining which of them is really important. Given the current project, which combines many types of data, this method is essential for the dataset refinement .

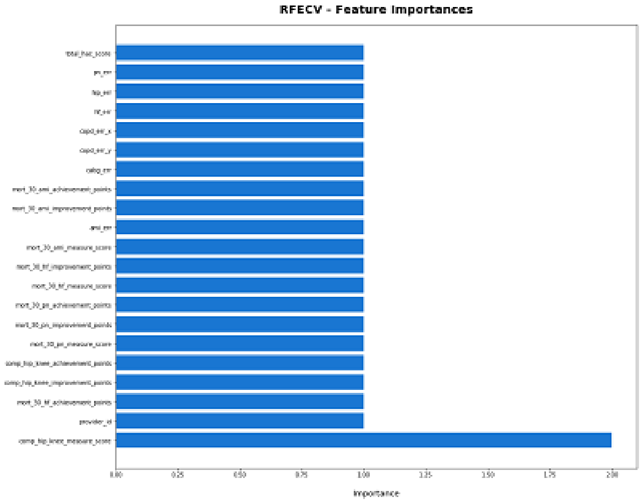


Fig. 11. .

In terms of simplification, it can be presented as follows. For example, if we have a dataset with many different factors, not all of them equally influence whether a patient will be readmitted to the hospital. Using RFE, it is possible to determine the most impactful by removing the least important feature each iteration new model training is performed. Therefore, in the described situation, it is done after the database merging, while preparing the dataset. The next step is the feature selection for the model training, and in that sense, RFE is used there. Probably, the provided code snippet illustrates how RFE is used, given the description: the model is train iteratively, the feature importance is calculated, and the least ones are removed. That way, the predictive model obtained is simpler, i.e., it only recognizes the features that are important in predicting the patient outcomes. Additionally, overfitting is reduced as well. Therefore, RFE aims to reduce the model's complexity so that the trained predictive model can make predictions about the patient outcomes while being interpretable.

XII. Conclusion

In conclusion, the analysis has added a fundamental understanding of the hospital performance and payment reduction association by implementing a logistic regression analysis

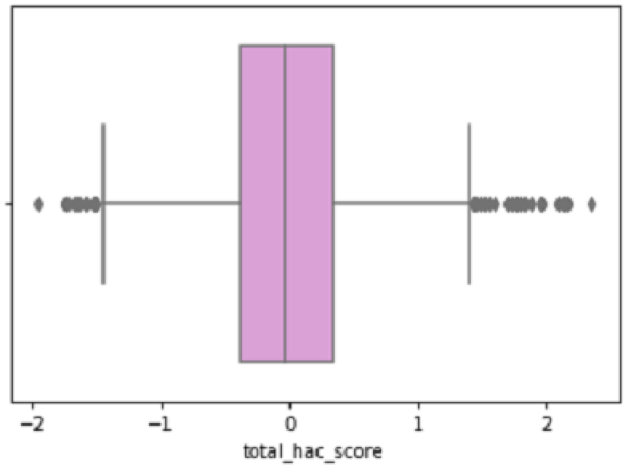
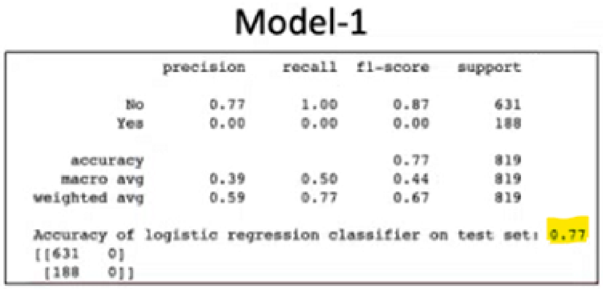


Fig. 12. .

that created the model and trained it using 70 percentage of the dataset to ascertain the likelihood of the payment reduction status of the remaining 30 percentage of the hospitals . Moreover, the model as computed, according to the discussed analysis, results in an accuracy of 68 percentage, high precision, recall, and f1-scores for both positive and negative cases and presents a strong case of cutting out hospitals depending on the payment reduction factor.



Model-1

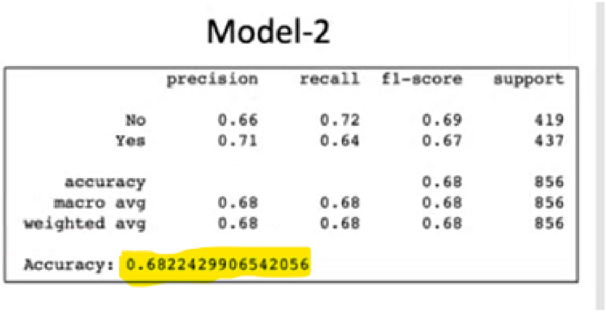
```
              precision    recall  f1-score   support

          No       0.77      1.00      0.87       631
         Yes       0.00      0.00      0.00       188

    accuracy                           0.77       819
   macro avg       0.39      0.50      0.44       819
weighted avg       0.59      0.77      0.67       819

Accuracy of logistic regression classifier on test set: 0.77
[[631   0]
 [188   0]]
```

Fig. 13.



Model-2

```
              precision    recall  f1-score   support

          No       0.66      0.72      0.69       419
         Yes       0.71      0.64      0.67       437

    accuracy                           0.68       856
   macro avg       0.68      0.68      0.68       856
weighted avg       0.68      0.68      0.68       856

Accuracy: 0.6822429906542056
```

Fig. 14.

Furthermore, the analysis has supported several primary predictors of payment reduction in hospitals, and hip-knee complication, heart attack excessive readmission, and total HAC score were ranked as the top predictors. The predictors show that there is a substantial correlation between the quality of hospital care and payment. Hence, hospitals with the least number of complications and readmission have the least payment reduction factor. Therefore, the analysis presents the basis of evidence for better ways of ensuring quality patient care. Further, the analysis would be better in the future using other healthcare data, including other health schemes which form robust variables and levels in the PCDS system and a thorough understanding of the factors contributing to payment reduction. The analysis eventually provides the basis for a more in-depth understanding of the health sector for the azimuth of payment and quality care.

References

[1] Han, J., Pei J., Kamber M. Data mining: concepts and techniques. Morgan Kaufmann; 2011.

[2] Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer science and business media; 2009.

[3] James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning . Vol. 112. Springer; 2013.

[4] Goodfellow I., Bengio Y., Courville A. Deep learning. MIT press Cambridge; 2016.

[5] Kohavi R., Provost F. Glossary of terms. Machine Learning. 1998;30(2-3):271-274.

[6] Murphy K.P. Machine learning: a probabilistic perspective . MIT press; 2012.

[7] Bishop C.M. Pattern recognition and machine learning. springer; 2006.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

[10] Raschka, S., and Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2 (2nd ed.). Packt Publishing.

.