Predicting Real Operating Room Occupation, an Interpretable ML Approach

Teresa Marcelino, teresamarcelino@tecnico.ulisboa.pt[1]

[1]Instituto Superior Técnico, University of Lisbon, Portugal

***Abstract*—Nowadays, the potential of using Machine Learning (ML) techniques to solve real-world problems is extensively explored, and many are the application domains such as cybersecurity, aviation and healthcare, where there is in-depth research into their applicability. With the amount of data currently gathered in the hospital environment, models capable of learning and improving automatically through the use of data might solve problems that endanger the proper functioning of hospitals. The Operating Room (OR) is a high-cost environment, and its usage must be efficient. Therefore, our presented solution focuses on developing interpretable prediction ML models for an OR decision support system to improve the prediction of surgical times, comparing them with traditional methods to aid the OR scheduling process. We implemented three different ML models, XGBoost, RuleFit and a neural network, and we compared and analyzed their performance, including both accuracy and interpretability. For each of these algorithms, we implemented three different strategies. Then, since surgical durations showed a significant imbalance and this is known to hinder the performance of accuracy-based ML algorithms, we trained a Gaussian Mixture Model (GMM) to learn the probability distribution on the minority values of our label enabling sampling to overcome the imbalance. The performance of the models on balanced and imbalanced datasets was compared using the Utility-Based Algorithm (UBA). This research work is an evidence that the proper implementation of interpretable ML technologies can significantly improve current standards of estimation, representing a cost reduction from an operation's perspective, maintaining the decision-makers' confidence in the system.**

***Keywords*—Operating Room, Machine Learning, Efficiency, Surgery Case Duration, Interpretable Models**

I. Introduction

THE schedule planning of operating rooms is one of the biggest challenges in the health sector since this service is a hospital key element, responsible for around 42% [1] of income but, simultaneously, due to high cost of use, most hospital expenses are related to the OR. Operating rooms are costly, ranging from \$30 to more than \$100 per minute [2]. Therefore, they represent a critical financial bottleneck and it is crucial to maximize the efficiency [3].

Besides financial criticality, this service is one of the biggest headaches in the hospital due to its extremely high complexity. The interactions between different healthcare professionals (such as surgeons, patients, nurses, and anesthesiologists), the difficulty of predicting the time in certain types of procedures due to unpredictable patient circumstances, the need for sterile material that depends on third parties, and the availability of beds in Post-Anesthesia Care Unit (PACU) are just a few reasons that help us understand the difficulty in managing this service [1].

Nowadays, the historical information on the OR operation is well annotated and there is a lot of information available, such as the surgical service performing the procedure, the duration of surgery, and the patient's information, which has a vast potential to optimize the OR pathway. However, these data are still not fully explored in most hospitals and forecasts of the surgery duration are made based on the experience and opinion of surgeons, that estimate the operating times that they consider necessary, or by using simple statistics on the conventional Electronic Health Records (EHR), the electronic collection of a patient's medical history where the historical average for each case duration can be performed. The study conducted by Laskin, Abubaker, and Strauss [4] with oral and maxillofacial surgeons showed that only 26% of surgeon estimates were accurate and there is an overestimation in 42% of the analysed cases. Overestimation occurs because various factors can influence the doctor's prediction, simply because complications arise during the procedures or sometimes the doctor may overestimate or underestimate the surgery depending on the number of appointments they have scheduled on that specific day.

Regarding the EHR sample means method also used by CUF, the healthcare provider whose data was analysed in this study, it allows predicting surgical time based on the average of historical data from a specific procedure or surgeon. However, this type of approach does not take into consideration other factors, such as patient and procedure-specific information, which can influence up to 30% of the total surgery duration [5]. Tuwatananurak et al. [5] used Leap Rail engine to show how can a machine learning algorithm improves the EHR predictions, getting a significant reduction of around 70% in the total scheduling inaccuracy, improving the estimation in approximately about 7 minutes per case regarding actual case duration. Moreover, in Rozario and Rozario [6] work the baseline time prediction was the surgeon's average procedure time of the last 10 cases. However, with the current method, case times follow a Gaussian distribution with an underestimation in 50% of the cases.

As these modest results evidence the challenging nature of the problem, they also encourage a machine learning approach, given the excellent results that machine learning methods have provided in natural language understanding [7], computer vision [8] or games [9]. We must, nevertheless, provide an interpretable prediction when informing human decision making, particularly in healthcare.

II. Related Work

The need for efficiency in planning and scheduling procedures has led to an increase in research in OR related problems since 2000, with a significant increase in publications since then [10]. In addition, since 2015, there has been an exponential growth in research in terms of the application of ML in the scope of medicine since the availability of big data and the growth of data science have contributed positively to the decision-making processes [2].

Firstly, statistical analysis of the variability of surgical durations has been studied for years [11], and techniques such as Lognormal Estimation and Bayesian statistical techniques were intensively explored. These approaches find the best fit in a family of distributions to predict surgical durations and characterize relationships between variables. Stepaniak et al. [12] fitted a 3-parameter lognormal model that improved the OR scheduling and reduced the mean over reserved OR time per case by up to 11.9 minutes. Strum, May, and Vargas in two studies [13, 14] compared the modeling of surgical procedure times with normal and lognormal distributions and concluded that lognormal models provide accurate predictions and fit better procedure times.

Moreover, models based on Gaussian Mixture Model (GMM) are also widely applied as a prediction model, even in the surgical area e.g. support patient flow models [15]. The Bayesian method obtained by Dexter and Ledolter [16] allowed improving predictions for cases where few or no historical data exist and concluded that GMM can be a reasonable choice when surgical times do not follow a lognormal distribution. Taaffe, Pearce, and Ritchie [11] also studied the application of Kernel Density Estimation (KDE) to model surgical durations. The results outperformed traditional methods such as lognormal and GMM when there is limited historical data.

Other studies also investigate the potential of using mathematical models to improve durations, showing an OR efficiency improvement by combining advanced mathematical and financial techniques [17, 18]. However, these approaches postulate a simplified model for the data distribution and this thesis takes a data-driven, machine learning approach, while keeping interpretability as a requirement. Although machine learning and statistics are closely related fields in terms of methods, their main goal is different. Lee and Yoon [19] summarized the differences between classical statistical analysis and big data medical analysis. While ML models are designed to make the most accurate predictions possible and find patterns in the data that can be generalized, statistical models are designed for inference about the relationships between variables and reach conclusions about populations or derive scientific insights from data. Thus, in ML, the algorithm learns from a considerable amount of data and generates the hypothesis from the data, while in statistical models, we need to commit on *a priori* assumptions based on various underlying probability distribution functions [2].

Even in the machine learning field, the high complexity of the OR environment allows and leads to different approaches to the problem and the use of different metrics by authors and researchers. Fairley, Scheinker, and Brandeau [20] defined as objective the minimization of maximum Post-Anesthesia Care Unit (PACU) occupancy, using constraints to control and maintain OR utilization. Thereby, to predict PACU recovery times for each patient, the author used a gradient boosting tree model as input in a program that formulates the schedule of procedures in the OR. Abedini, Li, and Ye [21] developed a blocking minimization model to reduce the number of blockings between OR and PACU, allowing the hospital to define the OR schedule for the next day, considering the current stage occupancy of the OR to ensure the availability of downstream resources, such as beds in PACU and Intensive Care Units (ICU).

The case duration accuracy is one of the most common approaches since to allocate the staff and maximize the use of OR accurately, it is important to predict the time required for each surgery with the smallest possible error. Bartek et al. [1] used a linear regression and two ML models to predict OR case-time duration, with the XGBoost [22] attaining the best performance. Besides these, service-specific and surgeon-specific models were considered, where each specialty and doctor were modeled individually. Tuwatananurak et al. [5] compared the duration of the predicted cases from the conventional method based on averaged historical means for case duration with cases duration predicted by the Leap Rail engine, a proprietary algorithm that combines different supervised learning algorithms. Rozario and Rozario [6] resorted to the Operations Research Tools from Google Artificial Intelligence (AI), an open-source software suite for optimization, and developed an algorithm to optimize efficiency in OR in the era of COVID-19 with the objective of minimizing overtime and undertime cases in an OR that has shown to be beneficial to reduce the long waiting lists generated during this period.

Regarding machine learning-based solutions proposed to accurately predict surgical durations, Martinez et al. [23] compared Linear Regression, Support Vector Machines, Regression Trees, and Bagged Trees. In general, the methods considered are beneficial for operating room scheduling, but Bagged Trees was the one that achieved the best overall performance to predict the surgical time duration. Furthermore, Hosseini et al. [24] developed a classical Least-squares Linear Regression and a Stepwise regression, showing both improvements compared to traditional methods. Lastly, Edelman et al. [25] performed linear regression models with data from six academic hospitals. Even with few variables, all are highly significant predictors and models presented a low error.

Researchers frequently use the approaches described above, however, other metrics can also be used with the goal of optimizing the operating room management. Lee, Ding, and Guzzo [3] performed an OR's efficiency review and mentioned methods such as identifying surgeries with high risk of cancellation and optimizing the turnover time between surgeries as frequent metrics used to evaluate and improve efficiency. Furthermore, Bellini et al. [2] presented a systematic review about the AI implementation in ORs where the majority of the studies use supervised learning techniques such as random forest and decision trees algorithms. Decision trees are powerful, intuitive data structures and easily interpretable, which allows them to be widely explored in the context of medicine, where it is essential to explain the predictions of the model, something difficult in ML because most predictive models are complex and challenging to interpret.

Moreover, several researchers address the features used as inputs in their optimization models. Bartek et al. [1] took greater account of procedures and personal data to the detriment of the patient's health status and described the primary surgeon as the most important feature to create variability. Fairley, Scheinker, and Brandeau [20] used a set of 10 features chosen based on discussions with health professionals, such as surgical service, patient information and the hospital unit the patient will go to after PACU recovery, where the most important feature was the procedure type with 0.41 of weight within the total of features. Tuwatananurak et al. [5] took into consideration more than 1,500 features, factors related to patients, providers, facility/room, procedures and prior events. Lastly, Rozario and Rozario [6] addressed that the machine learning algorithm held features such as frequency and distribution of procedure types, average case times and case times variability, highlighting the importance of the development of surgeon-specific models due to the variability that this feature can generate.

III. Methodology

In this section, we discuss the methodology of the work conducted. After the data collection, we perform an analytical exploration to summarize the data's main characteristics, a crucial initial step in data science. Three different algorithms are compared through Python: RuleFit [26], XGBoost [22] and a Feedforward Neural Network, an opaque algorithm.

A. Exploratory Data Analysis

CUF, the largest private operator of health care in Portugal, provided the data under study. Four anonymized datasets corresponding to the years 2017, 2018, 2019 and 2020 were made available, even as a dataset with the description of all types of hospital procedures described in medical association, the official Portuguese Order of Physicians table. Historical datasets provide the surgeries that have been performed at CUF in the past four years, so each row represents an episode. For each surgery, relevant data related to unit, patient, doctor and surgery performed were made available. Concerning patient information, it is provided age, gender and encrypted CUF ID (common to all units). About the surgery, data such as the surgical specialty, type of anesthesia, procedure types, the predicted and real used time inside OR and the recovery room time are given.

The dataset includes a total of 191,046 surgeries and 31 features containing surgeries performed in 15 hospitals and clinics. We used the official Portuguese Order of Physicians table to remove any mistakenly introduced procedure not listed on the table but presented on procedures columns in the historical dataset.

The dataset contains a similar proportion of male and female patients and 80% of them have only one surgery at CUF in the last four years. Regarding the type of anesthesia, it is noticeable that specific categories of anesthesia are associated with longer times of OR usage and the number of procedures is essential to estimate the final surgical time, since with the increase in the number of procedures, the average time within the OR is increasing.

Concerning specialties, CUF's dataset covers 26 specialties, of which 25 are valid surgical specialties for further analysis. Administration request was excluded as it was incorrectly recorded as a specialty and therefore should not be considered. Orthopedics, general surgery and ophthalmology are the specialties with more surgeries covering almost 50% of the total number of surgeries in ORs. However, although they are at the top in terms of the number of surgeries, obstetrics and gynecology is the one that contains more surgeons.

Lastly, it is critical to understand how the data is distributed over time and look for patterns, such as trends and seasonality, so we performed a time series analysis. August and December are the months in which we have a considerable reduction in the demand for surgeries, a fact consistent over the years and potentially related to summer, hospital staff vacation, and the end of the year. In addition, the reduction in demand for surgeries during the early phase of the COVID-19 pandemic is visible from April 2020, however, there is also a greater demand after the summer of 2020, probably related to the reduction of fear and demand for scheduling surgeries previously postponed, and therefore, in the annual total, there is no significant reduction in surgeries in 2020.

B. Data Engineering and Feature Selection

Before the feature selection and with the insights gained after investigating and exploring the data, it was clear that it would be interesting to develop some specific features. Feature engineering is related to the good utilization of domain knowledge in order to ably transform raw data into new additional features that improve the performance of machine learning models. Thus, some essential features, such as age, month, weekday and part of the day, had been generated through this process, directly from date and surgery time columns.

Moreover, due to the potential difficulty using procedures columns (I1, I2, I3, I4, I5, I6) since most of the columns have a considerable number of missing values, a column with the total number of procedures is created. Thus, if a surgery has I1, I2 and I3 values not null but I4, I5 and I6 with *NaN* values, 3 will be the value present in the additional column.

Additionally, the doctors' daily capacity and the total number of surgeries performed by the doctor in CUF may have an impact on its performance and, therefore, in surgery duration. Thus, a column was created to reflect the surgery order on a specific day and for a given doctor, and another column to reflect the doctor's experience level.

After the feature engineering process, the dataset contains a total of 41 features, however it includes a lot of information, some of it redundant. Thus, the columns that were considered relevant to generate the machine learning models were selected. This includes: specialty, CUF unit, anesthesia category, the total number of procedures performed, first procedure, surgeon, patient's gender and age, number of surgeries that a given doctor has performed so far, number of surgeries that the doctor has performed on that day, temporal data such as month, weekday and part of the day, the actual duration of surgery, planned time by CUF, type of hospitalization (outpatient or inpatient surgery) and planned or urgent surgery. The column with the time planned by CUF is not an input to the models but is kept to compare the current methods used. From these 17 columns, we removed all surgeries that contained surgeries with missing information. Thus, the final dataset contains a total of 169,772 surgeries.

Lastly, it is important to understand if the data needs to be transformed to be compatible with a specific model type. Therefore, features with two values were converted into a binary representation and categorical features with more than two values were encoded with target encoding [27]. With target encoding some considerations must be taken into account because we want to minimize target variable leakage in the new encoded feature. To prevent this problem, target encoding utilizes training data to fit the encoder and transform the new categorical data in both training and test sets.

C. Prediction Algorithms

The study used three machine learning algorithms to develop prediction models that help healthcare professionals decide the time required for surgery based on historical data. As a starting point for the development of models, we studied one interpretable and explanatory model, RuleFit [26], as well as two opaque models, XGBoost [22] and a Feedforward Neural Network, in order to comprehend the behavior of CUF data with black-boxes and the trade-off between predictive power and interpretability. Below, we present a brief discussion of these algorithms:

1) XGBoost: XGBoost is a gradient boosting algorithm that uses decision trees, that combines simple decision rules, as its "weak" prediction to predict a target variable accurately. XGBoost minimizes the objective function with Lasso (L1) and Ridge (L2) regularization to prevent overfitting and penalizing model complexity. Thus, during training, the algorithm will iteratively generate decision trees to predict the residual errors of previous trees, and then combine the result with the generated trees in order to get the final prediction.

2) RuleFit: RuleFit is an algorithm that combines tree ensembles and linear models to take advantage of tree ensemble's accuracy and linear models interpretability. This algorithm allows us to generate rules from a decision tree that create a set of "new" features from interactions between the original features. To circumvent the increase in dimensionality, Lasso, the L1 regularization technique, is called to assign weights to each decision rule since the current implementation of RuleFit can produce redundant features. By assigning a coefficient of 1 or 0 to the rules, Lasso will shrink the less important feature's coefficient and transform the input feature space into a smaller subset and easier to explain.

3) Feedforward Neural Network: The Feedforward Neural Network (FNN) is a set of structured neurons in a series of layers, with each neuron in a layer containing weights to all neurons in the previous layer. The name "Feedforward" is derived from the assumption that inputs and outputs are independent of each other and the corresponding decision that there are no feedback connections in which outputs of the model are feedback into itself [28]. The model is associated with a directed acyclic graph and represented by a combination of many layers of perceptrons. The first layer is the input layer and the rightmost is the output layer. Between them, there are a set of hidden layers with hidden units associated with often nonlinear activation function to preserve many of the properties that make linear models generalize well. In FNN the piecewise linear function, Rectified Linear Unit (ReLU), is the recommended activation function. The ability of the ReLU function set to zero values lower than zero, ensuring that the function is linear for values greater than zero, brings many advantages to the backpropagation process and the use of gradient-based methods.

D. Models Development

After analyzing similar works in the OR topic and considering the indication of CUF's stakeholders, the strategy chosen to move forward was the study of three types of approaches: a general, specialty-specific and surgeon-specific models. The last two specific models were based on the work developed by Bartek et al. [1], in which the authors generated specific models where surgeons are modeled individually and specific machine learning models for each specialty. The two specific approaches were developed for each surgical specialty or each surgeon with more than 100 surgeries in the training dataset to achieve a reasonable performance value in the test dataset. Therefore, we developed a total of 18 specialty-specific models corresponding to each surgical specialty and 381 surgeon-specific models for surgeons. The models of each approach will contain a different structure, its specific encoding and different observations in the target column.

The machine learning models were developed on the 80% training data and validated on 10% of the data. Another 10% of the data is preserved as a test dataset to compute the generalization error.

Regarding the choice of model parameters during model development, in FNN and RuleFit, it was not possible to introduce a custom evaluation function, thus, our choice fell on the Mean Squared Error (MSE) metric. Nonetheless, the XGBoost algorithm allows to define a metric of our choice, so, taking into account the approach taken by Bartek et al. [1], the scoring strategy combines both the Mean Absolute Percentage Error (MAPE) but also the percentage of within cases, where surgeries considered within present an error of less than 10%, concerning the difference between the actual duration of the surgery and the one predicted by XGBoost.

E. Data Enrichment and Model Selection

Imbalanced data is a common issue in learning problems mainly in classification problems where the ratios of each class are unbalanced and may lead the model to ignore minority classes. However, this problem is inherent in the real world as it is rare to have uniform distributions across several categories and we always end up observing skewed distributions in data labels.

Through the EDA, it was possible to identify that the data was imbalanced and the histogram of the actual values, as shown in Figure 1, is left-skewed. Consequently, models will probably have difficulty delivering optimal results for some surgeries presented in less predominant regions, thus could show considerable difficulty in predicting surgeries with a longer operative time. Therefore, we enriched three different datasets, one of each of the previous approaches, using the GMM to improve the prediction of surgeries by learning the distribution of the features in the scarce label regions and sample from it, to rebalance the dataset. This strategy aims to ensure that the models can predict surgeries in general without bias and attempt to predict surgeries outside the main range with a smaller error.

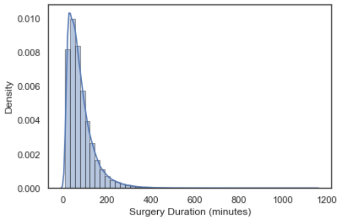


Fig. 1. Histogram and density plot of target.

First, for each model, a set was generated containing most of the data, around 75%, including the mode. The remaining surgeries were divided into two minority classes. Thus, we will be able to consider having three types of surgeries belonging to the class 1, 2 or 3, and treat the problem as a classification problem. In this way, taking the sets of the two minority classes, we will generate new synthetic samples using GMM.

As the basis for evaluating the model complexity and choosing the number of components, different types of covariance were analyzed, including spherical, diagonal, full and tied, using the Akaike (AIC) and the Bayesian Information Criterion (BIC). The GMM was applied to each minority class until the sum of the data points generated with the initial class data be equal to the total number of points in class 1.

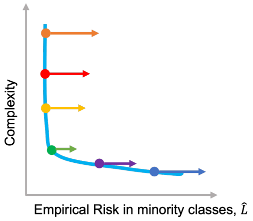


Fig. 2. The proposed curve illustrating the generalization ability of the elbow on lower left corner and the empirical risk effect of increasing model complexity.

Moreover, a new version of the Rashomon Curve [29], named Interpretability Curve and illustrated in Figure 2, will be designed to find the optimal model. The Interpretability Curve will be designed to better respond to the problem at hand. The curve will evaluate the relationship between complexity as a function of the empirical loss of the train set in the minority classes to assess the error in the class that the model has more difficulty in predicting. Unlike the Rashomon Curve, the arrows will represent the difference in Root Mean Squared Error (RMSE) in the minority classes between the test dataset and the training dataset to identify the transition between the model's underfit and overfit. This approach will be applied both to the model generated from imbalanced and balanced data and in both algorithms, XGBoost and RuleFit.

Regarding the y-axis, the complexity will be represented by the number of rules generated by RuleFit or the depth of the XGBoost trees. Thus, to generate the respective Interpretability Curve for each algorithm, a set of trees with different depths and different number of rules are initialized to generate the final curve. The Interpretability Curve will be used for the model selection and to choose appropriate complexity. As described by Semenova et al. [29] the elbow of the curve seems to be a reliable model selection criterion, an important selection from the interpretability point of view.

F. Performance Evaluation Method

The evaluation of a machine learning algorithm is a crucial step during the machine learning process. After getting the predictions, it is important to understand how close they are to the expected value and therefore, different metrics can be used. However, we will have to keep in mind that different metrics will lead to different results depending on our goal and data distribution, and that our model can get outstanding results on the training set, but behave poorly with the test set.

Firstly, during the development of approaches to develop predictive models the results obtained from each algorithm were compared taking into account the actual case-time duration. The metric used to compare the performance of models was the percentage of within cases, where these cases are those in which the forecast has a maximum error of 10%, which is the threshold chosen, and therefore the higher this percentage, the better the model's performance. Overutilization and underutilization are cases that have been estimated with an error greater than 10% in module, with a time shorter than the real one and with a time exceeding the actual case-time duration respectively. For the surgeon and specialty approaches, the performances of the different models belonging to each were concatenated, taking into account the percentage of each model to make a more direct comparison throughout strategies.

In the second phase of the work, imbalanced and balanced models were developed with the same metrics explained, however UBA imbalanced learning metrics [30] will be applied to compare the results of both datasets. The performance evaluation sometimes may require the use of special metrics as the most popular metrics are based on averages and are not prepared for unbalanced domains [31]. To address regression problems where extreme values are also important to predict accurately and where we can focus on key application cases, Torgo and Ribeiro [32] developed a regression algorithm in the non-uniform costs domain, which allows user to specify domain preferences and it also includes utility-based performance metrics, precision and recall metrics, often used in classification, but to be applied in regression tasks. The idea is to assign different importance to each surgery prediction provided by the model. For example, if it is more important for CUF to predict more accurately shorter procedure times than longer surgeries, or if it is preferable to underestimate the time rather than overestimate, as it does not affect the following surgeries, probably it makes no sense to use metrics that give the same rate of importance to each forecast and thus we will base our metrics on the application's target. The package provides various pre-processing functions to deal with classification and regression problems, and involves evaluating the utility (cost/benefit) of predictions.

IV. Results & discussion

A. Algorithmic Analysis

Firstly, we studied the three model approaches by applying the above-described algorithms to understand the behavior of the data. In the implementation phase, the parameter tuning was performed depending on the algorithm. For XGBoost, parameters such as maximum depth, eta and minimum child weight were optimized. In RuleFit, it was tuned parameters like the type of decision tree and its depth. And lastly, in FNN, the number of layers, learning rate and dropout. We evaluated performance considering the percentage of within surgeries. We also compared the model's predictions with the CUF estimates. The results achieved with each algorithm and approach are represented in the following Tables I,II and III.

TABLE I
Validation error obtained for each approach with XGBoost and from CUF model.

Model	Within	Overutilization	Underutilization
CUF	0.20	0.31	0.49
General Model	0.26	0.39	0.35
Specialty-specific models	0.27	0.41	0.32
Surgeon-specific models	0.33	0.41	0.26

TABLE II
Validation error obtained for each approach with RuleFit and from CUF model.

Model	Within	Overutilization	Underutilization
CUF	0.20	0.31	0.49
General Model	0.22	0.30	0.48
Specialty-specific models	0.24	0.29	0.47
Surgeon-specific models	0.26	0.30	0.44

TABLE III
Validation error obtained for each approach with FNN and from CUF model.

Model	Within	Overutilization	Underutilization
CUF	0.20	0.31	0.49
General Model	0.24	0.31	0.45
Specialty-specific models	0.24	0.34	0.42
Surgeon-specific models	0.24	0.33	0.43

According to previous Tables, all algorithms show better results about the current CUF estimates, and the surgeon-specific models still have an improvement in cases within the threshold in respect to the specialty model. Additionally, XGBoost presents a higher improvement than the rest of the algorithms, an expected outcome since this scalable implementation of gradient boosting behaves well under imbalanced data.

In both XGBoost and RuleFit, CUF predictions are less accurate than each approach and it is possible to verify that the less generic the dataset is, thus moving from the general model to the surgeon's models, the percentage of correct predictions increases.

Through the FNN method, we cannot verify a better forecast when we use specific models compared to the general model. These results may be related to the difficulty of the neural network to adapt to imbalanced datasets [33] since it works based on the calculation of errors and assumes equal costs. Therefore, neural networks end up adapting more to a particular class, in the case of classification, or to a range of more frequent labels in the regression case.

Later, it was also analyzed the importance of the features for the final output through insightful model interpretation such as Shapley values [34]. Interestingly, for the general and specialty approaches, the most important features are the first procedure, the doctor, the number of procedures performed during the surgery, and the type of hospitalization. Therefore, this result emphasizes the importance of a specific model for each doctor due to the relevance given to the doctor column.

B. Data Balancing Approach

In addition to the algorithmic analysis and the use of more robust methods like XGBoost for imbalanced datasets, we also adopt a data balancing approach. Our data is imbalanced and has difficulty in delivering optimal results for surgeries associated with longer times, as observed in Figure 3. Therefore, we enriched three different datasets using generative modeling techniques to improve the prediction of surgeries.

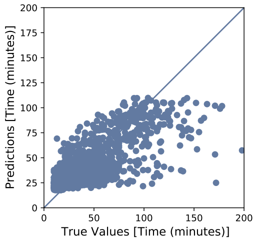
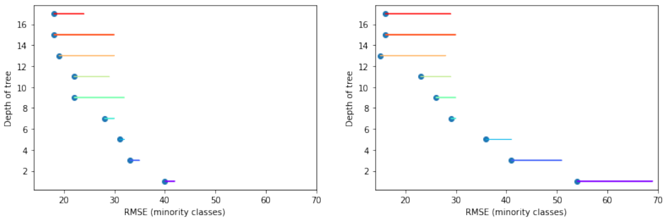


Fig. 3. Graph of true labels versus predicted labels for the ophthalmology specialty with RuleFit algorithm.

Then, the choice of the optimal model will take into account the behavior of the Interpretability Curve, including the moment when the complexity does not justify the minor reduction of RMSE. A curve is generated for imbalanced and balanced data and in both algorithms, XGBoost and RuleFit. Figure 4 shows the Interpretability Curve generated for ophthalmology using the XGBoost algorithm. As expected, we can observe a trend of increasing RMSE error in the training set as we have a lower tree profundity since we will have fewer splits and a higher tendency to underfit. Therefore, based on Figure 4, the selected XGBoost model for imbalance data has a depth of 7 and for the ophthalmology model with balanced data generated from GMM, the tree depth will be 9. In the case of balanced data, Figure 4 b) presents a minimal error when the depth is 12. However, not only 12 is a very high depth and hard to interpret, as the difference between the errors of the minority classes from training and test datasets is almost 50%.



(a) Proposed Interpretability Curve for ophthalmology model with imbalanced data.

(b) Proposed Interpretability Curve for ophthalmology model with balanced data.

Fig. 4. XGBoost model selection for ophthalmology.

In summary, Table IV presents the selected ideal model based on the Interpretability Curve for each approach and algorithm studied. It is possible to notice that the value of the balanced data is consistently higher than the value associated with the imbalanced data, coherent results considering that it contains a more considerable amount of data in the training set.

TABLE IV
MODEL SELECTION FOR EACH MODEL APPROACH AND ALGORITHM TYPE FOR IMBALANCED AND BALANCED DATA BASED ON INTERPRETABILITY CURVE. THE VALUE REPRESENTS THE TREE DEPTH FOR XGBOOST AND THE CHOSEN NUMBER OF RULES FOR RULEFIT MODELS.

Algorithm	Data	General Approach	Ophthalmology Speciality	Surgeon ID 96440008
XGBoost	Imbalanced	9	7	3
	Balanced	11	9	3
RuleFit	Imbalanced	67	42	28
	Balanced	79	47	20

Furthermore, we applied the UBA tool, an evaluation methodology that gives more importance to points that are more difficult to predict, and, at the same time, it allows providing different costs to regions that represent underutilization or overutilization. In our case problem, the penalization costs factor p was set to 0.90 since opportunity costs are considered more severe than false alarms. In other words, overutilization is more costly than underutilization of operating rooms. Thus, when the estimated time for a given surgery is longer than the real one, these false alarms are less punished and lower the cost. On the other hand, if the predicted time is less than the real one since this can cause congestion in the flow of surgeries, the cost associated with this region is higher. The accuracy of models with and without imbalanced data was measured with mean utility, recall, precision and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Tables V, VI and VII represent the results obtained after applying the UBA library in imbalanced and balanced datasets for the general model, ophthalmology specialty and surgeon ID 96440008, respectively. The results of the two interpretable models, XGBoost and RuleFit, can be compared to understand which algorithms better compress predictions and better estimate surgeries belonging to the classes that we consider part of the minority class.

In general, through the Tables presented, we can confirm that balanced models are an improved version of imbalanced models, particularly on observations with rare extreme values, and thus get better scores. Firstly, recall is one of the most important metrics because it evaluates the y points considered with a high variance and are well predicted, thus it estimates how good the model is at verifying that a certain value belongs to minority classes. Hence, a higher recall value means that our model is predicting better points considered highly relevant. Analyzing the imbalanced models with the ones from balance data, this metric has a consistent improvement. According to the algorithm, RuleFit can better estimate the set of relevant points than XGBoost, showing a maximum value in the RuleFit algorithm with balanced data in all model approaches.

TABLE V
RESULTS OF UTILITY METRICS FOR GENERAL MODEL WITH IMBALANCED DATA AND WITH BALANCED DATA FROM GMM.

Algorithm	Data	Mean Utility	Precision	Recall	AUC-ROC
XGBoost	Imbalanced	0.092	0.804	0.672	0.771
	Balanced	0.092	0.789	0.677	0.779
RuleFit	Imbalanced	0.099	0.784	0.686	0.778
	Balanced	0.113	0.788	0.704	0.791

TABLE VI
RESULTS OF UTILITY METRICS FOR OPHTHALMOLOGY SPECIALTY MODELS WITH IMBALANCED DATA AND WITH BALANCED DATA FROM GMM.

Algorithm	Data	Mean Utility	Precision	Recall	AUC-ROC
XGBoost	Imbalanced	0.071	0.759	0.661	0.765
	Balanced	0.071	0.751	0.676	0.772
RuleFit	Imbalanced	0.079	0.758	0.673	0.776
	Balanced	0.089	0.732	0.726	0.816

TABLE VII
RESULTS OF UTILITY METRICS FOR SURGEON '96440008' MODELS WITH IMBALANCED DATA AND WITH BALANCED DATA FROM GMM.

Algorithm	Data	Mean Utility	Precision	Recall	AUC-ROC
XGBoost	Imbalanced	0.010	0.486	0.505	0.686
	Balanced	0.071	0.711	0.674	0.794
RuleFit	Imbalanced	0.092	0.725	0.704	0.800
	Balanced	0.092	0.697	0.722	0.806

Regarding precision, we do not constantly have a higher value for this metric in balanced models. The value presented represents the proportion of points estimated as highly relevant by the model correctly predicted. In this context, most balanced models predict this parameter slightly lower than unbalanced models. Furthermore, the mean utility, the metric that we want to maximize, is expected to have a small value because most of the points in the data test belong to class 1. These points have a small utility score or even zero, so they do not influence and add a considerable value to the metric. Models developed with imbalanced data are expected to have a small mean utility because it is a method that predicts classes 2 and 3 very poorly, so presents the fewest points in the highest-scoring area. For most approaches, the mean utility value for balanced models is higher or equal to the imbalanced value. Lastly, AUC-ROC, which tells how much the model is capable of distinguishing between classes, was consistently inferior for the imbalanced data in the totality of the models, meaning a better balanced model performance in identifying minority classes.

Ultimately, Table VIII presents the results of RMSE in minority classes for each balanced model approach and respective algorithm, establishing their comparison with CUF predictions. The RuleFit algorithm outperforms XGBoost on all approaches when analyzing the error in minority classes. In line with what was found through the UBA library, RuleFit reveals fewer difficulties in learning from imbalanced data.

TABLE VIII
SUMMARY OF THE RMSE IN MINORITY CLASSES FOR EACH MODEL APPROACH AND ITS COMPARISON WITH CUF PREDICTIONS.

Approach	RMSE in minority classes		
	CUF	XGBoost	RuleFit
General Model	62.09	60.89	57.39
Ophthalmology Model	31.59	29.54	26.62
Surgeon ID 96440008 Model	27.78	18.37	16.71

C. Generalization error and results from an operation's perspective

With the study of different approaches, algorithms, and balanced techniques, we are in conditions to present the proposed model that allow better planning of CUF's operating rooms and evaluate generalization set predictions. We intend to develop an interpretable machine learning algorithm that can help CUF health professionals in estimating the time associated with each surgery and thus reduce the uncertainty and high errors correlated with the surgical times. For that reason, RuleFit will be our choice. This algorithm allows the creation of a set of easily interpretable rules with different importance, being easy from an explanatory point of view its application in the hospital environment. Furthermore, it presented very interesting results with balanced data, even presenting better recall values than XGBoost.

Earlier, with the analysis of balanced data, we concluded that GMM technique improves the results of the models and makes the forecasting method less susceptible to overestimation, a parameter that we intended to reduce. As we explained before, we used GMM to produce synthetic samples however this implementation was not possible for all models of each approach, but for only one model of each due to time constraints. The duration of the generation of the synthetic samples, the identification of the three classes of each model, the design of the Interpretability Curve for the choice of the best model and the guarantee of the same percentage of each class within different sets are limitations that lead us to present the final model in a theoretical concept.

Hence, the final model presented to CUF is a RuleFit algorithm that uses the balanced models of three different approaches depending on its data. If the initial features contain a doctor whose model already exists, that is, a doctor with more than 125 surgeries registered in CUF, the specific surgeon model will be used because it presents the smallest error among the approaches. On the other hand, if a doctor has no trained model, we will move to the specialty models. Following the same reasoning, we will use this model if there is already a specialty model for the input specialty. Ultimately, if none of the above conditions are possible, the general model will be used for forecasting the time needed. Algorithm 1 represents the entire process described but further studies are needed before incorporating machine learning-based decision support systems into clinical practice.

Algorithm 1 Operating room decision support system.

```
$F \leftarrow Features$
$S \leftarrow Specialty$
$N \leftarrow SurgeonNumber$
if $N$ has more than 100 surgeries in training set then
    $SurgeonModel(F, N)$
else if $S$ has more than 100 surgeries in training set then
    $SpecialtyModel(F, S)$
else
    $GeneralModel(F)$
end if
```

Finally, generalization errors are critical to understanding the performance of machine learning models, however as it was not possible to develop the final algorithm with all models and approaches with balanced data, it will not be possible to find this value. For this reason, we will use the unseen set to estimate the error for the three balanced models conducted to exemplify how we would have done it if it had been possible to develop all balanced models. The results are presented in Table IX and are consistent with the results presented in the development of the models.

TABLE IX
SUMMARY OF THE GENERALIZATION ERROR MEASURED BY RMSE IN MINORITY CLASSES FOR EACH MODEL APPROACH AND ITS COMPARISON WITH CUF PREDICTIONS.

Approach	RMSE in minority class	
	CUF	RuleFit
General Model	61.67	57.17
Ophthalmology Model	35.54	28.28
Surgeon ID 96440008 Model	29.34	18.21

Lastly, from the perspective of the final consumer, the hospital, we developed a cost function to explain more practically the benefits that the results of our proposal can bring. We consider it essential to contemplate the relative cost of overutilization and underutilization activities, which are changeable costs that will depend on the hospital's perspective.

Thus, to calculate the proposed solution's costs and establish a comparison with the cost previously supported by CUF, we generate a set of equations that take into account several factors.

The total cost for underutilization presented in Equation (1) considers the percentage of surgeries that falls 10% below the real time, the number of surgeries with undertime (#Under) divided by the total number of surgeries (#Surg), and the average loss of time in minutes. The assigned cost will be considered a cost per minute (C_{under}).

$$Underutilization_{cost} = C_{under} \times \frac{\#Under}{\#Surg} \times Avg(min)_{under} \tag{1}$$

Similarly, Equation (2) represents the total cost for overutilization, where now we consider the number of surgeries with overtime (#Over) and the average in minutes of this overuse.

$$Overutilization_{cost} = C_{over} \times \frac{\#Over}{\#Surg} \times Avg(min)_{over} \tag{2}$$

Finally, the total cost is given by

$$C_{total} = Underutilization_{cost} + Overutilization_{cost} \tag{3}$$

Both very long and very short time planning can lead to undesirable consequences for the organization of operating rooms. From the domain knowledge, we know that C_{over} is higher than C_{under} because the idle operating room produces underutilization costs, and indirectly, we are not maximizing the use of the room with a surgery that could be scheduled. In contrast, the overuse costs represent increases in the additional overtime payments and schedule reorganization costs [35]. Thus, we will assume the ratio between underutilization and overutilization costs as $C_{over} = r \cdot C_{under}$.

The operational cost is given by the difference between our proposed model and CUF baseline, where we desire to obtain a cost reduction as presented in the following Inequality (4). The chosen cost values will be based on the opinion of the hospital's stakeholders and will be kept as unknown variables as they may have slight variations depending on the purpose of its use. We will isolate these variables as much as possible to estimate their relationship by finding a minimum r value.

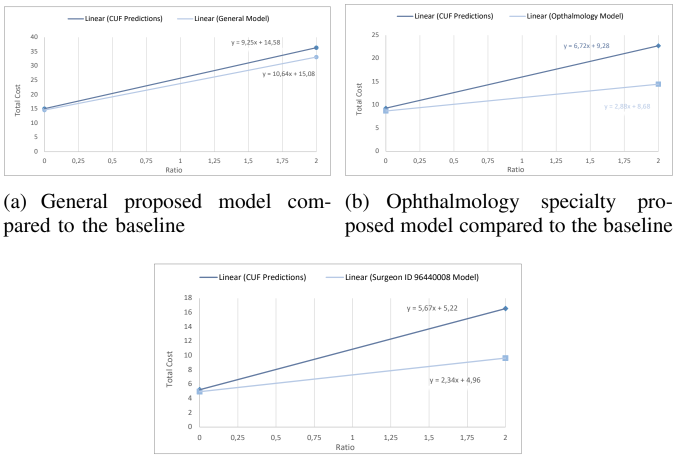
$$C_{model} < C_{CUF} \tag{4}$$

The inequality is applied to both CUF and model predictions in order to understand if, from an operational point of view, our model outperforms the current model. Consequently, the values obtained for r are presented in Table X. The proposed solutions are cheaper than current standards when $r > r_{min}$. The acquired r_{min} values are considered small since for any hypothetical overutilization and underutilization costs, our results overcome the current estimates, presenting a cost reduction compared to the CUF baseline. Moreover, in line with what we found previously, specific models have a lower cost when compared to the general model.

Figure 5 shows the cost comparison of the baseline and proposed solutions, where the blue line is associated with the CUF baseline, and the light blue corresponds to proposed solutions. We designed the graphs for a relationship between C_{over} and C_{under} at most twice, so r varies from 0 to 2. The objective is to get the light blue line below the current estimates line, with the largest possible gap for the proposed model to remain more cost-effective. As we can notice in the proposed solutions, we have a cost reduction compared to the baseline. Moreover, in line with what we found previously, specific models have a lower cost when compared to the general model.

TABLE X
RATIO BETWEEN PREVENTIVE COSTS FOR EACH MODEL IN RELATION TO CUF'S BASELINE COST.

	r_{min}
General Model	-0.36
Ophthalmology Model	-0.16
Surgeon ID 96440008 Model	-0.078



(a) General proposed model compared to the baseline

(b) Ophthalmology specialty proposed model compared to the baseline

(c) Surgeon ID 96440008 proposed model compared to the baseline

Fig. 5. Cost comparison of baseline and proposed solutions. Total cost in function of ratio. The proposed solutions are cheaper than the baseline in the three approaches. The objective is to be as much as possible below the current estimate line.

To conclude, it is essential to point out that in a deep analysis, possible indirect costs should also be analyzed and other metrics. The measurement tool developed was based on the type of results we obtain throughout the thesis.

V. CONCLUSIONS

We used explanatory algorithms to develop models that predict the surgical time required at the operating room through machine learning techniques associated with a regression problem. Our models can more accurately predict the time required to perform an operating room surgery than the CUF's current standards.

We developed three different approaches that merged into the same algorithm can be used depending on the context and available variables. The extensive work shows that specific models can bring advantages, especially models developed individually for each surgeon, having been this approach to obtain the highest percentage of within surgeries and at the same time the lowest RMSE error. Besides, we demonstrate that the use of techniques that generate large synthetic data from small data may help to improve the overall accuracy compared to the measures achieved using the original dataset.

The research work is an evidence that the proper implementation of technologies that use machine learning can significantly improve current standards of estimation, and maintaining staff and patients confidence on the system.

ACKNOWLEDGMENT

This document was written and made publically available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work described herein was performed at CUF (Lisbon, Portugal), during the period March-October 2021, under the supervision of Eng. João Leal and Eng. Daniela Burrinha. The thesis was co-supervised at Instituto Superior Técnico by Prof. Cláudia Soares and also by Prof. Qiwei Han, Professor of Data Science and Business Analytics at Nova SBE.

REFERENCES

[1] Matthew A Bartek et al. “Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration”. In: vol. 229. Elsevier Inc., Dec. 2019, 346–354.e3. DOI: 10.1016/j.jamcollsurg.2019.05.029.

[2] Valentina Bellini et al. “Artificial Intelligence: A New Tool in Operating Room Management. Role of Machine Learning Models in Operating Room Optimization”. In: *Journal of Medical Systems* 44 (Jan. 2020). DOI: 10.1007/s10916-019-1512-1.

[3] Daniel J. Lee, James Ding, and Thomas J. Guzzo. “Improving Operating Room Efficiency”. In: *Current Urology Reports* 20 (June 2019). DOI: 10.1007/s11934-019-0895-3.

[4] Daniel M. Laskin, A. Omar Abubaker, and Robert A. Strauss. “Accuracy of predicting the duration of a surgical operation”. In: *Journal of Oral and Maxillofacial Surgery* 71.2 (Feb. 2013), pp. 446–447.

[5] Justin P Tuwatananurak et al. “Machine Learning Can Improve Estimation of Surgical Case Duration: A Pilot Study”. In: (Jan. 2019). DOI: 10.1007/s10916-019-1160-5.

[6] Natasha Rozario and Duncan Rozario. “Can machine learning optimize the efficiency of the operating room in the era of COVID-19?” In: *Canadian Journal of Surgery* 63 (2020), E537–E529. DOI: 10.1503/CJS.016520.

[7] Ruhi Sarikaya, Geoffrey E. Hinton, and Anoop Deoras. “Application of Deep Belief Networks for Natural Language Understanding”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 778–784. DOI: 10.1109/TASLP.2014.2303296.

[8] Brian L. DeCost et al. “Computer Vision and Machine Learning for Autonomous Characterization of AM Powder Feedstocks”. In: *JOM 2016 69:3* 69.3 (Dec. 2016), pp. 456–465. ISSN: 1543-1851. DOI: 10.1007/S11837-016-2226-1.

[9] Leo Galway, Darryl Charles, and Michaela Black. “Machine learning in digital games: a survey”. In: *Artif Intell Rev* 29 (2008), pp. 123–161. DOI: 110.1007/s10462-009-9112-y.

[10] Brecht Cardoen et al. “Operating room planning and scheduling: A literature review”. In: (Mar. 2010). DOI: 10.1016/j.ejor.2009.04.011.

[11] Kevin Taaffe, Bryan Pearce, and Gilbert Ritchie. “Using kernel density estimation to model surgical procedure duration”. In: *International Transactions in Operational Research* 28.1 (Jan. 2021), pp. 401–418. DOI: 10.1111/ITOR.12561.

[12] Pieter S. Stepaniak et al. “Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: A multicenter study”. In: *Anesthesia and Analgesia* 109.4 (2009), pp. 1232–1245. DOI: 10.1213/ANE.0B013E3181B5DE07.

[13] David P. Strum, Jerrold H. May, and Luis G. Vargas. “Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models”. In: *Anesthesiology* 92.4 (2000). https://pubmed.ncbi.nlm.nih.gov/10754637/, pp. 1160–1167. ISSN: 0003-3022. DOI: 10.1097/00000542-200004000-00035.

[14] David P Strum et al. “Estimating times of surgeries with two component procedures: comparison of the lognormal and normal models”. In: *Anesthesiology* 98.1 (Jan. 2003). https://pubmed.ncbi.nlm.nih.gov/12503002/, pp. 232–240. ISSN: 0003-3022. DOI: 10.1097/00000542-200301000-00035.

[15] Elia El-Darzi et al. “Length of Stay-Based Clustering Methods for Patient Grouping”. In: *Studies in Computational Intelligence* 189 (2009). https://link.springer.com/chapter/10.1007/978-3-642-00179-6_3, pp. 39–56. DOI: 10.1007/978-3-642-00179-6_3.

[16] Franklin Dexter and Johannes Ledolter. “Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data”. In: *Anesthesiology* 103.6 (2005). https://pubmed.ncbi.nlm.nih.gov/16306741/, pp. 1259–1167. ISSN: 0003-3022. DOI: 10.1097/00000542-200512000-00023.

[17] Mark Van Houdenhoven et al. “Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling”. In: *Anesthesia and analgesia* 105.3 (Sept. 2007). https://pubmed.ncbi.nlm.nih.gov/17717228/, pp. 707–714. ISSN: 1526-7598. DOI: 10.1213/01.ANE.0000277492.90805.0F.

[18] Franklin Dexter and Rodney D. Traub. “How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time”. In: *Anesthesia and analgesia* 94.4 (2002). https://pubmed.ncbi.nlm.nih.gov/11916800/, pp. 933–942. ISSN: 0003-2999. DOI: 10.1097/00000539-200204000-00030.

[19] Choong Ho Lee and Hyung-Jin Yoon. “Medical big data: promise and challenges”. In: *Kidney Research and Clinical Practice* 36.1 (Mar. 2017). http://www.krcp-ksn.org/journal/view.php?id=10.23876/j.krcp.2017.36.1.3, pp. 3–11. ISSN: 2211-9132. DOI: 10.23876/J.KRCP.2017.36.1.3.

[20] Michael Fairley, David Scheinker, and Margaret L. Brandeau. “Improving the efficiency of the operating room environment with an optimization and machine learning model”. In: *Health Care Management Science* 22 (Dec. 2019), pp. 756–767. DOI: 10.1007/s10729-018-9457-3.

[21] Amin Abedini, Wei Li, and Honghan Ye. “An Optimization Model for Operating Room Scheduling to Reduce Blocking Across the Perioperative Process”. In: *Procedia Manufacturing* 10 (2017), pp. 60–70. DOI: 10.1016/j.promfg.2017.07.022.

[22] Tianqi Chen and Carlos Guestrin. “XGBoost : A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016). DOI: 10.1145/2939672.2939785.

[23] Oscar Martinez et al. “Machine learning for surgical time prediction”. In: *Computer Methods and Programs in Biomedicine* 208 (Sept. 2021), p. 106220. ISSN: 0169-2607. DOI: 10.1016/J.CMPB.2021.106220.

[24] N. Hosseini et al. “Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study”. In: *AMIA Annual Symposium Proceedings* 2015 (2015). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765628/, p. 640.

[25] Eric R. Edelman et al. “Improving the Prediction of Total Surgical Procedure Time Using Linear Regression Modeling”. In: *Frontiers in Medicine* 4.Jun (2017), p. 85. DOI: 10.3389/FMED.2017.00085.

[26] Jerome H Friedman and Bogdan E Popescu. “Predictive learning via rule ensembles”. In: *The Annals of Applied Statistics* 2.3 (2008), pp. 916–954. DOI: 10.1214/07-AOAS148.

[27] Daniele Micci-Barreca. “A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems”. In: *ACM SIGKDD Explorations Newsletter* 3.1 (2001), pp. 27–32.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[29] Lesia Semenova, Cynthia Rudin, and Ronald Parr. *A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning*. arXiv: 1908.01755. 2021. arXiv: 1908.01755 [cs.LG].

[30] Rita Paula Almeida Ribeiro. “Utility-based Regression”. PhD thesis. Faculdade de Ciências da Universidade do Porto, 2011.

[31] Nuno Moniz et al. *Evaluation of Ensemble Methods in Imbalanced Regression Tasks*. Tech. rep. http://www.kdd.org/kdd-cup [Accessed on 2021/06/15]. 2017, pp. 129–140.

[32] Luis Torgo and Rita Ribeiro. “Precision and recall for regression”. In: *International Conference on Discovery Science*. Springer. 2009, pp. 332–346. DOI: 10.1007/978-3-642-04747-3_26.

[33] Chong Zhang et al. "A Cost-Sensitive Deep Belief Network for Imbalanced Classification". In: (). arXiv: 1804.10801v2. arXiv: 1804.10801v2.

[34] Scott M Lundberg, Paul G Allen, and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. Tech. rep. https://github.com/slundberg/shap [Accessed on 2021/05/01].

[35] Andreas Fügener, Sebastian Schiffels, and Rainer Kolisch. "Overutilization and underutilization of operating rooms - insights from behavioral health care operations management". In: *Health Care Management Science* 20.1 (Mar. 2017), pp. 115–128. DOI: 10.1007/S10729-015-9343-1.