International Journal of Rough Sets and Data Analysis
Volume 6 • Issue 3 • July-September 2019

The Effects of Sampling Methods on Machine Learning Models for Predicting Long-term Length of Stay:
A Case Study of Rhode Island Hospitals

Son Nguyen, Bryant University, Smithfield, USA
Alicia T. Lamere, Bryant University, Smithfield, USA
Alan Olinsky, Bryant University, Smithfield, USA
John Quinn, Bryant University, Smithfield, USA

ABSTRACT

The ability to predict the patients with long-term length of stay (LOS) can aid a hospital's admission management, maintain effective resource utilization and provide a high quality of inpatient care. Hospital discharge data from the Rhode Island Department of Health from the time period between 2010 to 2013 reveals that inpatients with long-term stays, i.e. two weeks or more, costs about six times more than those with short stays while only accounting for 4.7% of the inpatients. With the imbalance in the distribution of long-stay patients and short-stay patients, predicting long-term LOS patients becomes an imbalanced classification problem. Sampling methods—balancing the data before fitting it to a traditional classification model—offer a simple approach to the problem. In this work, the authors propose a new resampling method called RUBIES which provides superior predictive ability when compared to other commonly used sampling techniques.

KEYWORDS

Bagging, Boosting, Classification, Imbalanced Data, Length of Stay, Rare Event, Resampling, Rhode Island Department of Health, Tree-Based Models, Under-Sampling

INTRODUCTION

Predicting a patient's length of stay (LOS) in a hospital setting has been widely researched (Panchami & Radhika, 2008; Walczak, Pofahl & Scorpio, 1998; Liu et al., 2010, Azari, Janeja, & Mohseni, 2012; Morton et al., 2014; Pendharkar & Khurana, 2014; Gentimis et al., 2017; Turgeman, May & Sciulli, 2017; Rojas et al., 2018; Yakovlev et al., 2018). Accurately predicting an individual's LOS can have a huge impact on a healthcare provider's ability to care for that individual by allowing them to properly prepare and manage resources. A hospital's productivity requires a delicate balance of maintaining enough staffing and resources without being overly equipped or wasteful. Key to maintaining this balance is the ability to accurately anticipate a patient's care requirements, a core aspect of which is each individual's LOS (Gustafson, 1968).

Of particular interest, though more difficult to predict, are long-term LOS. Several studies have shown that long-term LOS are associated with poor patient satisfaction (Farley et al., 2009, Kainzinger

DOI: 10.4018/IJRSDA.2019070103

et al., 2009; Yankovic & Green, 2011; Eggers et al., 2013). The billing structure within hospitals also results in little economic gain from long-term LOS—the majority of overhead and indirect costs are incurred within the first few days of stay. Consequently, long-term LOS often results in losses for hospitals in the form of so-called "opportunity costs" when new patients are unable to be admitted due to capacity constraints (Taheri, Butz & Greenfield, 2000). Hence, accurately predicting long-term LOS would allow hospitals to better anticipate bed availability and the associated costs.

In this study, the authors focus on the prediction of these long-term LOS at the time of admission in Rhode Island hospitals through discharge data obtained from the Rhode Island Department of Health over the time period of 2010 to 2013. Long-term LOS generally makes up a small percentage of patients, and as a result can be viewed as rare events. As such, their accurate prediction requires the use of resampling methods when training a model to account for imbalanced data. Working with data containing imbalanced events is an important area of research, with applications and challenges that extend beyond LOS datasets (Krawczyk, 2016; He & Garcia, 2009; Fernández, García & Herrera, 2011; Sun, Wong & Kamel, 2009; Visa & Ralescu, 2005). This work proposes a new resampling method called Randomly Under-sampled Bag-boosting for Imbalanced-Event Samples (RUBIES) that combines under-sampling, bagging and boosting. This study compares this algorithm's performance to four widely used methods: over-sampling, under-sampling, Random Over-Sampling Examples (ROSE), and Synthetic Minority Over-Sampling Technique (SMOTE). Specifically, the comparison will be made when they are combined with three commonly used classification models: random forests, decision trees, and Adaboost.

BACKGROUND

Critical to the prediction of long-term LOS is the handling of the imbalanced data concern. For this reason, the majority of this section is devoted to a discussion of resampling techniques. The tree-based classification models that were used to compare the authors' proposed resampling method's performance are also discussed.

Resampling Techniques

Approaches in dealing with imbalanced data can be grouped into two categories: algorithmic-level approaches and data-level approaches. In algorithmic-level approaches, the imbalanced data is kept unchanged and classical classification models that work with balanced data are modified to work with imbalanced data. In data-level approaches, the aim is to balance the data before applying conventional classification models.

Many algorithmic-level approaches have been proposed. In the case of support vector machines (SVM), Veropoulos et al. (1999) modified traditional SVM by implementing two cost parameters within the objective functions. Imam et al. (2006) modified SVM further by reducing the bias toward the majority class by adjusting the decision boundary. Wu et al. (2003) proposed another modification to SVM by extending the boundary region of the classes through the use of conformally transformed kernels. Other modifications and extensions of SVM can be found in the works of Batuwita et al. (2013), Cortes and Vapnik (1995), and Boser et al. (1992). Researchers have explored the modification of other classification methods as well. Maalouf et al. (2011) suggested the use of logistic regression through the implementation of the truncated Newton method in prior correction logistic regression with the addition of a regularization term. The use of deep neural networks has been explored by Wang et al. (2017), who proposed a new loss function for use while training the network. The performance of convolution neural networks has also been explored for imbalanced data in the work of Buda et al. (2017).

In general, data-level approaches can be grouped into two classes: under-sampling (in which the majority class is under-sampled, reducing its presence in the training dataset) and over-sampling (in which the minority class is over-sampled, increasing its presence). The simplest under-sampling

method is random under-sampling (RUS), in which the majority observations are randomly selected from the majority class to balance with the minority class. While maybe suffering from the disadvantage of removing important observations in the majority, the clear advantage of RUS is its fast running time as it results in a much smaller dataset. A modification of under-sampling removes noisy data points in the majority class through the use of Tomek Links (Elhassan et al., 2016), but impacts running time.

As for over-sampling techniques, observations must be added to the minority class to balance with the majority class. These added observations could be either the same minority observations, through the random over-sampling (ROS) technique, or artificially created observations. One of the most influential techniques to synthetically create observations is Synthetic Minority Over-Sampling Technique (SMOTE) developed by Chawla et al. (2002). This method uses linear interpolation to create new observations for the minority class. In SMOTE, new observations are created from a point, say A, by selecting random points lying on the lines connecting A with its neighbors. Han et al. (2005) proposed the modification Border-SMOTE, where SMOTE is applied only in the border region between two classes. Alternatively, the Random Over-Sampling Examples (ROSE) technique proposed by Menardi et al. (2012) uses the kernel density to generate new minority observations. There also exist several cluster-based techniques such as Cluster-Based Oversampling (CBO) by Jo and Japkowicz (2004) or the DBSMOTE algorithm which uses DBSCAN clustering proposed by Bunkhumpornpat et al. (2012).

Despite all of these advancements—due to their simplicity in idea and computation—RUS and ROS remain two of the most widely-used resampling methods. Consequently, the authors chose to focus on these methods, as well as the popular methods SMOTE and ROSE, for comparison.

Classification Models

The three classifiers chosen for this study to apply resampling methods for comparison are decision tree, random forest and Adaboost, which are all tree-based models. Tree-based models were chosen due to their versatility in working with categorical variables without encoding them and their proven effectiveness in working with imbalanced data (Chen, 2004; Cieslak, 2008). The Rhode Island hospital discharge data contains several variables with a large number of categories (for example, admitting diagnosis has 6641 categories). As a result, using distanced-based models such as support vector machines or neural networks would be too computationally expensive. These three tree-based models consisting of decision trees, random forests, and Adaboost are briefly discussed below in more detail.

Decision Tree

Decision tree structures a flow chart where at each node of the flow chart the incoming data is partitioned into multiple subsets based on a splitting rule. There are several criteria that can be used to decide both the splitting variable and the value at which the variable splits, including criteria about information (entropy) gain, Gini Index and logworth values.

The final nodes—where the splitting stops—are called the leaves and are used for prediction. The tree is usually fully grown to obtain the most complicated tree and may be pruned down to a simpler tree to avoid overfitting issues. Proposed by Breiman (1984), Decision Tree or CART (Classification and Regression Tree) can be used with both categorical targets (classification tree) and continuous targets (regression tree). Note that for this study, a classification tree was used, as the goal is to predict long-term versus short-term LOS.

Random Forest

Random Forest is an ensemble classification model that combines the results of multiple other classifiers (Breiman, 1984). In Random Forest, a set of decision trees is built on a subset of the original data. However, unlike in decision trees, at each split of a tree only a subset of randomly selected variables is considered for the split. Random Forest then makes a prediction by taking the

majority vote among the entire set of trees. For instance, if two out of three trees of a random forest predict long-term LOS, then the random forest will also predict a long-term LOS. The two main hyper-parameters (unknown quantities that are determined by the users) of random forest are the number of trees and the number of variables considered for splitting at each node, which impact the runtime and performance of the model.

Adaboost

Adaboost, or Adaptive Boosting, is a boosting method which is designed to improve the performance of a poor-performing classifier, or "weak" classifier (Freund, 1999). Adaboost contains a sequence of classifiers where each classifier in the sequence is trained on "weighted" observations. The weights are initially set to be equal among observations and then updated based on the classifier's prediction error for each observation, with larger errors leading to larger weights. The idea of boosting is that each iteration of the classifier in the boosting sequence focuses more on learning observations that were "difficult" to learn in the previous classifier. The final prediction of Adaboost is a weighted vote from all of the classifiers in the sequence with higher-performed classifiers having greater weights.

PROPOSED RESAMPLING METHOD

Description of the Algorithm

This work proposes a new method for dealing with imbalanced data. The method uses the ideas of under-sampling, bagging and boosting, and hence is called Randomly Under-sampled Bag-boosting for Imbalanced-Event Samples (RUBIES). Bagging is the approach where a model (for example, a decision tree) is run repeatedly with bootstrapped samples and then the results combined, which in the case of a tree, would be by voting (choosing the most frequent outcome for a given input). This new proposed algorithm first uses bagging and under-sampling techniques to train the sample model (base-model) on multiple under-sampled datasets. It then ensembles this set of base-models into a new model by majority voting. Finally, the algorithm uses this newly obtained model to update the weights of the majority observations so that observations with higher weights, i.e. higher predicting error, will be more likely to appear in the next round of under-sampling, which is the concept behind the boosting technique.

The RUBIES algorithm is as follows:

Step 0: Decide the base-model, B. We under-sample the data and train several classifiers on this balanced data. B is the classifier that performs the best. In our implementation, out of the decision tree, Adaboost and random forest. The random forest performs the best, so it is selected to be the base-model.

Step 1: Set $i = 1$.

Step 2: Train untrained instances of B in k under-sampled data and ensemble these k classifiers by majority voting to obtain B_i

Step 3: Perform B_i on the dataset of the majority class only to obtain predicted probabilities for all majority observations, normalize these predicted probability (divide by their summations) to obtain the weights W_i for the majority class. Also obtain voting power a_i, which is the out-of-bag (OOB) error of B_i.

Step 4: Let $i = i + 1$ and repeat Step 2 to Step 4, noticing that the k under-sampled data are under-sampled with the weights W_i.

Step 5: Stop when i equals a predetermined value N.

Step 6: Use $B_1, B_2 \ldots B_N$ with their corresponding voting power $a_1, a_2 \ldots a_N$ to vote for the final prediction.

Note that the choice of k and N will impact the performance of RUBIES. Larger values will cause an increase in computation time, while small values may not capture enough of the majority data. The authors recommend a value for k proportional to the portion of the dataset corresponding to the minority class. Generally, a value between 20-25 is recommended for N.

A detailed diagram demonstrating the steps of the RUBIES algorithm on a simple example with $k = 4$ and $N = 2$ can be found in Figure 1.

Data and Variable Selection

The dataset used in this paper was Rhode Island Hospital Discharge Data obtained from the Center for Health Data & Analysis and Public Health Informatics at the Rhode Island Department of Health (2014) and consisted of hospital discharge data for the years 2010 to 2013 from 14 Rhode Island hospitals. The dataset contained 539,395 observations of 134 variables. Missing entries in the data were handled by imputing either the mean of the corresponding variable (for numeric variables) or the mode of the corresponding variable (for categorical variables). Looking at the dataset, the authors chose to define long-term LOS as a LOS of 14 days or more. This definition for long-term LOS consisted of 4.7% of the observations in the dataset.

As this study desires to develop an accurate predictor for long-term LOS immediately upon admittance to the hospital, these 134 variables were reduced to only those observed at the time of admittance for the patients. This resulted in 20 variables for consideration. An initial exploration of these variables displayed differences in LOS trends. The distribution of age is shown in Figure 2, split by long/short term LOS. Clearly, there is an increase in long-term LOS above an age of approximately 50, as well as a significant drop between the ages of approximately 20 to 50. Figure 3 explores the impact of patient sex. While short term shows a larger number of female patients, interestingly long-term LOS is balanced between both sexes. To more formerly rank variable importance, the mean of each variable's importance in thirteen random forest models trained on under-sampled datasets was computed. The largest importance score amongst the 20 variables was then used to normalize all scores, resulting in the most important variable having a score of 1 while a variable with no importance having a score of 0. The results for all 20 variables are shown in Figure 4.

Looking at Figure 4, it can be seen that the most important variable is diag_adm, which is the diagnosis information of the patient at the time of admittance. Conversely, b_wt, which represented birth weight has a score of 0 and hence was removed from the analysis. All other variables were retained, resulting in a final set of 19 variables used for prediction. Detailed descriptions of the 19 variables retained for this study can be found in the Appendix.

Computational Results

In this section, the results of computing three predictive models (decision tree, random forest and Adaboost) on six datasets (the original imbalanced data, RUS data, ROS data, balanced data using SMOTE technique, balanced data using ROSE technique, and finally balanced data using our proposed RUBIES algorithm) are reported and discussed. First, the metrics used to evaluate the performance of each model-resampling method combination are explained.

Model Evaluation Metrics

Using misclassification rate or overall accuracy to measure the classification quality of models can be misleading when working with imbalanced data. A trivial model that predicts that all observations are in majority class would give a very high overall accuracy, despite misclassifying all observations in the minority class. It has been observed that the true positive rate, or sensitivity, of a classical model applied to an imbalanced dataset is usually very low and models that improve the sensitivity, given by

Figure 1. Diagram demonstrating RUBIES algorithm for a simple example where $k = 4$ and $N = 2$, where $m_1, \ldots, m_n$ represent observations from the majority class, $p_{i1}, \ldots, p_{in}$ represent the predicted probabilities for these observations based on model i.

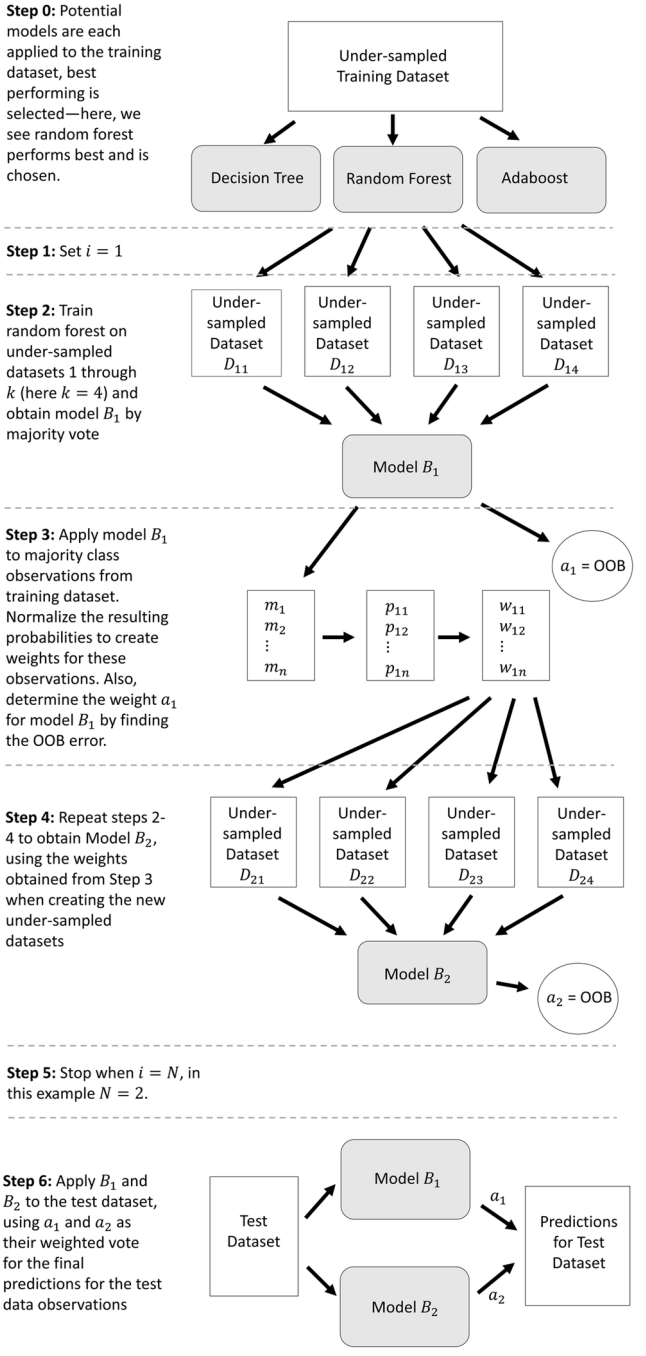


Figure 2. Age distribution of the long-term LOS patients versus short-term LOS patients

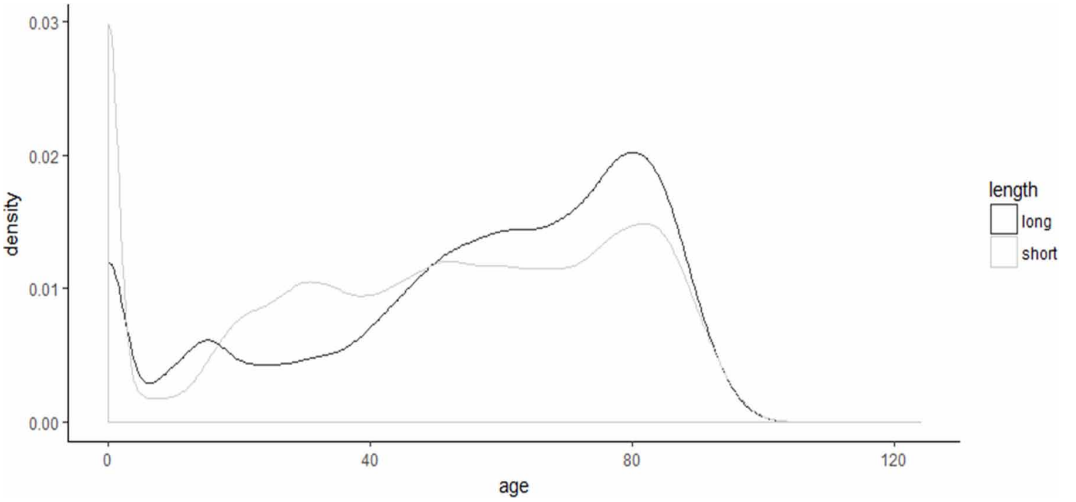
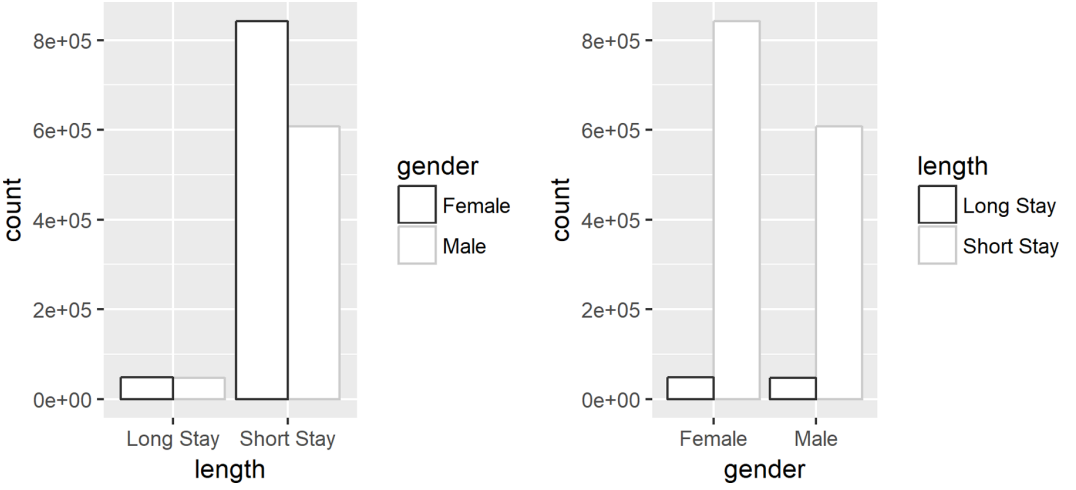


Figure 3. Distribution of Male and Female toward long-term LOS and short-term LOS patients

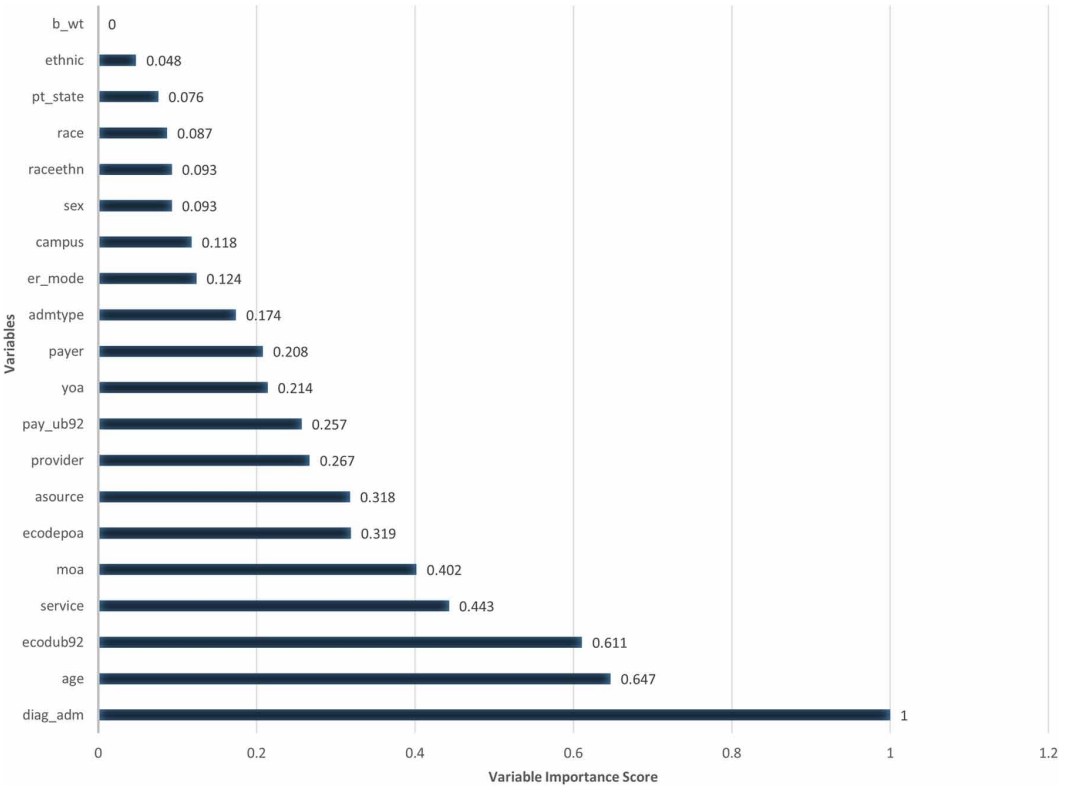


$$\left(\frac{TP}{TP+FN}\right),$$

often reduce the true negative rate, or

$$\left(\frac{TN}{TN+FP}\right).$$

Figure 4. Variable Importance by Random Forests trained on multiple under-sampled data for the 20 variables observed at the time of admittance for patients in the Rhode Island hospital data



This study seeks models that can improve the sensitivity while not significantly reducing the specificity. Thus, this study uses the average of these two quantities as the main evaluation metric of these models, called balanced accuracy, which is given by

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right).$$

By using this measure, the accuracy is now symmetric with respect to both classes (Brodersen et al., 2010). Table 1 shows the Confusion Matrix.

Results

In this section, the model performances at predicting long-term LOS in the Rhode Island hospital discharge data using the 19 variables available at the time of admittance are explored. The performances of three predictive models (decision tree, random forest and Adaboost) applied to five datasets including the original (imbalanced data) and data computed by four resampling methods (random under-sampling, random-over sampling, SMOTE and ROSE) in comparison to the proposed method, RUBIES, are discussed.

In all four comparison resampling techniques, the data were balanced so that each class accounts for approximately 50%. In SMOTE, 5 was chosen as the number of k-nearest neighbors. For the decision tree, a binary tree was used (a node can only have two branches) with splitting rules decided

Table 1. Confusion Matrix: TN, FP, FN, TP stands for True Negative, False Positive, False Negative, and True Positive, respectively

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

by the Gini Index. This was implemented with the R package "rpart" (Recursive PARTition) for the computation (Therneau, 2012). For the random forest, the "ranger" R package was used (Wright, 2015) and the number of trees was set to 30 and the number of variables at each node set to 5. Finally, Adaboost was applied to boost a decision tree using the R package "fastAdaboost" (Chatterjee, 2016). Due to the slow-running time of boosting, the number of classifiers in the sequence was set to 7.

When applying RUBIES to resample the hospital data, a value of $k = 13$ was chosen due to 4.7% of the data containing the long-term LOS observations. Also, a value of $N = 21$ was used for the number of iterations.

The computation process can be described as follows. First, the original data is partitioned into two datasets, training dataset (75%) and testing dataset (25%), ensuring that each maintains the original ratio of 4.7% long-term LOS observations. Then, the resampling method is applied to the training dataset. Finally, the model is built using the resampled training dataset and is evaluated based on its prediction of the testing dataset. The results of each resampling method were grouped by the model they were applied to for decision trees, random forest, and Adaboost, and are shown in Figure 5, Figure 6, and Figure 7, respectively.

The first observation is that all three models running on the original imbalanced dataset suffered from the imbalance issue as they struggled to detect the minority long-term LOS observations. Of the three models, Adaboost gave the highest true positive rate of only 1.67%, while the decision tree was unable to detect any positive observations and predicted all observations as not long-term LOS. This is a common issue of classical predictive models performing on imbalanced data.

This study also observes that RUBIES provides the best balanced accuracy, with its performance of 76.7%. While all comparison resampling methods did improve each model's ability to detect long-term LOS observations, the effectiveness depends both on the predictive models and the resampling method used. Overall, SMOTE appeared to be the least effective resampling technique while RUS, although simple, brought the most improvement in balanced accuracy of the comparison methods, achieving a balanced accuracy of 75.2% when combined with random forest. In contrast, the decision tree combined with SMOTE gave the lowest balanced accuracy of 55.3%.

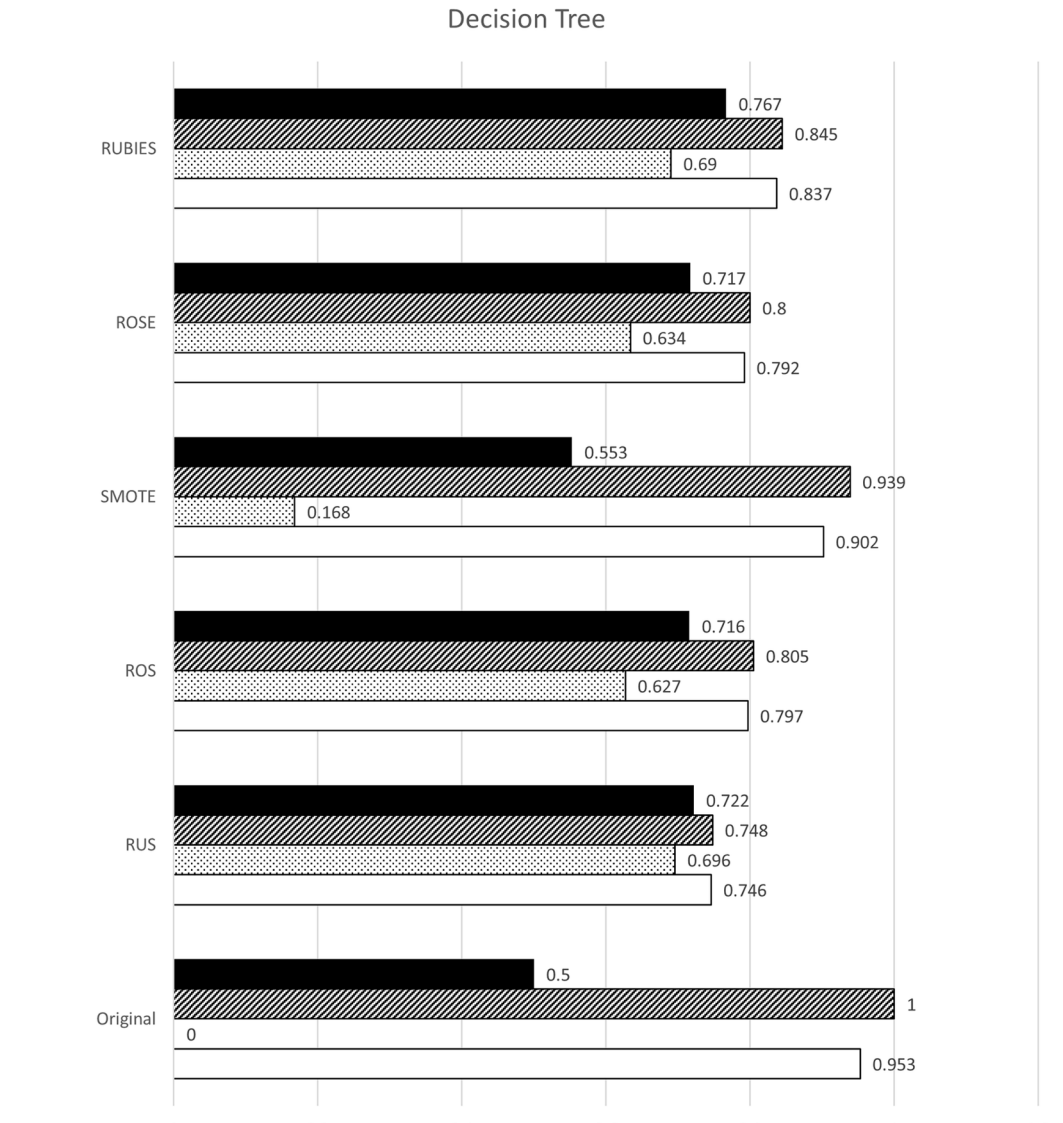
CONCLUSION

Predicting long-term LOS at the time of admission is an important issue for hospitals, and when performed accurately, can aid them in providing better patient care while improving the management of resources.

This work studied the performances of multiple classification models at predicting long-term LOS from an imbalanced dataset obtained from Rhode Island hospitals, focusing in particular on the effectiveness of using resampling techniques to improve the balance accuracy of predicting both positive (long-term LOS) and negative (short-term LOS) observations. The computational results demonstrated that all of the resampling techniques studied helped improve the overall performances of the classification models and that the random forest model combined with the RUS technique was the most effective and out-of-the-box combination, giving 75.2% balanced accuracy.

To further improve the performance of the classification models, the authors propose the use of the algorithm RUBIES which implements the idea of under-sampling, bagging and boosting with the

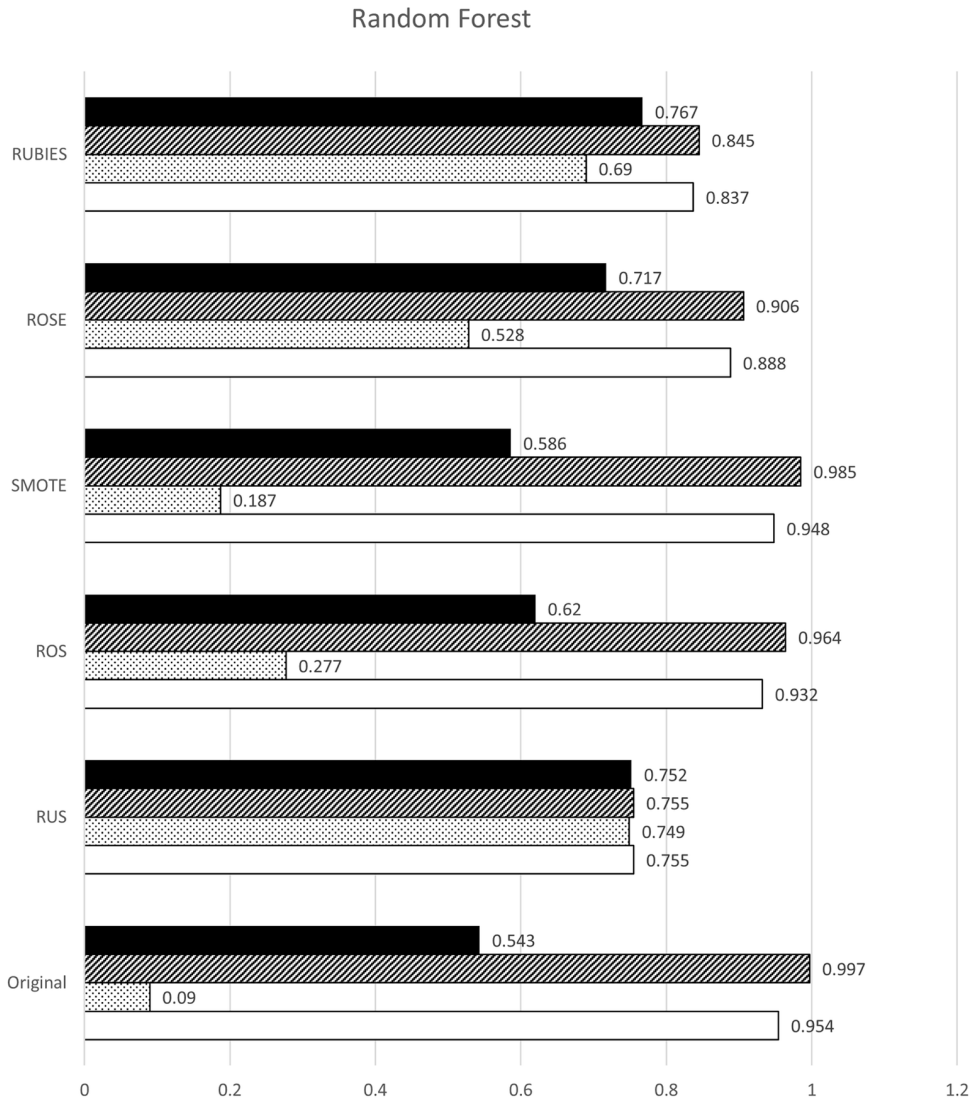
Figure 5. Performances of RUBIES and Decision Tree model trained on the original and resampled data in predicting long-term LOS for the Rhode Island hospital discharge dataset



	Original	RUS	ROS	SMOTE	ROSE	RUBIES
Balanced Accuracy	0.5	0.722	0.716	0.553	0.717	0.767
Specificity	1	0.748	0.805	0.939	0.8	0.845
Sensitivity	0	0.696	0.627	0.168	0.634	0.69
Accuracy	0.953	0.746	0.797	0.902	0.792	0.837

integration of the predictive power of the random forest. The proposed method improved the balanced accuracy to 76.7%. Although RUBIES has a longer running time when compared to the combination of random forest and RUS, it still runs much faster than other over-sampling type techniques such as SMOTE, ROS or ROSE and should be a valuable tool for predicting imbalanced datasets.

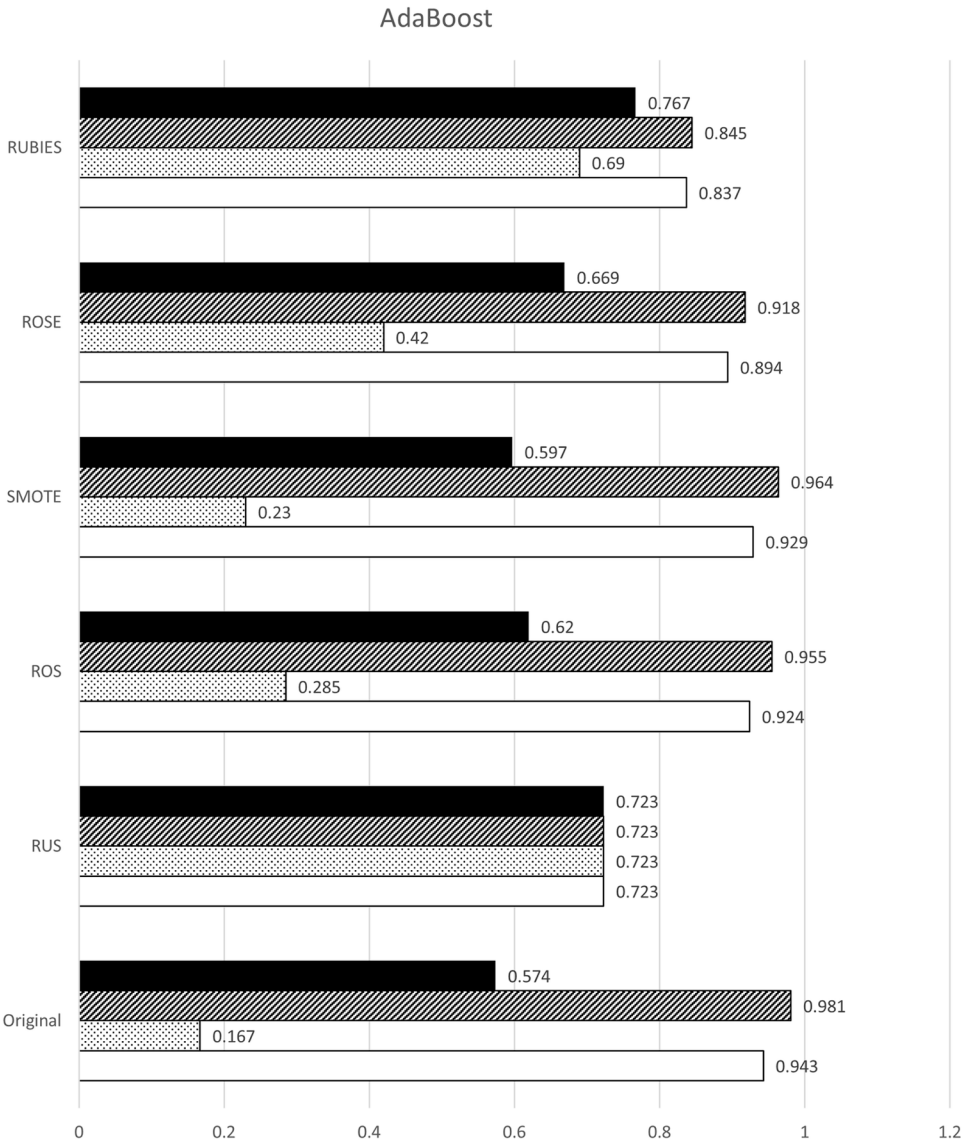
Figure 6. Performances of RUBIES and Random Forest model trained on the original and resampled data in predicting long-term LOS for the Rhode Island hospital discharge dataset



	Original	RUS	ROS	SMOTE	ROSE	RUBIES
Balanced Accuracy	0.543	0.752	0.62	0.586	0.717	0.767
Specificity	0.997	0.755	0.964	0.985	0.906	0.845
Sensitivity	0.09	0.749	0.277	0.187	0.528	0.69
Accuracy	0.954	0.755	0.932	0.948	0.888	0.837

Although our procedure produces a higher accuracy in identifying long stay patients and is ready to be implemented for any future hospital discharge data, we believe this procedure can still be improved. In our model, we manually select all and use most of the variables available at the admission. Although it simplifies the process by avoiding the variable selection stages, it may result in a more complex model in the end. It is worth looking at the applications of popular variable selection

Figure 7. Performances of RUBIES and Adaboost model trained on the original and resampled data in predicting long-term LOS for the Rhode Island hospital discharge dataset



	Original	RUS	ROS	SMOTE	ROSE	RUBIES
Balanced Accuracy	0.574	0.723	0.62	0.597	0.669	0.767
Specificity	0.981	0.723	0.955	0.964	0.918	0.845
Sensitivity	0.167	0.723	0.285	0.23	0.42	0.69
Accuracy	0.943	0.723	0.924	0.929	0.894	0.837

techniques for hospital discharge data. One of the first considerations is rough set based variable selection techniques, which have shown to be effective for improving performance of predictive models (Singh et al., 2016) and can be implemented easily for healthcare (Park, 2013). This serves as a starting idea for our next project.

REFERENCES

Azari, A., Janeja, V. P., & Mohseni, A. (2012, December). Predicting hospital length of stay (PHLOS): A multi-tiered data mining approach. In *Proceedings of the 12th International Conference on Data Mining Workshops (ICDMW)* (pp. 17-24). IEEE. doi:10.1109/ICDMW.2012.69

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152). ACM. doi:10.1145/130385.130401

Breiman, L., Friedman, J. H., Olsen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, (pp. 3121-3124). IEEE. doi:10.1109/ICPR.2010.764

Buda, M., Maki, A., & Mazurowski, M. A. (2017). A systematic study of the class imbalance problem in convolutional neural networks.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: Density-based synthetic minority over-sampling technique. *Applied Intelligence*, *36*(3), 664–684. doi:10.1007/s10489-011-0287-y

Carden, R., Eggers, J., Wrye, C., & No, P. (2013). *The relationship between inpatient length of stay and HCAHPS Scores*. HealthStream Discover Paper.

Chatterjee, Sourav (2016). *fastAdaboost: a Fast Implementation of Adaboost* (Version 1.0.0) [R package].

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California.

Cieslak, D. A., & Chawla, N. V. (2008, September). Learning decision trees for unbalanced data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 241-256). Springer. doi:10.1007/978-3-540-87479-9_34

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Elhassan, T., Aljurf, M., Al-Mohanna, F., & Shoukri, M. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Journal of Informatics and Data Mining*, *1*(2).

Fernández, A., García, S., & Herrera, F. (2011, May). Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In *Proceedings of the International Conference on Hybrid Artificial Intelligence Systems* (pp. 1-10). Springer. doi:10.1007/978-3-642-21219-2_1

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society for Artificial Intelligence, 14*(771-780), 1612.

Gentimis, T., Ala'J, A., Durante, A., Cook, K., & Steele, R. (2017, November). Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data. In *Proceedings of the Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence & Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 1194-1201). IEEE. doi:10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191

Gustafson, D. H. (1968). Length of stay: Prediction and explanation. *Health Services Research*, *3*(1), 12. PMID:5673664

Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing* (pp. 878-887). Springer. doi:10.1007/11538059_91

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. doi:10.1109/TKDE.2008.239

Imam, T., Ting, K. M., & Kamruzzaman, J. (2006, December). z-SVM: An SVM for improved classification of imbalanced data. In *Proceedings of the Australasian Joint Conference on Artificial Intelligence* (pp. 264-273). Springer. doi:10.1007/11941439_30

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, *6*(1), 40–49. doi:10.1145/1007730.1007737

Kainzinger, F., Raible, C. A., Pietrek, K., Müller-Nordhorn, J., & Willich, S. N. (2009). Optimization of hospital stay through length-of-stay-oriented case management: An empirical study. *Journal of Public Health*, *17*(6), 395–400. doi:10.1007/s10389-009-0266-5

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. doi:10.1007/s13748-016-0094-0

Liu, V., Kipnis, P., Gould, M. K., & Escobar, G. J. (2010). Length of stay predictions: Improvements through the use of automated laboratory and comorbidity variables. *Medical Care*, *48*(8), 739–744. doi:10.1097/MLR.0b013e3181e359f3 PMID:20613662

Lucas, R., Farley, H., Twanmoh, J., Urumov, A., Olsen, N., Evans, B., & Kabiri, H. (2009). Emergency department patient flow: The influence of hospital census variables on emergency department length of stay. *Academic Emergency Medicine*, *16*(7), 597–602. doi:10.1111/j.1553-2712.2009.00397.x PMID:19438415

Maalouf, M., & Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, *55*(1), 168–183. doi:10.1016/j.csda.2010.06.014

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. doi:10.1007/s10618-012-0295-5

Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., & Kakadiaris, I. A. (2014, December). A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In *Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA)* (pp. 428-431). IEEE. doi:10.1109/ICMLA.2014.76

Panchami, V. U., & Radhika, N. (2014, February). A novel approach for predicting the length of hospital stay with DBSCAN and supervised classification algorithms. In *Proceedings of the Fifth International Conference on Applications of Digital Information and Web Technologies (ICADIWT)* (pp. 207-212). IEEE. doi:10.1109/ICADIWT.2014.6814663

Park, M. (2013) Data Mining in Uniform Hospital Discharge Data Set Using Rough Set Model. In: V. Toi, N. Toan, T. Dang Khoa, & T. Lien Phuong (eds) In *Proceedings of the 4th International Conference on Biomedical Engineering in Vietnam (IFMBE)* (pp. 332-334). Springer. doi:10.1007/978-3-642-32183-2_82

Pendharkar, P. C., & Khurana, H. (2014). Machine Learning Techniques for Predicting Hospital Length of Stay in Pennsylvania Federal and Specialty Hospitals. *International Journal of Computer Science & Applications*, *11*(3).

Rhode Island Department of Health. Center for Health Data & Analysis and Public Health Informatics (2014). *Rhode Island Hospital Discharge Data* [data file]. Available from http://health.ri.gov/data/hospitalization/discharge/

Rojas, J. C., Venable, L. R., Fahrenbach, J. P., Carey, K. A., Edelson, D. P., Howell, M. D., & Churpek, M. M. (2018). Predicting Hospital Length of Stay After Intensive Care Unit Discharge with Machine Learning. *American Journal of Respiratory and Critical Care Medicine*, A4287–A4287.

Singh, D. A. A. G., Leavline, E. J., Priyanka, E., & Sumathi, C. (2016). Feature Selection Using Rough Set For Improving the Performance of the Supervised Learner. *International Journal of Advanced Science and Technology*, *87*, 1–8. doi:10.14257/ijast.2016.87.01

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719. doi:10.1142/S0218001409007326

Taheri, P. A., Butz, D. A., & Greenfield, L. J. (2000). Length of stay has minimal impact on the cost of hospital admission1. *Journal of the American College of Surgeons*, *191*(2), 123–130. doi:10.1016/S1072-7515(00)00352-5 PMID:10945354

Therneau, T., Atkinson, B. & Ripley, B. (2012). *Rpart: Recursive Partitioning* (Version 3.0) [R package].

Turgeman, L., May, J. H., & Sciulli, R. (2017). Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. *Expert Systems with Applications*, *78*, 376–385. doi:10.1016/j.eswa.2017.02.023

Veropoulos, K., Campbell, C., & Cristianini, N. (1999, July). Controlling the sensitivity of support vector machines. In *Proceedings of the International Joint Conference on AI* (p. 60). Academic Press.

Visa, S., & Ralescu, A. (2005, April). Issues in mining imbalanced data sets-a review paper. In *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference* (pp. 67-73). Academic Press.

Walczak, S., Pofahl, W. E., & Scorpio, R. J. (1998, May). Predicting Hospital Length of Stay with Neural Networks. In *Proceedings of the FLAIRS Conference* (pp. 333-337). Academic Press.

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2017). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(12), 2591–2600. doi:10.1109/TCSVT.2016.2589879

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R.

Wu, G., & Chang, E. Y. (2003, August). Class-boundary alignment for imbalanced dataset learning. In *Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, (pp. 49-56). Academic Press.

Yakovlev, A., Metsker, O., Kovalchuk, S., & Bologova, E. (2018). Prediction of in-hospital mortality and length of stay in acute coronary syndrome patients using machine-learning methods. *Journal of the American College of Cardiology*, *71*(11), A242. doi:10.1016/S0735-1097(18)30783-6

Yankovic, N., & Green, L. V. (2011). Identifying good nursing levels: A queuing approach. *Operations Research*, *59*(4), 942–955. doi:10.1287/opre.1110.0943

APPENDIX

Description of variables from discharge data obtained from the Rhode Island Department of Health that were included in the final models:

admtype: type of admission, categorized as emergency, urgent, electric, newborn, court committal, trauma, or N/A
age: reported age of patient
asource: source of admission, categorized as physician referral, clinic referral, HMO referral, trans-hospital, trans-nurse facility, trans-healthcare, emergency room, court/law enforcement, or N/A
campus: geographical location of hospital, coded for hospitals with more than one campus
diag_adm: admitting diagnosis, presented as ICD-9-CM codes
er_mode: mode of arrival to emergency room (if applicable), categorized as rescue service/ambulance, helicopter, law enforcement or social services, personal or public transport, other, or N/A
ecodepoa: external cause of injury present upon admission
ecodeub92: external cause of injury supplied by hospital, presented as ICD-9-CM codes
ethnic: ethnicity reporting Hispanic, not Hispanic, or not reported
moa: month of admission
pay_ub92: expected source of payment, categorized as Medicare Fee for Service, Medicare Managed Care, Medicaid Fee for Service, Rite Care, Out-of state Medicaid Managed Care, Blue Cross, Coordinated Health Partners Inc, United Healthcare, Commercial insurance (other than listed), Champus, Worker's Compensation, Other, Self pay, Missing, or Error.
payer: expected source of payment, categorized as Medicare, Medicaid, Worker's Compensation, Blue Cross, Commercial Insurance, Self pay, Other, Champus, United Healthcare, Coordinated Health Partners Inc, Rite Care, Neighborhood Health Plan of RI, Insurance error, Missing, or Unknown.
provider: healthcare provider, coded as Newport, St. Joseph Health Services of RI, Memorial, Miriam, Rhode Island Hospital, Roger Williams, South County, Kent County, Westerly, Rehab of RI, Landmark Medical Center, Women and Infants, Bradley, Butler
pt_state: patient's state of residency, categorized using state abbreviations, Unknown, or Not Applicable (for outside US)
race: patient's race, categorized as White, Black, Asian, American Indian, Hispanic, Other, Unknown, or N/A
raceethn: patient's race/ethnicity, categorized as White (not Hispanic), Black (not Hispanic), Asian not Hispanic), American Indian (not Hispanic), Native Hawaiian (not Hispanic), Other (not Hispanic), Hispanic, Unknown, or N/A
service: service, categorized as Pediatrics, Medicine, Cardiology, Psychiatry, Surgery, Ophthalmology, ENT, Oral Surgery, Orthopedics, Urology, Gynecology, Abortio, OB – Not Delivered, OB – Delivered, Newborn, or Rehabilitation
sex: patient's sex, categorized as male or female
yoa: year of admission

Son Nguyen earned his master's degree in applied mathematics and doctoral degree in mathematics, statistics emphasis, both at Ohio University. He is currently an assistant professor in the department of mathematics at Bryant University. His primary research interests lie in dimensionality reduction, imbalanced learning, and machine learning classification. In addition to the theoretical aspects, he is also interested in applying statistics to other areas such as finance and healthcare.

Alicia T. Lamere received her master's and doctoral degrees in Applied and Computational Mathematics and Statistics from the University of Notre Dame, where she was a Schmitt Fellow. She is now an assistant professor in the Mathematics Department at Bryant University. Her primary research interests lie in working with non-Gaussian data and network analysis, with a focus on applications for RNA-Sequencing data. She is also interested in imbalanced data problems and resampling techniques, with applications in healthcare and social science data.

Alan Olinsky is a professor of mathematics and computer information systems at Bryant University. He earned his PhD in Management Science from the University of Rhode Island and holds an MS in Mathematics Education and a BBA in Public Accounting from Hofstra University. His research interests include statistics, management science, and data mining. He has published articles on these topics in professional journals including the "Journal of American Academy of Business," "Journal of Mathematical Education in Science and Technology," "European Journal of Operational Research," "Interfaces," and "Advances in Business and Management Forecasting." He is Co-director of the Bryant University Advanced Applied Analytics Center and a member of the Northeast Decision Sciences Institute. In addition to his research interests, Dr. Olinsky is committed to pedagogical issues at Bryant University as well as on a national scale. He also has appeared several times as an expert witness in statistical matters at hearings and trials.

John Quinn is a Professor of Mathematics at Bryant University and has been teaching there since 1991. Prior to teaching, Professor Quinn was a mechanical engineer at the Naval Underwater Systems Center (now the Naval Undersea Warfare Center) in Newport, R.I. He received his Sc.B. degree from Brown University in 1978, and his M.S. and Ph.D. degrees from Harvard University in 1987 and 1991, respectively. Professor Quinn has had articles published in multiple areas. He has done previous research in mathematical programming methods and computable general equilibrium models in economics. He currently does research in data mining applications, social networks and simulation models.