Prediction of Patients’ Length of Stay at Hospital During COVID-19 Pandemic

Jianing Pei¹, Xin Lin², Qixuan Chen³,*

¹ School of Actuarial Science, University of Waterloo, Waterloo, ON N2L3G1, Canada
² School of Computer Science, University of Washington, Seattle, WA 98105, USA
³ Suzhou North America High School, Suzhou, Jiangsu 215104, China

*Corresponding author. Email: xlin7799@uw.edu

Abstract. Machine learning has been extensively used in diverse healthcare settings since the 21st century. Statistical models are proven to be powerful in detecting early disease symptoms and could potentially aid decision-making in the healthcare system. To help improve medical resource allocation during COVID-19 pandemic, we aim to develop machine learning models that predict each patient’s length of stay (LOS) in hospital. Three machine learning models, namely, K-nearest Neighbors Algorithm, Logistic Regression and Random Forest are implemented and optimized on the same healthcare dataset. The final accuracy of each model is 0.3442, 0.3524 and 0.3541 respectively, which are not very high. Our subsequent correlation analysis on the healthcare dataset shows the patients’ features used do not provide sufficient information for accurate LOS prediction. Yet, machine learning approaches could potentially yield much better results if the data quality can be improved by including additional relevant patient features and breaking LOS into more appropriate intervals. More detailed healthcare data should be obtained to make the LOS prediction useful for healthcare management.

Keywords: Machine Learning; Prediction; Length of stay; K-nearest Neighbor; Support Vector Machine; Random Forest Classifier.

1. Introduction

The world has been exposed to the COVID-19 epidemic since December 2019[1]. The coronavirus disease is found to be highly contagious, causing serious lung infection and potentially respiratory failure[2]. The sudden outbreak of COVID-19 has tremendously overloaded the hospital system and there is a huge excess in demand for hospital beds[3]. Till now, the virus is still spreading rapidly, and better management of healthcare and treatments are required to deal with the continuing crisis[4]. Many work has been done to improve the management of healthcare previously, including some ways to reduce Health Disparities for Priority Populations[5], the Quality Improvement Initiative for Nursing Facilities[6]. Although there are many cases of using data technology in the health system[7], the length of stay(LOS) in hospital is an important indicator to monitor and predict the health management
process. Thus, aiming to promote efficient medical resource allocation during COVID-19 pandemic, we propose using machine learning approach to predict patients’ LOS according to relevant patient features.

Machine learning models are proven to be helpful in improving many aspects of the healthcare system[8]. Tanuja et al. used multi-layer network(MLP), Naive Bayes, K-NN and decision tree to predict patients’ LOS at admission. Their results showed that MLP and Naives Bayes had the highest classification accuracy of around 85%, and K-NN performed poorly on the dataset with only 63.6% accuracy[9]. Several other studies performed LOS regression task, using support vector machine[10], classification and Regression Tree[11], and Poisson Regression[12].

In our project, we employed K-NN, SVM and Random Forest to classify COVID-19 patients’ LOS in hospitals. In the dataset of AV: Healthcare Analytics II[13], the stay length is separated into 11 classes. The first 10 classes correspond to 0-10 days, 11-20 days...91-100 days respectively, and the last class is more than 100 days. For each model, we compared different data preprocessing techniques, such as encoding methods and principle component analysis(PCA). After model optimization and hyperparameter tuning, the final results showed that the three models had very similar classification accuracy of 0.3442, 0.3524 and 0.3541 respectively. We then further performed correlation analysis on the dataset and found that the patients’ features used were weakly correlated with their LOS label. If the data quality can be improved by including more relevant patient features and using reasonable LOS intervals, the classification accuracy of our models is likely to be greatly boosted to enforce better medical resource management.

2. Method

2.1. KNN

KNN (K-nearest neighbor) was used to predict the length of stay in hospital of each person according to their different features. To make the dataset be well trained by KNN, some preprocessing methods were used. As the dataset contains categorical data, which cannot be recognized by KNN, one-hot encoding and simple encoding were considered to be used. After normalizing the data by min-max normalization:

$$y = \frac{x - \min}{\max - \min}$$

[14], the function of the different encoding was compared with the same set of hyperparameters. Since the accuracy of simple encoding was higher, it was chosen finally.

As the dataset, which contained more than 400,000 samples, was too large to be run out with KNN, 50,000 samples were chosen randomly and divided into the training set and test set, and the percentage of them were 90% and 10% respectively. There were three hyperparameters for this model: distance measure (Manhattan distance: $$\sum_{i=1}^{d} |x_i - y_i|$$ [15] and Euclidean Metric: $$\sqrt{(x_1 - y_1)^2 + (y_1 - y_2)^2}$$ [16], different ways of processing depending on correlation and K(number of neighbors). Different combinations of hyperparameters were implemented to obtain the best model, which had the highest accuracy.

2.2. Support Vector Machine

SVM was used to analyze the dataset. The data contains two types of data, numerical and categorical. Multicollinearity, however, occurs when two or more independent variables in the dataset are correlated with each other, so to avoid it, I drop the first column after one-hot encoding[17]. As a result, the dimensions of the train and test are 116. According to the principal component analysis, we choose 68 features to analyze.

Since the original training set included more than 100,000 samples, according to the official document, first I used linear SVM instead of kernel SVM. Secondly, I still want to use the kernel function, since the dataset is nonlinear. However, the original training set samples were too large to analyze. I changed the sample size for the training set and test set. There are 40000 samples in the new training set and 10000 samples in the new test set. I can’t do a good job of analyzing 68 features based on 50,000 data. So, I changed to use simple label encoding, and then I found 11 features to consider. There are three different kernel functions: Gaussian radial basis function, Polynomial function and
Sigmoid function[18]. And when we use different kernel functions, we need to find the optimal value of C and gamma for each function. Where parameter C controls how much you want to punish your model for each misclassified point for a given curve, and gamma defines how far the influence of a single training example reaches.

2.3. Random Forest Classifier

Random Forest Classifier is an ensemble learning method used to classify samples from different classes[19]. In the model, a number of decision trees are constructed, and the majority class predicted by these decision trees is considered to be the classification result. In each tree node, a random subset of features is selected and among which, the best predictor is chosen according to a loss function, such as Gini impurity.

To implement this model, the healthcare dataset is first divided into training set (80%), validation set (10%), and test set (10%). During data preprocessing, different encoding methods, either one-hot encoding or simple encoding, were used to encode the categorical features in each sample. Principle component analysis (PCA) was then applied to reduce the dimensionality of the feature space and get rid of irrelevant features. In PCA, only the top n components were selected to explain 85% variance in the feature space. Therefore, the four preprocessing regimes are one-hot encoding, one-hot encoding with PCA, simple encoding, simple encoding with PCA. These four preprocessing methods were evaluated based on their corresponding random forest classification accuracy, and the one with highest classification accuracy on the validation set was chosen to be the final preprocessing regime.

Four hyperparameters in the random forest classifier are then tuned using the training set and validation set. The number of estimators is the number of decision trees in the random forest, where each tree gives a classification result and votes for the most likely class. The maximum depth of each tree is another parameter that controls the model complexity and bias-variance tradeoff. Two rest two parameters are the minimum number of samples at each leaf node and the minimum samples required to split a node. For each of the parameters, a wide range of values were experimented, and the best values were selected based on model accuracy management.

3. Results and Discussion

3.1. KNN

Different distance measures with the same dataset: 50,000, K \( \in (1,20) \), 11 features.

Manhattan distance and Euclidean Metric figure 1

![Figure 1. Hyperparameter change of distance measure](image)

(a) Manhattan distance \( k=13 \) accuracy(max)=0.2762  (b) Euclidean Metric \( k=17 \) accuracy(max)=0.3056

Because the accuracy of Euclidean Metric was higher than the Manhattan distance obviously, it was assumed that the models which used Euclidean Metric were higher than those used Manhattan Distance.
With fixed simple encoding and Euclidean Metric, the features were adjusted according to their correlation figure 2

![Figure 2. Hyperparameter change of features](image)

(a) 17 features $k=40$ accuracy(max)=0.3166  
(b) 12 features $k=19$ accuracy(max)=0.31  
(c) 6 features $k=26$ accuracy(max)=0.3442  
(d) 1 feature $k=16$ accuracy(max)=0.3166  

The final model uses simple encoding, Euclidean Metric, 6 features with correlation $>0.05$, and had its highest accuracy 0.3442 at $K=26$.

3.2. *Support Vector Machine*

Firstly, I used Linear SVM. As the dataset has 11 classes, we use one-versus-one method. Finally, the accuracy is 0.35003. Then I tried to use different kernel functions. For Gaussian radial basis function (rbf):
According to the Figure 3, $C=1$ and $\gamma=\text{scale}$. Then the accuracy is 0.352. For polynomial function: According to the Figure 4, $C=0.1$ and $\gamma=\text{scale}$. Then the accuracy is 0.3516. For Sigmoid function: According to the Figure 5, $C=0.001$ and $\gamma=0.001$. Then the accuracy is 0.3509.

Figure 3. An image of C and $\gamma$ in rbf

Figure 4. An image of C and $\gamma$ in poly
Figure 5. An image of $C$ and $\gamma$ in sigmoid

Gaussian radial basis function is the best, according to Figure 6, since the accuracy is largest. And then we need to check whether it is over-fitting. According to Figure 7, the model does not overfit because as the sample size of data increases, the performance of the model doesn’t change. Overall, the best model is Gaussian radial basis function with $c=1$ and $\gamma=\text{scale}$. And the final accuracy is 0.3525.
3.3. Random Forest Classifier

The classification accuracies of random forest classifier using different preprocessing techniques are very similar. As shown in Figure 8, while the classification accuracy for each preprocessing method increased with the number of decision trees, there was less than 2% accuracy difference across encoding methods. Simple encoder with PCA was chosen to be the preprocessing regime because it yields the highest classification accuracy even when the number of estimators is low.

![Figure 8. Comparison of different preprocessing regimes](image)

During hyperparameter tuning, the model is found to be insensitive to different parameter values. Figure 9 shows that the amount of variation in classification accuracy was lower than 3% for all four hyperparameters tuned. In the final model, number of estimators = 100, max depth = 10, min sample split = 20, and minimum samples leaf = 20. The test accuracy is 0.354 and the train accuracy was 0.356. Therefore, there is no overfitting observed.

![Figure 9. Hyperparameter tuning](image)
3.4. Discussion

According to Table 1, we find that the accuracy of the three models is about 0.35. When we tuned the hyperparameters, we found that the adjustment of the hyperparameters could not improve the accuracy very well. At best we can only improve accuracy by four percent. Also, we tried different ways to encode, one-hot encoding and simple label encoding, there was no significant change in accuracy.

| Model          | KNN     | SVM     | Random Forest |
|----------------|---------|---------|---------------|
| Accuracy       | 0.3442  | 0.3525  | 0.3541        |

We think that the accuracy of this dataset can only reach about 0.35, and we cannot significantly improve the accuracy. We think this is due to some features of the dataset. According to Figure 10, we can find only one factor that has a strong correlation with stay and other factors have very little relevance to it. Also, we know this dataset is unbalanced, based on the table 2. Class 0 has a huge amount of data, while class 6 and class 9 contain very few samples. From what has been discussed above, we believe that the factors contained in this dataset are not the main factors. And the classification of stay is not reasonable, we should not divide stay into 11 classes.

![Figure 10. Correlation matrix](image-url)
Table 2. Class size

| Class | 0   | 1   | 2   | 3   | 4   |
|-------|-----|-----|-----|-----|-----|
| Size  | 160661 | 78139 | 87491 | 55159 | 11743 |
| Class | 5   | 6   | 7   | 8   | 9   | 10  |
| Size  | 35018 | 2744 | 10254 | 4838 | 2765 | 6683 |

4. Conclusion

In this project, machine learning is utilized to solve real-world medical resource allocation problem during COVID-19 pandemic. By predicting each patient’s estimated stay at hospital as a small 10-day interval, doctors and governors could make plans about medical equipment and human resources accordingly.

However, the highest classification accuracy of 35.41% given by our models is a bit too low to be useful as a diagnosis tool in practice. We propose several ways to optimize the prediction model in the future. Firstly, since the classification classes are highly imbalanced, with most of the patients spending less than 40 days in hospital, the classification labels could be reformatted into fewer classes concentrating on patients’ stays less than 40 days. Secondly, the feature space in the raw dataset could be further expanded to include more relevant patient features such as body temperature. Extra features could potentially be better predictors for patients’ length of stay, since Coronavirus infection symptoms are still understudied due to the novelty of the disease.

Acknowledgements

We thank Dr. Victor Adamchik for giving insightful guidance on our project. We also acknowledge Siyi Wu for generous help and useful discussion.

References

[1] Chi Zhang, J. L. K. T. W. Y. H. Z.-G. W., Zhao Wu. (2020) ischarge may not be the end of treatment: pay attention to pulmonary fibrosis caused by severe covid-19. DAWN. Retrieved from https://www.dawn.com/news/1585323
[2] Mohapatra, R. K., Pintilie, L., Kandi, V., Sarangi, A. K., Das, D., Sahu, R., & Perekhoda, L. (2020). The recent challenges of highly contagious covid-19, causing respiratory infections: Symptoms, diagnosis, transmission, possible vaccines, animal models, and immunotherapy. Chemical Biology & Drug Design.
[3] COVID, I., Murray, C. J., et al. (2020). Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. MedRxiv.
[4] Junaidi, I. (2020). Minister warns of increase in covid-19 cases. DAWN . Retrieved from https://doi.org/10.1002/jmv.26634
[5] Improving cultural competence to reduce health disparities for priority popu- lations. (2014). AHRQ. Retrieved from https://effectivehealthcare.ahrq.gov/products/cultural-competence/research-protocol/
[6] Jo A. Taylor, M. P. P. P. H. B. M. A.-B. J. O. M., R.N. (n.d.). The falls management program: A quality improvement initiative for nursing facil- ities. AHRQ. Retrieved from https://www.ahrq.gov/patient-safety/ settings/long-term-care/resource/injuries/fallspx.html
[7] Dela Cuesta, J. B. (2018). Research proposal form example: Health re- search. NursingAnswers. Retrieved from https://nursinganswers.net/ essays/research-proposal-form-example-health-5513.php
[8] Callahan, A., & Shah, N. H. (2017). Machine learning in healthcare. In Key advances in clinical informatics (pp. 279–291). Elsevier.
[9] Tanuja, S., Acharya, D. U., & Shailesh, K. (2011). Comparison of different data mining...
techniques to predict hospital length of stay. Journal of Pharmaceutical and Biomedical Sciences, 7(7).

[10] Houthooft, R., Ruysinck, J., van der Herten, J., Stijven, S., Couckuyt, I., Gadeyne, B., . . . others (2015). Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artificial intelligence in medicine, 63(3), 191–207.

[11] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

[12] Clark, D. E., & Ryan, L. M. (2002). Concurrent prediction of hospital mortality and length of stay from risk factors on admission. Health services research, 37(3), 631–645.

[13] Prabhavalkar, N. (n.d.). Av: Healthcare analytics ii. Retrieved from https://www.kaggle.com/nehaprabhavalkar/av-healthcare-analytics-ii

[14] Normalization. (2020). Codecademy. Retrieved from https://www.codecademy.com/articles/normalization

[15] E.SzaboPhD, F. (2015). The linear algebra survival guide. Academic Press Publishing.

[16] Weisstein, E. W. (2020). Euclidean metric. Retrieved from https://mathworld.wolfram.com/EuclideanMetric.html

[17] Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[18] Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS technical report, 14.

[19] Liaw, A., Wiener, M., et al. (2002). Classification and regression by random- forest. R news, 2(3), 18–22.