**Hospital Length of Stay Prediction with Ensemble Methods in**

**Machine Learning**

Ling Zheng, Jiacun Wang and Alex Sheriff

Monmouth University

West Long Branch, NJ 07764, USA

Xuemin Chen

Texas Southern University

Houston, TX, USA

***Abstract*** The primary mission of hospitals is to meet the demand for care by moving patients through care pathways while improving coordination of care, patient safety, and health outcomes. The patient length of stay (LOS) in a hospital is used to gauge the efficiency of a healthcare facility. The goal of this thesis project is to develop a machine learning system to predict LOS at the admission phase of patients, with initial diagnosis and test results. Both regression models and classification (including deep learning) models are implemented and tested. The NY Hospital Inpatient Discharge in 2015 Dataset is used for model training and validation. Results are compared and discussed. The imbalanced nature of the dataset is discussed and addressed as well.

***Keywords***: Length of stay, machine learning, classification, feature engineering.

# Introduction

Patients may experience extensions in hospitalizations due to delays in decision-making by providers while they wait for results, schedule diagnostic tests, conduct discharge planning, or wait for consultation because of inadequate access to consultants and specialists [1]. Patient length of stay (LOS) is an important measure of the efficiency of hospital management. Hospitals have limited resources, requiring efficient use of beds and clinician time. When all other things are equal, a shorter stay will reduce the cost per discharge and shift care from inpatient to less expensive post-acute settings. This notion is even more important for the time being while the pandemic is ongoing. Of course, it is also in patients’ best interests not to spend longer than necessary in hospitals.

Researchers and healthcare providers have adopted many ways to reduce LOS for patients. As a branch of artificial intelligence, machine learning has been on the rise in several areas for the last decade following improved computational power, availability of data and improved algorithms. It has found great success in applications such as image segmentation and classification, machine translation and recommender systems. Machine learning has shown promise in segmenting and classifying radiology images [8], diagnosing and identifying high-risk patients [3] and predicting LOS [9]. If upon a patient’s check in and initial diagnosis and test results the LOS of the patient can be predicted, the hospitals can better analyze the factors that influence length of stay the most. Predicting a patient's LOS will also benefit the patients and patient’s families as they can have an idea of how long they can expect to stay upon being admitted and so they can better plan everything for the days that are forthcoming.

The objective of this study is to develop a software system to predict LOS of patients when their initial diagnosis and test data are available. Over the past decades a lot of machine learning algorithms have been developed and found successful applications [10][14][11]. Various machine learning algorithms are considered in this study, in both categories of regression and classification, and including both traditional algorithms and recently developed deep learning algorithms. The dataset used for training and testing those machine learning algorithms are the NY Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File, which contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges.

It is our intention to try out different yet popular models, compare their results and figure out which model performs better than others for this particular problem. Some authors chose to use only linear models to solve LOS prediction problems, e.g. [5]. We believe a nonlinear model should produce more accurate results, and our experiments proved this hypothesis. Other researchers applied machine learning methods to LOS prediction for patients with specific problems. For example, [4] studied LOS prediction in an acute care medical psychiatric inpatient service, [2] focused on patients with diabetes. In [7][13] the authors used machine learning models to predict the LOS of patients with heart problems at the time of admission. Jamei et. al presented an approach to predicting all-cause risk of 30-day hospital readmission using artificial neural networks in [12].

All algorithms are trained with the NY state inpatient dataset and thus can be used directly for the LOS prediction of patients as long as their data entry matches the dataset in format. The algorithms are general though, and thus they can be trained with different datasets and used to solve different prediction problems.

The rest of the paper is organized as follows: Section 2 introduces the NY Hospital Inpatient Discharge Dataset, analyzes the features of the dataset, and pre-processes the dataset. Section 3 briefly introduces machine learning and some of the popular algorithms that are used in this project, including regression algorithms, classification algorithms and deep learning. Section 4 presents the results from various algorithms and their comparisons. Section 5 provides a summary of the study and future work.

# Dataset and Feature Engineering

## The NY Hospital Inpatient Discharges in 2015 Dataset

The publicly accessible “NY Hospital Inpatient Discharges in 2015” dataset from the New York State Government health data website was used to conduct this thesis study. This dataset contains rows of patient data, including profiles, conditions, medications, facilities, prices, and charges for over 2.3 million patients whose information has been de-identified in accordance with HIPAA regulations. The dataset has 34 distinct categories or features and they are listed on Table 1.

Table 1. List of Features of the Dataset.

| **Feature Name** | **Type** |
| --- | --- |
| Health Service Area | String |
| Hospital County | String |
| Operating Certificate Number | String |
| Facility Id | Float 32 |
| Facility Name | String |
| Age Group | String |
| Zip Code – 3 digits | String |
| Gender | String |
| Race | String |
| Ethnicity | String |
| Length of Stay | String |
| Type of Admission | String |
| Patient Disposition | String |
| Discharge Year | Float 32 |
| CCS Diagnosis Code | Float 32 |
| CCS Diagnosis Description | String |
| CCS Procedure Code | Float 32 |
| CCS Procedure Description | String |
| APR DRG Code | Float 32 |
| APR DRG Description | String |
| APR MDC Code | Float 32 |
| APR MDC Description | String |
| APR Severity of Illness Code | Float 32 |
| APR Severity of Illness Description | String |
| APR Risk of Mortality | String |
| APR Medical Surgical Description | String |
| Payment Typology 1 | String |
| Payment Typology 2 | String |
| Payment Typology 3 | String |
| Birth Weight | Float 32 |
| Abortion Edit Indicator | String |
| Emergency Department Indicator | String |
| Total Charges | Float 32 |
| Total Costs | Float 32 |

It should be pointed out that this dataset is significantly imbalanced. Although the LOS ranges from 1 to 120 days and beyond across the dataset, 14.26% of samples involve one-day of stay, 24.16% involve two-day of stay, 19.33% involve three-day of stay. Altogether these three classes of samples account for 57.75% of the dataset. As shown in Fig. 1, the frequency of samples over 3 days drops dramatically. The figure only shows the frequency distribution in the range of 15 days. The total frequency of LOS over 15 days only accounts for 1.88% of the dataset.

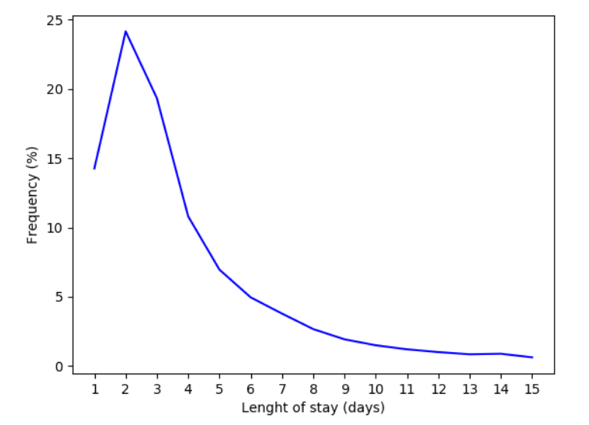


Fig. 1. Distribution of frequency of samples with LOS with 15 days.

## Feature Analysis and Selection

All 32 features in the dataset are not useful in the LOS prediction. For example, the health service area and hospital county are obviously unrelated with the LOS. We can visualize the relationship of each feature with the LOS. As an example, Fig. 2 shows the relationship between the APR severity of illness code and the LOS. It indicates that the higher the APR severity of illness code, the longer the LOS. Therefore, the feature of severity of illness affects the LOS and it should be considered in the LOS prediction. After analyzing 14 unique features of the dataset, we were able to determine which features were most influential and integral to a high predictability. These features, starting from most influential, are as follows: APR DRG Code, CCS diagnosis code, CCS procedure code, patient disposition, APR MDC code, APR risk of mortality, APR severity of Illness, age group, emergency department indicator, type of admission, APR medical surgical description, gender, race, and ethnicity. Features like patient disposition, although highly influential and important, are not included as they would only be documented after the patient had been discharged from the hospital. At the end, we selected 6 features in the prediction model training: APR DRG code, APR severity of illness code, CCS diagnosis code, APR risk of mortality, CCS procedure code, and APR MDC code.

Data preprocessing is also necessary, which includes converting all string types of data to float type, and replacing illegal data with valid ones.

## C. One Hot Encoding

5 of our features were determined to have non-ordinal relationships. Using one hot encoding, we ensure that our models do not place added “value” to the feature levels with higher integers. Depending on the sample size, our dataframe will increase dramatically from 6 features to over 800.

# Method

There are two approaches for the LOS prediction. One is to use regression algorithms, which attempts to predict exactly how many days a patient will stay in the hospital. The other one is to apply classification algorithms and predict a range of days for the patient. Due to the insufficient information in the dataset about patients, the regression algorithms would not produce prediction results with acceptable accuracy. Therefore, we try to predict a range of days of LOS. This way we turn the problem to a classification problem.

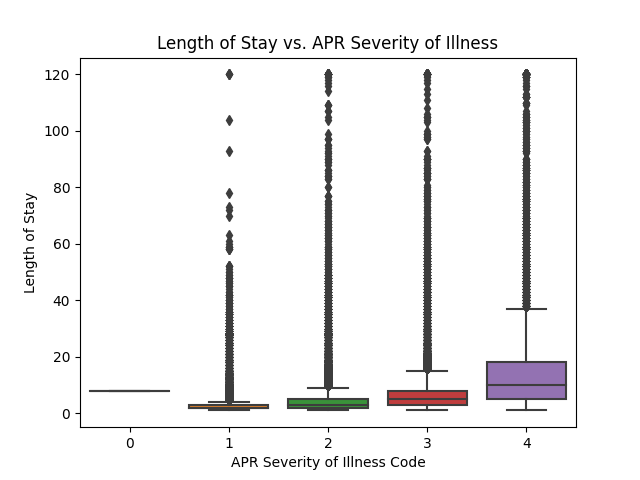


Fig. 2. LOS vs. APR Severity of Illness.

We used two types of machine learning algorithms in this study: ensemble algorithms, including random forests, gradient boosting and Ada boosting, and deep neural networks.

## Ensemble Methods

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [15]. Classifiers are different due to different sampling of training data, or randomized parameters within the classification algorithm. The objective of ensemble learning is to take a simple weak algorithm and transform it into a super classifier without requiring any fancy new algorithm. Bootstrap aggregating (bagging) and boosting are among common types of ensembles.

Random forests are an ensemble learning method for classification. A random forest is a series of decision trees, each based on a different subset of training samples that overlap. A decision tree constructs classification or regression models in the form of a tree structure. For classification, it employs an if-then rule set that is mutually exclusive and exhaustive. The rules are learned one by one, one by one, using the training data. In a random forest, the node size, number of trees, and sample size must all be predetermined before generating the forest. Each split node of each tree is a different subset of descriptors that is evaluated for the best split until it has been decided and the tree has been formed. As a result, a complex forest of different decision trees is created, with the average prediction of all trees serving as the final prediction for a new sample.

Adaboosting and gradient boosting are two boosting methods in machine learning. A boosting method incrementally constructs an ensemble by training each new model instance to focus on the training samples that previous models misclassified. The assumption is that when the best possible next model is combined with previous models, the total prediction error is minimized. Both Adaboosting and gradient boosting algorithms work for boosting the performance of a simple base-learner by iteratively shifting the focus towards problematic observations that are challenging to predict. The main difference is, the former achieves this by up-weighting observations that were misclassified before, while the latter identifies the difficult observations by large residuals computed in the previous iterations.

## Deep Learning

Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning from data. Deep learning utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning, which enables machines to process data with a nonlinear approach. Multi-layer perceptron (MLP) is a special type of deep learning, which is fully connected and normally only has one hidden layer. A more general deep neural network may have more than one hidden layer.

In this study, we used a 4-layer fully-connected learning network. To prevent over-fitting from happening too soon, we also added two drop-out layers.

# Results

We used two measurements to assess the performance of various learning algorithms, namely classification accuracy and confusion matrix.

The classification accuracy is measured as ratio of the number of correctly classified testing samples over the number of total testing samples. In many cases, classification accuracy is not sufficient to measure the performance of an algorithm; we also need to examine the confusion matrix.

A confusion matrix shows for each class of testing samples how many of them are classified correctly. The following figure is an example. There are four classes (labels). The first row shows that for label 0, 58% of testing samples are classified correctly, 39% of cases are classified as label 1 by mistake, 2.4% as label 2, and 0.011 as label 3. For label 2, 64% are classified correctly, which is not bad. However, for label 3, only 22% are classified correctly, which is bad. A confusion matrix indicating a good learning performance should at least have the biggest value along the diagonal. Ideally, the values on the diagonal should be close to 1, or the number of samples in the class.

| Chart, bar chart  Description automatically generated |
| --- |

Fig. 3. A confusion matrix.

To turn the LOS prediction to a classification problem, we need to decide groups of LOS. We did experiments with two different ways of grouping. In the first grouping, we divided the LOS to two disjoint ranges or classes:

1-3 days, 4+ days

The second grouping divided the LOS to three disjoint ranges:

1-3 days, 4-8 days, 9+ days

In addition to the ensemble algorithms and DNN mentioned above, we also experimented with linear regression (LR) and decision trees (DT). Table 2 lists the classification accuracy from each algorithm and way of grouping. The best results are produced by the random forest algorithm. In the binary classification case, it achieved an accuracy of 76.82%. In the three-class case, the accuracy is 66.61%. Notice that in the three-class case, the GBC and ABC algorithm appear to have higher accuracy over RF. However, because we are dealing with an unbalanced dataset, it is important we further check their confusion matrices. Comparing the two confusion matrices of the classification results from RF and GBC shown in Fig. 4, GBC has better classification accuracy on cases in the class of 1-3 days, which is the majority class. However, for the other two minority classes, the RF’s performance is much better than the GBC classifier.

(a) (b)

Fig. 4. (a) Confusion matrix of RF. (b) Confusion matrix of GBC.

Table 2. Accuracy of classification with raw data

| Ranges: 1-3, 4- | | Ranges: 1-3, 4-8, 9- | |
| --- | --- | --- | --- |
| Classifier | Accuracy | Classifier | Accuracy |
| LR | 0.7420 | LR | 0.6389 |
| DT | 0.7659 | DT | 0.6602 |
| RF | 0.7682 | RF | 0.6661 |
| GBC | 0.7593 | GBC | 0.6846 |
| ABC | 0.7548 | ABC | 0.6703 |
| DNN | 0.7545 | DNN | 0.6448 |

Out of the 200000 samples we used for training and testing, the distribution over the first grouping is as follows:

1-3 days: 115570 samples

4+ days: 84430 samples

The distribution for the three-class grouping is as follows:

1-3 days: 115570 samples

4-8 days: 52890 samples

9 and above: 31540 samples

It is obvious that the classes are imbalanced. To improve the classification accuracy, we can try to make classes balanced. To the end, we dropped about 30,000 samples from the class of 1-3 days and trained and verified each model again. The new results are listed in Table 3. It shows that for the binary classification case, there is about a one percent improvement across all classifiers. However, the change for the three-class case is very little. The reason is that even after we remove 30K samples from the class of 1-3 days, the three classes are still significantly imbalanced. Of course, we can try to make all the three classes balanced and improve the classification accuracy.

Table 3. Accuracy of classification with adjustment to the dataset.

| Ranges: 1-3, 4- | | Ranges: 1-3, 4-8, 9- | |
| --- | --- | --- | --- |
| Classifier | Accuracy | Classifier | Accuracy |
| LR | 0.7422 | LR | 0.6169 |
| DT | 0.7580 | DT | 0.6481 |
| RF | 0.7689 | RF | 0.6521 |
| GBC | 0.7636 | GBC | 0.6594 |
| ABC | 0.7624 | ABC | 0.6503 |
| DNN | 0.7644 | DNN | 0.6481 |

# *Discussion*: One might expect a classification accuracy at the level of 90%, but even when we formulate the LOS prediction as a binary classification problem, we only achieved a 76.82% of accuracy at the best. The major reason is with the quality of the data set. Because these data were collected from initial diagnosis at patients’ visit to hospitals, they don’t provide sufficient information to support accurate LOS prediction.

# On the other hand, the purpose of this study is to offer healthcare providers a tool that could help them with resource planning within a certain level of confidence. From that perspective, the classification accuracy achieved in this study is acceptable.

# Conclusion

In this study, we applied classification algorithms to predict the LOS. Logistic regression, decision tree, random forest, gradient boosting and deep learning algorithms are applied to model training and validation. Experiments show that among these classification algorithms, the random forest classifier achieved the best performance. Its classification accuracy is 76.82% when we divided the samples into two classes, less than (including) three days and more than three days. When dropped some samples from the first class to make the two classes more balanced, the classification accuracy was increased by one percent. When we divided the dataset to three classes based on the LOS, the classification accuracy was about 66%.

The overall prediction accuracy is not high. The main reason is the low quality of the dataset. LOS is determined by many factors. The features of the dataset can only provide a basis for a rough estimate of the LOS. Another reason is that the dataset is significantly imbalanced, which affects the machine learning performance.

We tried to under sample from the majority classes (i.e. classes 1-3 and 4-7) such that all classes have nearly the same number of data samples. This way we have a balanced dataset and it will help improve the quality of the trained models. This approach is doable because there are 1.3 million samples in the dataset. Another approach is to over sample the minority classes. One such technique is called SMOTE – synthetic minority oversampling technique. Its idea is to create synthetic (instead of simply duplicating) data points of the minority classes.

Another consideration would be to one hot encode our features. We converted features with string values to floats, however in doing so we created an inherent ordinal relationship between data values. If we instead one hot encoded our features, we could expect a higher LOS prediction accuracy.

# References

1. AHRQ, Length of stay, AHRQ Publication No: 090041, March 2009.
2. A. Alahmar, E. Mohammed, and R. Benlamri. Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes". In: Aug. 2018, pp. 38-43. doi: 10.1109/Innovate-Data.2018.00013.
3. David W. Bates et al. “Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients". In: Health A\_airs 33.7 (2014). PMID: 25006137, pp. 1123-1131.
4. Mark A. Blais, John Matthews, Robin Lipkis-Orlando, Erin Lechner, Michelle Jacobo, Robert Lincoln, Christina Gulliver, John B. Herman, and Alyson F. Goodman. “Predicting Length of Stay on an Acute Care Medical Psychiatric Inpatient Service.” in Administration and Policy in Mental Health and Mental Health Services Research. October 2003.
5. Catherine Combes, Farid Kadri, Sond\_es Chaabane. Predicting hospital length of stay using regression models: application to emergency department. Proceedings of 10th Francophone Conference on Modeling, Optimization and Simulation- MOSIM'14, Nov 2014, Nancy, France.
6. Georgios Drakos. “Decision Tree Regressor Explained in Depth.” *GDCoder*, GDCoder, 23 May 2019, [gdcoder.com/decision-tree-regressor-explained-in-depth/](http://gdcoder.com/decision-tree-regressor-explained-in-depth/). Accessed 27 Apr. 2021.
7. T. Daghistani, R. Elshawi, S. Sakr, A. Ahmed, A. Al-Thwayee, M. Al-Mallah, “Predictors of in-hospital length of stay among cardiac patients: A machine learning approach,” *International Journal of Cardiology*, 288, 2019, 140-147.
8. H. Greenspan, B. Van Ginneken, and R. M Summers. “Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique". IEEE Transactions on Medical Imaging 35.5 (2016), pp. 1153-1159.
9. L. He, S. Madathil, G. Servis, M. Khasawneh, “Neural network-based multi-task learning for inpatient flow classification and length of stay prediction,” *Applied Soft Computing*, 108, 2021.
10. X. Heng, Z. Wang, and J. Wang, “Human activity recognition based on transformed accelerometer data from a mobile phone,” International Journal of Communication Systems, October 2014.
11. B. Hu and J. Wang, "Deep Learning Based Hand Gesture Recognition and UAV Flight Controls, " International Journal of Automation and Computing, 2019.
12. M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, E. Liu, “Predicting all-cause risk of 30-day hospital readmission using artificial neural networks”. *PLoS ONE*, 12(7), 2017.
13. L. Turgeman, J. May, R. Sciulli, “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, 2017.
14. W. Zhang and J. Wang, "Dynamic Hand Gesture Recognition Based on 3D Convolutional Neural Networks Models," *IEEE International Conference on Network, Sensing and Control*, Banff, Canada, May 9 - 11, 2019.
15. T. Dietterich, Ensemble methods in machine learning, Multiple classifier systems, 2000.