2023 2nd International Conference on Futuristic Technologies (INCOFT)
Karnataka, India. Nov 24-26, 2023

Prediction of Length of Stay in Hospitals using Ensemble Machine Learning Models

Varun H Ramurs
Dept of CSE
PES University
Bengaluru, India
varun.hramurs@gmail.com

Shreyas C
Dept of CSE
PES University
Bengaluru, India
shreyaschikkanna@gmail.com

Raghu A Bangalore
Dept of CSE
PES University
Bengaluru, India
raghubarao@pes.edu

***Abstract*—Length of Stay(LOS) refers to the duration of time that an individual spends in the hospital. LOS plays a pivotal role as a key indicator, offering valuable insights into resource allocation, service planning, and operational efficiency. Understanding and effectively managing LOS can lead to improved operational efficiency, cost reduction, and enhanced customer experience. In this paper, we implemented and compared 5 Ensemble Machine Learning Models namely Logistic Regression (LR), Random Forest (RF), Gradient Boosting Classifier (GBC), CatBoost (CB) and XGBoost (XGB) on the New York's state Hospital Inpatient Discharges (SPARCS De-Identified): 2015 dataset. We have used both regression and classification to train these models. For classification we have used 3 different bin labels and calculated the performance metrics and plotted the comparison metrics graph.**

***Index Terms*—Length of Stay, Machine Learning, Ensemble Machine Learning Models, Classifier, Regressor**

I. Introduction

Hospitals always look for ways to make efficient use of their resources. The length of stay (LOS), which is defined as the interval between hospital admission and discharge, is an important measure of the efficiency of hospital management. It serves as an indicator of hospital resource utilization and determines the effectiveness of bed management in hospitals.

Estimation of LOS at the time of admission of the patient not only enables hospitals to create better plans for care activities, but also optimize resources, reduce expenses, and serve additional patients. Managing the length of stay in hospitals is considered to be one of the most important tactics for proper management of resources int the hospital, especially during times of crisis like Covid-19 and other pandemics. Predicting the length of stay allows patients to get an idea of their expenses based on the predicted duration and enables them to plan better for the forthcoming days.

With machine learning gaining importance, various industries and fields are utilizing its potential to improve their facilities and products. Healthcare facilities are also leveraging the power of machine learning. Tasks such as detecting tumors through scanned images and predicting the diseases patients are suffering from have become seamless through machine learning. Identifying the factors that affect the length of stay (LOS) is crucial in predicting its value. Removing irrelevant features from the data is essential as it prevents the model from providing inaccurate results and also saves time by preventing the model from training on unnecessary information. Data mining and several machine learning techniques are used to select these factors.

The main objective of this paper is to use the prowess of machine learning and implement a machine learning algorithm that predicts the length of stay of a patient upon admission by only studying their initial diagnosis report and test data. Referring to previous studies on this topic many machine learning algorithms have been considered, and suitable models have been selected and trained after trying out many others. We experimented with both regression and classification techniques on our dataset and compared the results, including traditional algorithms from scikit-learn and the recently developed CatBoost. Upon literature survey, we noted that some studies considered features which had unusually high correlation with LOS, which eventually led to data leakage. Few studies neglected some features and did not encode the categorical features. We selected our features after going through their influence on the target variable and also scaled our data so that it would not be influenced by outliers.

The dataset used in this study consists of approximately 2 million values, and our target variable, 'Length of Stay,' is skewed and unbalanced, with the mean value being around 5. To address this imbalance, we under-sampled our data and trained our model. The categorical columns of the dataset were one-hot encoded, and the columns with numerical data were scaled. Logistic Regression, Random Forest, Gradient Boost are the general algorithms used to train the model. CatBoost is also used to as it is more efficient and easier to use as compared to other general models. It enables us to work without encoding our categorical columns which saves us a lot of time in training the model. The method proposed in this paper fares comparatively better than the other methods proposed on this specific dataset, providing better results.

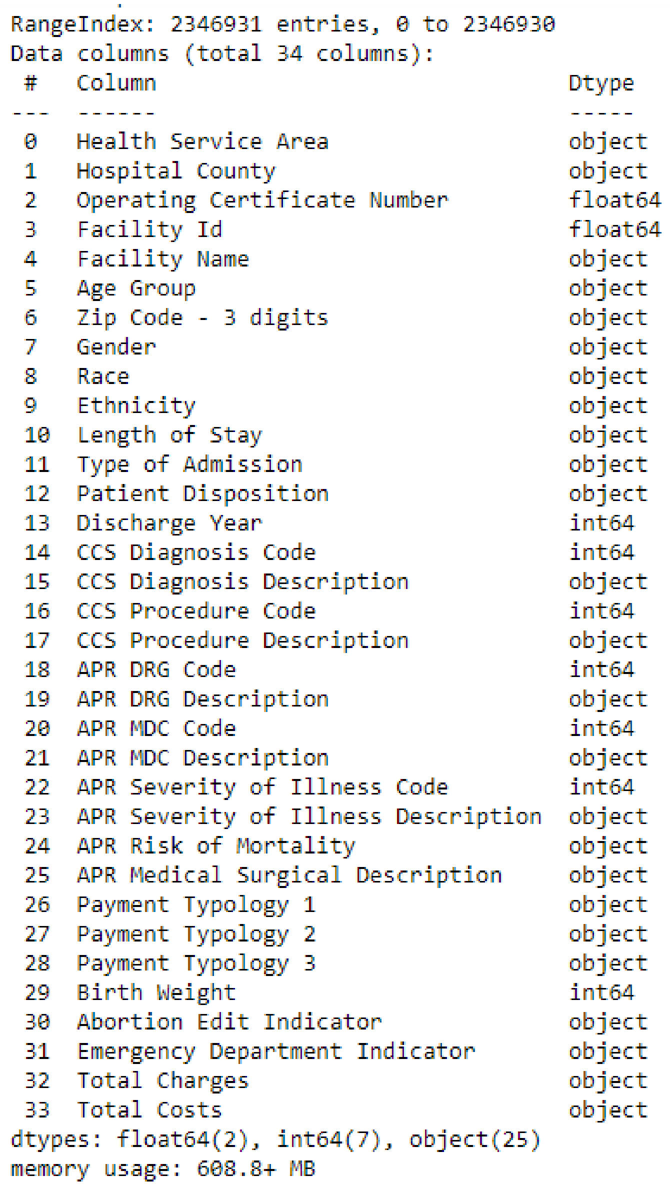
The rest of the paper is divided into the following categories: Section 2 presents the steps involved in preprocessing and preparing the dataset. Section 3 presents the models used and their importance in this study. Section 4 familiarizes the performance metrics used and the results achieved from respective models.

979-8-3503-0884-6/23/$31.00 ©2023 IEEE

II. Dataset Description and Data Preparation

A. New York Hospital Inpatient Discharge 2015 Dataset

This dataset is easily accessible to the public at HEALTH.DATA.NY.GOV website [1]. It contains 34 unique features and has around 2.3 million records. Fig. 1 shows us the information about the dataset. The dataset doesn't contain any data that is Protected Health Information (PHI). For more information about the dataset visit [1].



```
RangeIndex: 2346931 entries, 0 to 2346930
Data columns (total 34 columns):
 #   Column                              Dtype
---  ------                              -----
 0   Health Service Area                 object
 1   Hospital County                     object
 2   Operating Certificate Number        float64
 3   Facility Id                         float64
 4   Facility Name                       object
 5   Age Group                           object
 6   Zip Code - 3 digits                 object
 7   Gender                              object
 8   Race                                object
 9   Ethnicity                           object
 10  Length of Stay                      object
 11  Type of Admission                   object
 12  Patient Disposition                 object
 13  Discharge Year                      int64
 14  CCS Diagnosis Code                  int64
 15  CCS Diagnosis Description           object
 16  CCS Procedure Code                  int64
 17  CCS Procedure Description           object
 18  APR DRG Code                        int64
 19  APR DRG Description                 object
 20  APR MDC Code                        int64
 21  APR MDC Description                 object
 22  APR Severity of Illness Code        int64
 23  APR Severity of Illness Description object
 24  APR Risk of Mortality               object
 25  APR Medical Surgical Description    object
 26  Payment Typology 1                  object
 27  Payment Typology 2                  object
 28  Payment Typology 3                  object
 29  Birth Weight                        int64
 30  Abortion Edit Indicator             object
 31  Emergency Department Indicator      object
 32  Total Charges                       object
 33  Total Costs                         object
dtypes: float64(2), int64(7), object(25)
memory usage: 608.8+ MB
```

Fig. 1. Information of the dataset

B. Experimental Data Analysis

On analyzing the dataset, we can see that it is very unbalanced. The dataset's target variable, LOS, which represents the length of stay, spans an extensive spectrum ranging from 1 to 120 days. Nonetheless, a significant portion of the dataset (approximately 66%) comprises samples with LOS values of 1, 2, 3, or 4 days. In contrast, instances with LOS exceeding 4 days are notably less frequent. Fig. 2 shows the LOS distribution. We got an average LOS of around 6 days.

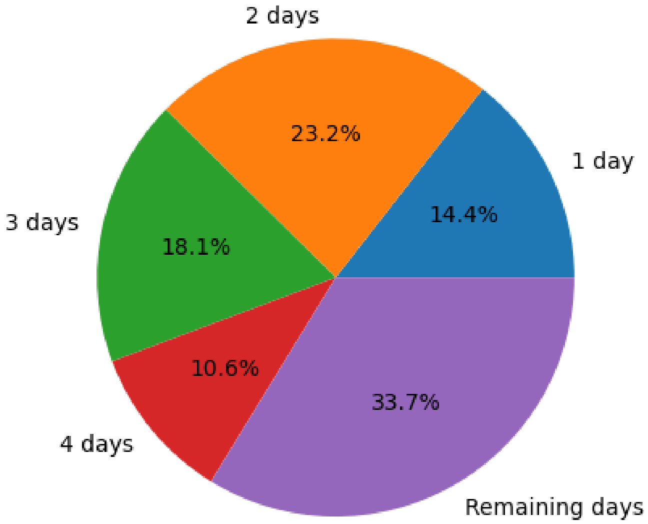


Fig. 2. LOS Distribution

C. Data Preparation

1) One-Hot-Encoding: Some machine learning models, like Random Forest, Gradient Boost, and Logistic Regression, can't handle categorical data. To use these models, we need to encode the categorical data into numerical values. This is done in two steps. First, we convert the categorical values into integers. Then, we convert the integers into binary representations.

There are a few reasons why we do one-hot-encoding. First, it allows machine learning algorithms to understand the meaning of categorical data. For example, the machine learning algorithm would know that ”red” is different from ”green” even though their values are integers. Second, one-hot-encoding helps to prevent the curse of dimensionality. As the number of features in a dataset increases, the dimension of the data fed into the machine learning algorithms also increases making it difficult for the algorithms to learn the relationships between the features. One-hot encoding can help to reduce the dimensionality of a dataset by creating new features that are not as correlated as the original features.

2) Feature Selection: Since the dataset contains 34 unique features, not all of them will influence LOS. There are some features like 'Total Charges' and 'Total Costs' which are highly correlated with our target variable but considering those 2 features will lead to data leakage as charges majorly depend on the duration of length of stay in hospitals. The columns: Facility Name, Race, and Ethnicity have no affect on LOS whereas columns like Age Group and Type of Admission impact LOS greatly. We can see that As age increases, LOS also increases from Fig. 3. Similarly ”Urgent”, ”Emergency” and ”Trauma” Type of Admissions have significantly high LOS as well.

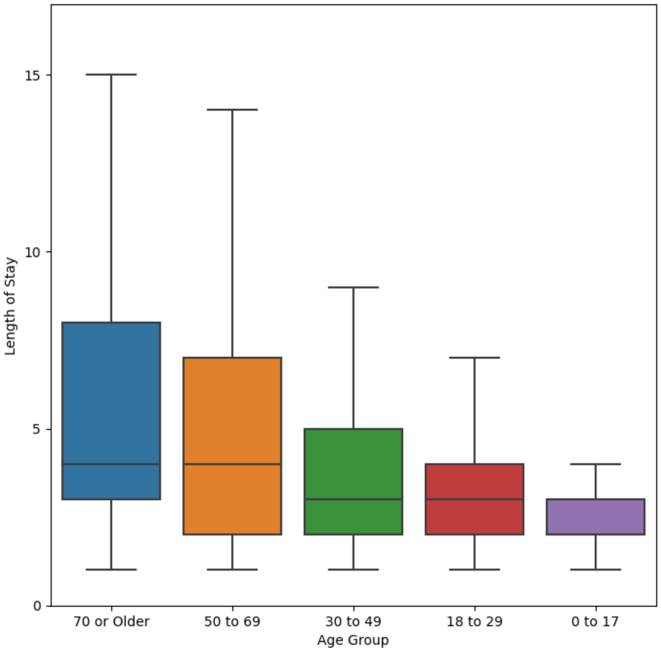


Fig. 3. LOS vs Age Group

To identify the columns that do influence LOS, we plotted graphs of all columns against LOS. We chose a total of 11 features to train the model:

- Age Group
- Length of Stay
- Type of Admission
- CCS Diagnosis Code
- CCS Procedure Code
- APR DRG Code
- APR MDC Code
- APR Severity of Illness Code
- APR Risk of Mortality
- Payment Typology 1
- Emergency Department Indicator

We partitioned the dataset into training and testing sets, ensuring a 7:3 proportion, by employing the train-test split functionality provided by Scikit-Learn. The results are calculated on the testing set.

3) Data Standardization: After selecting the features that influence LOS, we normalize the data using sklearn’s StandardScaler function which uses mean and standard deviation to center the values. Standardizing is useful as it is not sensitive to outliers and retains the relationship between the variables.

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

D. Under Sampling

Since the average LOS is around 6 days, it is skewed and there is a class imbalance, with most of the entries between 0–8 days. This leads to overfitting, and the model predicts the lesser values most of the time. This gives us high accuracy with this dataset, but it may not be the case when some other dataset is used. To avoid overfitting, we have also undersampled our data to see how it compares against the model with unbalanced data.

To undersample the majority classes of our data, we have used imbalanced learn’s Random UnderSampler. The table below shows the data before and after undersampling for various bins where 2 represents the bins (0-6, 6-120+), 3 represents (0-6, 6-30, 30-120+) and 4 represents (0-6, 6-14, 14-30, 30-120+).

TABLE I
UNDER-SAMPLING VALUES

Bins	**Before Under-Sampling**	**After Under-Sampling**
2	1282981, 359793	112307, 80000
3	1282691, 334744, 25339	80000, 50000, 25339
4	1282698, 2518243, 82882, 25351	100000, 80000, 50000, 25351

III. MODELS

This section highlights the methods applied for this problem. There are two types of algorithms that can be used to predict LOS: regression algorithms and classification algorithms. Regression models forecast the precise duration of a patient’s hospital stay, whereas classification algorithms provide estimates for a span of days. Referring to previous studies and comparing the results they achieved, we used Random Forest Regressor, Gradient Boost Regressor, CatBoost Regressor and XGBoost as our regression algorithms. Regression algorithms achieve best results when the dependent variables of the data extensively correlate with the target variable. However, since the dataset lacks such features, we tried out classification algorithms.

To apply classification algorithms, we divided our dataset into three bin classes (0-6, 6-120+), (0-6, 6-30, 30-120+) and (0-6, 6-14, 14-30, 30-120+), corresponding to weekly, biweekly, monthly and many days. Dividing the length of stay like this allows hospitals to get a clear idea of the ranges of lengths of stay that patients will have in terms of weeks. We trained our models on three different sets of bin classes and applied Logistic Regression as our preliminary model. After achieving acceptable results in Logistic Regression, we also trained Random Forest Classifier, Gradient Boost Classifier, and CatBoost Classifier, expecting better results.

We make use of ensemble machine learning models to carry out the predictions. Ensemble Learning combines multiple weak learners to produce strong learners which improve the performance. Ensemble Models are of two types:

- Bagging - Generates several model variations through sampling the training data with repetition. This approach implies that certain data points could be present in multiple models, while others might not be featured in any. The predictions from the individual models are then combined to create a final prediction. Ex: Random Forest

- Boosting - Boosting techniques function by progressively constructing a series of minor models, each aimed at rectifying the inaccuracies of its predecessors. This process is repeated until the errors are minimized, and the dataset is predicted correctly. Ex: Gradient Boost

A. Logistic Regression

Logistic regression stands out as one of the widely acclaimed machine learning algorithms, offering predictions for the outcomes of categorical dependent variables, which means that the output is a probability between 0 and 1. It is used in cases when we have to predict whether a given test object belongs to a particular class or not. Logistic regression uses the sigmoid function to map the predicted values to probabilities.

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2}$$

B. Random Forest

In the context of a random forest model, each decision tree is constructed using a random subset of data points and features. This procedure is iterated to produce numerous decision trees. Subsequently, the forecasts from these individual trees are merged to generate the ultimate prediction.

For classification objectives, the ultimate prediction is determined through a majority consensus among the individual decision trees. In regression tasks, the ultimate prediction is obtained by averaging the prognoses of the individual decision trees.

By creating multiple decision trees and amalgamating their forecasts, this technique aids in curbing both variance and bias within the model, leading to more precise predictions.

Algorithm 1: Random Forest Algorithm

```
Data: Training dataset
Result: Ensemble of decision trees
while Number of decision trees $\leq$ Total number of trees do
    Randomly select a subset of training data;
    Train a decision tree on the selected subset;
    Add the decision tree to the ensemble;
if Classification problem then
    Use Majority Voting to determine the final output;
if Regression problem then
    Take the average of individual decision tree outputs as the final prediction;
```

C. Gradient Boost

Gradient boosting is a sequential model-building technique where each successive model aims to rectify the errors of its predecessor. This iterative process continues until the desired level of accuracy is attained.

Two variations of gradient boosting exist: gradient boosting regressor and gradient boosting classifier. The sole distinction lies in the employed loss function. This function gauges the model's efficacy in predicting the target variable. Mean-squared error is often chosen for regression tasks, while log-loss is favored for classification challenges.

$$L(y,p) = -\sum_i [y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{3}$$

The central goal of gradient boosting is to diminish the loss function by incorporating weak learners through gradient descent. Weak learners are simple models that are not very good at predicting the target variable. However, when they are combined, they can form a strong model that is able to predict the target variable with high accuracy.

D. CatBoost

CatBoost is a fairly new Machine Learning technique based on Gradient Boosting. Its uniqueness lies in its ability to effectively manage both categorical and numerical features, eliminating the need for feature encoding.

CatBoost exhibits notable strength in two significant aspects:

- It can attain cutting-edge outcomes without demanding extensive data training, which is commonly essential for other machine learning techniques. This renders it an invaluable asset for enterprises facing data constraints or lacking the capacity to gather and train vast datasets promptly.
- Additionally, CatBoost offers robust, immediate compatibility with diverse data formats that often arise in business scenarios. This versatility enables the analysis of data that might prove challenging for other machine learning approaches, encompassing text, images, and audio.

While many machine learning algorithms necessitate data to be in numeric format for training, CatBoost uniquely handles categorical features directly, negating the need for encoding. To employ CatBoost with categorical features, the data should be transformed into CatBoost's specialized Pool datatype, accomplished using the **Pool()** class. Furthermore, the categorical feature names should be specified within the cat_features parameter.

Algorithm 2: CatBoost Algorithm

```
Data: Training dataset, hyperparameters
Result: Trained CatBoost model
Initialize CatBoost model with hyperparameters;
if Categorical features exist then
    Create a CatBoost Pool from the training set with categorical features;
Train the model using the CatBoost Pool;
```

If the dataset exclusively comprises numerical features, conversion to the Pool datatype is unnecessary. In such instances, data can be directly provided to CatBoost in the form of a numpy array or Pandas DataFrame. Depending on the target variable, either CatBoostRegressor or CatBoostClassifier can be employed.

E. XGBoost

XGBoost is based on Gradient Boosting but it is different from other boosting algorithms in a few ways. First, XGBoost uses a more efficient algorithm for training the trees. This makes it much faster to train XGBoost models than other boosting algorithms. Second, XGBoost uses regularization techniques to prevent overfitting. This helps to ensure that XGBoost models generalize well to unseen data. Finally, XGBoost is more flexible than other boosting algorithms. It can be used for a wider variety of machine learning tasks, and it can be tuned to improve performance on specific datasets.

IV. Results

Cross validation plays a pivotal role in mitigating overfitting concerns by partitioning the training data into several folds. The model is subsequently trained on a subset of these folds and assessed on the remaining ones. This iterative procedure is conducted multiple times, with the outcomes averaged to derive a more precise evaluation of the model's efficacy on previously unseen data. We chose the value of the fold to be 5. The mentioned results represent the averaged accuracy obtained through 5-fold cross-validation. These results closely align with those achieved by manually splitting the data using train-test split.

A. Regression

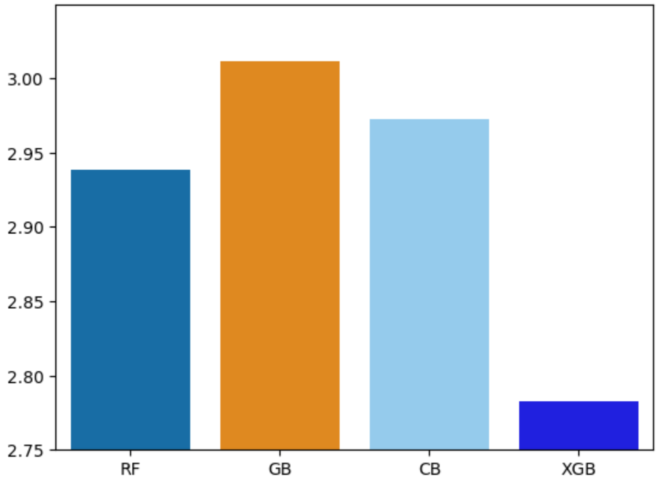


Fig. 4. MAE of various regressor models

We chose Mean Absolute Error (MAE) as the performance metric for the regression models because of its simplicity in comprehension and interpretation. It quantifies the average of absolute errors between predicted values and actual values. The mean absolute error (MAE) is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

The average coefficient of determination of the regression model came out to be approximately 0.33, which explains the lack of correlation between the various features of the dataset. This prompted us to use classification models.

B. Classification

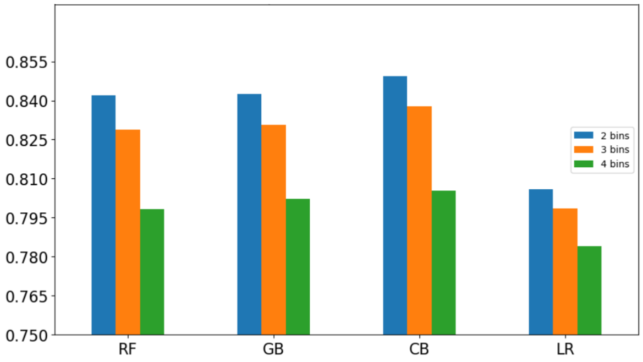


Fig. 5. Accuracy of various classification models

We employed two evaluation metrics to assess the performance of the algorithms implemented in this research: accuracy and confusion matrix. Accuracy represents the ratio of accurate predictions made by a model to the total number of predictions. Given the classification nature of our data, we also utilized a confusion matrix to display the correct classification count of data samples. This matrix aids in identifying potential overfitting of the training data by the model.

Optimal results emerged when the model dealt with a binary classification scenario involving two classes. The CatBoost model exhibited superior performance, demonstrating an average accuracy of 0.83 across all three considered categories of bins within this study.

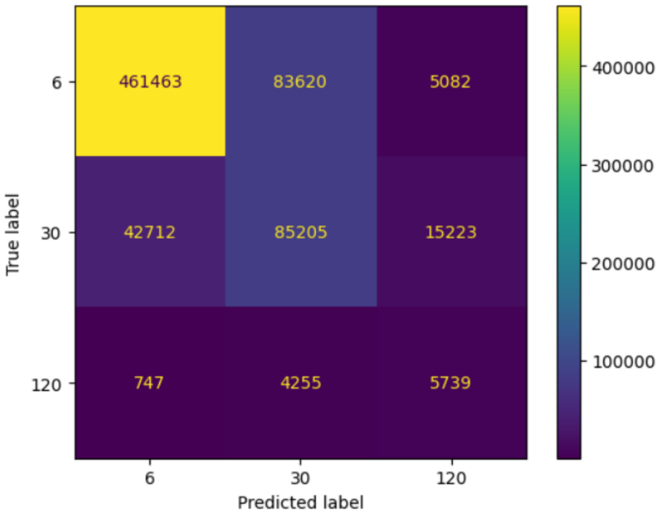


Fig. 6. Confusion matrix for 3 bins undersampled data

The table below shows the cross-validation accuracy, in column 2 and the undersampled accuracy for all three bins in column 3, of the Random Forest Classifier model.

TABLE II
UNDER-SAMPLING ACCURACY VALUES

Bins	Before Under-Sampling	After Under-Sampling
2	84.11%	80.70%
3	82.77%	78.46%
4	79.81%	73.01%

The reduced accuracy after undersampling data is due to irregularity in the values considered for training the model.

CONCLUSION

We applied both regression and classification models to predict length of stay (LOS). In the regression models, XGBoost gave the best result with the lowest mean absolute error (MAE) of 2.783. We observed that there was no correlation between the columns, so we also tried the classification models.

We split the dataset into three different bin classes. Since the average LOS was around 6 days, we split the bins into three classes: (0-6), (6-120+), and (0-6, 6-30, 30-120+). CatBoost gave the best result overall with an accuracy of 83%. We also tried cross-validation of the results, but we still ended up with the same accuracy. Doing under-sampling to eliminate the imbalance in the dataset reduced the overall accuracy of all models by 5%.

ACKNOWLEDGEMENT

We extend our heartfelt appreciation to our mentor, Raghu A Bangalore, for granting us the privilege to engage in this project. Our gratitude also goes to the institution for furnishing us with the necessary resources essential for the successful completion of this endeavor.

REFERENCES

[1] https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8

[2] Zheng, L., Wang, J., Sheriff, A. and Chen, X., 2021, December. Hospital length of stay prediction with ensemble methods in machine learning. In 2021 International Conference on Cyber-Physical Social Intelligence (ICCSI) (pp. 1-5). IEEE.

[3] Mekhaldi, R.N., Caulier, P., Chaabane, S., Chraibi, A. and Piechowiak, S., 2020, April. Using machine learning models to predict the length of stay in a hospital setting. In World conference on information systems and technologies (pp. 202-211). Cham: Springer International Publishing.

[4] Suha, S.A. and Sanam, T.F., 2022, July. A Machine Learning Approach for Predicting Patient's Length of Hospital Stay with Random Forest Regression. In 2022 IEEE Region 10 Symposium (TENSYMP) (pp. 1-6). IEEE.

[5] Alabbad, D.A., Almuhaideb, A.M., Alsunaidi, S.J., Alqudaihi, K.S., Alamoudi, F.A., Alhobaishi, M.K., Alaqeel, N.A. and Alshahrani, M.S., 2022. Machine learning model for predicting the length of stay in the intensive care unit for COVID-19 patients in the eastern province of Saudi Arabia. Informatics in Medicine Unlocked, 30, p.100937.

[6] Ayyoubzadeh, S.M., Ghazisaeedi, M., Rostam Niakan Kalhori, S., Hassaniazad, M., Baniasadi, T., Maghooli, K. and Kahnouji, K., 2020. A study of factors related to patients' length of stay using data mining techniques in a general hospital in southern Iran. Health information science and systems, 8, pp.1-11.

[7] Rahman, Md Mahbubur, Dipanjali Kundu, Sayma Alam Suha, Umme Raihan Siddiqi, and Samrat Kumar Dey. "Hospital patients' length of stay prediction: A federated learning approach." Journal of King Saud University-Computer and Information Sciences 34, no. 10 (2022): 7874-7884.

[8] Chrusciel, J., Girardon, F., Roquette, L., Laplanche, D., Duclos, A. and Sanchez, S., 2021. The prediction of hospital length of stay using unstructured data. BMC Medical Informatics and Decision Making, 21(1), p.351.