**Classification and Prediction of Hospitalization for HPN**

Shuyao Wang

**Abstract https://github.com/Ginawsy/HHP.git**

To mitigate the issue of unnecessary hospitalizations and related costs, I have developed algorithms using random forest, lasso regression, and ensemble models to classify patients who are more likely to require extended hospital stays and the duration of their stay. The model's accuracy in predicting whether a patient would require hospitalization or not is significantly higher than predicting the length of their stay. I found that the random forest algorithm performed better than lasso regression, resulting in a smaller root of the mean squared error (RMSE). The results of my model suggest that the majority of patients will require fewer than two days of hospitalization, though bias might exist considering a large number of missing values in the original dataset. The decision to admit a patient to the hospital is primarily influenced by the types of services and group conditions, while the severity of the illness is the main factor that determines the length of stay. To address this issue, I recommend that hospitals provide special attention to the elderly and allocate additional funds to rehabilitation and laboratory departments. Furthermore, it may be beneficial to reach out to patients who have exceeded a certain number of claims or medication assumptions to educate them about the potential risks and severity of their condition ahead. Special attention should be paid to people in the evaluation and management procedure group and others who assume large quantities of drugs and lab tests as well to address the issue of payment delays. From iRF, missing values also show pair effects with payment delay which might give us a reasonable explanation for the missing data at the end.

**Introduction**

The Heritage Health Prize Contest was a competition that challenged participants to develop an algorithm that could predict how many days a patient would spend in a hospital within the upcoming year, based on their electronic medical records. The question focused on solving avoidable hospitalizations by identifying patients at high risk of hospitalization and providing them with appropriate treatment before they require hospitalization.

Given the facts of the limited capacity of hospitals to accommodate patients, it is critical to identify individuals who are at a high risk of requiring hospitalization ahead first. By doing so, hospitals can better plan and allocate their resources and provide patients with the necessary care and support, thereby improving their mental well-being. In this project, I have also employed a classification method to identify the patients who are more likely to need hospitalization in the next year and single or pairs of features that have the most significant impact on the likelihood of hospitalization. By identifying these factors and their interactions, healthcare providers can take proactive steps to manage patients' conditions and minimize the risk of hospitalization.

**Data Processing**

One of the significant challenges encountered in processing the dataset was the handling of a substantial number of missing values. To address this issue, I first calculated the percentage of missing values in each column and focused on addressing the length of stay column, which has the highest number of missing values, named LengthOfStay. To do so, I removed 0.4% of the data in which the LengthOfStay value was null and SupLOS, the sign of suppression or not, was 1. I then created a new label to represent the remaining 97.31% of missing values in the SupLos column and filling it with 0 in the LengthOfStay column to differentiate these two attributes. Then, I employed the mcar\_test and missingno matrix (Fig1) for the remaining missing values in other columns' labels. However, this approach was not feasible for the category of MCRA, as the results of the test and the missing values pattern of ProviderID and Vendor precluded the possibility of dropping the values directly. Therefore, I converted each column containing missing values into a binary matrix and used the chi-square test to examine whether there was any relationship between the missing values and the observed variables. The results did not support the assumption of Missing at Random (MAR) since no significant relationships were found, except for Year and PCP. Based on these findings, I developed a strategy to create new labels to represent missing values in the dataset.

The dataset also contains some top-coded values that required examination. I first checked each feature to see if it exceeded the 99.5th percentile cut-off (Emam K, 2008), and only deleted values that exceeded the bound to retain more complete information. This resulted in only retaining rows where CharlsonIndex, the measure of the effect diseases have on overall illness’, was '5+'. For other top-coded features, I removed the “+” character and converted all non-discrete variables to numeric ones. For level variables, I replaced the values with the average number within the range. Next, I grouped the data by MemberID and Year and the histogram showed that the values 43 and 44 exhibited an unusual trend (Fig2). Thus, I truncated the data where the number of claims exceeded 43. I then counted each unique variable for discrete features and summed variables for numeric features. I transformed the remaining discrete variables into individual columns for each dataset and merged them by MemberID. Finally, I divided the merged dataset into two datasets recording the first-year and second-year data separately and merging them with the following year's targeted DaysInHospital values for further analysis.

Chart, line chart

Description automatically generatedChart, histogram

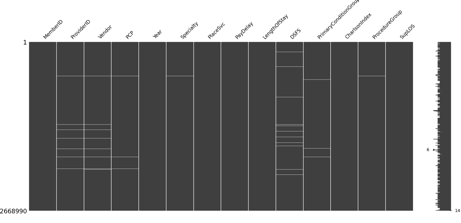
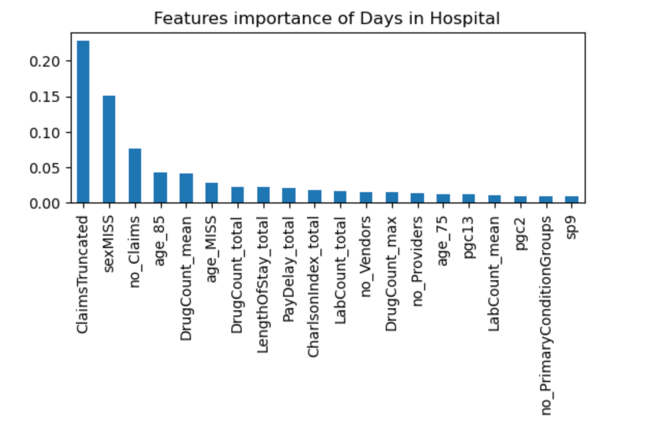
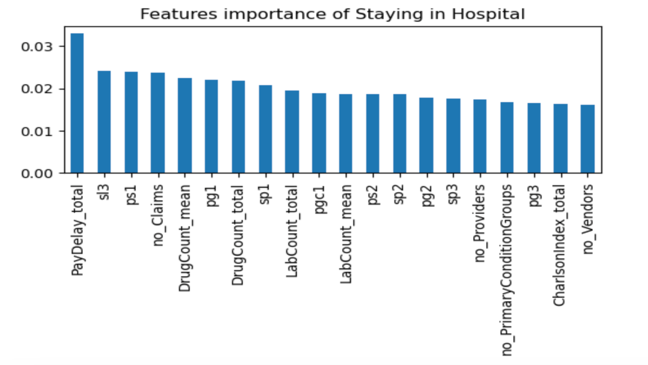
Description automatically generatedThe resulting data frames include 127 variables, including MemberID and Year, and comprise 71412 rows of data for year 1 and 76037 rows of data for year 2.

Figure1: missingno matrix for missing values matrix Figure2: Histogram of the number of claims Figure5. Density plot of hospitalization in year 3 by three models

**Data Analysis**

Reflecting on the Milestones, the successful teams employed various innovative techniques in their approach to the problem. Gradient Boosting Machines emerged as a popular choice, but other methods such as Neural Networks, Bagged Trees, and linear models were also used by the winning teams. Additionally, some teams employed customized models and varied combinations of datasets for model building. Ensembling was used by some teams to combine multiple models into a single solution, and algorithm modifications were made for the subsequent milestones. Cross-validation was also utilized to evaluate the Root Mean Square Error (RMSE, which is our objective function for models to minimize) of each model, aiding in determination of their accuracy and efficiency.

Chart, treemap chart

Description automatically generatedFigure3. Histogram of Features Importance of Hospitalization in Year 3 Figure4. Heat Map of Pearson correlation for Hospitalization in Year 3 Figure5. Histogram of Feature Importance of Hospital Length in Year 3

The focus of this project is on lasso regression and random forest methods, which were chosen due to the presence of a large number of variables and outliers in the dataset. Lasso regression was selected for its ability to automate feature selection and variable elimination, while the random forest was chosen for its robustness to the impact of outliers. For the future classification question, the number of days patients stayed in the hospital that was greater than zero was transformed into one, and a random forest classifier was used to identify the top 20 important features. The number of claims, drug assumptions over one year, the number of lab tests, and the number of primary conditions were found to have strong feature importance for the length of stay, as expected (Fig3). The types of groups and services were found to play a significant role in determining whether a patient needed to stay in the hospital or not, with patients requiring laboratory and drug services being more likely to require hospitalization. The number of providers and vendors was also found to be informative in predicting the length of stay. Another interesting observation was that patients who did not report any information on suppression or length of stay were among the top contributors, leading to the suspicion of important missing information. Pearson correlation was calculated for each pair of features, and the results were visualized in a heat map (Fig4). After further implementing the iterative random forest model, I find that there are also a couple of pairs that result in high stability scores from the iterative random forest model such as drug assumption and payment delay, number of claims, and payment delay. Hospitals should also keep in mind the mild pair effect between payment delay and evaluation and management procedure group and lab number separately; and the number of claims and drug assumption. Interestingly, the payment delay also indicates the pairing effect from the missing values significantly which might give us a possible reason for that lost information.

In order to make further predictions regarding the length of hospital stays, a test subset comprising 30% of the data was partitioned and an evaluation matrix was used to determine resulting root mean square error (RMSE) after standardizing the data. The results indicate that the RMSE for the first and second years are both approximately 0.49, with only a 0.06 difference observed between the RMSE values obtained from random forest model. For lasso regression model, the RMSE values obtained for the two years are 0.486 and 0.4985, respectively. After applying the year 2 data to the model for prediction, results were grouped by DIH and a density graph was produced to visualize trends (Fig5). Furthermore, an attempt was made to blend the random forest and lasso regression models to obtain better predictions by averaging the prediction values. However, it is interesting to note that lasso regression and blended models produced some truncated values from the density graph, resulting in predictions below zero, whereas random forest models did not exhibit this behavior. The feature importance results reveal that several common features, such as number of claims and drug, appear in both graphs, albeit with slightly different ranks. Age also appears to be an important factor contributing to the duration of hospitalization (Fig6).

**Discussion**

Based on the model predictions, the vast majority of patients were not expected to stay in the hospital, and among those who did stay, most of their stay was less than two days based on our models. The classification problem gave more accurate results than the specific prediction, as evident from the smaller RMSE. Among the three algorithms we tested, random forest performed the best and had all data in the density graph above zero. During our investigation, we discovered that the number of claims and drug assumptions had a significant impact both on the likelihood and length of hospitalization. However, the types of services and group conditions played a crucial role in determining whether patients stayed in the hospital or not, while the severity of illness was more critical for the length of stay. Patients who required laboratory and drug services were more likely to stay in the hospital, and those with conditions such as chest pain, metabolic disorders, and undergoing rehabilitation therapy were more likely to have a longer stay. Furthermore, the number of truncated claims from the previous year also played a critical role in the length of stay. The probability of hospitalization can also be determined by observing patients' providers and vendors as well. To achieve the goal of reducing healthcare costs and unnecessary hospitalization, I suggest that hospitals pay special attention to the elderly and allocate more resources to rehabilitation and laboratory-related departments. Furthermore, it would be beneficial to reach out to patients who have reached a certain number of claims and drug assumptions to make them aware of the severity and risks of their condition. Moreover, to deal with the issue of payment delay, I would suggest they especially notice the patients with a large number of claims and people who are in the evaluation and management procedure group if they want to loan some money. Remind people who have a large number of lab tests to pay on time ahead, not at the end of the treatment.

In order to improve the accuracy of the model, several considerations must be taken into account. During data processing, it was difficult to determine when to exclude high-risk patients because previous studies suggested that patients with certain illnesses, such as HIV, abortion, abuse, psychosexual disorders, mental retardation, or plastic surgery, should be removed since their conditions could be inferred from data patterns (El Eman, 2012). However, these diseases are part of a subset at a specific level, and the percentage of data specifically related to the diseases of interest is unknown, so further exploration of the dataset is required before data processing can proceed. Additionally, my algorithm's results might contradict the conclusion of Jonathan's research a little which stated that ensemble methods are typically more effective than any individual model (2012). This may be due to the particular ensemble strategy I used, and I would consider changing other ensemble methods, such as stacking and blending, to see if the results change. If so, my findings could indicate the significance of the chosen ensemble methods. Furthermore, we must devise more creative solutions to address the large number of missing values, which are among the most important features according to the results. Strategies to address missingness, such as modeling the missing values themselves, should be considered. Additionally, we would also employ feature selection to identify the most impactful features first, then perform predictions to compare the results.

**Reference**

El Emam K. Heuristics for de-identifying health data. IEEE Secur Priv. 2008;6(4):58–61. doi: 10.1109/MSP.2008.84.

El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J (2012). De-identification methods for open

health data: the case of the Heritage Health Prize claims dataset. J Med Internet Res.14(1):e33. doi: 10.2196/jmir.2001

Erickson, N., Shing, X., Sharpnack J., Smola A. (2022). “Multimodal Automl for Image, Text and Tabular Data.” Proceedings of the 28th

ACM SIGKDD Conference on Knowledge Discovery and Data Mining, <https://doi.org/10.1145/3534678.3542616>.

Stroud, J., Enverga, I., Silverstein, T., Song, B., Rogers, T. (2012). “Ensemble Learning and the Heritage Health Prize.”https://www.math.uci.edu/icamp/summer/research\_12/student\_research/ensemble\_learning/ensemblelearning.pdf.