Group 4

Dannica Concepcion, Kelsey Manglicmot, Kemil Herath, Ryan Kelly, Xiangyu(Shine) Lou

11/30/2020

MIS 776 - Executive Summary

**Hospital Admissions**

**INTRODUCTION**

The current covid-19 pandemic has raised major concerns about healthcare management and the efficiency of hospital admissions. It is important for hospitals to analyze their admission data to understand the patterns and be able to organize their admissions accordingly. This is especially important for room and bed allocation planning. It would allow hospitals to determine whether they can withstand capacities. As of November 23, 2020, the utilization of inpatient beds occupied across the United States is 72.77%. Because of these concerns, Group 4 chose to do an analysis of a dataset about hospital admissions.

**DATASET OVERVIEW**

The hospital admissions dataset was downloaded from Kaggle.com (<https://www.kaggle.com/nehaprabhavalkar/av-healthcare-analytics-ii>). The dataset consists of admission data from 32 hospitals, 11 cities, 92,017 unique patients, and has 318,438 entries. It contains 18 variables of which, 3 are numerical, and 15 are categorical.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Case\_id | categorical | Case\_id registered in Hospital |
| Hospital\_code | categorical | Unique code for the Hospital |
| Hospital\_type\_code | categorical | Unique code for the type of Hospital |
| City\_code\_hospital | categorical | City Code of the hospital |
| Hospital\_region\_code | categorical | Region code of the Hospital |
| Available\_extra\_rooms | numerical | Number of extra rooms available |
| Department | categorical | Department overlooking the case |
| Ward\_type | categorical | Code for the ward type |
| Ward\_facility\_code | categorical | Code for the ward facility |
| Bed\_grade | categorical | Condition of bed in the ward |
| Patient\_id | categorical | Unique patient ID |
| City\_code\_patient | categorical | City Code for the patient |
| Type\_of\_admission | categorical | Admission Type registered by the Hospital |
| Severity | categorical | Severity of illness at the time of admission |
| Visitors\_with\_patient | numerical | Number of visitors with the patient |
| Age | categorical | Age of the patient |
| Admission\_deposit | numerical | Deposit at the time of admission |
| Stay | categorical | Stay days by the patient |

The Case\_id was dropped because it represented the row number and was unnecessary. The City\_code\_patient column was dropped because it contained too many null values (4532). 113 null rows for Bed\_grade were also dropped. Conversions had to be done to the categorical variables to allow for modeling. Ordinal variables (Stay, Age) were converted to a numeric order, and categorical variables (Hospital\_code, Hospital\_type\_code, Hospital\_region, Department, Ward\_type, Ward\_facility\_code, Admission\_type) were one-hot encoded.

**BUSINESS QUESTION**

Group 4 wanted to use the dataset to predict how long a patient will stay in a hospital.

This will help hospitals identify patients of high length of stay (LOS) risk at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimized to minimize LOS and lower the chance of staff/visitor infection. This prediction can also aid in logistics such as room and bed allocation planning.

Other insights that Group 4 wanted to identify are what variables contribute the most to a patient’s LOS as well as what model(s) would be best to analyze the data and predict LOS.

**EXPLORATORY DATA ANALYSIS (EDA)**

**Numerical EDA**

* The average number of available extra rooms in a hospital is 3.2 (3-4), with a minimum of 0 and a maximum of 24.
* The average number of visitors with a patient is 3.3 (3-4), with a minimum of 0 and a maximum of 32.
* The average admission deposit is $4880.75, with a minimum of $1880.00 and a maximum of $11,008.00.
* According to the pairplot constructed, there is a lack of correlation among numerical variables. Therefore, there is no evidence of multicollinearity.

**Categorical EDA**

* Most cases fall into the 11-20 and 21-30 day range.
* The majority of cases are overlooked by the gynecology department.
* Median stay across all departments is 21-30 days.
* Age ranges 31-40 and 41-50 have the most cases.
* Patients 81 and older have a higher median for stay (31-40 days).
* Most cases are considered moderate at the time of admission.
* The upper quartile of ‘Extreme’ severity cases stay 51-60 days.
* Median stay across all severity levels is 21-30 days.
* Hospital city code 1 has the most patients.
* There are no cases for hospital city code 8 or 12.
* Median stay across all hospital city codes is 21-30 days.

The target variable that was used was Stay. Stay is divided into 11 different categories ranging from 0-10 days to more than 100 days.

**PREDICTIONS**

According to the cumulative EDA, the model Group 4 believes would be the best fit to predict the length of stay of a patient at the time of admission is a Classification model. This kind of model would generate the best results to predict Stay, a categorical variable. The variables Group 4 believes will contribute the most to the model are severity, age, and admission type. The more severe an illness is, it is predicted that the patient will have a longer stay. The EDA shows that the patients older in age (81+) stay longer. The type of admission (emergency, trauma, urgent) is also predicted to have an impact because patients that have a lot of trauma, tend to need more care and a longer LOS.

**MODELS**

Our target variable “stay” is divided into 11 unique categories, and our task was to predict the stay class given the features. Therefore, the problem we are dealing with is a classification problem. Hence, we have decided to use classification algorithms such as Decision trees, Random forest, clustering and neural networks.

**Clustering**

Initially, we assumed that due to the categorical nature of data that using a clustering algorithm would improve our ability to model and predict the length of stay. Based on our research using the k-means technique would provide the best clusters for our data. This resulted in 2 distinct clusters with only patient id being the difference between them. However, this is misleading and likely due to the way patient id is captured by the algorithm, as it clusters based on low and high patient ids. We trained the model using the random forest classifier. The accuracy was 77% and the accuracy for each cluster was 78%. However, these results held no value, so we decided not to move forward with this model.

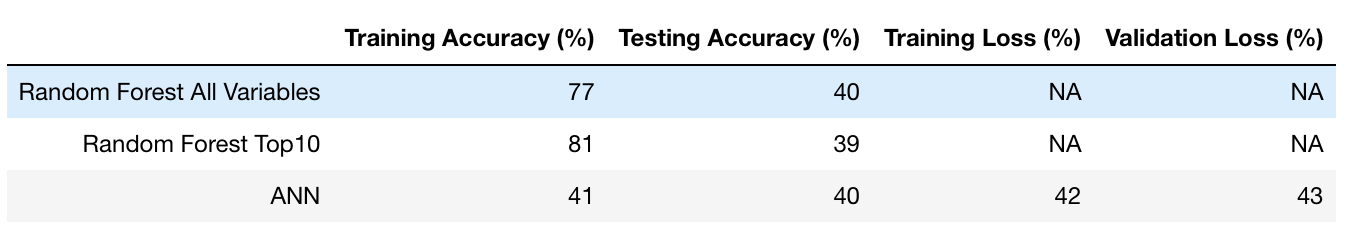
**Neural Network**

Given the classification problem at hand, we decided to fit a neural network to predict the stay variable. Our initial assumption was given the problem at hand: the dataset in question a neural network would be a better model to fulfill the task. With that in mind, we have trained 10 to 12 neural networks and contrary to our belief we discovered that using neural networks did not provide better accuracy or lower loss. In all our models the accuracy scores were around 40% and the loss did not improve much as well. One interesting factor regarding neural networks was that on all occasions they seem to generalize well (accuracy scores and loss scores were identical in training and validation sets).

**Random Forest Classifier**

We decided to use a Random Forest Classifier as well given the Classification problem at hand. Our initial intention with the Random Forest Classifier was to use it to find our feature importance. However, to our surprise, this model turned out to be the best performing out of all and it highlighted features such as deposit, patient id, and number of visitors, which turned out to contribute the most, and the rest of the features fell behind.

With Random Forest models we were able to get training accuracies around 80% and testing accuracies around 40%. We then trained multiple models using the important features and got various accuracy results.



Given the results, we decided to use Random Forest as the main model for predictions.

**CONCLUSION**

Being able to accurately predict the length of stay (LOS) for each patient upon check-in holds importance with or without a pandemic. Although the pandemic has brought to light how important it is to predict a patient's LOS, it has implications in the areas of profits, efficiency, availability, and legal situations. It has been published that increasing a patient's LOS beyond necessity will improve that patient's chances of acquiring a disease or infection. Being able to predict LOS allows the hospital to effectively address patient care and make sure that the hospital has the proper resources available when they are needed. By continuously monitoring patients’ progress, bed/room availability will be easier to accommodate as LOS will shorten. This will then be able to be applied to wings and departments to ensure that they do not get overwhelmed. Lastly, by analyzing and predicting LOS, the overall efficiency of the hospital will improve and prevent the hospital from keeping patients longer than necessary, thus keeping patient costs down and avoiding potential legal situations that could negatively impact the hospital’s public relations, reputation, and financials.