1 December 2019

Project Report

Predicting Patients length of stay in Hospital

Soham Ashtekar

1) Introduction

The most challenging part in managing hospital is managing its resources, and the important factor which affects resources and makes it challenging is the time span for which patient stays in the hospital. If the length of stay of a patient can be better estimated, it will assist in scheduling the usage of wards, management of beds scheduling surgeries, scheduling physician's appointments. With the right quality of data, we can also predict the upcoming admissions.

As per the current literature, the methods used for predicting length of stay are not robust, some of the studies do use the Neural Networks which is going to be one of the main pillars of our study, although the features used to train the algorithm are not clear or seem questionable. There are few to none studies which explain the problem well or have the instructions to recreate the solution or lack some sort of basic analysis.

This study improves upon the existing literature, and it has strong support of practical work with can be represented with some simple steps and can be further improved using different point of views.

A model which can better estimate the actual length of stay using the features we may have on hand while the hospital admissions, we can benefit the HCO's to reduce the workload and improve the quality of healthcare provided. It is always desirable for the hospitals to optimize amount of beds and provide the best healthcare possible while reducing the expense. Clinical models can provide the support needed to make these decisions. By relating the admissions and accounting of the hospital we can create a cost analysis model which can reduce the cost while providing the best healthcare. To maintain the health of the population we need to analyze the available data and create models which predict the upcoming challenges, also solve the existing ones. Able to provide better healthcare using predictions will improve value-based services.

Historically healthcare records used to be non-digitalized, but in last decade we have seen a huge increase in EHR (Electronic Healthcare Records). This revolution helped many machine learning based approaches.

In our study we are using MIMIC III (Medical Information Mart for Intensive Care III)[1] database which is an openly available dataset developed by the MIT Lab for

Computational Physiology. The access to this database is free, upon application with proper information the access will be provided within 2-3 business days.

We used Python to clean the data and do some basic visualizations and then we used Tableau[2] for in depth visualizations. Data cleaning consists of creating some new columns and imputing null values and deleting some rows. Neural network was created using Tensorflow and Keras, we used a sequential model consisting of multiple hidden layers. The network was trained more than 500 times with a validation set.

2) Related Work

There are couple of studies which tried to predict length of stay of a patient, one of which only had a subset of patients from having a certain kind of disease, and other study included some of the columns with may leak the data into the model and can make the accuracy questionable. Some of the examples from our review are stated below.

Hospital records were analyzed in a paper by Maria Kelly [3] to decide if any variables could predict hospital length of stay (LOS) and readmission by linear regression. Their choice of data included combination of National Cancer Registry and Hospital In-Patient Enquiry scheme. Her study concluded that one fourth patients stay in hospital for more than 25 days after colorectal resection. The study also helped in knowing the factors for the higher costs associated with the Longer Length of stay. This research did not include a general model which can predict length of stay by diagnoses or considering procedures which our model can do.

The LOS of patients with cardiac problems was the focus of the paper by Peyman Rezaei Hachesu [4], one of the few to use machine learning techniques, he implemented Artificial Neural Networks (ANN). The data had health records of 4,948 patients who had coronary artery disease (CAD) and the data was retrieved over term of 5 years. There were 36 different attributes in the data and three different machine learning models were used. The best performer was SVM(Support Vector Machines) in terms of accuracy.

One more study which implemented similar techniques on the MIMIC dataset we are using is done by Daniel Cummings[5], He used Admissions, Patients, Diagnoses_ICD, ICUSTAYS tables from the mimic dataset. His choices for algorithms were Random Forest, KNeighbors, Linear Regression, Gradient Boosting, SGD Regressor of which for his dataset Gradient boosting regressor gave the best output. Top 3 features which help getting those results were prenatal, respiratory, injury.

The difference between the approaches above were, the dataset used, different machine learning techniques and some studies were focused on a segment and some were general. Healthcare is a huge field, there can be thousands of different factors and each organization may collect different type of data which can solve different problems.

Our literature indicates that numerous and complex factors can influence Length of Stay and that these factors are still not fully understood and that there is currently a lack of a robust predictive model.

3) Methods

As mentioned above we are going to predict the length of stay of a patient from the time of admission till the time of discharge. We are going to use the MIMIC III database, I particularly will be getting access of MIMIC from googles big query. On a personal computer installation of MIMIC III is not a viable option as Installing complete dataset takes around 40GB of space.

The Database: - To gain the access of MIMIC III you must complete following checklist and follow the guide in given link. You need to complete some mandatory courses, which are required to learn about data privacy and some healthcare terms. Then you need to give a proper description why you want the access and information about your institute and the professor you are working under. https://mimic.physionet.org/gettingstarted/access/ follow steps which are given in this web address.

Data Extraction: - After gaining access to the database we asked for access through googles big query, which is very helpful if you are a individual researcher, Yes! It does limit the amount of data you ca extract but for this study we extracted about 1.8 million rows which the Big Query was able to extract. Big Query saves hassle of installing a huge database on personal computers, and for smaller projects like this it is the best option. We tried multiple combinations of tables and columns and then after lot of testing we took information from five different tables.

The five tables are: -

- Admissions: - The ADMISSIONS table gives information regarding a patient's admission to the hospital. Columns: - **HADM_ID, SUBJECT_ID, ADMITTIME, DISCHTIME, ADMISSION_TYPE, ADMISSION_LOCATION, insurance, MARITIAL_STATUS**.
- Patients: - Identifies single patient's information. Columns: - **gender**
- ICUstays: - Information abut stay in ICU of a patient. Columns: - **los**
- Procedures_ICD: - Contains ICD procedures for patients, most notably ICD-9. Columns: - **ICD9_CODE, SEQ_NUM**
- Diagnoses_ICD: -Contains ICD diagnoses for patients, most notably ICD-9 diagnoses. Columns: - **ICD9_CODE, SEQ_NUM**

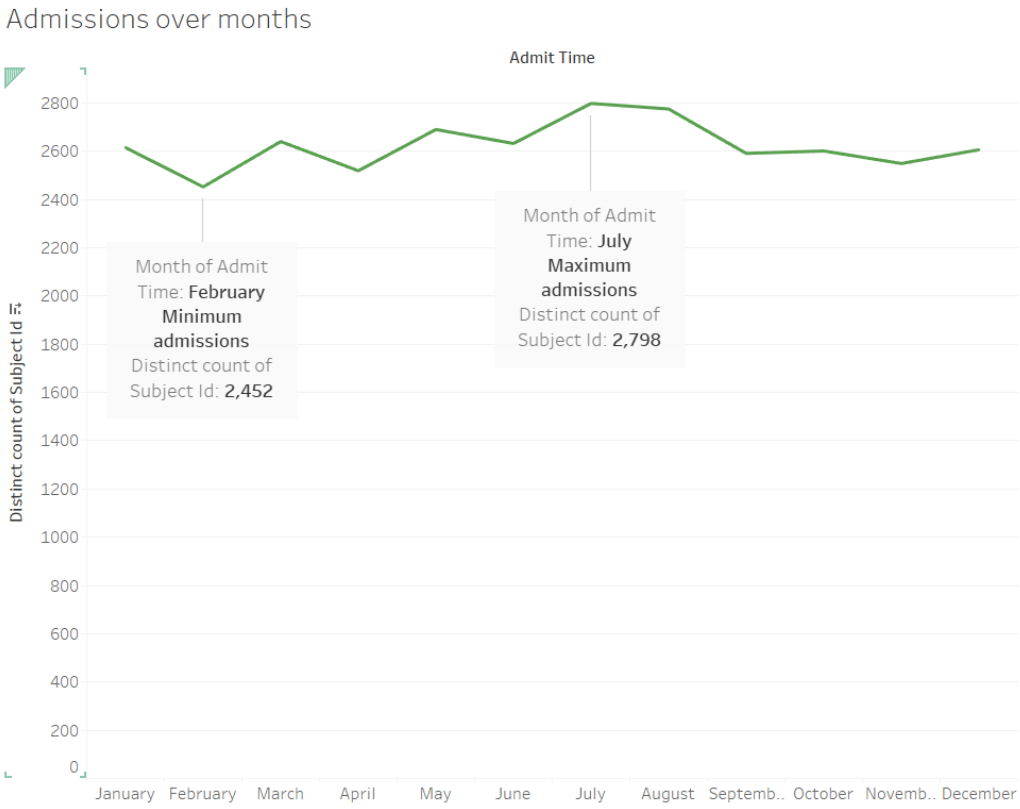
Data Preparation: - After data extraction we did the data cleaning and preparation for making a predictive model.

We did all our data cleaning in python using pandas and NumPy libraries, we did multiple visualizations while cleaning the data to look for minute details. We changed some categories, so they are more understandable, we removed the outliers, so they do not affect the actual model. We combined similar values and much more. The exact code and details about each step will be provided in the code section of the git hub repository.

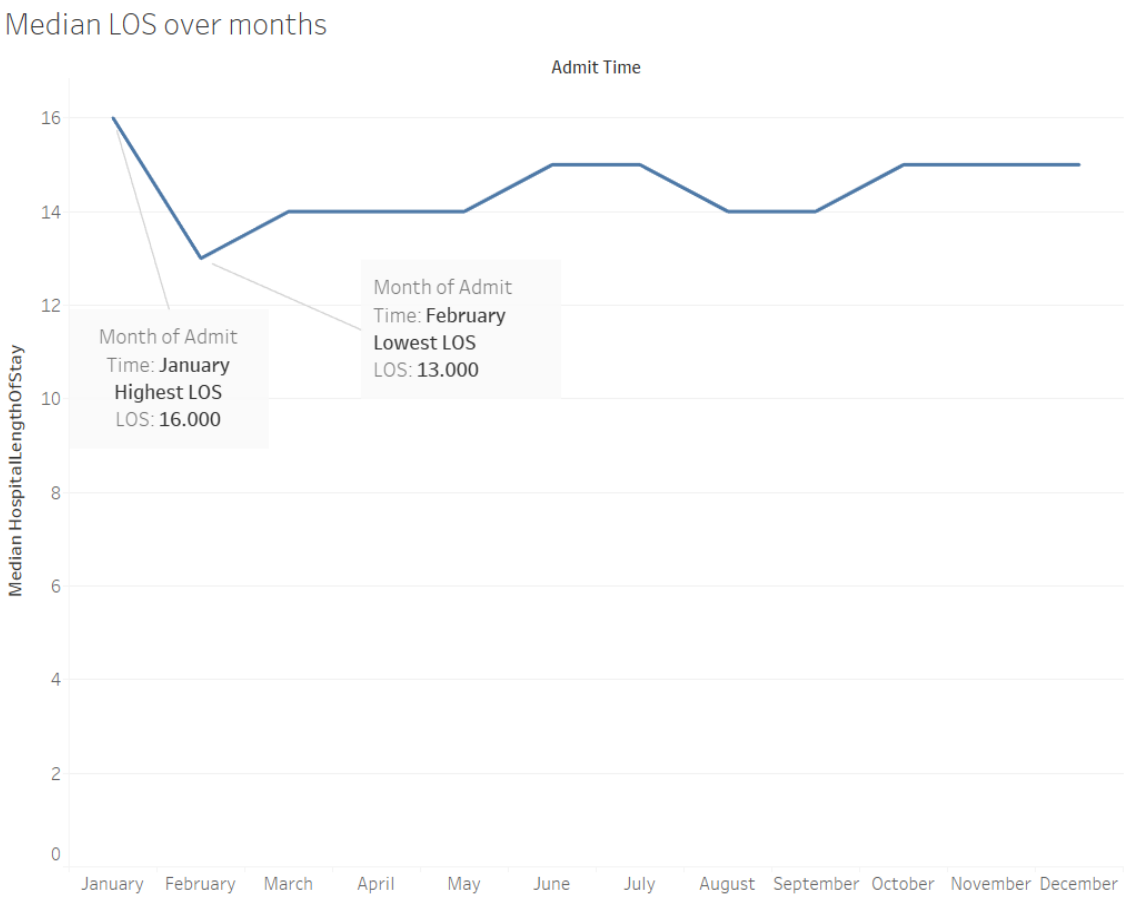
The main takeaway from this is, MIMIC III database is hugely anonymized and it is separated in multiple tables for easy understanding of the data, but it impacts negatively to some people as you have to read description of each of the tables in detail.

Analysis: - Analysis was the important of all, to extract minute details out of the pile of data is a task. For analysis of data we started plotting some tasks and looking through the data using some filters.

First thing we noticed is amount of admissions, we potted admissions over months we can clearly see there are more admissions in month of august and September, when usually the fall or cool weather starts.

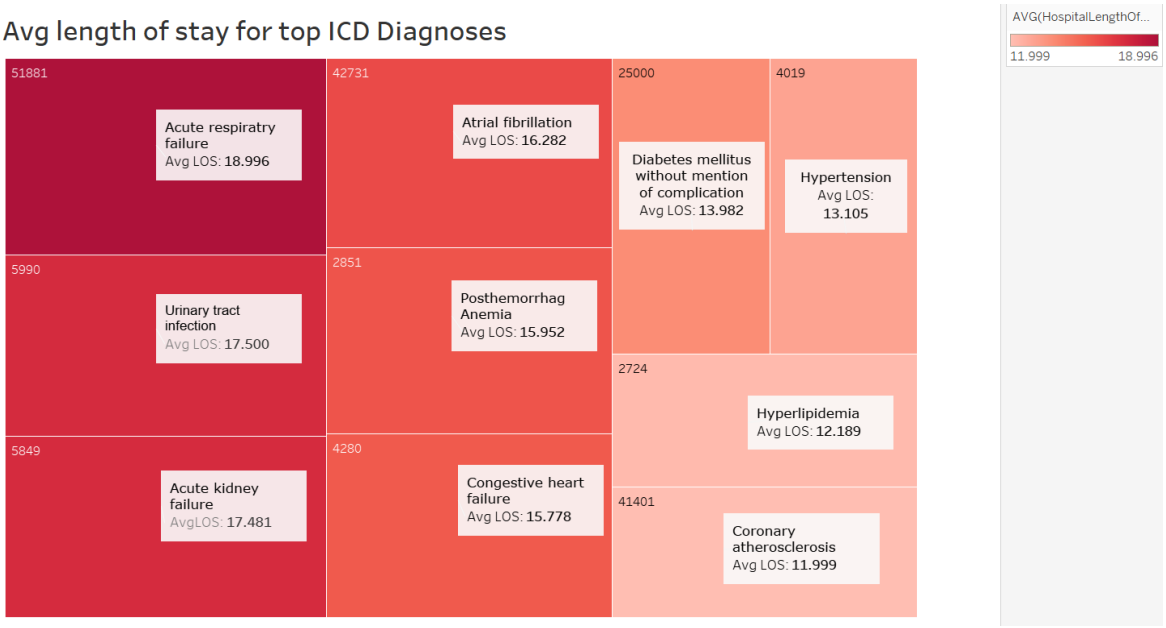


But the most important thing we saw is the Median Length of Stay was not that high for those months, highest LOS was for the month of January, as we can see in the graph below.

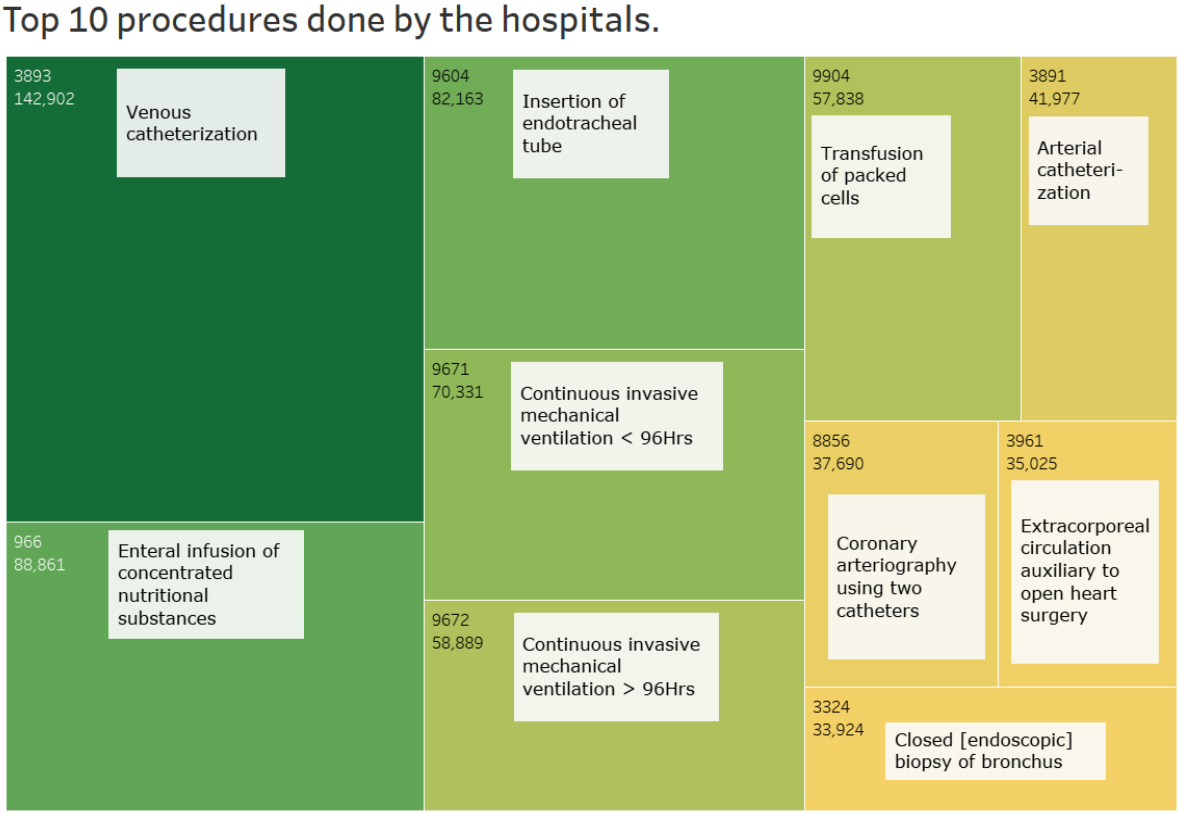


The we checked the Procedures and Diagnoses which occurred the most in the HCO to know the top Procedures and Diagnoses. The number one Diagnoses which takes the longest is Respiratory Failure, there are two types of Respiratory Failures and both happen because body couldn't balance Oxygen and carbon dioxide in arteries.

In the graph below you can see how other diagnoses fair to each other in terms of Length of Stay.



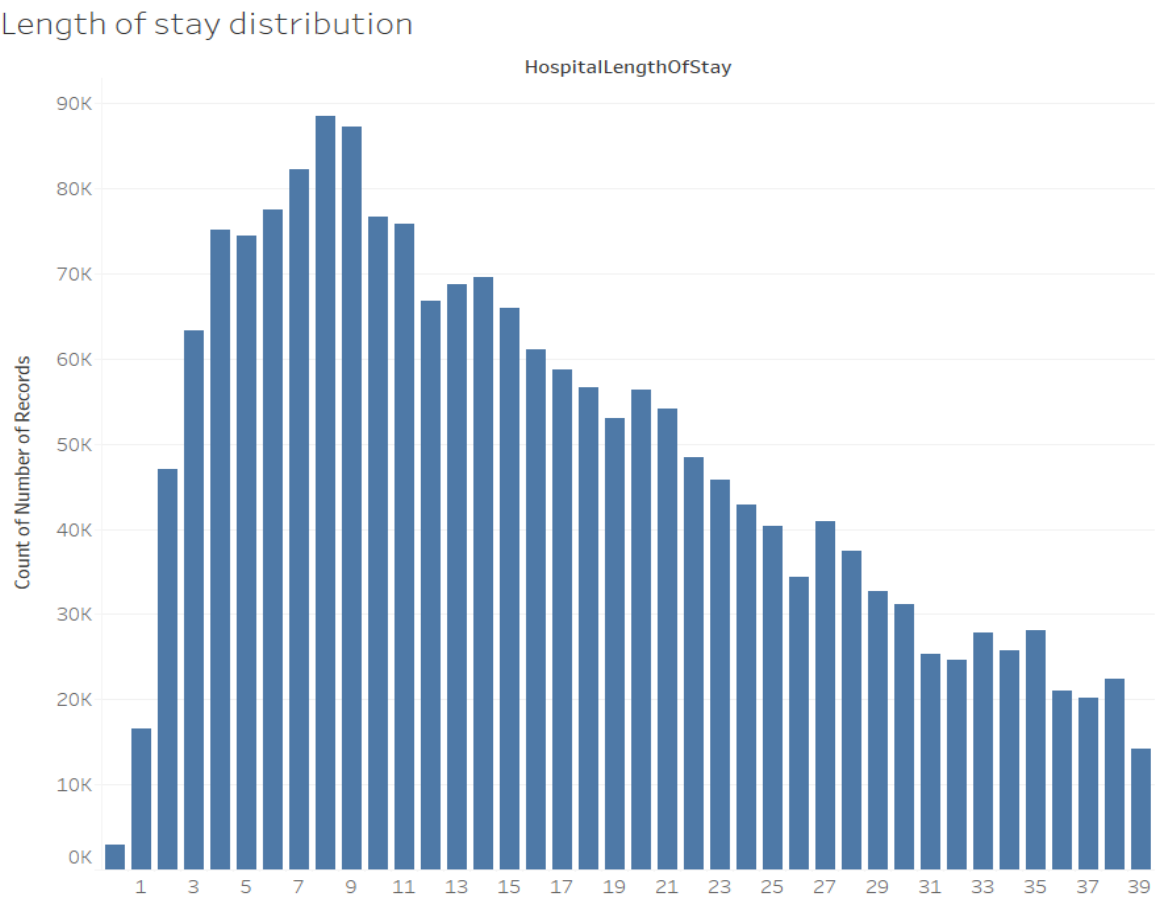
The next is top 10 procedures which we do recurrently on patients.



Now that we have all this information, it is time to create a predictive algorithm, we are going to try multiple algorithms to predict the LOS. We are going to implement predictions by two different methods.

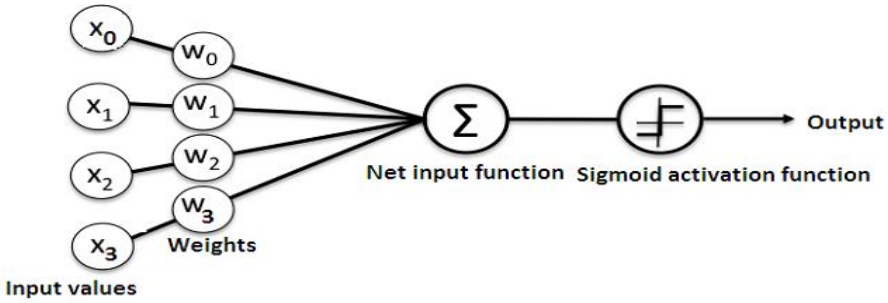
Method 1: - Regression problem, in this we will try to predict LOS using regression algorithms such as KNeighborsRegressor, DecisionTreeRegressor, SGDRegressor, RandomforestRegressor and lastly Neural Networks. Regression means we will be predicting LOS with a exact number lets say 1.5 days. Which means this approach will be tougher than any classification approach.

Method 2: - As mentioned above that classification approach is easier and will give you a output in category like length of stay from 0 to 2 days, the reason we are trying to use both the methods is to pin point target and also see an estimate which can be more reliable. In the healthcare industry estimating length of stay with a simple figure is though, because LOS is affected by 100 different measures even in the same exact HCO. Estimating length of stay in categories like Short LOS, Average LOS, Long LOS and Very Long LOS. We can categorize this as 0 to 3 days for short, 3 to 8 for average and 8 to 18 for long and 18 and beyond as very long. We created this categories by looking at the distribution of our data.

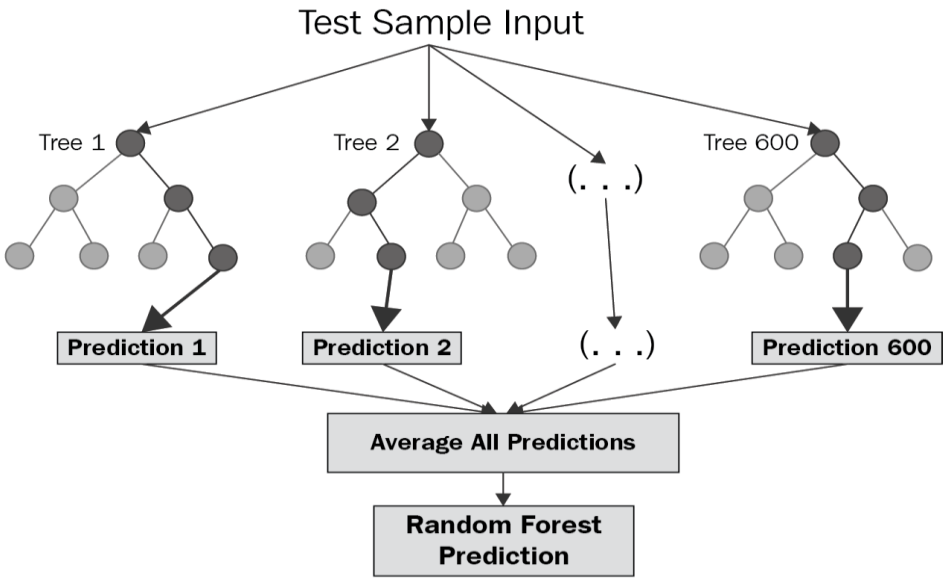


Background: - Lets discuss two of the most powerful machine learning algorithms and how they work, also why they are useful.

- **Neural Networks**: - Neural network is a collection of algorithms that attempt to understand the underlying relationships in a data set through a mechanism that imitates how the human brain works. Neural networks are very powerful when crunching the tabular (Structured data) but the most important part is, It works much better when working with unstructured data such as images and textual data. Below we have a simple neural network. [6]



- **Random Forest**: - Random forest is a type of Decision Tree which is created by taking the useful part of multiple decision trees, It splits an observation into a yes/ no question and that split is called a node, there can be thousands of different nodes which pass decision forward and give the required output. A simple Random forest looks like[7]



4) Results

Our goal was to predict Length of stay with acceptable accuracy, in healthcare the acceptable accuracy is around 88-93%. After training the model and doing the validation we performed as expected and the average accuracy for both the methods was above 90%.

With Random Forests we constantly got accuracies over 97%, but Random Forests tends to over fit the data and they return very high accuracy, so with our neural networks we got accuracy score of around 90% which is reliable enough considering the variance in data. We started with 20% accuracy for our first neural net and we were able to improve it to around 90%.

5) Discussion

If we compare our results with other studies previously done we can clearly see that we have more robust results. Compared to studies which focused Length of stay of patients with diseases we have more general model which only needs to know diagnoses codes and their sequence.

What we did differently here is we heavily invested in tasks like data cleaning and pre-processing which lead us to better results. We also invested a lot of time in training and tuning neural network.

What we could have done is, dig deeper into dataset and encode each sequence as a step and each icd code to know more about diagnoses. So in the result when we pass diagnoses sequence and id's to the model it will output predicted length of stay.

6) Conclusion

To conclude we can say that our goal was achieved, but we can further improve this study by doing more analysis of data such as Combining sequences and diagnoses so we can do multiple predictions and improve the healthcare provided.

We can take some ideas from other researchers and improve the data gathering process which will lead to quality results, as we cannot create quality results with garbage data.

7) References

[1] https://mimic.physionet.org/

[2] https://www.tableau.com/

[3] https://www.researchgate.net/profile/Maria_Kelly3

[4] https://scholar.google.com/citations?user=HainsJMAAAAJ&hl=en

[5] https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598

[6] https://www.kdnuggets.com/2018/10/simple-neural-network-python.html
[7] https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f

8) Appendices

All the data, ipynb and requirements are given in this link
https://drive.google.com/drive/folders/1dqQmMjhdAZ3eRQ-sqdPRKFbteIYxtfyY?usp=sharing