Columbia University

IEOR E4999 - Summer 2018 Fieldwork

Internship Report

Capacity Planning Model

Students:
Sélim Amrouni
Thibault Duplay
Yixuan (Erin) Ge

Supervisors:
Martha Bailey
Pr. Ali Hirsa

September 4, 2018

Abstract

We spent 12 weeks at New York University Langone Medical Center for our summer internship. We carried out a project under the supervision of Martha Bailey, the Senior Director in the Comprehensive Transfer Center and Pr. Ali Hirsa, our Deep Learning professor.

The subject of our internship focused on Capacity Planning Model, predicting the flow of patients and optimizing them. This model is widely adopted by health-care institutes and is beneficial for hospital operations. The goal of our model is to predict the bed availability within the next 72 hours for the Tisch hospital, Kimmel Pavillon, and Hassenfeld Children's Hospital.

The first part of our internship was to get familiar with how the hospital works and study the data. We did a literature review on the hospital's patient flow and Length of Stay prediction, during which we understood what is the hospital's operation process. Our supervisor gave us a tour of the hospital, which helped us understand NYU hospital's mechanism. We have been in contact with the data providers, who extract ".csv" files from databases for us. We discussed with them about the data fields and values.

The second part of our internship was to build the model. We divided the project into three parts:

- The prediction of the Length of Stay for the patients currently in the hospital
- The reading of the schedule for the patients going to the Operating Room and the simulation of their Length of Stay
- The simulation of the patients arriving at the Emergency Room and their Length of Stay

We tried to build a probabilistic model more attractive mathematically. This model mainly concerns the prediction for the Length of Stay of the current patients. It did not work well because of the lack of data.

Following the suggestion of the Comprehensive Transfer Center team, we used the expected discharge time given by physicians for now, and we will try the probabilistic model again in the fall, once we get access to the databases.

We got the statistics of the emergencies for each unit from the data provided by the Discharge Department of the hospital, and we sampled unscheduled arrivals from it. We also simulated their Length of Stay from the historical data.

We also simulated the Length of Stay of the scheduled patients using the historical data. The hospital provided us with a patient placement scheme and the schedule of the upcoming surgeries. From this information, we had to assign the scheduled patients to their nursing unit after their surgeries. We needed to use some natural language processing methods to decode the patient placement scheme that was written in raw text. The placement is often updated so the reading has to be robust to the changes.

The third part of our internship was to build a Graphical User Interface (GUI). We used the Python package Tkinter to create the GUI, which allowed users to update the databases and get nursing units occupancy status displayed in a few clicks. Then, we created a ".bat" file, which allowed users to interact with our model without the installation of Python.

Beyond what we learned from a technical point of view, this internship enabled us to learn more about the health-care industry and experience how it feels to work as a data scientist. It peaked a desire to pursue a possible career in data science after graduation.

Contents

1 Introduction

1.1 Internship presentation

We spent 12 weeks at New York University (NYU) Langone Medical Center (LMC) in order to complete our Summer 2018 Curricular Practical Training. Capacity Planning, the prediction, and optimization of bed occupancy in a hospital is a popular field. Its goal is to enhance the operation process and improve efficiency in the health-care industry. Several hospitals, such as Cincinnati Children's Hospital and Yale New Haven Hospital, have already improved their operation system thanks to the Capacity Planning.

As a hospital in a major city of the world, New York, NYULMC experienced several shortage of beds (saturation of the bed availability). These situations are extremely annoying that is why predicting in advance the bed occupancy and then using these predictions to optimize the patient placement is an interesting and useful problem. We wish to perform a mathematical model able to predict the bed demand, therefore, the operation team would be able to optimize and adjust the Operating Room (OR) schedule in order to maximize the bed availability.

1.2 People

We were supervised by Martha Bailey, she is the Senior Director of Comprehensive Transfer Center (CTC) at NYULMC and Pr. Ali Hirsa our Professor of Deep Learning at Columbia University. This is a very interesting management scheme. Indeed, Martha Bailey and Pr. Ali Hirsa have complementary skills and areas of interest. Martha Bailey has a highly valuable knowledge of the hospital operations while Pr. Ali Hirsa brings his solid background in Applied Mathematics. However, this management scheme always forced us to adapt our language to our targeted audience. This was heavily instructive for us: as future scientific workers, communication and adaptability is a key skill to navigate in the professional ecosystem.

More generally, we also had the opportunity to communicate with almost the whole CTC Team and data providers of the hospital. In the first two weeks, we worked closely with the operation team and other data providers to get familiar with how the hospital works and study its structure in order to model it. We are very grateful to Martha Bailey, Jennifer Kessler, and Gina Salovic, who helped us get familiar with the hospital and set up several

meetings with data providers from other departments as well as a complete visit of the hospital. We would also like to thank Ricardo Ringor and Michael Wuchovich, who provided a lot of insights of the datasets.

1.3 Goal

Our goal has been to predict the bed availability in the next 72 hours (for each day). The model consists essentially of predicting the Length of Stay (LoS) of the hospital patients. The model must be re-calibrated as the data arrives: we do not know what will happen today, but tomorrow we will know everything about what happened, so the model has to constantly re-fit its prediction with the updated data.

2 Our contribution

Located at the heart of the largest city in the United States, NYU Langone Health experiences bed shortage periods all the time. Driven by the duty of improving patient care operations and motivated by the need to reduce shortage periods, NYULMC hired us to provide a short-term bed availability tool directly usable by the operational teams. Knowing the bed status in advance, the management team is able to forecast shortage periods, increase the efficiency of the hospital, accommodate more patients and deliver better services. It is very important for the operational team to get a precise prediction of bed status, so they can arrange inpatients (patients that occupy a bed for at least one night) and operations accordingly in order to avoid overcrowded Emergency Room and patients stuck in PACU (Post-Anesthesia Care Unit - where the patients recover from their surgeries before going to their bed) due to the shortage of nursing units beds. The CTC team was used to gather and interpret all the information manually on an irregular basis without the latest data entries. Indeed, each week, a nurse is assigned during six hours to create this manual occupancy report. We succeeded to automate the processing of all the hospital data in order to provide the operational team with the most up-to-date forecasting. Currently, the operational team is able to use a finished product. The user interface we built gives a good summary of the hospital's occupancy status and is employable even by non-technical professionals. Also, thanks to the product, now the nurse can be assigned to care patients. All the work is done by simply clicking a few

buttons.

We are glad the hospital is interested to pursue the work this Fall semester. The phase one will focus on improving the current forecasting model. In a second phase, we hope to work on the optimization of the scheduled surgeries in order to maximize to the efficiency of the hospital.

3 Typical day

We worked full-time this summer at NYU Langone Tisch Campus. We were physically located in a dedicated office within NYU Langone premises. Usually, we started at 9:00 AM and we left at 5:00 PM. We also worked on the Columbia campus where we met with Pr. Hirsa to talk about mathematical modeling and discuss the next steps.

During the first two weeks, at NYU Langone campus, we interacted with the hospital's staff frequently. Usually, we studied the datasets and prepared our questions in the morning. We got in touch with the data providers at noon and processed what we learned from the meeting in the afternoon. In the meetings with the data providers, they explained to us what each data field represents and the meaning of the values, especially for categorical data. These first two weeks were really important because we learned a lot about the hospital operations and data management. The providers explained to us how a procedure is scheduled and the routine of a procedure. We also met with physicians to improve our domain knowledge. We daily checked-in with our supervisor, Martha Bailey, to talk about our progress. As the hospital's data is stored in several different databases, we needed to talk to several professionals to gather all the data sources. The operation team also gave us a tour of the hospital. We visited all of their nursing units (NU), the Emergency Department and the operation room. This tour gave us deeper understanding on how the hospital works and why capacity management is needed. During the following 10 weeks, we worked more independently on the program. Usually, we scheduled weekly meetings with Martha Bailey and Jennifer Kessler to present our current status and our plan for the following week. We also met several times Professor Hirsa at Columbia. We would talk about technical difficulties such as the optimization of hyper-parameters, unbalanced samples, etc. We would go over our work first and ask for Pr. Hirsa's insights. Then we discussed our plan for the following week.

4 Personal benefits of the internship

We are three students interested in Applied Mathematics and wanted to find an internship to develop hands-on skills in Analytics.

- Sélim has a background in Electrical Engineering and Energy Economics, he graduated from the French "Grande Ecole" CentraleSupelec and the Institut Français du Pétrole. After one year (July 2016 - July 2017) working as a Deputy Project Manager at Engie, he discovered that he was really interested in Artificial Intelligence. He decided to pursue his studies, after multiple application in several universities he chose the MSOR program of study at Columbia University.

- Thibault has a background in Applied Mathematics, he studied until June 2017 at the French "Grande Ecole" ENSAE ParisTech. He has not graduated yet from this institution, the MSOR program at Columbia University is replacing his last year at ENSAE Paristech. He needs to complete his program at Columbia University to graduate from both institutions at the same time.

- Erin has a background in Finance, she graduated from the University of International Business and Economics in China. She developed an interest in Quantitative Finance during her undergrad and decided to pursue a Master's degree in the field. She chose the MSOR program at Columbia University after applying for multiple programs.

This project has been a great immersion in the corporate world. We learned how to process real-world messy data, which is a real need in the industry. We had to model our ideas, and the ones we took from scientific publications. We faced real problems in the industry such as communicating on our work with a whole team containing non-technical employees. We provided a product accessible to everybody in the CTC team. We also had to contact several different services and teams to aggregate as much data as possible. This experience will extremely facilitate our journey in the Data-Science world.

- By the intermediate of this internship, Sélim realized how much Data Science can be leveraged in the health-care domain. He is passionate about the use of machine learning tools and wants to be involved in

the numeric revolution of the medicine. This Fall, aside from his part-time Research Engineer position at NYU Langone, he will work as a Volunteer Research Assistant in the Tatonetti Lab at Columbia University New York-Presbyterian Hospital in Biomedical Informatics. Sélim hopes to secure a full-time position for January 2019 in the Health Industry or pursue his academic path through a Ph.D. in Biostatistics or Bioinformatics.

- Thibault is really interested in Applied Mathematics in general. He loves the fields associated with Statistics, Theory of the Probability and Optimization. That is why the Artificial Intelligence is a perfect fit for his tastes. This internship has helped him to secure a Full Time in Deep Learning starting next January in Los Angeles. He will continue to work in the Health industry. The start-up that hired him as a Research Engineer is applying Deep Learning methods to the Pharmaceutical area.

- Erin wants to pursue a career in Quantitative Finance after graduation. From this internship, she enhanced her skills in cleaning, processing raw data, and Analytics. She feels that this experience will be very valuable for her future career development.

5 Overview of the Hospital

5.1 NYU Langone

"NYU Langone Health is one of the nation's premier academic medical centers. The trifold mission to serve, teach, and discover is achieved daily through an integrated academic culture devoted to excellence in patient care, education, and research. Since 1841, NYU Langone has taken care of the NY and world-wild population and is currently dispatching its patients among 1,069 beds. Located in the heart of Manhattan, with additional facilities throughout the New York City area (200 facility locations), NYU Langone consists of six inpatient locations:

- *Tisch Hospital, the flagship acute-care facility*

- *Kimmel Pavilion, the newly opened (06/24/2018), state-of-the-art healthcare facility Rusk Rehabilitation, ranked as one of the top 10 rehabilitation programs in the country by U.S. News & World Report since 1989, and recently awarded a three-year accreditation from the Commission on Accreditation of Rehabilitation Facilities (CARF)*

- *NYU Langone Orthopedic Hospital, formerly known as Hospital for Joint Diseases, a dedicated inpatient orthopedic hospital*

- *Hassenfeld Children's Hospital at NYU Langone, which provides pediatric inpatient care, outpatient care, procedural and surgical services, the KiDS Emergency Department, and multiple ambulatory services*

- *NYU Langone Hospital—Brooklyn, formerly known as NYU Lutheran Medical Center, a full-service teaching hospital and Level 1 Trauma Center located in Sunset Park, Brooklyn"*

Source: `https://nyulangone.org/our-story`

5.2 Comprehensive Transfer Center

The Comprehensive Transfer Center (CTC) is a dozen-members-team which manages all patient-flow activities throughout NYU Langone Health. It manages transfers into and within NYU Langone Health in collaboration with leadership and staff as well as interdisciplinary teams. The CTC is often compared to the "Control Tower" of the hospital. The team is doing on a daily basis an amazing job to ensure the perfect management of all the bed's facilities in NYU. The team is led by Martha Bailey who manages roughly 10 operators dispatching the patients within NYU facilities.

5.3 Hospital units

As seen before, NYU Langone owns 200 locations and 6 of them care for inpatients. The scope of our project is limited to 3 of the inpatient units: Tisch Hospital, Kimmel Pavilion & Hassenfeld Children's Hospital. Before going further, we first need to define a nursing unit (NU). A NU is a physical subdivision of the hospital. Each NU corresponds to a nurse desk.

In the scope of the project we will consider only the following locations:

5.3.1 Tisch Hospital

Tisch Hospital is NYU's flagship for inpatient care. However, please the reader can note that all the NU are not in-scope. Indeed, our project scope does not include the Obstetrics, Rehab, and Psychiatric. Finally, 138 beds spread between 6 NU belong to our project scope in the Tisch Hospital.

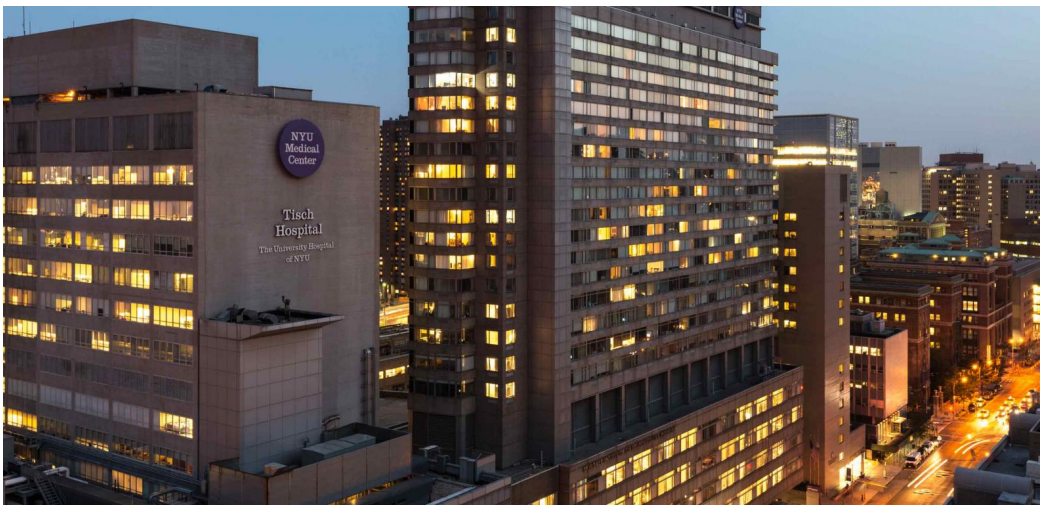


Figure 1: Tisch Hospital

Picture: `https://med.nyu.edu/research/office-science-research/`

5.3.2 Kimmel Pavillon

Kimmel Pavillon is the newly opened location of NYU Langone. It is a brand new hospital with the latest technologies such as robotic nursery tools. The scope of the project incorporates all the 311 beds and 9 NU of this location.

5.3.3 Hassenfeld Children's Hospital

This location is physically situated in the same building as the Kimmel Pavillon. The scope covers all the 113 beds and 6 NU of the location.

5.4 Summary

All the other rooms of the hospital are considered out of scope. The 51 operating rooms (where the patients have their surgeries) and the PACU are also out of scope.

The project scope counts around 562 beds distributed in 21 NU, each NU is also subdivided in 3 according to the different level of care: (from lowest to highest) Acute, Step Down & Intensive Care. The patients can, therefore, be allocated according to their service and the level of care they need. Usually, there are three to four NUs available, the placement is aiming to increase the operation efficiency of the hospital.

6 Hospital patients flow

In order to predict bed demand, it is important to understand the hospital patients flow. First of all, we need to create a simple scheme to understand in which state a certain patient can be. This scheme can then be easily modeled with Mathematics and be adjusted with the historical data.

The next figure illustrates the simplified mechanism of the hospital we chose for our study. The real diagram is more complicated but the changes that are induced are negligible in the scope of our project.

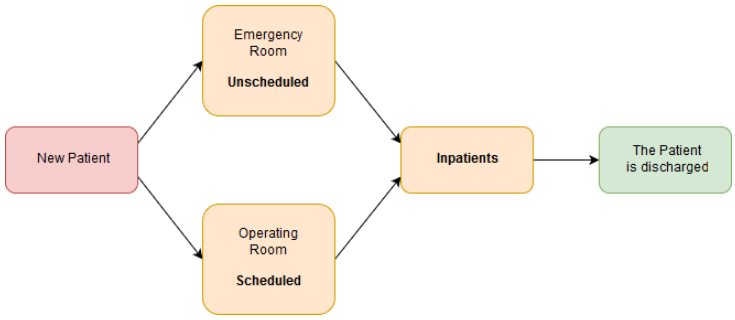


Figure 2: Diagram of the patients flow

There are two kinds of patients at the entrance: the scheduled (the ones that have an appointment at the hospital, mainly surgeries) and the unscheduled (mainly the emergencies). After being seen by the employees of the hospital, these patients can occupy or not a hospital bed and become

inpatients until being discharged. Since the goal of the project is to predict the bed availability, the patients that do not occupy a bed are not taken into account.

In order to estimate the future bed availability, we need to separate the following cases:

- Current inpatients: We have access to a current snapshot of the inpatients, based on this data we wish to assign a particular Length of Stay (LoS) to each of the inpatients.
- Upcoming scheduled: We have access to the schedule of the next surgeries. We have some information on these "scheduled" patients. Based on this schedule, we will try to adjust the number of incoming scheduled inpatients and assign them a LoS.
- Upcoming unscheduled: By nature these patients are random. Based on historical data we will try to sample the upcoming arrivals in the Emergency Room. With these simulations, we wish to adjust the number of incoming unscheduled inpatients and assign them a LoS.

We have access to the data in real time. Basically being today at noon we cannot deterministically know the state of today at 4:00 PM, but later we will have access to this information. For this reason, the model has to be re-calibrated in real time considering the freshest information.

7 Data

7.1 Hospital data management

NYU Langone does not have a centralized data management team. Understanding and gathering all the data sources was a real challenge for us. We had to speak to a lot of different teams in order to get, interpret and link the data. It was a true chance for us, indeed, we learned a lot about the preprocessing of real-world data and had the opportunity to strengthen our skills in this particular field. Due to confidentiality concern, the hospital was reluctant to provide us direct access to databases. That was our main slowdown during the internship. Indeed, we had to contact the data providers to extract data for us on a regular basis.

7.2 Data used for the project

Among all the provided datasets, four sources have been used:

- Current Census: It is a snapshot of all the inpatients in the hospital at the moment it has been pulled.
- Discharges: This is a dataset of the 2-last-months discharges, it is useful to make statistics on the historical data.
- OR Schedule: It is a schedule where we can access upcoming surgeries within 15 days, we have access to the intensity of the care, the type of medicine, and the hours of the surgeries. The NU is not assigned yet.
- Placement Matrix: This is a matrix often updated by the CTC team, associating particular a type of medicine and a level of care to a NU.

Other datasets were available (not detailed in this report) but did not bring additional information.

7.3 Requested data for future work

Given the difficult access to databases, it was a huge difficulty to pull and gather the data. However, thanks to the CTC team, we hope this Fall to get a full direct access. Then, we hope we will be able to use biomedical data and medical orders (future care ordered by the physician) to improve.

Moreover, the biggest issue with the data we had is the overwriting of past data with last entries. When we pulled the data from the Current Census, every entry of a given patient was overwritten with the last entry of this given patient. We finally came up with a solution in agreement with the I.T. team: four times a day, a snapshot view of the Current Census is stored. Then, we hope this fall being able to deliver a meaningful analytical solution for better prediction of current inpatients length of stay.

8 Model

8.1 General explanation

8.1.1 Input

The inputs are the four datasets listed in the previous section.

8.1.2 Output

The outputs are little tables for each nursing unit separated into three different sources: "scheduled", "unscheduled" and "current inpatients" that can be easily aggregated. The columns of the tables separate six-time slots. Each time slot represents 12 hours (principally for morning and afternoon). The timeline is starting today at 0:00 AM and is ending after tomorrow at 12:00 PM. The rows of the tables separate the admissions, the discharges and the total number of patients in the NU (coming from the considered source) for each time slot.

NU: xxx , Source: Unscheduled						
	0-12 h	12-24 h	24-36 h	36-48 h	48-60 h	60-72 h
Admissions	13	25	15	28	6	17
Discharges	0	0	4	11	23	15
Total	13	38	49	66	49	51

Table 1: Example of an output for the unit xxx, sourced with the unscheduled patients (the name of the unit is masked for non-disclosure agreement)

8.1.3 Use of the three models

In order to have a prediction of the next 72 hours for a particular NU ignoring the source, we just have to sum the tables from the three sources. If we want the prediction for a complete location (Tisch Hospital, Kimmel Pavillon or Hassenfeld Children's Hospital) we just have to sum the tables of the corresponding NU.

8.2 Inpatients Model

8.2.1 Preliminary

Inpatients are already into NYU Langone premises. As they are already admitted, the unique goal is to predict their discharge date. Then, by summing the discharge within each NU, we can retrieve the number of patients within the NU.

Two approaches can be implemented to predict the discharges of the patients. First, we can try to forecast the patient LoS directly at the moment of the admission (example: 3 days, 4 days...). However, the discharge of the patient is very unpredictable and the estimated discharge date changes several times during the patient care process. Indeed, the patient's body reactions are uncertain and nobody knows exactly how long the patient will stay, what will be the future treatments, etc... So, a second approach is to predict every day what is the remaining LoS of the patient. This is the approach we adopted.

Predicting the discharge date can be seen both as a regression or a classification problem.

In the regression problem, the goal is to predict the right amount of time remaining in the hospital. However, due to the granularity required by the CTC team, we opted for a classification approach. Indeed, for an operational point of view, the importance is to know every day the occupancy of the hospital but they do not care about the exact time of the discharge. After several reviews with the team, in addition to the day, we realized that an important information to know for them is if the discharge will happen before noon (12 PM). Indeed, knowing if a bed is a Discharge Before Noon (DBN) enables the team to directly re-attributes this bed during the afternoon. Then, there is no under-occupancy of the bed which will host of the patient during the night. That is why we turned the regression problem to a classification one. We split the time in seven twelve-hours slots going from "0-12h" to "+72h". For example, if a patient belongs to the class "24-36h", it means his remaining LoS is between 24 and 36 hours. So, the problem of predicting the number of inpatients in the hospital is turned into a supervised seven classes classification problem.

8.2.2 Features used

The dataset used for this task is the Current Census. It is the snapshot view of the hospital with about two hundreds of features. However, as explained before, due to data leak of future data in the past data (the past data of a patient are overwritten by the last entries), we were not able to use these features. Hopefully, after several meetings with the I.T. team and thanks to their amazing work, we started the capture and storage of this Current Census dataset four times a day. Thus, we hope we will be able to try and test our methodology this Fall.

At first glance, because we don't have access to historical data, we were only able to use one feature for our inpatients model: the estimated discharge data given by the physician. Indeed, the physician during its daily patient tour updates the estimated discharge date of the patient.

8.2.3 Processing of the data

Given the estimated discharge date by the physician and the current time, we can compute the remaining Length of Stay of the patient.

Then, knowing the remaining amount of time the patient will stay, we can attribute each patient to the category it belongs.

The last step is counting, within each NU, the number of patients belonging to each category in order to know the number of discharges.

8.2.4 Results

This scatter plot of the Actual Remaining LoS vs. Expected Remaining LoS demonstrates the inefficiency of the physician prediction. Based on this observation, we tried to develop a supervised learning classification algorithm.

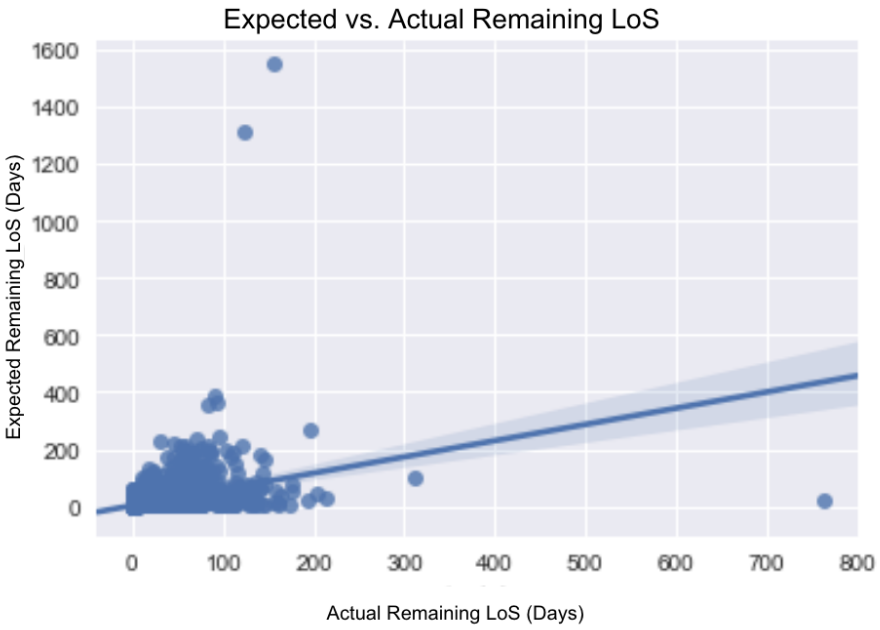
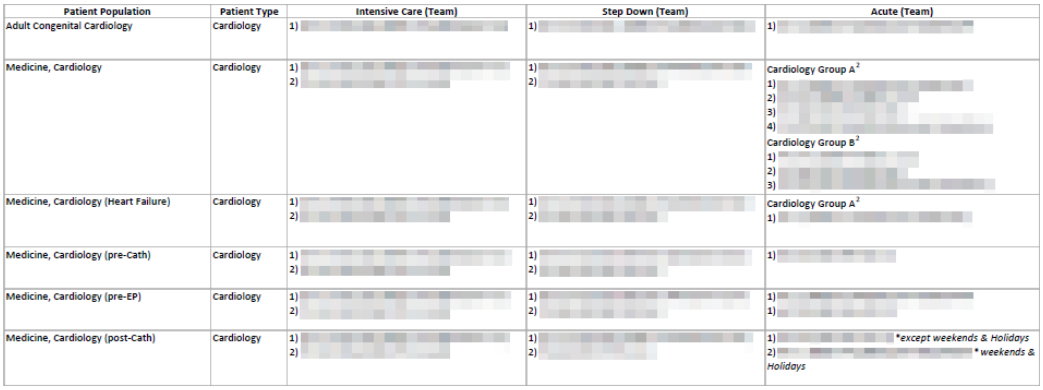


Figure 3: Scatter Plot of the Actual vs. Expected Remaining LoS

8.3 Scheduled Model

8.3.1 Placement Matrix

The Placement Matrix assigns a type of medicine and a level of care to a NU.



Patient Population	Patient Type	Intensive Care (Team)	Step Down (Team)	Acute (Team)
Adult Congenital Cardiology	Cardiology	1)	1)	1)
Medicine, Cardiology	Cardiology	1) 2)	1) 2)	Cardiology Group A[2] 1) 2) 3) 4) Cardiology Group B[2] 1) 2) 3)
Medicine, Cardiology (Heart Failure)	Cardiology	1) 2)	1) 2)	Cardiology Group A[2] 1)
Medicine, Cardiology (pre-Cath)	Cardiology	1) 2)	1) 2)	1)
Medicine, Cardiology (pre-EP)	Cardiology	1) 2)	1) 2)	1) 1)
Medicine, Cardiology (post-Cath)	Cardiology	1) 2)	1) 2)	1) *except weekends & Holidays 2) * weekends & Holidays

Figure 4: First rows of the Placement Matrix (details have been blurred for non-disclosure agreement)

We can see that in some cases there are listings of NU (1), 2) . . .). This describes an order of NU filling, if "1)" is full then put the patient in "2)" and so on. We can also observe that sometimes depending on the team of surgeons or the type of the day (weekdays, weekends or holidays) the placement can change.

The format is not thought to be used by machines but by the human employees. We had to interpret raw texts. Sometimes the names of the NU differ a bit for the same NU. We needed to find a method that dealt with these difficulties, we could not code a simple mapping for that particular Placement Matrix. Indeed the Placement Matrix is often changing, we needed to perform an algorithm robust to these changes.

We decided to make a list of unchanged names of NU, then using a word distance we associated the NU suggested by the Placement Matrix to the one that minimizes the distance in our list.

For this purpose, we used the following distance: Let $ch1$ and $ch2$ be two chains of characters. The distance between $ch1$ and $ch2$ noted $D(ch1, ch2)$ is the lowest number of elementary operations on $ch1$ needed to transform it into $ch2$. There are three elementary operations:

- Interchanging two characters (1) (ex: dog $\rightarrow$ god)
- Inserting a character (2) (ex: god $\rightarrow$ gold)

- Removing a character (3) (ex: gold → old)

For example $D(\text{“dog”}, \text{“glove”}) = 5$.

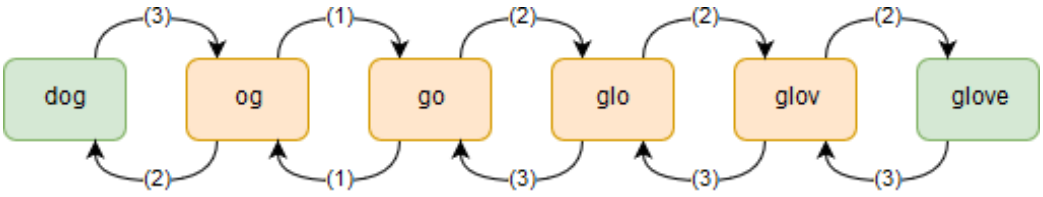


Figure 5: How to calculate the distance between “dog” and “glove”

This distance has been inspired from the famous Levenshtein Distance. We added the operation “interchanging”. In our case, the chains of characters are in fact sentences. The interchanging allows flexibility if words are similar but not in the same order in the two sentences.

We can notice that the distance verifies the widest mathematical definition of a distance:

- $D(ch1, ch2) = 0 \leftrightarrow ch1 = ch2$
- $D(ch1, ch2) = D(ch2, ch1)$
- $D(ch1, ch3) \leq D(ch1, ch2) + D(ch2, ch3)$

The results were really satisfying, none of the NU names were wrongly mapped.

8.3.2 Admissions

The admission was straightforward once we have been able to handle the Placement Matrix correctly. We chose the hypothesis that the patient enters his NU in the time slot following the one of his surgery. We did this choice in order to take into account the length of the surgery and the time spent in the PACU. For example, if a patient is scheduled for tomorrow afternoon, therefore in the slot “36-48h”, we will assume that he will arrive at his NU in the time slot “48-60h”.

8.3.3 Discharges

Since we do not have yet any other information on the patient (he is not in the hospital yet), we cannot affect him with a satisfactory accuracy a LoS. According to his NU, we can sample its LoS based on the dataset of the discharges. This dataset allows making some analysis on the historical data. The next figure gives us the average LoS of the patients for each unit.

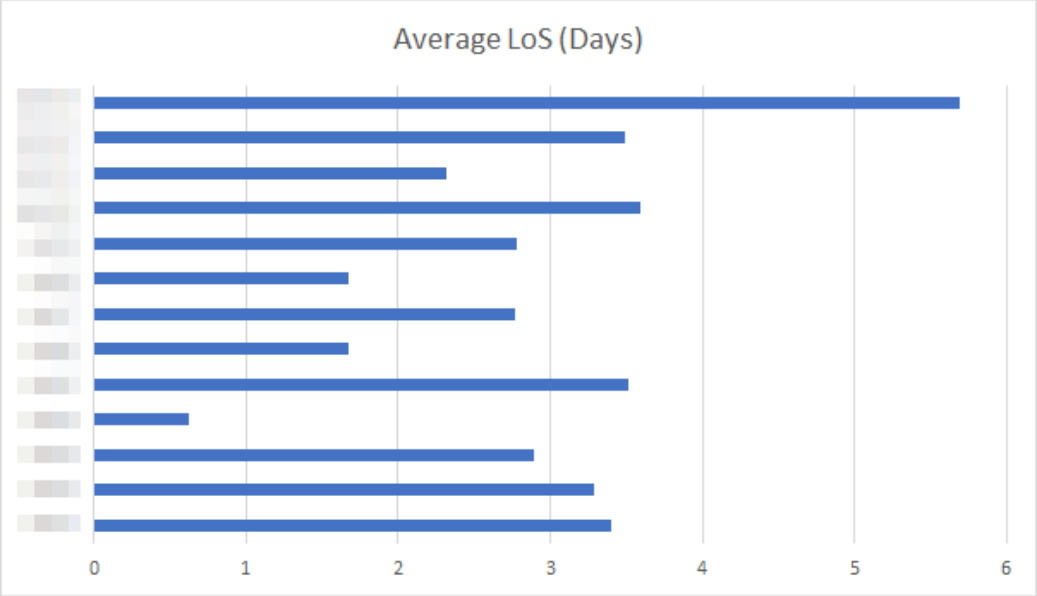


Figure 6: The average LoS for each unit (details have been blurred for non-disclosure agreement)

8.4 Unscheduled Model

For the unscheduled patients, we needed to sample the arrivals in the Emergency Room using the historical data.

8.4.1 Admissions

The principle is simple, in the dataset of the discharges we can derive the average number of arrivals per day, taking in account day-of-the-week-seasonality. The figure below gives the precision about this particular sea-

sonality. The seasonality due to the period of the year (there are some peaks, especially during the summer or the winter) is supposed to be directly taken into account. Indeed the dataset of the discharges contains only the 2 last months. Therefore the statistics made consider only the current period of the year.

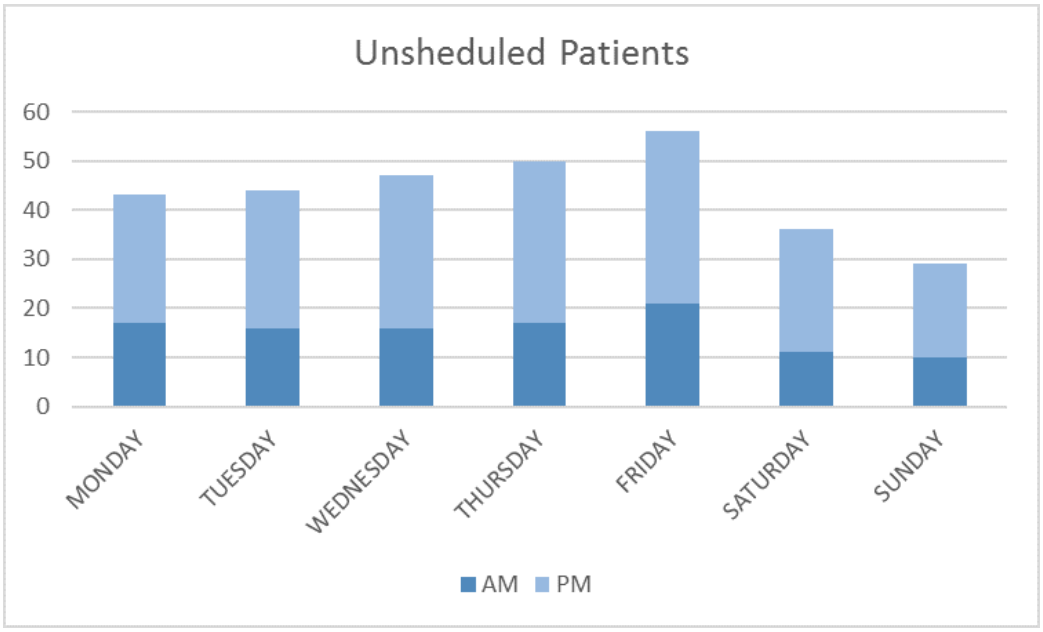


Figure 7: Day-of-the-week-seasonality for the unscheduled arrivals

Once we made statistics on the days, we know how much unscheduled patients we need to simulate today, tomorrow and after tomorrow. We are even more precise: we take into account whether it is the morning or the afternoon to fill our time slots correctly. Then, if we know we have to sample "n" unscheduled arrivals in a particular time slot, we can also assign to this "n" different patients a NU. We did this assignment using the historical data, the next figure is illustrating the share of the NU for the unscheduled arrivals. Based on this distribution we can easily sample our unscheduled patients (it is just a Multinomial Distribution). We used this model to fill our inputs for the Emergency Room arrivals.

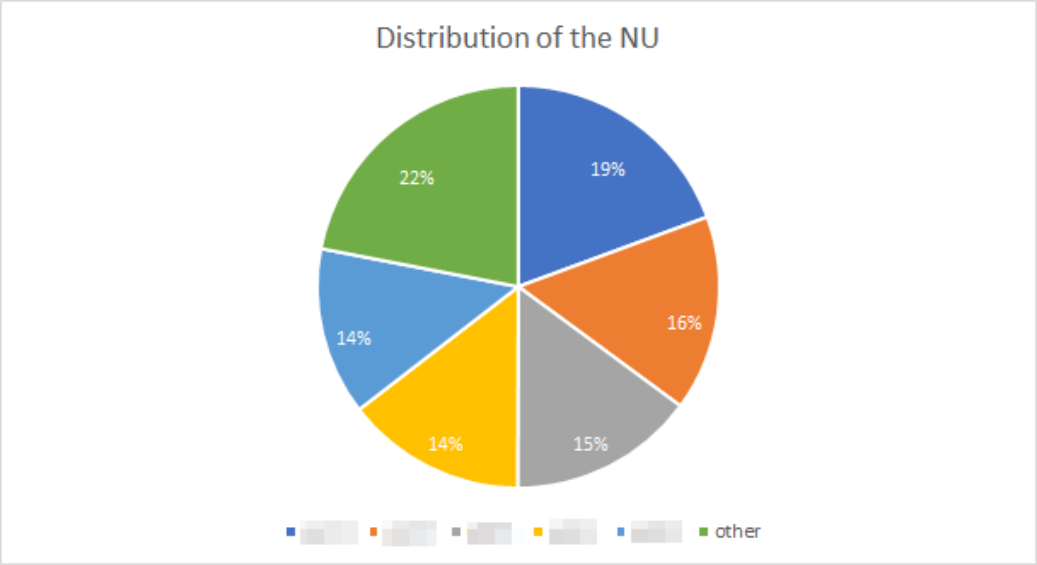


Figure 8: The distribution of the NU concerning the unscheduled arrivals (details have been blurred for non-disclosure agreement)

8.4.2 Discharges

As we did for the scheduled we sample the LoS of the unscheduled patients. Then we can derive the discharges dates of the unscheduled arrivals.

9 Graphic User Interface

9.1 Motivation

Because the product has to be used by non-technical teams, we had to build a GUI (Graphic User Interface). This GUI enables everyone to easily use the application with just a few clicks and with no-coding skills.

9.2 Technology

Due to its simplicity of use when it comes to Machine Learning projects, we chose to run the back-end of the product with Python. So, we naturally used the Python package named Tkinter to create the GUI.

All the product's files and Python application are wrapped in one folder which lives in the NYU Langone's server system. Then, anyone who has access to the folder can download it and has just to push it on his own Desktop.

Then, the user can launch the product by clicking on a dedicated icon which launches a ".bat" file (Microsoft Windows script file).

9.3 Product details

We created a three-page GUI. This GUI allows users to update data, exhibit overview of the hospital's occupancy status and show details about a specific unit in the hospital. A notice is available for the user which explains every step

9.3.1 Page 1: Configuration Page

The first page of the GUI is the Configuration Page. It is the page displayed at the launching of the page which enables the user to update the data inputs. As we are not authorized to directly work with the NYU database, the user has to manually update the required ".csv" files.

Because the back-end calculations take about a minute to be entirely run, we designed it to be run only if a data input has been updated. If the user re-opens the product and click on "Go to Hospital Overview" button without updating the files, then the app goes directly to the second page without running the back-end.

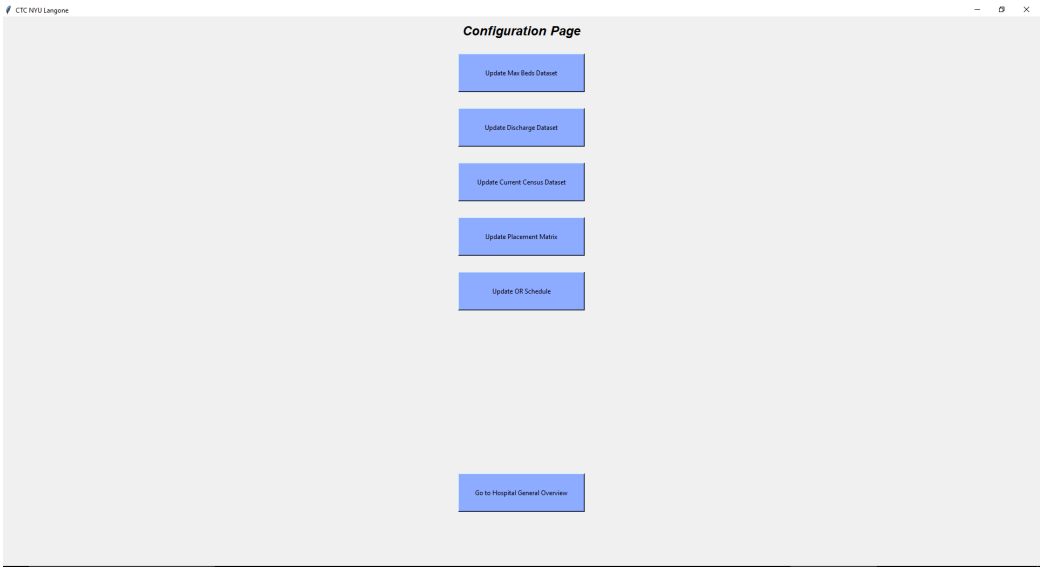


Figure 9: Configuration Page

9.3.2 Page 2: Hospital General Overview

The second page is the "Hospital General Overview". It gives the general occupancy of each of the in-scope NU at the selected time slot. Users can switch time slots by selecting in the drop list at the top of the window.

The top row of buttons corresponds to the aggregation of certain NU (NYU Langone, Tisch Hospital, Kimmel Pavillon and Hassenfeld Children's Hospital). Then, the other buttons represent the NU. The displayed colors set corresponds to the level of occupancy and is defined following the health industry layout.

By clicking on a NU button, the user can go to the "Details Page" of this unit.

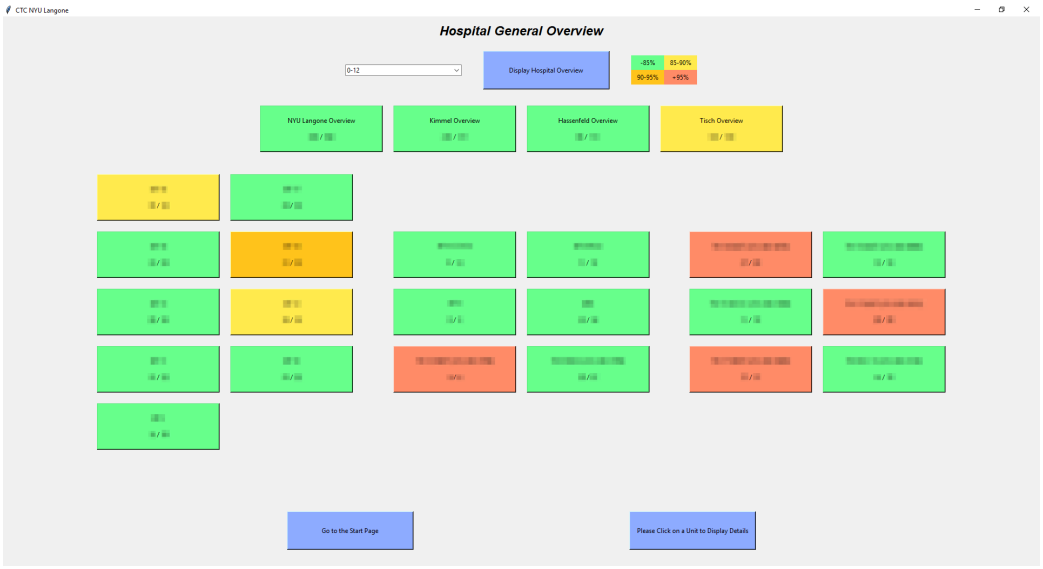


Figure 10: Hospital General Overview Page (details have been blurred for non-disclosure agreement)

9.3.3 Page 3: Unit Details

This page details the patient flow within the selected unit. After selecting the type of patient (All, Inpatient, Scheduled, Unscheduled), the user can observe the admissions, discharges and unit occupancy due to this type of patient.

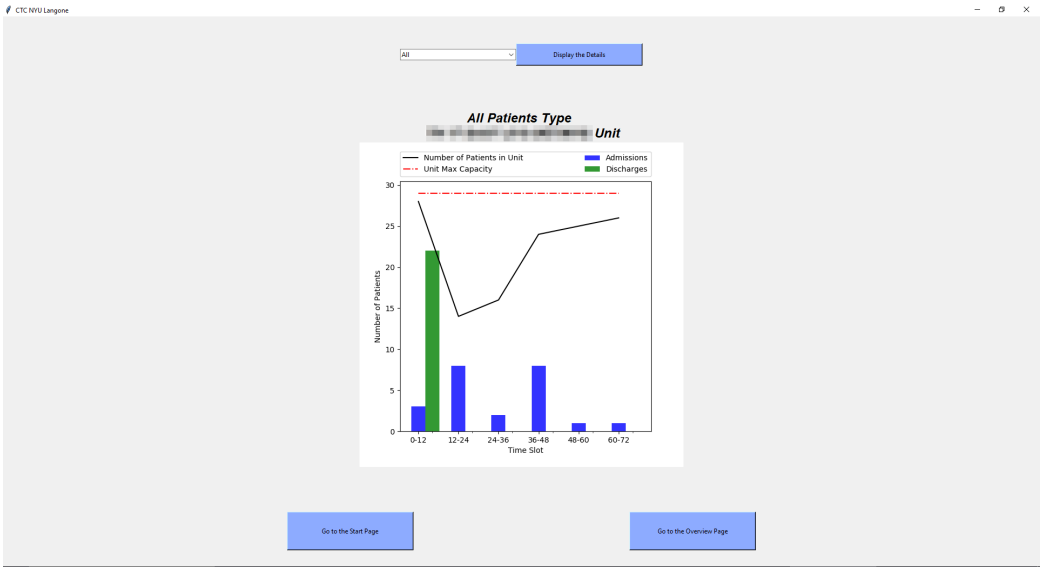


Figure 11: Unit Details Page (details have been blurred for non-disclosure agreement)

10 Improve the results: Probabilistic Inpatients Model

10.1 Preliminary

As explained before, we are not satisfied with the inpatients model. Indeed, because of a technical issue, we had to use the prediction realized by the physician and not the prediction inferred by an automatic algorithm for our final product. The technical issue is due to the I.T. system at NYU Langone which overwrites the past data coming from the Current Census table with the last entries of the patient. This is a problem due to the pulling of the data requested in the Oracle database.

We, with the help of the NYU I.T. team, came up with a solution during our summer internship. Every day, 4 snapshots of the Current Census are taken and stored in a new database. Then, after several weeks of capture, the CTC team will have a Current Census dataset without any overwriting issue.

However, during the research phase and before realizing that the past data were overwritten by the last entries, we started the analytical and coding part of the probabilistic inpatients model. In this part, this model will be explained as all the work is done. The only cautionary advisement is about the results we got since they were established with incorrect data. By the middle of the Fall we are aiming to try and test our hypothesis with better quality data.

10.2 Classification problem & Discharge probability distribution

As explained in the Model paragraph, the task to predict the remaining LoS of patients is a Supervised Classification Problem aiming to infer the number of discharges within each NU. Indeed, given the information we have about past patients, we want to know in which category the current patients belong. Then, it will be possible to retrieve the number of discharge.

However, in the probabilistic approach, instead of trying to predict the exact category a patient belongs, we will try to predict the probability distribution of the discharge (in each time slot) for each patient.

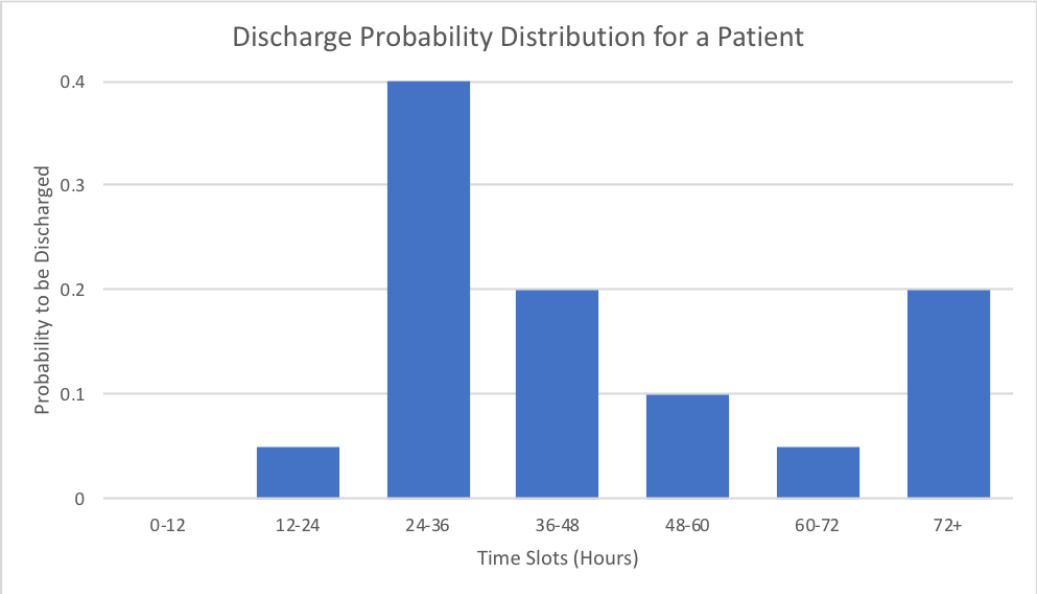


Figure 12: Probability distribution of being discharged for a particular patient

In order to infer these distributions, we tried several different machine learning models.

10.3 From a probability distribution to a prediction of the number of discharge within each NU

Once we have predicted the discharge distribution for all the patients of the hospital, then, the task is to infer the number of discharges within each NU of the hospital.

So, each time a new census is available and the CTC team desires new hospital capacity predictions, the algorithm is run:

- Definition:
  - P: A particular patient
  - L_P: List of all the inpatients

– NU: Nursing unit
– L_{NU}: List of all the nursing units
– TS: Time slot
– L_{TS}: List of all the time slots
– $p(P, TS)$: Probability of patient P to be discharged during time slot TS
– $D(P)$: Probability distribution of the patient P
– $EDN(NU, TS)$: Expected number of discharges in NU, at time slot TS
– $O(NU, t)$: Expected occupancy of NU at time t

$$L_{TS} = \begin{pmatrix} 0-12 \\ 12-24 \\ 24-36 \\ 36-48 \\ 48-60 \\ 60-72 \\ 72+ \end{pmatrix} \tag{1}$$

Equation 1: Time slots

$$D(P) = \begin{pmatrix} p(P, TS_0) \\ p(P, TS_1) \\ p(P, TS_2) \\ p(P, TS_3) \\ p(P, TS_4) \\ p(P, TS_5) \\ p(P, TS_6) \end{pmatrix} \tag{2}$$

Equation 2: Probability distribution of the patient P

$$EDN(NU, TS_i) = \sum_{P \in NU} p(P, TS_i) \tag{3}$$

Equation 3: Expected Number of Discharge in NU at TS

$$O(NU, t) = \sum_{TS_i \geq t} \sum_{P \in NU} p(P, TS_i) = \sum_{TS_i \geq t} EDN(NU, TS_i) \tag{4}$$

Equation 4: Expected Occupancy of NU at time t ($TS_i \geq t$ means $\max\{TS_i\} \geq t$)

Algorithm 1 Current Inpatients Probabilistic Algorithm

Probability Distribution
for P in L_P **do**
    Compute $D(P)$
end for

Expected Number of Discharges
for NU in L_{NU} **do**
    for TS in L_{TS} **do**
        Compute $EDN(NU, TS)$
    end for
end for

Expected Occupancy
for NU in L_{NU} **do**
    for t in L_{TS} **do**
        Compute $O(NU, t)$
    end for
end for

10.4 Statistical Model - Theory for Classification

10.4.1 Introduction

Since patients need a probability distribution as an output, our team needs statistical models in order to get this distribution.

Because we work in the health-care field, we have quickly decided to choose the simplest algorithm to understand as a baseline. Not only we needed to design an accurate algorithm, but also it was required that non-technical people are able to understand the Mathematics behind it. During the meetings with the operational teams, we had to explain the way algorithms work and that was why we naturally opted for the Decision Tree algorithm.

10.4.2 Decision Tree Algorithm

The Decision Tree is one of the most famous and easier machine learning algorithms. It is based on the principle of a succession of questions to answers which leads to the label of the data point. The easiest way to understand is through an example:

Figure 13: Example of a simple decision tree

In this graphical description of the tree, we understand that at each step (in red and orange), a decision is made thus answering the question. Then,

a "leaf" (in green) of the tree is reached, this "leaf" returns the label of the data point.

However, for modern usage in data science projects, the scope of features is important and it is not possible to come-up manually with the different decision steps. The succession of questions needs to be automatically set up by looking at the labeled data.

The method is mainly based on two ideas:

1. A recursive partitioning of the space induced by the variables
2. Pruning of the created trees using the validation dataset

The principle of the algorithm is as follows:

- Definition:
  - QIF: Indicator that quantifies the information contained in the features considered. We use one of the most famous: the Gini coefficient (we invite the reader to learn more about it).
  - L_f: List of the features

Algorithm 2 Decision Tree Algorithm

1: **Pick** L a sublist of L_f containing at most the square root of the elements of L_f.

2: **Find** the feature in L that maximizes/minimizes (it depends on the choice of the QIF, some need to be maximized and reversely).

3: **Split** the span of the input into leaves and remove the feature considered from the L_f.

4: **Repeat** recursively the algorithm independently on each leaf (partitioning) until the QIF is considered to be too high/low (pruning) with the updated L_f.

However, Decision Tree algorithms are subject to over-fitting. That is why we decided to use an evolution of the Decision Tree - the Random Forest.

10.4.3 Random Forest Classifier

Description

A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \theta_k),\ k = 1,\ \ldots,\ \infty\}$ where the $\{\theta_k\}$ are independently and identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. One of the valuable features of a random forest classifier, as an ensemble technique, is it has less tendency to over-fit. The accuracy of a random forest depends on the strength of the individual tree classifier and a measure of the dependence between them.

Breiman and Cutler (2008) give an algorithm for constructing a random forest, which we adopted in our model:

Algorithm 3 Random Forest Algorithm

1: **If** the number of cases in the training set is N, sample N cases at random, but with replacement, from the original data. This sample will be the training set for growing the tree.

2: **If** there are N input variables, a number $m \ll M$ is specified such that at each node, n variables are selected at random out of the N and the best split on there n is used to split the node. The value of n is held constant during the forest growing.

3: Each tree is grown to the largest extent possible. There is no pruning.

Feature Importance

Because the Random Forest is composed of a great number of individual Decision Trees, it becomes impossible to display an aesthetic decision diagram. This is a drawback in term of understandability, especially in the health-care industry. Indeed, lives are at stake, so professionals need to understand (at least having a baseline comprehension) of the algorithms before utilizing them. This is why we needed a methodology to display which features the algorithm uses.

For each feature in the random forest, we can compute a score referring to the importance of this feature. Breiman and Cutler (2008) suggested a way to compute the score.

For every tree in the forest, put down the testing data and count all the votes to the correct class; then, randomly permute the values of a variable, and put the perturbed data down the trees again and count the correct votes. The raw importance score of a tree is the number of correct votes of the original testing data minus perturbed data. The average of the raw importance score of all trees is the raw importance score of the variable. Finally, we compute the standard errors of the score and divide the raw score by the standard error to get a z-score.

Choice of the Random Forest

We choose the Random Forest Classifier because of the following advantages:

- Random Forest Classifier runs efficiently on large databases. As we are considering a day-patient model for the inpatients, the training dataset increases at a considerable pace, i.e. $365 \times 562 = 205,130$ data points per year.
- Random Forest Classifier handles thousands of input variables without variable deletion. It also gives an estimate of variables importance. It allows us to display the importance of the features, and then we were able to show it to non-technical teams. Thus, it gives an acceptable sense of the way the algorithm processes the patient data.
- The dataset is imbalanced, indeed the labels are not represented in the same proportions. Random Forest Classifier has method "cost sensitive based balancing". The idea is to weigh the cost of misclassification with a weight equal to the inverse of the proportion of the label.

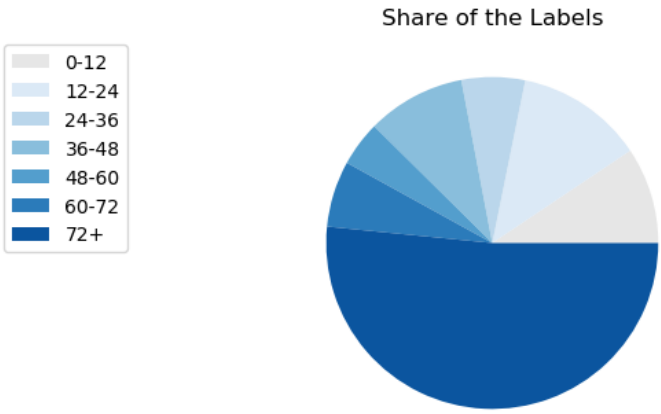


Figure 14: Illustration of the imbalanced labels

10.5 Feature Engineering

We used DataFrame in pandas as our data holder. Pandas is an open source BSD-licensed library. It provides several data structures and data analysis tools. We put all our data in DataFrame and did preprocessing like changing the data type, handling of missing values and merging datasets using pandas functions.

Typically, we had four types of features to engineer:

- Numerical features (ex: `Patient Age`, `Cmi Score`, `Anesthesia Score`, etc...)
- Ordinal: categorical features with an assumed order (ex: `Insurance Type`, etc...)
- Nominal: categorical features with no assumed order (ex: `Service`, `Ethnicity`, `Marital Status`...)
- Datetime Data: Originally stored as strings, then converted as datetime objects (`Elapsed LoS`, `Admission Day of Week`, `Current Day of Week`, etc...

Numerical features underwent two processings in this case. First, the missing values were imputed using the median of the feature. Then, we scaled the feature using Standard (Z) Scaling. Below, the reader can find the distribution of the two main numerical features `Patient Age` and `CMI federal`. We also engineered numerical ratio features, i.e., for each service of the hospital and for each day, we computed the following ratios. `Ratio of Open Beds`, `Ratio of Blocked Beds` and `Ratio of Closed Beds`.

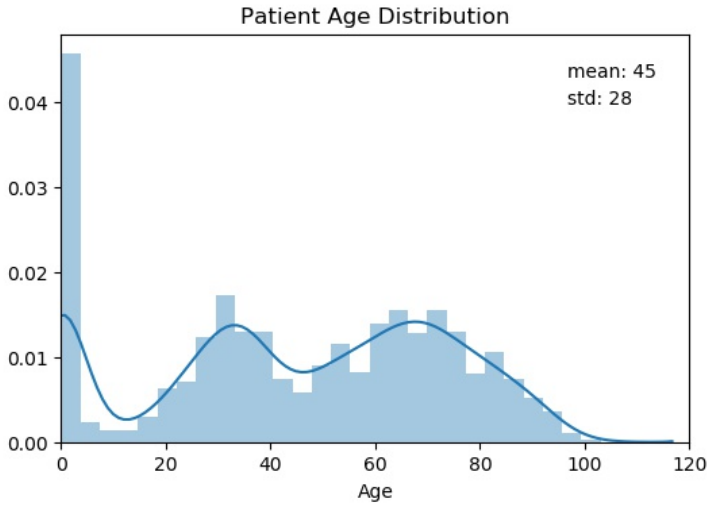


Figure 15: Patient Age Distribution

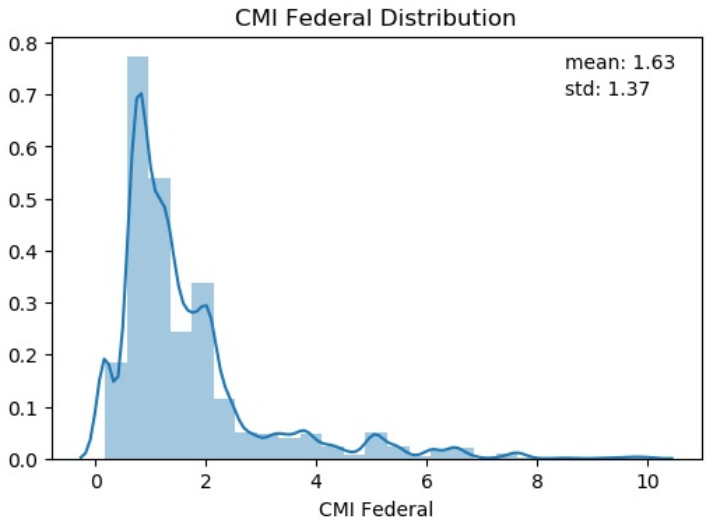


Figure 16: CMI Federal Distribution

These missing nominal features were input using the majority class as imputation class. The nominal features were encoded using "One-Hot Encoding" methodology. It is the most widespread approach for encoding categorical variables in machine learning models. One-hot encoding creates a new binary column for each value, and enter 0 if the value is taken for the instance and 1 otherwise. We dropped the first column after encoding to avoid information redundancy. Below, the reader can find two examples of nominal features:

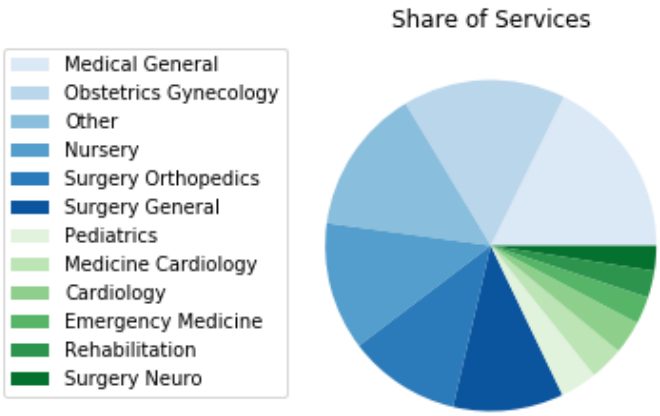


Figure 17: Share of services

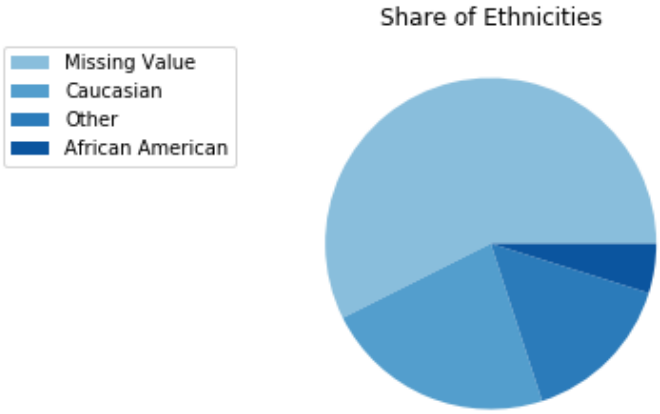


Figure 18: Share of ethnicities

For the ordinal features, we labeled data into number 0, 1, ... , N–1. This coding is used to retrieve the order between the classes of a feature. The reader can find below an example of an ordinal feature. Indeed, the `Insurance Type` has an assumed order between the different categories.

Someone Uninsured is assumed to be less affluent than someone under Medicare/Medicaid. Moreover, those under Medicare/Medicaid are supposedly less wealthy than individuals under a Private Insurance.

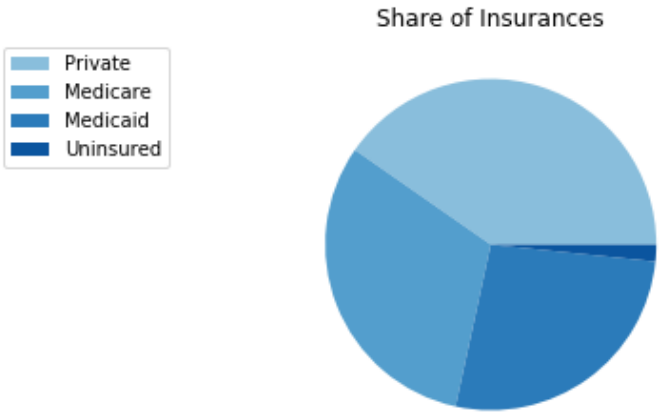


Figure 19: Difference in insurance providers

For Dates and Times, we transferred all data to DateTime object under python package DateTime. This enabled us to get information like the day of the week, month, year, etc...

One of the DateTime features we engineered was the `Elasped LoS`: it is the delta-time the patient has already spent since its admission. The reader can find the distribution of the `Elasped LoS`. We also engineered the `Admission Day of Week` and `Current Day of Week`. Also, we created a binary variable `AM` and `PM` differing whether the admission was in the morning or afternoon.

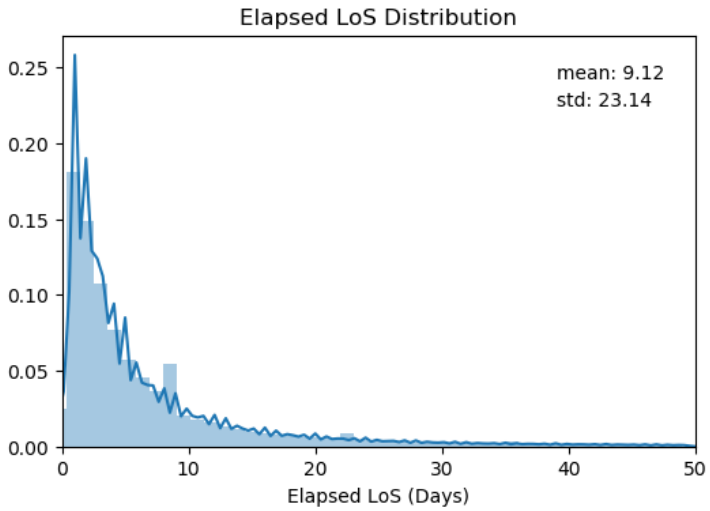


Figure 20: Elapsed LoS Distribution

10.6 Results

Bear in mind, that these results are temporary. Indeed, as the capture of the data started late this summer, it is still required to capture more data in order to have an acceptable sense of the model's metrics.

10.6.1 Confusion Matrix

The confusion matrix is a commonly used tool to assess the quality of a predictive algorithm. The rows correspond to the occurrence of the predicted labels while the columns correspond to the occurrence of the true labels. The confusion matrix has been widely used during our internship. Indeed, because we were working with non-technical teams, it was a great way to show our results.

Given our results, we are thrilled to pursue future work. Indeed, due to the imbalanced dataset, the classifier has a noticeable trend to classify the data point in the majority label, i.e., "72+". We want to proceed in 2 steps to deal with this issue:

- Identify the outfitter patients in the training dataset, i.e., the patients who are in the hospital for long-term care. Indeed, these patients should be removed because they are here for very specific and involved treatments. Our objective is to predict the hospital occupancy within 72 hours, we think we can obtain better results by removing them.

- Dealing with the imbalanced data by trying the following different methods:

  - Under-sampling: reduce the size of the abundant class by drawing samples from this class without replacement.
  - Over-sampling: increase the size of the minority class by drawing from this class with replacement.
  - Customize the cost function: use a customized cost function to optimize other metrics. We want to try to optimize the average $F1$ score of all the labels except the "72+".

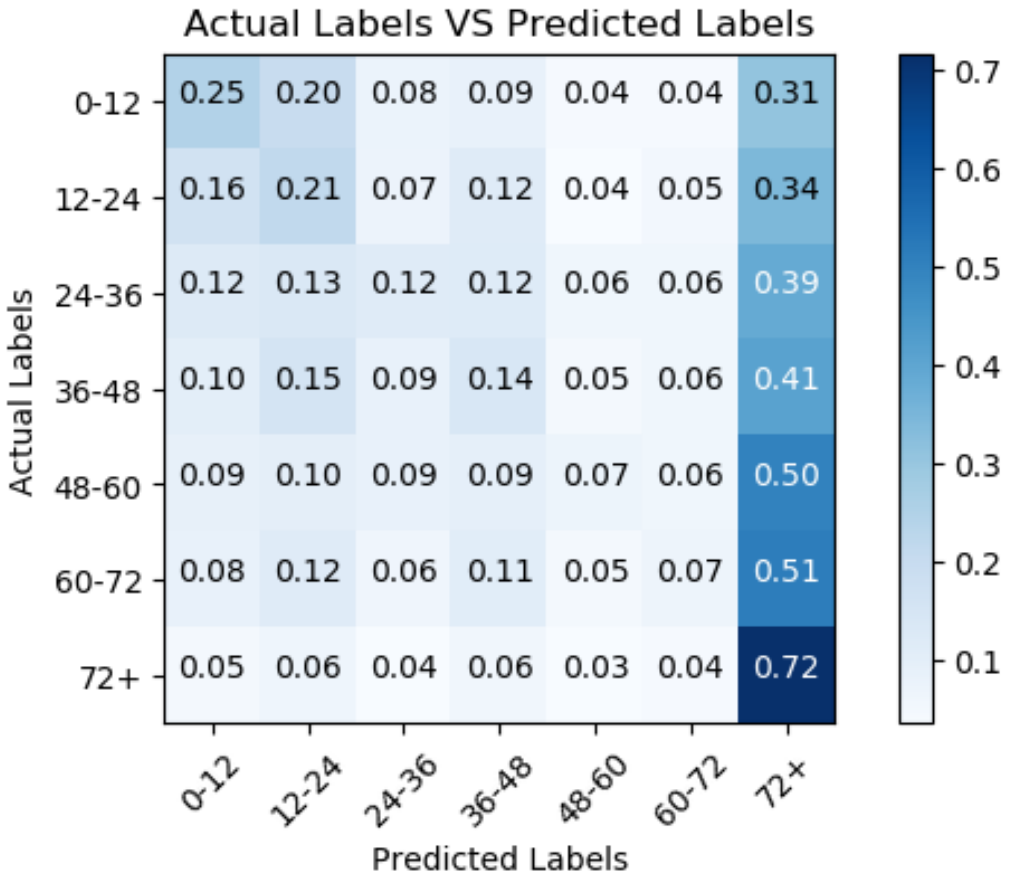


Figure 21: Confusion Matrix

10.6.2 Feature Importance

As explained before, feature importance is a beneficial tool of the Random Forest used to get a sense of how the algorithm makes its decisions. Indeed, by observing the scores, even non-technical teams were able to understand and have confidence in the algorithm.

In our specific case, we see the prevalence of the numerical features

`Patient Age`, `CMI Federal`, `Ratio of Open Beds` and `Elapsed LoS`.

The categorical features have a weaker predictive power, they are all several times weaker than the numerical features.

The surprise was the non-importance of the feature `Service` and was especially a surprise for the CTC team. Actually, the result is understandable: a `Service` can handle multiple types of care and surgeries. Then, we can easily realize that the `Service` is not a satisfying predictor of the patient's discharge.

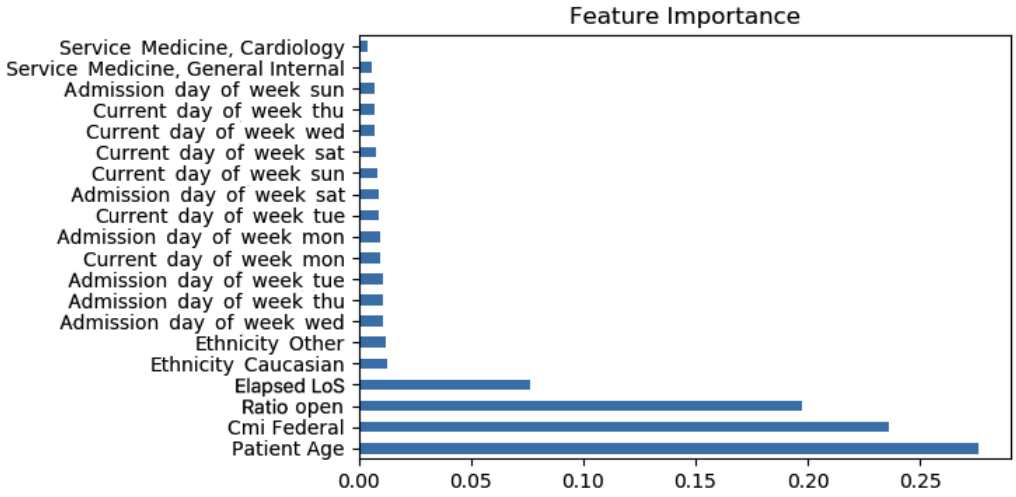


Figure 22: Feature Importance Score

11 Conclusion

During our summer internship, we produced a capacity planning tool that displayed the NYU hospital's bed occupancy status. Our product will generate improved results for the operation processes of NYU Langone Health, also saving money and resources.

The NYU hospital has a nurse calculating the occupancy status manually for now. It consumes a lot of time for the nurse to convert raw data into a readable display of the occupancy status. Our product automatizes this

process and makes it so that the nurse will no longer have to spend time completing such an onerous daily task. Additionally, our product provides a better view of the occupancy status, conveying convenience to the CTC team, allowing them to arrange surgeries better. This helps to increase the efficiency of the NYU hospital, allowing it to accommodate more patients, and deliver better attentiveness towards each individual.

Our team felt that overall we did not get access to enough data. Therefore, it was impossible for us to build the probabilistic model of patients' length of stay. However, this internship is still a valuable experience - we had the opportunity to work directly with real-world data. We put forth a large effort towards cleaning and processing the data to acquire useful information.

Fortunately, the NYU hospital is willing to provide us with direct access to databases in the fall, so that we can get access to more data fields and improve our product by making more accurate predictions for the patients' length of stay. In the future, our team may be able to move to phase two of the project, which is building an optimization model for scheduling surgeries. Our team is very interested in the idea and excited to continue the project in the fall.

References

[1] Breiman, L., *Machine Learning*, 45: 5., 2001, `https://doi.org/10.1023/A:1010933404324`.

[2] Miranda, D., Moreno, R. & Iapichino, G., *Intensive Care Med*, 23: 760, 1997.

[3] Green L., *Queueing Analysis in Healthcare. In: Hall R.W. (eds) Patient Flow: Reducing Delay in Healthcare Delivery. International Series in Operations Research & Management Science*, vol 91, Springer, Boston, MA, 2006.

[4] Mackay, M. & Lee, M., *Health Care Manage Sci*, 8: 221, 2005.

[5] Afilal M, Yalaoui F, Dugardin F, Amodeo L, Laplanche D, Blua P., *Forecasting the emergency department patients flow*, J Med Syst., 2016, 40:175.

[6] X. l. Zhang, T. Zhu, L. Luo, C. z. He, Y. Cao and Y. k. Shi, *Forecasting emergency department patient flow using Markov chain*, 2013 10th International Conference on Service Systems and Service Management, Hong Kong, 2013, pp. 278-282.

[7] Kolker, A, *J Med Syst*, 32: 389, 2008, https://doi.org/10.1007/s10916-008-9144-x

[8] Olugboji O., Camorlinga S.G., Faria R.L.., Kaushal A., *Understanding the Emergency Department Ecosystem Using Agent-Based Modeling: A Study of the Seven Oaks General Hospital Emergency Department*, In: Sturmberg J. (eds) Putting Systems and Complexity Sciences Into Practice., Springer, Cham, 2018.

[9] Littig, S.J. & Isken, M.W., *Health Care Manage Sci*, 10: 47, 2007, https://doi.org/10.1007/s10729-006-9000-9

[10] Steven Walczak, Walter E. Pofahl, Ronald J. Scorpio, *A decision support tool for allocating hospital bed resources and determining required acuity of care*, Decision Support Systems, Volume 34, Issue 4, 2003, Pages 445-456, ISSN 0167-9236.

[11] Renata Konrad, Kristine DeSotto, Allison Grocela, Patrick McAuley, Justin Wang, Jill Lyons, Michael Bruin, *Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study, Operations Research for Health Care*, Volume 2, Issue 4, 2013, Pages 66-74, ISSN 2211-6923.

[12] Eoin O'Mahony and David B. Shmoys, *Data analysis and optimization for (citi)bike sharing*, In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15), AAAI Press 687-694, 2015.

[13] Microsoft open source project, *Predicting Hospital Length of Stay*, https://github.com/Microsoft/r-server-hospital-length-of-stay