ORIE 4741 Midterm Report

Hospital Admissions

Yubin Kim (ytk3) & Maggie Liu (ml958)

1 Description of Dataset

Our dataset is MIMIC III from the MIT Lab for Computational Physiology. The data gives us information on a patients admission into a hospital. The data reveals what they were diagnosed with the specific times or admission and discharge.

We are still exploring what we want to predict/would like to define as a re-admission. For example, it is possible to predict the probability of a patient being re-admitted, the total number of stays expected for patients, or even a binary "will the patient come back or not within the year". Each different output has its own advantages and drawbacks with different use cases. We will need a different statistical model for each e.g. logistic regression vs. linear regression vs. classification.

We will do an 80-20 split on our data and calculating the train/test error to see how our models perform not only with different features, but with different statistical structures through the appropriate error measure for each type of model (miss-classification, MSE, etc.).

There are 19 raw features in the dataset with 58,976 rows of data. Some columns such as admission time and discharge time are fully populated while others such as religion and language contain many missing and corrupted/uninterpretable values. An example of an uninterpretable value in language is "** T". We have decided to transform these columns to a more useful format such as a binary English vs. non-English. For this particular column, around 40% of the data is missing.

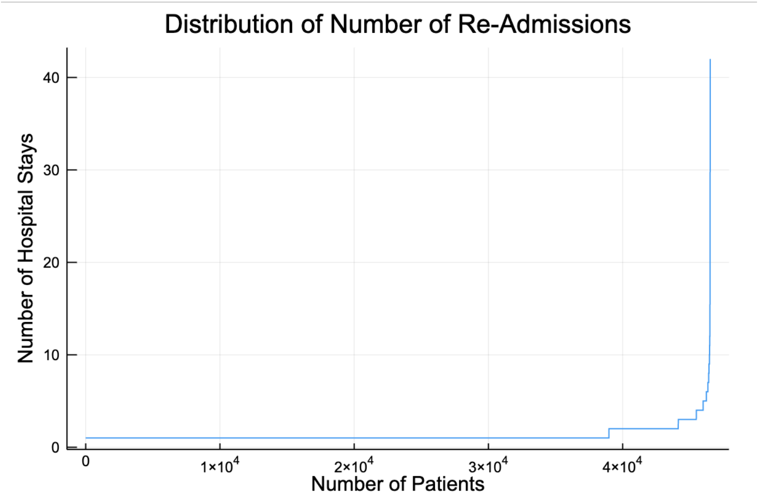

Figure 1: Distribution of the number of hospital re-admissions.

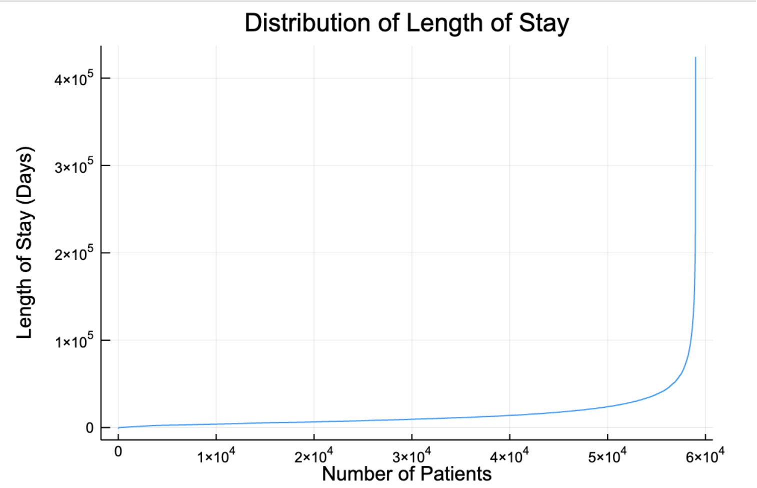

Figure 2: Distribution of the length of hospital stays.

2 Feature Transformation

Given that our data has much fewer features than the number of data points, we are less concerned with over-fitting our data than with under-fitting. To prevent under-fitting, we will be converting the existing data features into more useful columns to give us a more meaningful and insightful model. For example, there are the two data features for the time of admission and discharge. These two features are meaningless in context without each other so we convert these data features into one feature calculating the length/duration of stay. Similarly, we have a feature with the specific diagnoses with thousands of unique values. Instead of converting this to a one-hot vector, we have decided to use values from another table that has given each diagnosis a ordinal priority ranking.

Some feature transformations that we have used for the admissions data:

- **Length of stay:** transformation to get the difference between the features, admission time and discharge time.
- **Month of stay:** addition of a the data feature, date, to determine the month the patient stayed at a hospital.
- **One-hot admission type:** There are four existing admission types: Emergency, Elective, Newborn, and Urgent. We transformed these into a one-hot vector as they are categorical and nominal.
- **Convert language into binary:** We converted the language feature to have be either English or non-English. With the 76 different languages, making this one-hot didn't seem very necessary because the hospitals are in the US and therefore perform services in one language, English. We have also chosen this method because there seemed to have a lot of corrupted and missing values. Also, converting this into a one-hot vector would create a lot of features which may cause the model to be overfit.

3 Preliminary Analyses

Within our preliminary analyses, we used linear regression models to fit the features. Although, we know that this may not be the best representation, we aimed to use these analyses to develop some intuition about the data set.

3.1 Simple Linear Regression

We ran a simple linear regression on the data to predict the total number of hospital re-admissions using the features that we have transformed so far: length of stay, admission month, discharge month, admission type, language, expiry, and offset. We are using the values of the coefficients to determine which features are more influential and to guide us on what features to add next. The w obtained was $w = \{2.13{\times}10^{-7}, 0.00088, -0.003097, 0.779, 0.162, -0.14568, 0.39105, 0.424, -0.6502, 1.567\}$. With these coefficients we can see that admission types, expiry have significant effects on the number of admissions of a given patient.

3.2 Simple Logistic Regression

Since it shows that the length of stay is exponential. The small w show that our model may be under-fit. We tried to fit a logistic model instead to see if there is a better fit. The $w = \{1.12 \times 10^{-6}, -9 \times 10^{-5}, 0, 0.1987, 0.0497, -0.17, 0.104, 0.126, 0.086, -0.2, 0.18\}$ With these coefficients we see that there is consistent results from the previous model that admission types and the expiry have significant effects.

4 Future Work

As mentioned earlier, we plan to further explore the different model types that are feasible for our data and the different questions that they will help us answer. We have a few error metrics in mind for each type of model, but would also like to further explore the balance between false positives and false negatives and the consequences of each.

With the data, we have some questions about the timestamp where in place of the year, there is an arbitrary 4 digit code. We are reaching out to the curators of the data set to gather more information. Also based on our plots for the number of stays (a potential output to predict), we will consider transforming the data so that we can fit a linear model/perform regression.

We are also considering splitting our data further by sub-population (admission type, diagnosis) to see if different populations will behave different and therefore need differently tuned models.