94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang

Problem Framing

Readmission is a prevalent issue for hospitals and often translates directly to additional costs. Moreover, high readmission rates also give insight into the reduced quality of life and care of the patients. Therefore, understanding the characteristics of readmission can help a hospital (doctors, in particular) develop appropriate plans of action for patients with a high risk of readmission. Consequently, improving how doctors currently identify high-risk patients can reduce costs and improve the quality of life among patients.

In the U.S., there are health disparities related to race, gender, age. Data shows that racial and ethnic minority groups experience higher rates of illness across rates of health conditions - including diabetes.[1] Additionally, there exist gender differences in health and the use of health services.[2] Moreover, these health disparities may become more prevalent among the elderly. Consequently, we hypothesize that these demographic features - in addition to medical history (emergency visits) and current medical condition (number of medications, admission type) - are significant predictors for readmission.

Data Sources

Medical records are often not public as it includes private information. Consequently, we will be using a publicly available dataset provided by The University of California Irvine (UCI). The data consists of information about clinical care and integrated delivery networks from 130 US hospitals from 1999-2008. The dataset contains 50 features and 101,766 data records.

Data considerations

An area of concern is an inconsistency between the data description provided on the UCI website and the data file. More specifically, the website mentions there are 55 attributes, but the dataset available only contains 50. This discrepancy raises the question about which features are missing. Furthermore, the data discrepancy raises the question of data collection and the data cleaning process. Did these 130 hospitals all use the same data collection infrastructure? Questions about data collection and cleaning are important because they speak to the quality of the data. While these questions are valid, obtaining another dataset to investigate readmissions would be difficult. Therefore, we trust that the quality of the data is good.

---

[1] "Racism and Health." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, July 8, 2021. https://www.cdc.gov/healthequity/racism-disparities/index.html#:~:text=The%20data%20show%20that%20racial,compared%20to%20their%20White%20counterparts

[2] Cameron, Kenzie A, Jing Song, Larry M Manheim, and Dorothy D Dunlop. "Gender Disparities in Health and Healthcare Use among Older Adults." Journal of women's health (2002). Mary Ann Liebert, Inc., September 2010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2965695/

Another data consideration is about the data being publicly available. While there doesn't seem to be any identifying information, patient records are private. Were the patients consulted? Was consent required? While the questions mentioned above address data quality, these questions address data privacy - a critical aspect to consider. Furthermore, another ethical consideration relates to our label. Aside from the cost, reducing readmission rates also improves the quality of care of the patients. However, existing literature has well documented the greater barriers to healthcare access faced by people of color.[3] As a result, the readmission cases of patients of color in our data set might be an underestimation - potentially biasing our model.

Finally, other factors that may be strong predictors of readmission are not in the dataset. One of these predictors may be socioeconomic status. Individuals of low socioeconomic status may not have the resources to maintain the recommended healthy lifestyle. Thus, this may increase their risk of readmission. Overall, this limits the power of our model - a tradeoff we have to make.

Exploratory Data Analysis

Exploratory data analysis is a critical process to help gain a better understanding of the data. More specifically, exploratory data analysis helps identify trends, patterns, outliers, anomalies, relationships among features, check assumptions and test our hypotheses.

Basic Data Exploration:

1. Obtain an overview of the feature names and a glimpse of the values stored in the dataset
2. Understand the size and dimension of the dataset
3. Understand data types and formats.
    a. Clean data for analysis. Strive for consistency. For example, in the raw data sets missing values are represented in various ways: "NULL", "?", "Unavailable/Invalid".
4. Check for missing values, anomalies, and outliers.

Handling missing values: After initial data exploration, we discovered that there were a total of 7 columns with missing data. There are two ways we can deal with missing data. The first option

[3] "Health Equity Considerations and Racial and Ethnic Minority Groups." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. Accessed October 13, 2021. https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html.

is to drop the rows/columns with missing data. The second option is to impute the value of missing data.

Potential steps

- We considered two ways of dropping missing data. First, we can drop the rows with missing data, or second, we can drop the columns with missing data.
- We thought of 5 ways to fill in missing data: fill in using the average value, fill in the most frequent value, fill in following the existing data value distribution, fill using KNN[4], or fill with 0.

Steps taken

- We decided to drop the column "weight", as it is the column with more missing observations.
- We planned to fill in the columns using KNN as we have the data needed to figure out the most likely value for the missing cells, and this method would have allowed us to keep every row in the dataset. However, given limited technical skills and time constraints we didn't go through with these imputations and decided to drop any columns with missing values. We acknowledge that this limits the power of our model moving forward, and is something to address in version 2 of our project.

Data Visualizations

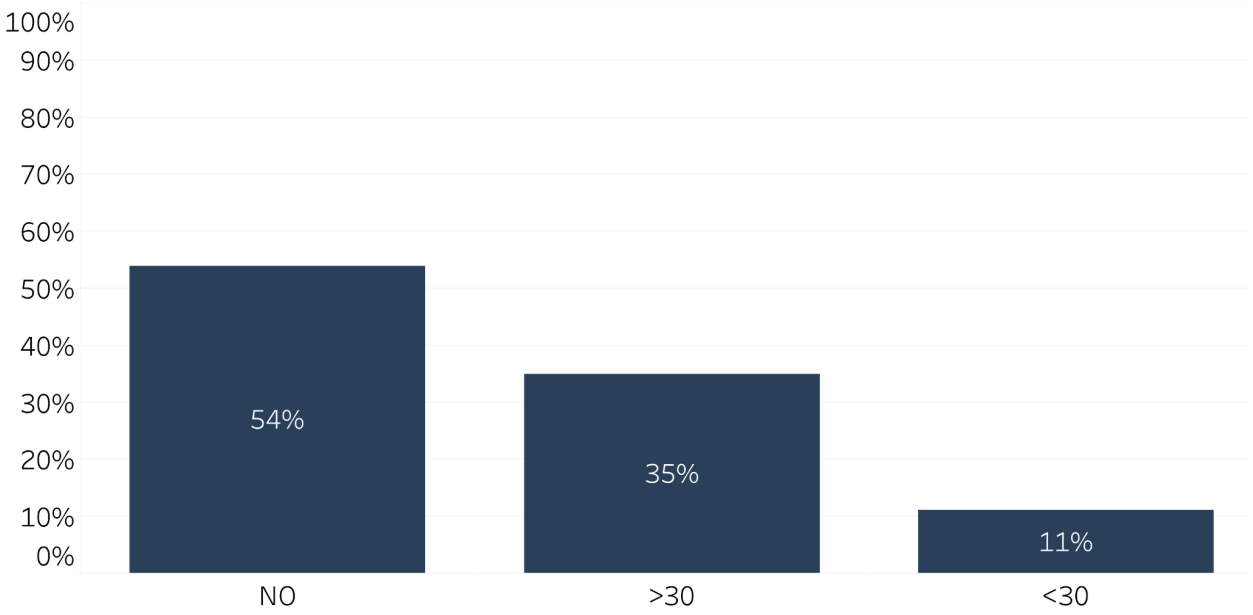
Figure 1. allows us to see that over half of the patients (54%) fall within the category of no readmission. We also see that among those readmitted, patients are more likely to be readmitted after 30 days. We decided to consider both ">30" and "<30" as readmitted to obtain a more balanced label for analysis (54% : 45%).

[4] kNN Imputation for Missing Values in Machine Learning, Jason Brownlee, https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/

94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang

Figure 1. Readmission Distribution

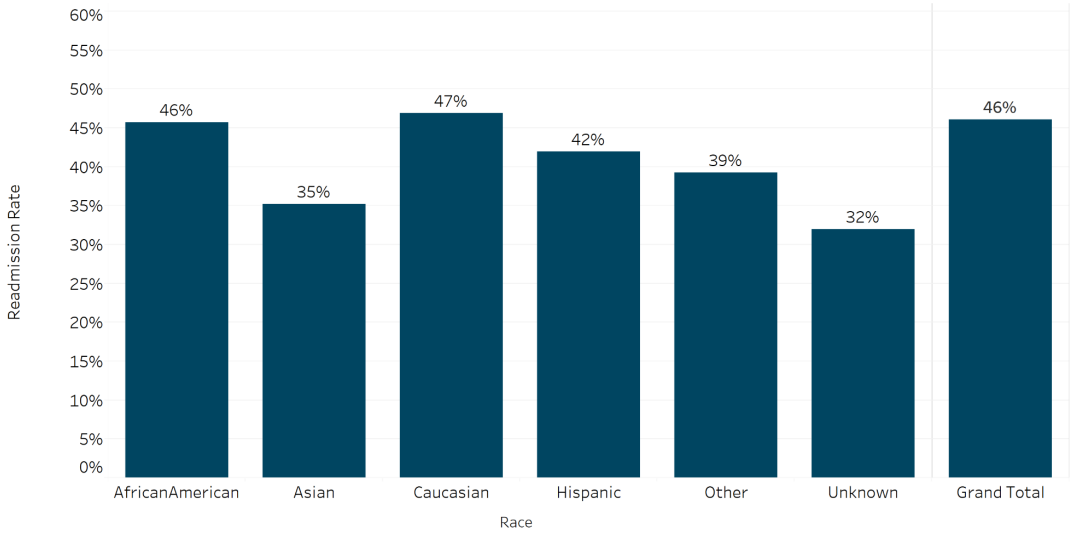


N (total records) = 101,766

Figure 2. shows a breakdown of readmission rates by age (where age is grouped into bins of 10 years). Not surprisingly, we see that younger patients (0 to20 years of age) have the lowest readmission rate. There is a 20 percentage point difference between patients that fall within the age bin of 0-10 and 10-20. This difference increases to 27 percentage points when comparing patients that fall in the bins 0-10 and 20-30. There is some variation in readmission rate between patients 20 and older, but this variation falls within 2-4 percentage points. Another observation to highlight is that patients 90 years and older experience a lower readmission rate compared to patients between 20-90 years of age. This lower readmission rate, however, may be due to patients passing away. Based on **Figure 2**., we hypothesized that age is a significant feature when predicting readmission cases.

Figure 3. provides a breakdown of readmission rates by race. For data exploration, we re-coded missing race categories as "Unknown". Overall, we observe differences in readmission rates, with Asians experiencing lower readmission rates than any other group. African Americans and Caucasians have the highest rates of readmission. It is also worth mentioning that 75% of all records are of Caucasian patients. With this finding, we hypothesized that race would not be a strong predictor because there is an over-representation of certain groups.

Figure 3. Race Vs Readmission Rate



% of population	18.9%	0.6%	74.8%	2.0%	1.5%	2.2%	100.0%
Count	19,210	641	76,099	2,037	1,506	2,273	101,766

The gender distribution in our dataset is 53.8% female and 46.2% male, which is relatively balanced (**Figure 4**.). The readmission rate by gender is 46.9% among females and 45.1% among males. Though females are experiencing higher readmission rates, the differences between the two genders are small. With this finding, we hypothesized that gender would not be a significant variable when predicting readmission cases.

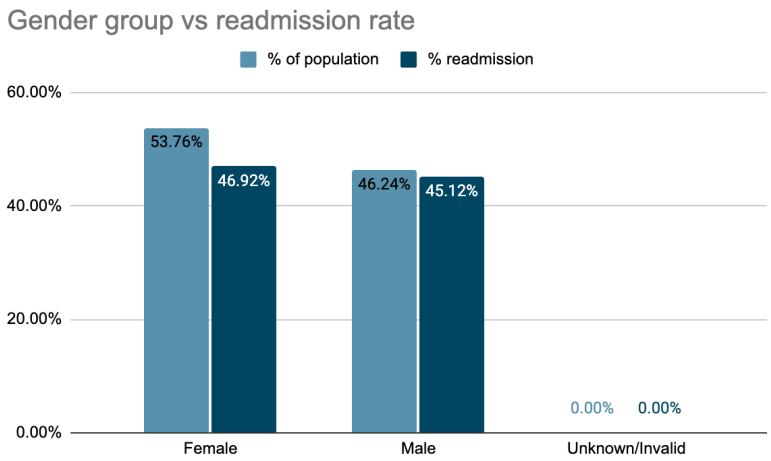


Figure 5. illustrates the readmission rate by the number of medications. We see a normal distribution between 2-50 medications - with a peak of around 50% readmission rate for 24-26 medications. When the number of medications is more than 66, the likelihood of readmission is over 50%. However, these may be outliers. We hypothesized that the number of medications is a strong predictor of readmission rate. Moreover, we will experiment with creating categorical buckets for this feature.

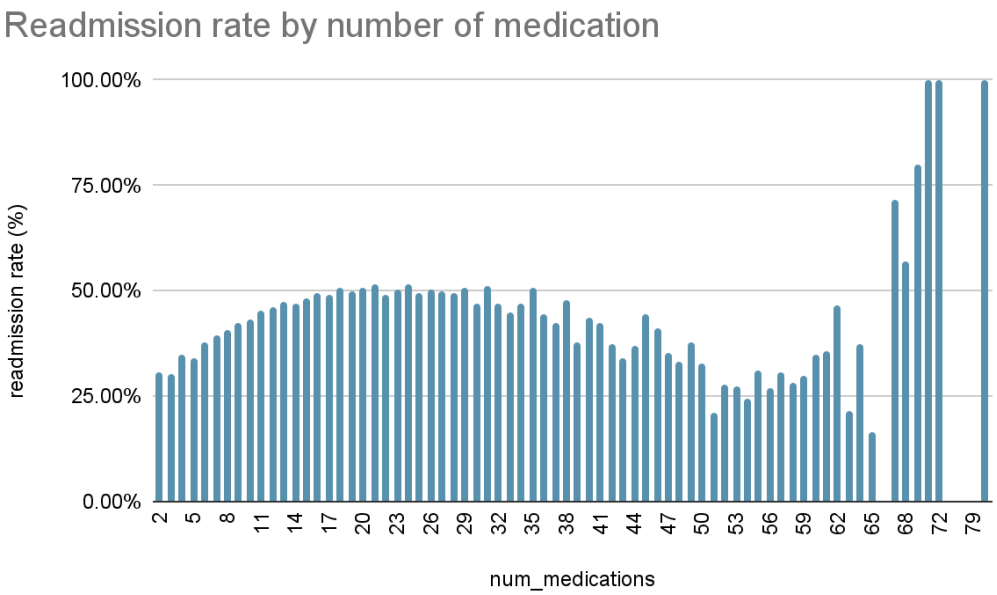
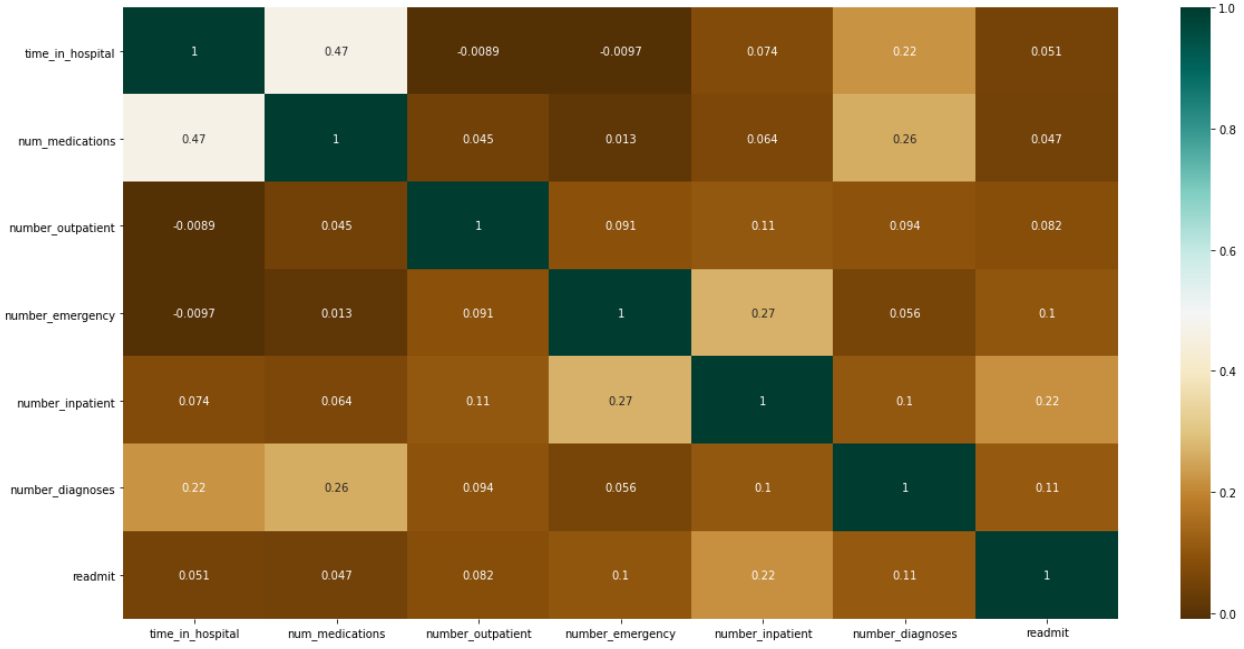


Figure 6. allows us that there is an overall weak correlation between numerical variables in our dataset. The largest correlations that we see are between num_medications and time_in_diagnoses (with a correlation coefficient of 0.47). We also see that number_impatient, number_diagonsis, and number_emergency are the features more strongly correlated with readmission.



94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang

Analytic Modeling

For version 1 of this project, we only used a single data source. Consequently, we didn't need to deal with data integration. Besides the five variables with missing values that were removed. We performed some feature engineering - aside from the data cleaning highlighted above. After the steps below, we ended up with 67 features used as training variables.

- **Transforming data types:**
    - **Numerical to Binary value**: some columns use numeric values although they should be categorical. In order to prevent mis-ranking, we decided to apply one-hot-encoding and change the data type of "*admission_type_id*" and "*admission_source_id*" from numeric to binary.
    - **Categorical to numerical value**: for the column "age", we take the mean of each 10-year range and transform it into a numerical value. For example, [30-40) becomes 35. Therefore, the age feature could be given weight when applying regression models.
    - **Categorical to Binary value:**
        - There were a series of categorical columns having values "No", "None", "Steady", "Up", or "Down". We decided to transform "No", "None", and "Steady" to 0, and "Up" and "Down" to 1 to represent whether there was volatility of the medical observation.
            - The following 22 Columns were transformed: 'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide', 'glimepiride', 'glipizide', 'glyburide', 'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'insulin', 'glyburide-metformin', 'tolazamide', 'metformin-pioglitazone','metformin-rosiglitazone','glimepiride-pioglitazone', 'glipizide-metformin', 'troglitazone', 'tolbutamide', 'acetohexamide', 'max_glu_serum'.
        - For the following 3 categorical columns: 'gender','race','A1Cresult', we decided to apply one-hot-encoding and transformed the data type to binary features.

Model Fitting

The goal of this project is to help hospitals identify patients with a high risk of readmission. More specifically, the project goal is to identify a category of patients that share similar specific characteristics and therefore draw conclusions on what type of patients are at higher risk for

readmission. Consequently, the type of analytic problem that matches this decision is classification. Diving into types of classification techniques, we considered the following:

- **Logistic regression**
  Considering that our label is binary (readmitted or not readmitted), Logistic Regression is a good starting place. The model will provide the likelihood of readmission, thus helping us to determine a given probability of what type of patients are likely to readmit. Most of the variables in the dataset are binary (True/ False). Thus, we needed to convert most of the variables into numeric types by encoding. Moreover, we also wanted to identify the relative significance of each feature.

- **Decision tree**
  Since we are to test which category (subgroup) of patients are most likely to readmit (dependent variable), and the data set contains different types of input variables with low correlation, these satisfy the criteria of using decision tree algorithms. Also, the number of variables in this data set allows us to run multiple times on different subsets to obtain robust results.

- **Random forests**
  The data set meets the requirement to perform on the decision tree described previously. Therefore, we can implement random forests as well. With random forests, we can lower the variance and prevent overfitting - potentially getting a better result.

Model Selection

In the section above, we listed several analytics techniques that we considered for this classification problem. While model fitting tends to be straightforward, model selection can prove to be more of a challenge. This challenge often arises because the model needs to meet stakeholders' specific requirements.

Model Complexity

In our case, one of the decision constraints is limited model complexity. Physicians must understand our model to make use of the insights that it provides to tackle readmission rates. A limited model complexity will also better position physicians to provide feedback regarding the model. In line with model complexity, the selected model needs to have a fast runtime. Actions taken by hospitals have a time component: the sooner physicians identify patients with a high

risk of readmission, the sooner the physician can offer additional care. Thus, overall improving the quality of care for the patient. When evaluating the runtime of our models, they all performed within 10 seconds.

Given the fast overall runtime, we do not think the slight differences significantly differentiate our models. Therefore, the ability to interpret and understand the model output becomes the priority. With this in mind, tree-based models, which could provide more transparency in the result-generating process, are preferred.

Model Performance

Other considerations in our model selection include precision, recall, and F1-score. Precision focuses on our predicted individuals for a high risk of readmission. Thus, it tells us of all our predicted high-risk patients, what percentages were correct. High precision is critical because there is a cost associated with additional care. Thus, incorrect predictions can represent a loss for hospitals. Aside from the cost associated with misclassifications, not correctly identifying patients with a high risk of readmission limits the reach of the model - translating to some patients not receiving the additional care they need. Consequently, our model should have a high recall rate. In other words, our model should identify a high percentage of patients with a high risk of readmission.

Furthermore, considering both recall and precision are valuable, looking at our model's F1 score provides another measure of performance to examine. F1-score balances the strengths that both recall and precision have by taking the harmonic mean. The F1-score can be useful, for example, when one model has a higher recall while another model has higher precision.

Table 1. Model Performance Comparison

	Precision	**Recall**	**F1 Score**
Decision Tree	**0.62**	**0.49**	**0.55**
Random Forest	**0.61**	**0.51**	**0.56**
Logistic Regression	**0.64**	**0.39**	**0.49**

Comparing model performance in Table 1, we can see that Logistic Regression has the highest precisions. Examining recall, Random Forest slightly outperforms our Decision Tree model and Logistic Regression. Finally, for F1 scores, we see our Decision Tree and Random Forest models

have similar performance - both outperforming Logistic Regression. Consequently, we opted for our Decision Tree model - a model with less complexity than Random Forest but comparable in model performance.

Model Implementation

Model Insights

For hospitals and physicians to make improved decisions, we need to communicate to stakeholders the features that are strong predictors of readmission. More specifically, we want to establish the relationship between these predictors, allowing a physician to use these predictors to determine whether a patient is at high risk of readmission or not. Consequently, it is critical to have transparency around how the model classifies individuals as high risk of readmission.

From the Decision Tree output as shown below , we could see that the first question being asked is about the number of inpatient visits <= 0.5, then if “yes”, the second question is about whether the number of diagnoses <= 6.5. If both questions were answered with Yes, then there is a high probability that a patient would be readmitted. With these in mind, when a doctor is treating a patient with similar characteristics, they could provide these patients with more careful examination to make sure the patients’ issues were being addressed, therefore preventing readmission in near future.

Figure 7. Decision Tree output

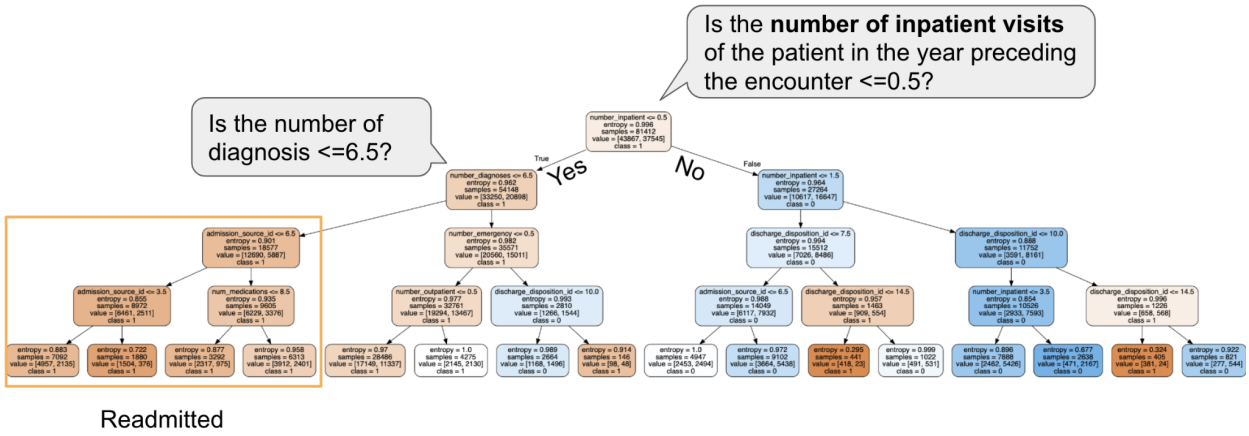
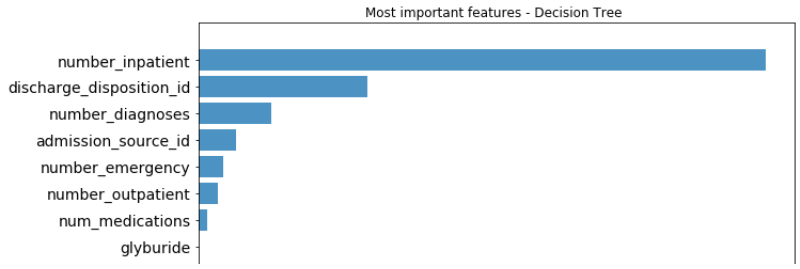


Figure 8. Feature Importance of Decision Tree

94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang



No model is a perfect model. No model will perfectly identify all individuals that have a high risk of readmission. Consequently, it is key to communicate to hospitals and physicians model performance (e.g., accuracy, recall rate). This model should be used to inform and improve their decision-making, but recognizing the limitations of our model will convey the importance of having human input. Furthermore, by communicating model performance, stakeholders can indicate whether this model is "good enough" or whether the model needs to be refined. Refining the model to have better performance would require feedback and input from the physicians themselves - again highlighting the importance of having a model with limited complexity that is easy to interpret.

Method of Communication: Visualizations

To communicate with doctors, who may not be familiar with data manipulations, making the process more interactive and simple is our goal. Therefore, we would like to propose three phases that could make use of the analytics project output.

1. The first phase would be a questionnaire/survey for the doctor to fill in a patient's health information. The questionnaire would follow the decision tree branches (Figure 1) and once there is a predicted result, the doctor would be informed whether this patient is predicted to be readmitted or not.
2. The second phase would be a well-developed chatbot. A chatbot is an interactive tool with a decision tree embedded so that the doctors could easily interact and get the predicted results; it would also provide a better user experience, as each step of the decision process would be clear and traceable.
3. The last phase would be implementing the decision tree model with voice recognition devices; we imagine the device actively listens to the conversations between patients and doctors (with the patient's permission) and could catch keywords that are related to the decision tree models. At this phase, the doctor would not have to take additional actions, such as typing in health records, to get the prediction of whether the patient that he/ she is talking to is likely to be readmitted or not.

These would be the applications and visualizations that would be useful for the doctors as the decision-generation process could be demonstrated and understood.

Stages and Milestones

In order to have a better blueprint and ensure the process of the project, it is important to identify what are the activities needed to reach goals along the process and also the potential milestones resulting from each stage of activities. With a clear deliverable, we can make sure that the goals are aligned with the stakeholders' and make sure the results meet the expectation of the client at each stage. The following is a table of milestones and deliverables that we planned to achieve:

	Milestones	Deliverables
Business Understanding	1. Have conversations with stakeholder aimed to (a) Identify problem (b) Identify project goals & relevant timeframe (c) Identify other possible data sources Understand how the model will be deployed and used. Are there any resource constraints? Should we identify all patients that are at risk of readmission, if so, does the hospital have the resources to take action on all of these patients?	☐ Proposed Project Plan
Data Exploration	1. Data Access & Data Cleaning 2. Feature Generation/Selection	☐ Clean and complete source data set ☐ Data Dictionary ☐ Correlations/Trends among key variables ☐ Analytical Formulation[5] ☐ Evaluation metrics ☐ Baseline Comparisons

[5] Analytical formulation would map the scope of the project to a machine learning problem and analytical approach guided by how the model will be deployed and used.

94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang

Data Modeling and Analytics	1. Develop classification models 2. Validate & Evaluate Models	☐ Final classification model ☐ Performance evaluation (evaluation metrics & baseline comparison and visualizations)
Communication & Next Steps	1. Translate model results into insights, avoiding the use of technical language 2. Discussion deployment efforts (software needed, model update frequency)	☐ Final visualizations ☐ Final report

Project StakeHolders

There are two major stakeholders: Doctors and Hospitals. Since the doctors are in first contact with the patients along with their professional expertise, they are ranked with high influence on the project and we expect to have frequent conversations with them. Hospitals are the ones that are taking up the cost and providing the basic infrastructures and therefore hold a high impact of the project. However, they have less input for the project and therefore the meetings with them can be less frequent.

Stakeholder	Role on the Project or Within the Organization	Contribution to the Project	Project Influence	Communication Plan (Frequency and Method)	Person Responsible
Doctors	The doctors have the role of implementation. Doctors will receive the list of at high risk patient and choose right plan of action	Doctors' prior knowledge will prove valuable for getting feedback on our models. In essence they can provide a “sanity” check of our models.	High	Weekly meetings, in-person	Chief Medical Officer

Hospital's Administrative Unit	Organization itself is responsible for having the data infrastructure needed to store and maintain data.	The infrastructure that each hospital has, will affect the method of deployment for our model. Moreover, the organization itself will be responsible for the added cost of maintaining and updating our model.	High	Bi-weekly, in-person	Chief Data Officer

Infrastructure (Hardware and Software) Resources

- Hardware and costs
    - Computers for staff: no additional cost (already have - hospital computer or devices)
    - Devices to record and enter patients information: no additional cost (already have - hospital computer or devices)

- Cloud services: As the model and analytics result would be used by multiple doctors, we suggest using cloud services to manage datasets and machine learning models. Also, the doctors could access the tool when they are moving around in the hospital. The data stored in the cloud would be anonymized before uploading and being used to train the model.
    - Amazon Web Services (AWS):

Service	Purpose	Cost
S3[6]	Data storage	$100 per year
SageMaker[7]	building, training, and deploying a machine learning model	$450 per year

Staffing Plan

There are four different positions that need to be hired in order to make sure the project can be done:

[6] Assume 256GB of storage needed; https://aws.amazon.com/s3/pricing/
[7] Estimated using "ml.t3.medium" instance; https://aws.amazon.com/sagemaker/pricing/

94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang

Roles	Description	Skills required	Utilization
Project Manager	● Communicate with external stakeholder and internal team to manage project scope and task priority ● Proactively identify and resolve potential issues that may negatively impact team members and the project to mitigate risks ● Lead process improvements to improve productivity and quality ● Provide knowledge in hospital operations and treating patients ● Explain and communicate medical terms to non-med team members ● Define hypothesis and collaborate with data analyst for testing	Excel, Powerpoint, Project Planning tools and methods, medical background	Full-time; 40 hrs per week
Data Scientist	● Ability to analyze dataset and make recommendations on model selection ● Identify and aggregate data for analysis ● Apply and update machine learning models to improve business objectives ● Develop reliable, scalable, and robust code for production ● Identify potential ways to Improve model performance and develop relevant code	SQL, Python, Excel	Full-time; 40 hrs per week
Data Analyst	● Recommend data manipulation and feature engineering possibilities	SQL, Python, Excel, Powerpoint	Full-time; 40 hrs per week

	• Conduct exploratory data analysis and further define & test hypothesis • Interpret the results and present summary of analysis to key stakeholders • Design and develop data visualization tools to communicate the results to stakeholder		
Data Engineer	• Data cleaning and validation • Manage data pipeline and ensure data quality • Work with existing infrastructure • Deploy model into cloud services to allow broader use of project result • Developed version control for the models and dataset being used	Python, Database Management, Cloud services platforms (e.g. AWS, Azure, GCP), Version control tools	Part-time; 20 hrs per week

Risk Management

Every project has different risks. It is important to anticipate potential risks so that we can come up with responses if any of those risks are to happen. Below are the potential risks associated with the project:

Description of Risk	Likelihood	Severity	Response	Responsibility
Doctor training: After the model has been developed, It will require additional communication effort to ensure all concerns are addressed. In addition, as it is a new way of working for doctors, training sessions and workshops should be held to communicate with new	High	Medium	Prepare a series of workshops, first target the “early adopters” and gather feedback on the model and prediction. Later, conduct workshops for more new doctors after	Project Manager

doctors to have them onboard.			tackling common concerns and doubts. Also, provide booklets and flyers describing detailed rationale about the data and features used to develop the model so that those who are interested in the details could access such detailed information.	
Require maintenance to continuously improve the model: Once the model has been deployed, new patients data should also be integrated and used as new input to the model. It will require efforts from data engineer and data scientist if the new data collected have different data format or structure.	Medium	Low	Clearly document data input format and steps to take in case there are changes in data format, so that the data pipeline could continuously operate and improve.	Data Scientist, data engineers
Project staff may change and the risk of a new member having the capability to pick up from scratches	High	Medium	Document critical domain-specific information and rationale behind decisions made. Schedule meetings between past members and new members to help answer and clarify if there are any questions.	Every team member

Success Measurements and Project Impact

94881 MAP F21 - Final Paper

Angel Lee, Lidia Ortiz Zamora, Yun-Yung Wang

The purpose of our analysis is to identify high-risk patients that result in hospital readmission so that the doctors can take precaution of the high-risk patients and lower the readmission rate.

Firstly, since we applied multiple categorical machine learning model we should examine the effectiveness of our model by

1. Predictive accuracy and recall rate
2. AUC (Area Under The Curve) & ROC (Receiver Operating Characteristics)

Also, the corresponding success measurements should be centered around whether the readmission rate does decline. Therefore, the relevant metrix of identifying success are the following:

1. The percentage change in readmission rate over time
2. The cost reduction associated with the decline of readmission rate over time

Besides, other than a hospital prospect, it is crucial to make doctors respond as an important indicator of success measurement as well because doctors are the ones that have the clearest picture of the condition of the patient and the ones that give direct professional advice and instructions to the patients. In addition, with the quantified prediction result, it can also improve the communication between the doctor and the patient and therefore yield better instruction for patients' self-management and increase the credibility of doctors' instructions. Therefore the following are also metrics that we should take into determining the success of the project.

1. Efficiency of doctors making patient's condition judgement
2. The clearness of communication between the doctor and the patient
3. The confidence level of the patient to doctor's instructions

Lessons Learned

There are a few things we learned throughout the project:

1. Understanding of the dataset

    Before starting the exploratory data analysis process, we did not clarify what all the columns meant (eg. the type/ purpose of the medicine) and therefore we tried to explain the feature impact backwards after seeing the exploration results. In order to make better planning and a more precise hypothesis, we should do further research on the column features that we are not familiar with. It would have been better if we had more time to conduct exploratory data analysis to generate and test different hypotheses before further developing a model as we would be able to understand the business problem more comprehensively.

2. Understanding the need of our stakeholder (doctor and hospital)

   In order to create a model that is most relevant to the decision makers, it is important that we fully understand the needs of the doctors. We should have a better picture of what are the criteria that the doctor cares the most while giving suggestions to a patient. Therefore, having a deeper conversation with a doctor to understand their perspective is also an important topic while doing any analysis and modeling.

Limitations and Future Research

- Data quality: we cannot put aside the fact that a large number of patients with complex health problems have a higher probability of being misdiagnosed and are given the incorrect medication and therefore provide inaccurate data of the patient. Also, it is likely that different doctors give inconsistent diagnosis on patients with the same condition.
- Insufficient data: There are many other factors that may highly affect (e.g., demographics) whether a patient would get readmitted but will not be kept recorded in medical history. For example, elderly patients who live alone compared to those who live with others may be more likely to have inappropriate home care and therefore may result in readmission to the hospital. We should continue to seek datasets or include new features to help with the analytical project.