Predicting Health

Mining Patient Data To Gain Invaluable Insights

Anthony Le
CSPB 4502 Summer 2024
University of Colorado Boulder
Boulder, CO, USA
anle4634@colorado.edu

Problem Statement/Motivation

The healthcare industry is constantly seeking ways to improve patient care and operational efficiency. Predictive modeling in healthcare analytics offers a promising avenue for achieving these goals. By leveraging large datasets and advanced machine learning techniques, predictive models can help healthcare providers anticipate patient outcomes, estimate treatment costs, predict hospitalization durations and more, helping augment the patient experience and lessen any uncertainty the patient may have about their prognosis.

For my data mining project, I aim to build a predictive model using the SyntheticMass dataset (further explanation and analysis of this dataset in a further section). As mentioned previously, there are a variety of interesting questions that can be answered. How accurately can we predict patient hospitalization durations based on health indicators and demographics? What are the key factors influencing treatment costs, and how can they be predicted?

By answering these questions, I hope to build insight into the growing field of healthcare analytics and build a better understanding of the mechanics of predictive models that can help aid in providing powerful and useful insights for both healthcare providers and patients alike.

Literature Survey

The field of healthcare analytics has undergone significant transformation due to advancements in machine learning and artificial intelligence. These technologies have helped enable more precise and predictive models, augmenting healthcare providers' ability to offer personalized and efficient care.

I have identified three pieces of literature that cover different studies and initiatives that have helped me align my approach towards this project.

1 AI models predicting patient response to immunotherapy

GE Healthcare & Vanderbilt University Medical Center (VUMC) partnered together to develop AI models that can be utilized to predict how a patient would respond to immunotherapy. This study demonstrated the potential machine learning has in being able to help develop personalized treatment plans for patients, with this study diving more into patients undergoing cancer treatment. The predictive models developed in this study utilized a variety of patient-specific data that ranged from genetic information and more to help improve that accuracy of the predictions being made by the models.

This study underscores the importance of using detailed and personalized data and how it can help in predictive modeling. For my project, I can identify the benefits of incorporating data related to patient demographic and other healthcare indicators as variables that can help provide a good basis to train my model on. The success of the approach taken by the team involved in this study informs my decision to use a wide range of patient-specific features and to not

2 Cloud-Based Predictive Modeling in the UK

A small team in the UK was able to develop a cloud-based predictive modeling solution in a relatively timely manner to enhance healthcare services. This initiative aimed to improve operational efficiency and patient care by leveraging scalable, cloud-based technologies.

This article highlights the scalability of predictive models and the importance of factoring and leveraging various technologies, such as cloud-based technologies, to help build out a successful solution. Although my project will be executed primarily locally, keeping in mind the principles of scalability and efficient resource management is crucial. The large size of the dataset of interest requires efficient handling and processing - finding and researching various technologies and libraries out there will help build a good focus on robust data preprocessing and model optimization techniques to utilize for the project

3 Predictive Analytics and Social Determinants

Research has shown that social determinants of health, from socioeconomic status to education and even access to healthcare, significantly impact patient outcomes. Predictive analytics can help identify those determinants and their overall effects, allowing for more targeted interventions and allow proactive measures to be taken to help ensure patients have success in navigating finding the care they deserve. This study demonstrated how incorporating social factors into predictive models can lead to more comprehensive and effective healthcare solutions.

Incorporating a wide range of factors, including social determinants, can be crucial for building accurate and meaningful predictive models. Incorporating a dataset that includes demographic variables and other relevant health indicators can help build and ensure a holistic approach to prediction - this study in particular informs my selection process, emphasizing the need to consider various determinants of health outside of what one would typically anticipate.

By building on the methodologies and findings of these studies, I aim to develop a robust predictive model that leverages my dataset of interest - incorporating personalized patient data, scalable processing techniques, and a comprehensive feature set can help enable me to address key healthcare questions and gain better insight on how contributions are built within the field of healthcare analytics.

Data set

Given the nature of HIPAA (Health Insurance Portability and Accountability Act) regulations, it is very difficult to obtain dataset that has patient electronic health record (EHR) - in this case, a lot of professionals that are want to explore healthcare analytics outside of a hospital setting typically employ the use of datasets that has numerous synthetically generated patient EHRs. Despite the synthetic nature of these datasets, they can still offer invaluable insights into healthcare analysis and patient records.

Keeping that in mind, the dataset I am choosing to work with for this project is supplied by SyntheticMass. This dataset is an extensive collection of synthetic healthcare data, comprising 1,000,000 records. All of these records are divided into 12 output folders, each containing various data formats such as JSON and CSV. For this project, I will focus on the CSV files, which provide structured data suitable for machine learning models.

Focusing on the CSV files, there are 10 CSV files, named as such: “allergies”, “careplans”, “conditions”, “encounters”, “immunizations”, “medications”, “observations”, “patients”, and “procedures”. This dataset is rich in providing various synthetic data related to various aspects of patient and patient care that can be beneficial to explore to identify what will be helpful in processing for a predictive model.

Given the synthetic nature of the data set, there are a lot of considerations I’ll keep regarding missing and inconsistent values, which will be better outlined in the following section. To give a brief however, when it comes to CSVs with rows containing null or missing values, it is beneficial to consider how many values are affected - if there are not too many, it can be easy to try and replace them with an appropriate approximation, but if a row has too many affected values, dropping the row may be better as a whole as that row, which represents a patient, will have much more “simulated” values vs. a row with less affected values.

Proposed Work

To achieve the goals of my project, I will undertake the following steps:

1 **Data Collection and Preprocessing**

Dataset will be downloaded from SyntheticMass - as this is a dataset that contains 1,000,000 synthetically generated patient records, that download will come in a large zip file that will be unzipped to give access to all of the files inside. Given that the dataset comes in a wide variety of different file formats, such as JSON, CSV, and more, I will go in and delete files not needed as I am interested in working with the CSV files, allowing me to trim out the other unnecessary files. Further preprocessing will be needed with diving into the CSV files and identifying missing or null values; depending on the quantity of such values, I’ll either implement imputation to replace these values with either the mean, median, or mode of the column to help preserve the 1,000,000 record count or deletion of the record which could reduce the overall number of available records. Moving beyond this, I will need to identify and remove any duplicate records to help ensure data integrity. Standardization of the format of values can help with ensuring consistent capitalization and correcting any misspellings that may be present that can introduce inaccuracy to the predictive model (cancer vs. cancr, DIAbetes vs. diABetEs, etc.). Outliers can be present within the dataset that can largely skew predictions and will need to be accounted for; both the z-score and IQR can help provide a range of values that can be considered within the norm and any values identified as outside of that range can be either removed or transformed through imputation or other methods. As the last step to consider in the preprocessing, to aid in feeding data to the model, I could apply one-hot encoding to categorical values to give them a tangible integer value that is logical for processing by the model. Optionally, there may be some CSV files that may not be utilized and as such, could be trimmed out after all of these steps or even before to help save on time spent and overall memory space requirements on my local machine (“careplans” may not come about in the questions I aim to ask and answer; the goal should be not to delete anything as the model could be adapted to answer other questions if trained on all available data).

2 Model Development

I will explore into more literature that is available on predictive models and choose the appropriate algorithms that can apply for my project, whether it be logistic regression, decision trees, or random forests - I'll need to keep my own expectations and aspirations for my model within a workable range where I don't try to overextend myself too much with building out something too complicated to complete. After exploring literature, I can begin to build and train a predictive model using the preprocessed SyntheticMass dataset. Additional optimization of the model's performance can come through the utilization of hyperparameter tuning and cross-validation methods.

There are a variety of different algorithms that exist with literature supporting them that can be worth exploring for my project, such as linear regression, random forest, and gradient boosting.

Linear regression models the relationship between a dependent variable and one or more independent variables, assuming a linear relationship between inputs and outputs. This type of model is usually easier to understand, interpret, and implement and can come across as being computationally less intensive compared to other more complex models. These traits can help with making predictions on hospitalization durations or treatment costs where the end output is to estimate a numeric value. However, a shortfall would be that if a nonlinear relationship exists in our data, the model wouldn't be effective in that case.

Random forest is built upon multiple decision trees that are merged to achieve better accuracy and stability related to making predictions. This algorithm is rather robust and can handle overfitting (the phenomenon where an algorithm can fit too closely to the training model) better than that of a single decision tree. There is some versatility where the algorithm can be used well in tasks related to classification and regression and for the project, can be helpful in predicting treatment costs and patient outcomes through handling complex interactions between features. However, they can be harder to interpret compared to a single decision tree and take up a decent amount of computational resources to run.

Gradient boosting is an algorithm that builds models sequentially and iteratively - each new model that is generated corrects errors made by the previous model and through this process, works to combine the predictions of multiple weaker learners to produce a strong learner that is capable of making good and accurate predictions. This algorithm has a variety of advantages with it that range from offering higher predictive performance to being flexible in the tasks it can handle (similar to random forest). For the project, it can be suitable for predicting patient outcomes and treatment costs where high prediction accuracy is much more paramount. A caveat though is that training time can be rather slow and for larger datasets, this can be a huge slump in the project progression as a hangup.

Choosing the right algorithm will depend on the specific needs I have for the project, taking into account the complexity of the data, importance of the interpretability of the results, and the computational resources I have available on my own local machine.

3 Comparison with Previous Work

One of the biggest differences between my work and previous work is the usage of synthetic data vs. real data - these previous works were in a space where they had ample access to real data to help in the development of their own models but for my particular use case, I wasn't in a space where I can get access to real patient data without violation of various regulations and laws in place to protect patient privacy and compliance. These previous works also produce results for a more targeted or specific application; my model would have a broader scope that can go beyond just predicting patient outcomes but also estimating treatment costs, determining hospitalization duration and more that wasn't answered in the previous literature mentioned. By leveraging synthetic data and addressing multiple healthcare questions, my project builds on existing research with innovative data handling and modeling techniques - this approach aims to contribute valuable insights and tools that help give more in depth insight into the methodology and work required at improving patient care and healthcare operations.

Evaluation Methods

To evaluate the performance of my predictive model, I will use a variety of quantitative metrics provided by the scikit-learn library:

- Accuracy: The proportion of correct predictions made by the model
- Precision: The ability of the model to correctly identify positive instances
- Confusion Matrix: A table that describes the performance of the model through comparison of predicted and actual values
- F1 Score: A metric that balances precision and recall that can be beneficial in imbalance datasets

Additional evaluation methods may need to be employed that goes beyond the metrics above. Since my dataset is synthetic in nature, consideration needs to be taken in about how this dataset was generated and if the population generated is representative of a population that can be encountered in reality. Comparing my results against findings from real healthcare data published in research papers can help provide a useful reference point for evaluating the model's accuracy that can help highlight discrepancies that come from being synthetic. To address the limitations of the synthetic data, I will focus more on relative performance rather than absolute accuracy and highlight the potential areas where synthetic data can fall short and diverge from real-world scenarios.

Considering the metrics mentioned above, there are some thresholds I could consider to help better classify and understand the performance of my model and how the project is going overall:

- Accuracy >= 80%
    - 80% can generally be considered good but there is still a complex of what constitutes an acceptable accuracy threshold; for things related to health this could be higher but for the purpose of exploration and analysis for the course, I find some leniency here can be beneficial
- Precision >= 75%
- F1 > 0.7

Selecting the right metric and threshold for these metrics can be difficult in practice, as generally, the better and higher metrics we strive for, the better confidence we can have in utilizing the model we built. However, for the purpose of this class, giving myself some leniency and give here is beneficial. Approaching this from a standpoint of having an actual deliverable in a production or work setting may see tighter constraints to ensure the most accurate and precise predictions are being made to ensure patients and other professionals are not led astray.

Tools

To help with the development of my project, I plan to employ the following tools:

- Python: A versatile programming language that has access to an extensive collection of libraries to aid in data manipulation and machine learning
  - Pandas and NumPy are particularly popular and powerful libraries that give access to data manipulation methods and numerical operations
  - Scikit-learn will help enable the ability to implement machine learning algorithms and evaluation metrics much more efficiently
  - Matplotlib and Seaborn can help with data visualization and exploratory analysis
  - TensorFlow/PyTorch can be considered for more advanced machine learning model implementation if I choose to go for models and algorithms that go beyond what I initially wish to explore and utilize
- Git/GitHub: For repository management, helping ensure my work is backed up and accessible from anywhere and allowing regression to previous versions should anything go wrong in development

Milestones Guideline

To ensure steady progress, I have defined the following milestones as markers to hit through development:

- Weeks 1-2: Data collection and initial exploration
- Weeks 3-4: Data preprocessing and cleaning
- Weeks 5-7: Model development and initial testing
- Weeks 8-9: Evaluation and comparison with existing solutions
- Weeks 10-11: Final report and presentation preparation

While these are great goals to keep in mind, being flexible can help ensure stress doesn't boil over as unpredictable events or other complexities can occur that skew the progression timeline. Adhering to this timeline can help with the systematic build and evaluation of my predictive model and help culminate in a comprehensive analysis of the SyntheticMass dataset and the overall potential application of synthetic datasets in healthcare analytics.

Milestones Completed

Thus far, I've completed milestones related to:

- Weeks 1-2: Data collection and initial exploration
- Weeks 3-4: Data preprocessing and cleaning
- (Partial) Weeks 5-7: Model development and initial testing
  - Have some headway done head in terms of looking at algorithms and devising which one to try and implement

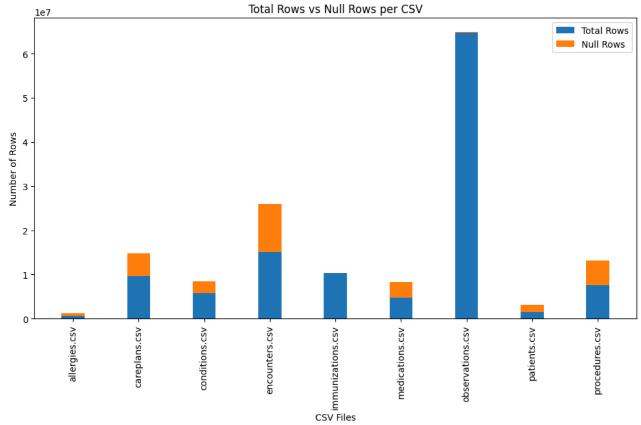
Milestones To-Do

There is still a decent amount left to do, but most of the bulk work felt like it was tied with Weeks 1-4, relating to preprocessing and processing of the dataset

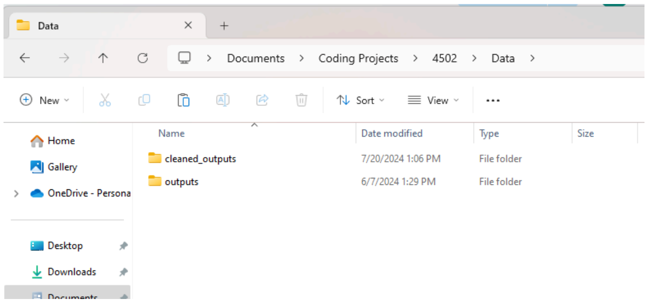
- (Partial) Weeks 5-7: Model development and initial testing
- Weeks 8-9: Evaluation and comparison with existing solutions
- Weeks 10-11: Final report and presentation preparation

Results so far

Don't have too much in the way of results to show, but I do have some visualizations of the dataset in its initial state



This is a stacked bar chart that is the result of iterating through the data set and running a count of how many rows were found with null values (in orange) and how many total rows there are (in blue). Running the calculations, it was estimated that roughly ~25% of the data in CSVs contained null values. This isn't even diving in to see the actual data itself to see how the consistency is with rows that have values in comparison to other rows. This simple visualization helped me to get an estimate of the work I'll need to do to dive into everything.



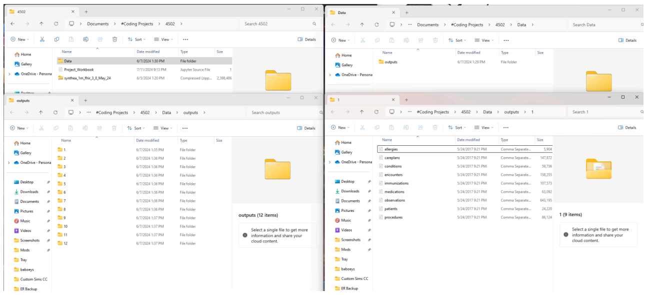
Name	Date modified	Type	Size
cleaned_outputs	7/20/2024 1:06 PM	File folder	
outputs	6/7/2024 1:29 PM	File folder	

I have set up in my notebook cells that will help handle the processing and cleaning of the data, taking the "unclean" data in the original "outputs" folder and generating new CSVs with the "cleaned" data into a new "cleaned_outputs" folder which can act as a new point of reference of training data further along.

Potential Complications and Considerations

As with any project, there are inevitable points during progress where things can go less than stellar for a variety of reasons:

- Complexity of the data set: the synthetic data set was not as straightforward as I would have hoped (which honestly probably reflects the reality of most and all datasets professionals work with) - the 1,000,000 records in the CSVs is split into a variety of different output files that itself is contained within a variety of different output folders



- Life complications also came up that hindered some of the time I wanted to have to dedicate towards the project - accepted a new job opportunity that came with a decently rigorous on-boarding experience of dedicated readings, lessons, and shadowings that is a 6 week long process along with some other personal matters in the family life
- Studying for the other material in the course

With these complications, it can feel like the end goal of this project is feeling more and more stretched and after reviewing everything I have so far, I may need to consider pivoting and build a new focus for this project; I'm putting in time to source an alternative data set for use of this project. The structure of the data set make it more complicated than I'd imagined originally at the beginning - multiple CSVs covering various attributes of interest (allergies.csv, careplans.csv, conditions.csv, etc.; 9 CSVs in total that are found in 12 output folders in total that divide and compartmentalize the 1,000,000 patient records of interest). Although this is a hit on my confidence, I'm still sticking to finding a dataset that is related to health so that I am not stretching too far from my original vision and passion. Two data sets I've found so far that can be an interesting application to a healthcare related predictive model is:

- Data set covering chronic disease indicators that can be utilized to predict whether some one may be disposed to developing a particular disease or, if I dive more into analytics, uncovering interesting trends related to demography and disease outcome
- Data set covering various characteristics related to medical appointments that can be utilized to help build a model for professionals to predict if a given patient may end up being a no-show appointment and if additional out-reach to these patients is needed to ensure they show up to their appointment
- Additionally, I understand that there is a lot of datasets out there that pertains to COVID-19 that could be used and applied in a variety of ways; predicting how a patient might respond to treatment related to curing/alleviating the diseases, how severe a patient with the disease might feel, and more

This can be a rather large undertaking to pivot in such a way but since I'm able to recognize this and have some additional research from before that I can follow-up on again, it shouldn't be too difficult to really dedicate time to find a new and simpler data set to utilize for the project.

REFERENCES

[1] Vanderbilt University Medical Center. 2024. GE Healthcare, Vanderbilt Public Data on AI Models Predicting Patient Response to Immunotherapy. Retrieved from https://news.vumc.org/2024/03/05/ge-healthcare-vanderbilt-publish-data-on-ai-models-predicting-patient-response-to-immunotherapy/

[2] Amazon Web Services Public Sector Blog. 2024. One Small Team Created Cloud-Based Predictive Modeling Solution to Improve Healthcare Services in the UK. Retrieved from https://aws.amazon.com/blogs/publicsector/one-small-team-created-cloud-based-predictive-modeling-solution-improve-healthcare-services-uk/.

[3] Health IT Analytics. 2024. Improving Social Determinants of Health with Predictive Analytics. Retrieved from https://healthitanalytics.com/news/improving-social-determinates-of-health-with-predictive-analytics.