ggk tech

Data Science to Improve Healthcare Outcomes

Predict possibility of patient readmission

WHITE PAPER

Table of contents

Executive Summary

Healthcare industry has immense opportunities to leverage patient data to improve the quality of care provided. The cost of providing care has gone up significantly over the last couple of decades in the US. All entities in the ecosystem such as Payers and Providers have to figure means to reduce their costs. Electronic Health Records present such an opportunity. They represent the historical evolution of care delivery records by health professionals from paper-based handwritten formats to electronic formats. They contain information which can be classified into categories like admissions, lab measurements, care-giver details, demographic, etc.

The Whitepaper covers how Machine Learning on attributes from electronic health records can help derive useful insights and how these insights can help Providers identify those patients who are at risk of getting readmitted and take appropriate measures.

Lead Authors

Manas Pant Manager
L Antony Shajin Sr. Software Engineer

Contributors

P Prathyusha Data Scientist
K Lakshya Data Scientist
L Akshay Chandra Data Scientist

About GGK

GGK Tech is a global IT consulting and services firm that provides Software development, Product engineering, Cloud, DevOps, IoT, AI and Blockchain services to enterprise clients around the world.
Our Vision is to be the world's preferred technology company known for superlative software quality and commitment. Our passion for leading edge technologies and our unwavering commitment to deliver exceptional business value makes us the partner choice to several global customers since 2004. As a testimony to our exponential growth, for three consecutive years, GGK Tech has been named among the fastest growing technology companies in India & APAC by Deloitte.

Introduction

United States spending on health care as percentage of GDP is currently 21%. It is highest in the world. But the quality of population's health hasn't proportionately improved. Despite spending more on health care, Americans had poor health outcomes, including shorter life expectancy and greater prevalence of chronic conditions. It ranks 30th in average life expectancy which is lower than countries like Costa Rica which spends fraction of US per capita spend on healthcare. According to OECD data, US has a higher percentage of people who go without health insurance than at least 12 other comparable countries. For every 100,000 people who died in the US before age 75, 112 people died from complications or conditions that could have been avoided with timely and effective care. A study published by Journal of American Medical Association revealed that chronic diseases like diabetes, ischemic heart disease and low back and neck pain accounted for maximum spend year on year by disease category. In the year 2013, ischemic heart disease accounted for the second-highest amount of health care spending with spending of $88.1 billion.

For every 100,000 people who died in the US before age 75, 112 people died from complications or conditions that could have been avoided with timely and effective care.

EHR a quick introduction

GGK Technologies has over a decade's experience in US healthcare space, handling various functions for Payers and Providers. One of the key artifacts of healthcare operations at Payers and Providers is electronic medical data. Electronic Medical Records (EMRs) are digital versions of the paper charts in hospitals. EMRs contain notes and information collected by and for the clinicians in that hospital and are mostly used by providers for diagnosis and treatment. The information captured in these records could be classified into following categories - admissions, services, caregiver details, lab measurements, demographics, medical history, medication and allergies, immunization status, vital signs, personal statistics like age and weight, and billing information. EMRs are more valuable than paper records because they enable providers to track data over time, identify patients for preventive visits and screenings, monitor patients, and improve health care quality. The same information when interconnected with the data from all clinicians relevant for a particular patient, is called Electronic Health Record (EHR).

Over the years we have built a team of subject matter experts who are part of the Healthcare Centre of Excellence (CoE). Armed with both domain knowledge and the technical know-how to manage EHR data, teams in GGK have designed, delivered and managed a number data pipelines for a variety of analytical use cases.

Those can be broadly categorized into two types:

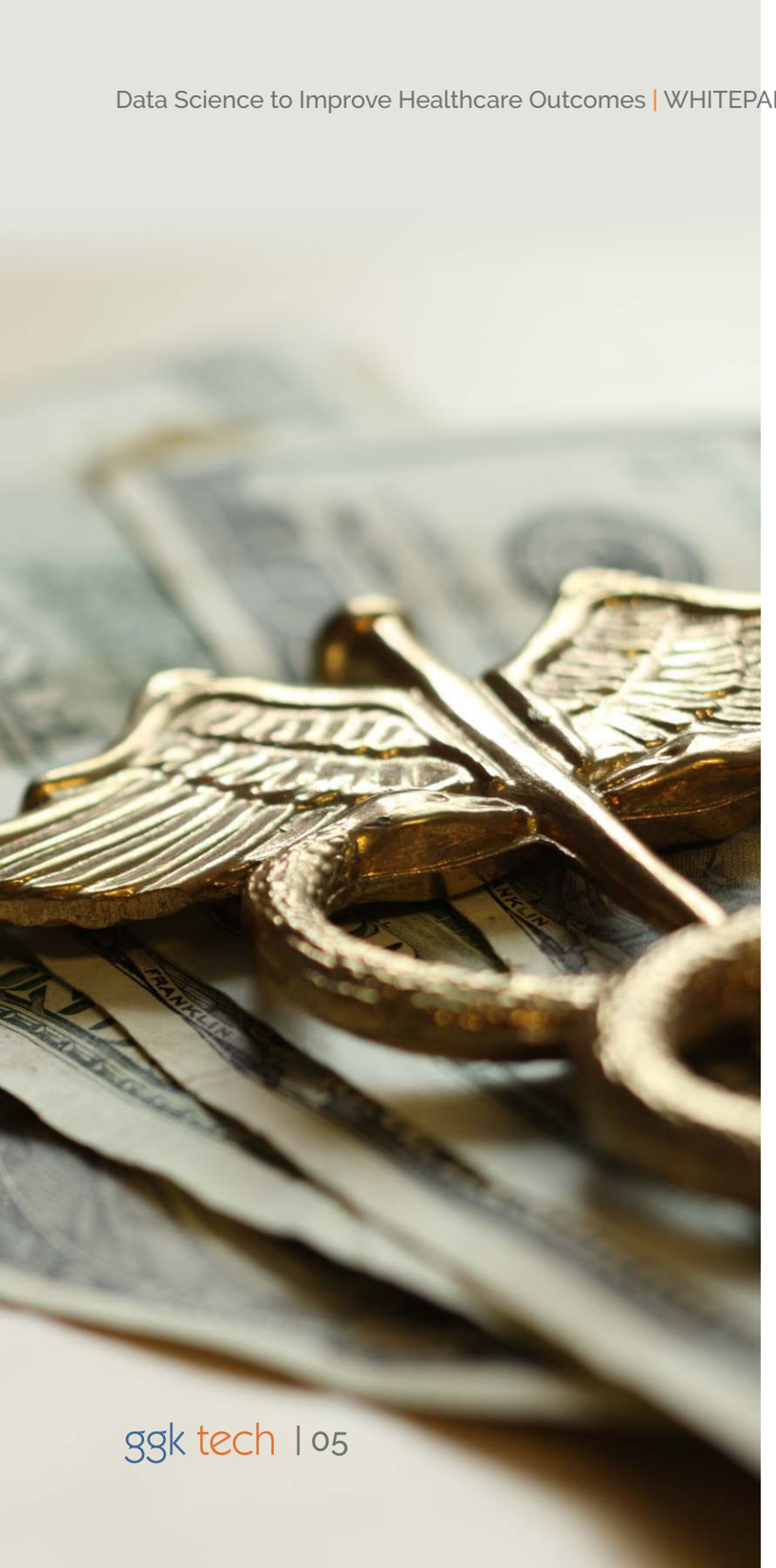
- Benchmarking and decreasing cost of care
- Benchmarking and improving quality of care

The same information when interconnected with the data from all clinicians relevant for a particular patient, is called Electronic Health Record (EHR).

The use cases can be broadly categorized into two types

Benchmarking & decreasing cost of care

Benchmarking & improving quality of care



Our team had significant prior experience implementing descriptive key performance indicators for both types of use cases. Electronic Medical Records, as they have become ubiquitous, present an opportunity to do predictive analytics with use cases such as predicting which patients could go septic, allocate risk score to patients, help maintain apt staffing ratios, predict revenue cycle based patient segments etc. The Medicare Payment Advisory Commission once said that readmissions cost the Medicare program $15 billion, $12 billion of which could be avoided. In such scenario, predictive analytics could be used by hospitals to cut down readmission rates. It can help identify exactly which patients need to be reviewed. Therefore Medicare implemented incentives to reduce hospital readmissions. One of those is the Hospital Readmission Reduction Program (HRRP), which financially penalizes hospitals with relatively high rates of Medicare readmissions.

We narrowed our scope to following problem statements where we predict the following for heart patients:

- Possibility of patient readmission – Post-discharge, will the patient be readmitted to hospital in the next 0-30 days
- Length of stay – Number of days a patient may stay in the hospital during his or her current visit
- List of diagnosis the patient may go through during a revisit – If a patient is indeed re-admitted to the hospital, which additional diseases or ailments could be found

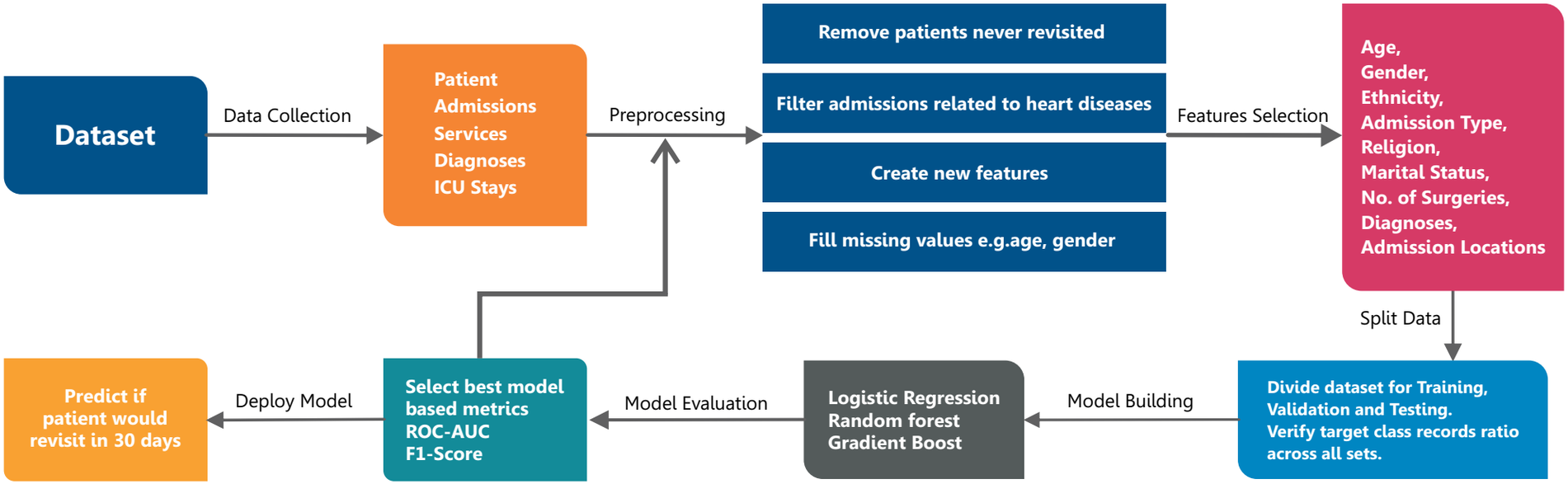
We used de-identified dataset which has records of 40,000 patients. It consists of structured and unstructured data with a number of attributes describing patients, admissions, ICU stay, services, medications, test results, procedures performed and free text notes.

We did the following to build an ensemble of machine learning models...

Predict possibility of heart disease patient readmission:

We went through metadata of the dataset and selected files patients, admissions, services, ICU and diagnosis which covered more relevant information for predicting patient readmission and then loaded all files using Python Pandas library into individual dataframes. Dataframe is an abstraction similar to a SQL table. We applied various filter criteria on admissions data such like removal of admissions which are not related to heart diseases and removal of patients who never revisited. Combining all other dataframes with final admission records, we created a single view for all columns/features. This is a machine learning classification problem, since we wanted to predict if patient would be readmitted or not. Classification is a method in machine learning to determine category, class or type of an item or row of data e.g. labelling email as spam or non-spam, identifying sentence sentiment as positive or negative. We calculated number of days between two admissions and then created our target variable READMISSION which represents if a patient was readmitted within 30 days or not. 1 - Readmitted or 0 - Not Readmitted. We also created a set of new features like Age to capture the difference between admission date and date of birth, number of visits to ICU and admission hour of the day, etc. Then we selected the features which were indicators for patient readmission in this dataset, like admission type, admission location, gender, ethnicity, marital status, length of stay, religion, insurance, number of surgeries, last care unit, diagnosis.

We divided the final dataset into training and testing in 80%, 20% ratio and then we did 10 fold cross validation. At times, the machine learning models work better with training data set but would fail to achieve good accuracy once deployed on production. Cross validation is a technique for validating stability of machine learning model for unseen data. As our target variable had two classes (1 or 0), we validated all the two sets such that they have enough ratio of records for both classes to evaluate stability of training model. We chose an ensemble of classification algorithms which included Logistic Regression, Random Forest and Gradient Boost to build machine learning models. We fed the training dataset with various input parameters using GridSearchCV to build multiple models and then selected the best model based on Accuracy and ROC-AUC values. We are able to predict 70% of admissions correctly. We then deployed the best final model on a server using Flask web framework to integrate the functionality with hospital application which gives feedback in real-time.

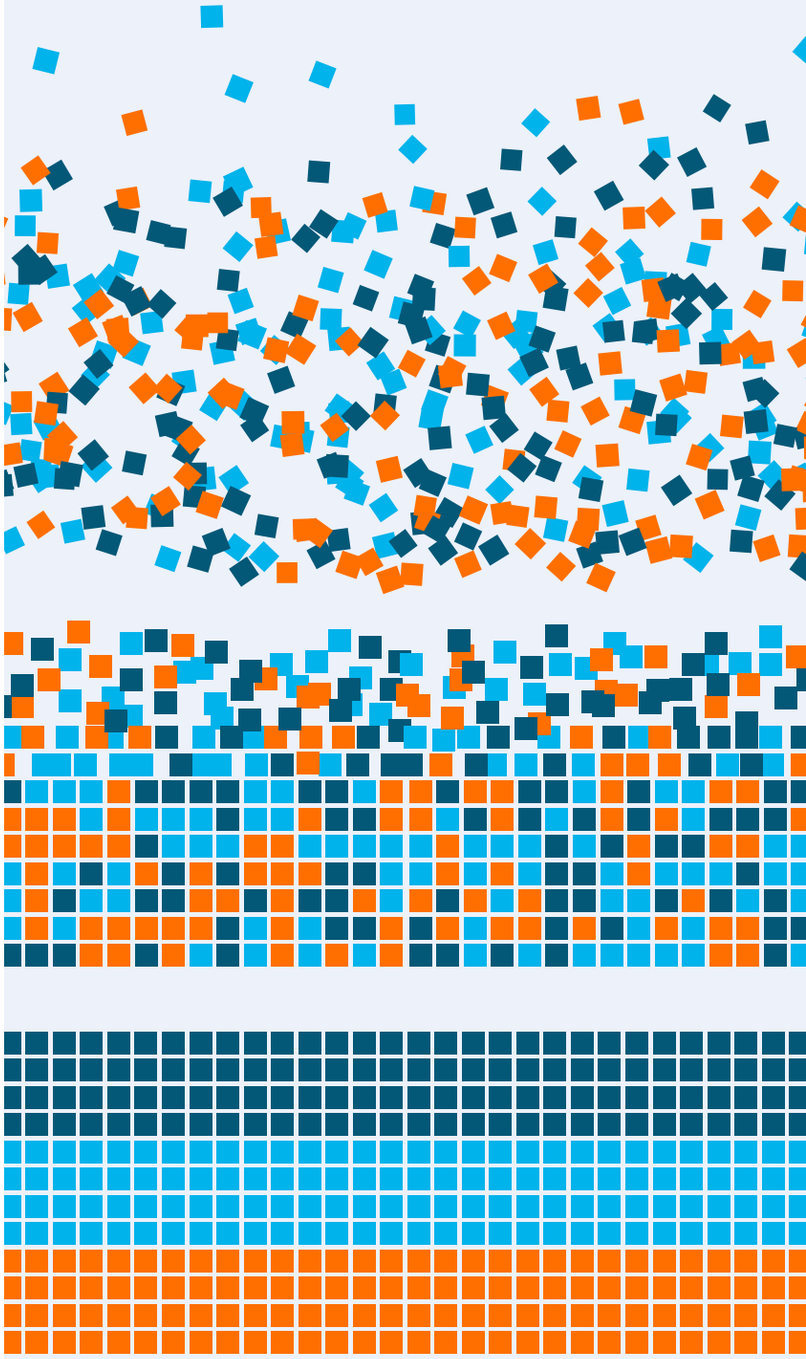


Predict length of stay:

We built another model which predicts number of days a patient may stay in hospital for the current visit. We performed the same procedures as described above with changes in preprocessing techniques, number of features/columns and training algorithms. We considered all the admissions except of type 'new born' and selected features which are captured in hospitals at the time of admission. We used stacking ensemble with Ridge Regression and Gradient Boost algorithms.

Predict list of diagnosis for revisit:

We built another custom model to predict list of diagnoses a patient may go through during his or her next revisit. We combined concepts and algorithms from Recommender System and Rule Based data mining. The entire algorithm is broken into two parts, Grouping and Prediction. As part of Grouping, we used an intelligent filtering algorithm that uses different distance/similarity metrics defined by us namely, Age Similarity, Gender Similarity and ICD Similarity between any two patients. All similarity metrics range from 0 to 1 with 0 being no similarity. Here, the ICD Similarity is nothing but how similar the medical histories of two patients are, and it is calculated by implementing Tree based Nearest-Common-Ancestor algorithm. The Prediction is done by applying Apriori, a Data Mining Algorithm which generates Association Rules (A => B), to the entire diagnoses data (diagnoses list of each admission).



We use the Apriori result to predict the ICD codes in a naive way. We simply use the most recurring ICD codes in the filtered group of patients and get their closest associated ICD codes from the Apriori result. We defined an accuracy metric, which considers predicted ICD as true prediction if at least two of the predicted codes are correct. We ran the model for different thresholds of Age, Gender and ICD used in Grouping. Our model was able to get accuracy of 60%.

We set up our Python 3 environment for machine learning with Anaconda distribution and used Spyder IDE for development. We used python libraries such as Scikit-learn, Pandas, NumPy, Matplotlib and Seaborn for various exploratory data analysis, machine learning model building and testing. We used Flask web development framework to deploy the model on Amazon Web Services.

These models are built as accelerators which any hospital could use to increase its operational effectiveness. Improving hospital operational efficiency through data science boils down to applying predictive analytics to improve planning and execution of key care-delivery processes. These include resource utilization that addresses infusion chairs, operating rooms, imaging equipment, and inpatient beds. It also includes staff schedules, patient admittance and discharge. When this is done right, providers see an increase in patient access which means being able to accommodate more patients sooner, and revenue, lower cost, increased asset utilization, and an improved patient experience.

When this is done right, providers see an increase in patient access which means being able to accommodate more patients sooner, and revenue, lower cost, increased asset utilization, and an improved patient experience.