ALLWYN

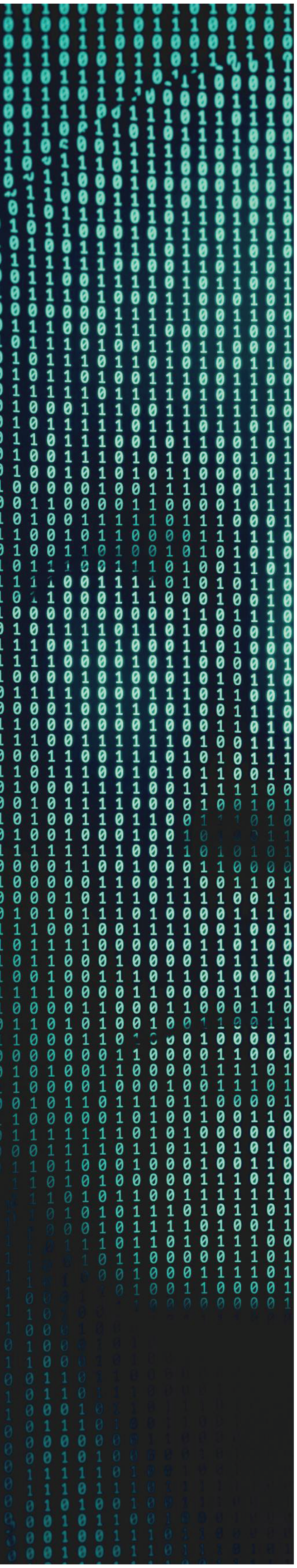
White Paper

Machine Learning and Lung Cancer

Prediction of Probability and Risk Factors of Patients Undergoing Lobectomy

datascience@allwyncorp.com

CONTENTS



Introduction

Readmissions are defined as a patient being admitted to any hospital and for any reason within 30 days of discharge from their hospital. Hospital readmissions for re-occurring problems has been a failure in the healthcare system. Readmissions are often costly and finding the likelihood and the factors driving them can be greatly beneficial for both the public and the healthcare industry.

The Hospital Readmissions Reduction Program (HRRP) by the Centers for Medicare & Medicaid Services (CMS) supports the national goal of improving healthcare for Americans. It is a Medicare program that penalizes hospitals with excessive readmissions. The payment reduction can be as high as 3% for healthcare institutions with high-readmission rates.

Readmission after pulmonary lobectomy is a frequent challenge for hospitals, healthcare plans and insurance providers.

This project focused on studying

- The probability of a patient’s readmission
- Underlying risk factors

The result can be used by various organizations such as hospitals or healthcare companies to take proactive measures and circumvent readmissions.

Our main goal is to reduce readmissions by taking data-driven preventive actions prior to the lobectomy procedure.

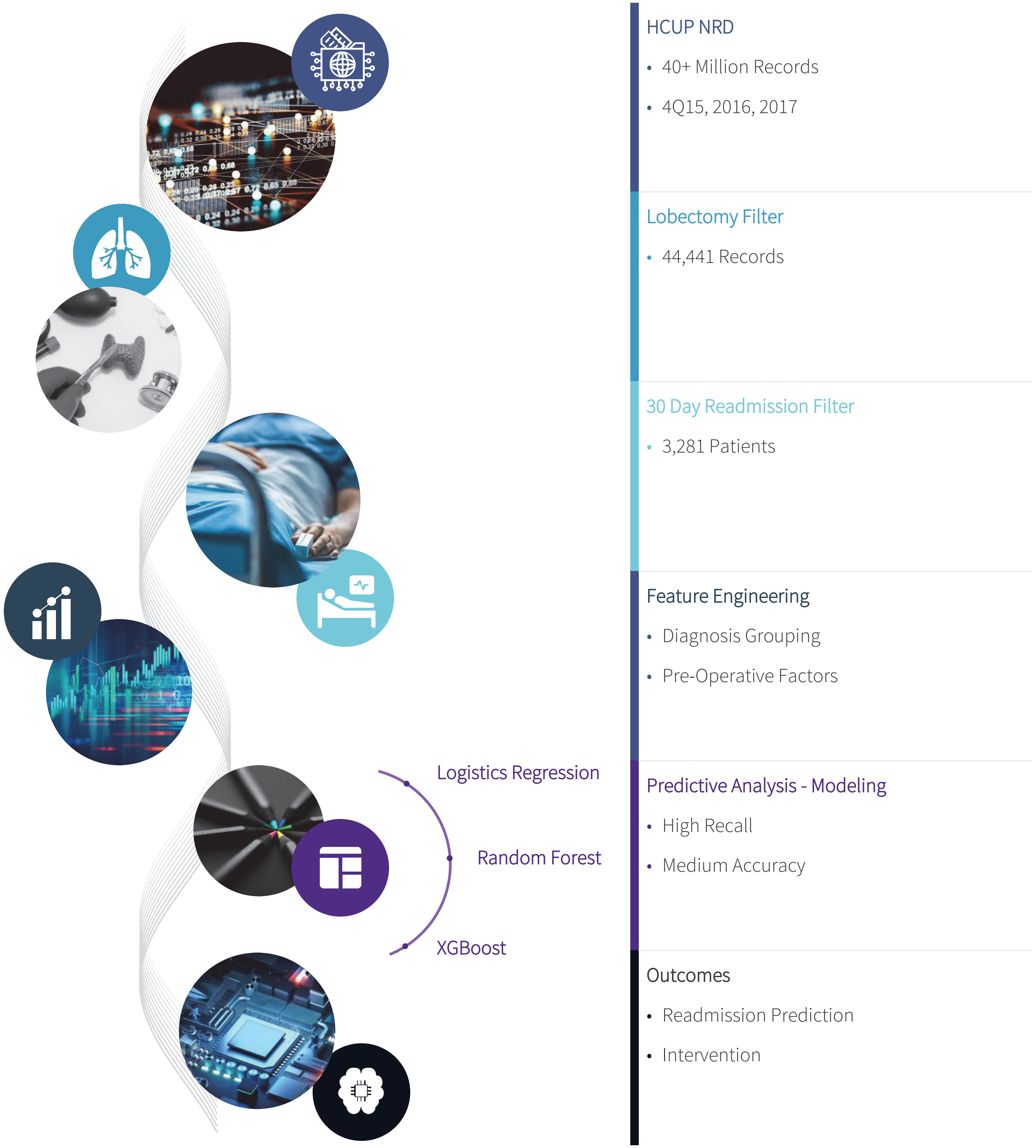
Data Exploration and Engineering

Finding a suitable dataset for machine learning to predict readmission was the first challenging task we had to overcome. In the current state of the healthcare world, available datasets can be either dirty and unstructured or clean but lacking in terms of the information contained. Most patient level data are not publicly available for research due to privacy reasons.

With these limitations in mind, after researching multiple data sources including SEER-MEDICARE, HCUP, and public repositories, we decided to choose the Nationwide Readmissions Database (NRD) from Healthcare Cost and Utilization Project (HCUP).

The HCUP databases are created by the Agency for Healthcare Research and Quality (AHRQ) through a Federal-State-Industry partnership and NRD is a unique database designed to support various types of analyses of national readmission rates for all patients, regardless of the expected payer for the hospital stay.

Our research involved using machine learning and statistical methods to analyze NRD. Data understanding, preparation, and engineering was the most time-consuming phase of this data science project, which took nearly seventy percent of the overall time.



HCUP NRD

- 40+ Million Records
- 4Q15, 2016, 2017

Lobectomy Filter

- 44,441 Records

30 Day Readmission Filter

- 3,281 Patients

Feature Engineering

- Diagnosis Grouping
- Pre-Operative Factors

Predictive Analysis - Modeling

- High Recall
- Medium Accuracy

Logistics Regression

Random Forest

XGBoost

Outcomes

- Readmission Prediction
- Intervention

Figure 1. Data analysis and machine learning process on HCUP NRD

By using big data processing and extraction technologies such as Spark and Python, we filtered near 40 million patient records (Q4 2015, 2016, 2017 years combined) for only the ones who have at least undergone a lobectomy procedure once.

The filtered data was later put through best data quality check processes and cleaned while imputing missing values. We explored more than 100 input variables to find out redundancies, analyze correlations with the outcome, and understand the demographics of our target group. Many of these features were categorical that required additional research and feature engineering.

NRD dataset mainly consists of three main files: Core, Hospital, Severity

Core file mainly includes the patient level medical and non-medical factors. Age, gender, payment category, urban/rural location of patient, residency of the patient in the same state as care provider, total charge, and many more are among the socioeconomic factors of the core file.

Medical factors include detailed information about every single diagnosis code, procedure code, their respective diagnosis related groups (DRG), time of those procedures, yearly quarter of the admission, etc.

Allwyn data engineering practices included analyzing every single feature, researching, and creating data dictionaries and feature transformation to see which features contribute to our prediction algorithms.

With an average age of 65 for lobectomy patients, the data showed that women had more lobectomies than men, more men were readmitted than women.

44,441 Total Population	65.39 Average Age	7.38% Readmitted

54% Female

46% Male

23,818 Population	65.38 Avg. Age	6.72% Readmitted	20623 Population	65.39 Avg. Age	8.15% Readmitted

Figure 2. Distribution of male and female lobectomy patients.

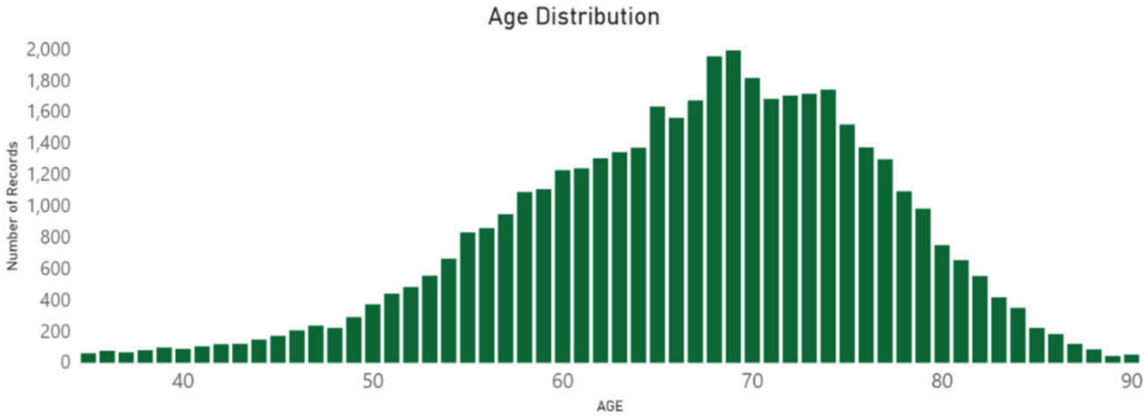
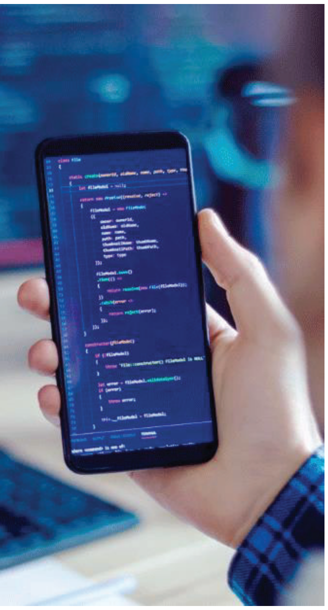


Figure 3. *Distribution of lobectomy patients where age is greater or equal to 35 years.*

Most of the patients who used Medicare to pay were aged 60 and older while most private insurance users were under 70.

Ages	Medicare	Medicaid	Private Insurance	Self	Free	Other
0-9	0	6	13	0	0	1
10-19	0	5	10	0	0	0
20-29	3	10	15	3	0	2
30-39	3	9	27	2	0	0
40-49	15	20	53	3	0	3
50-59	99	108	256	5	4	7
60-69	630	70	278	7	0	24
70-79	1,140	9	81	1	1	16
80-90	329	4	7	0	0	2
Total	2,219	241	740	21	5	55

Table 1. *Payment categories of readmitted patients*

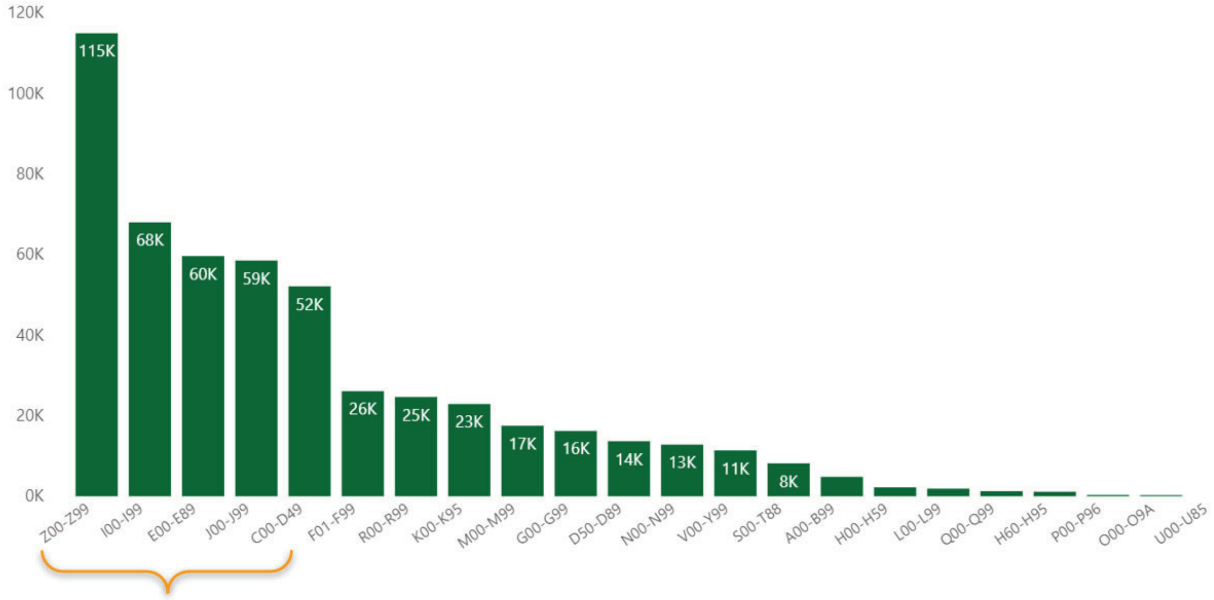


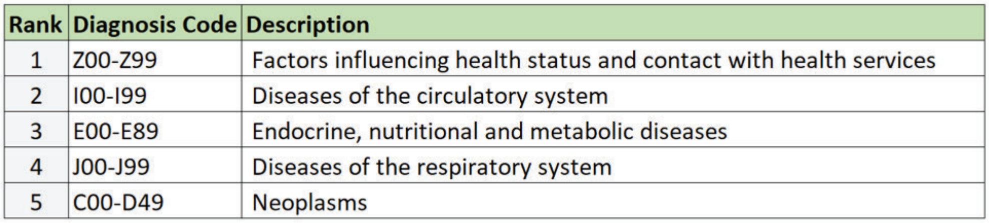
Severity file further provided us the summarized severity level of the diagnosis codes and the **Hospital** dataset presented us information with hospital level information such as bed size, control/ownership of the hospital, urban/rural designation, teaching status of urban hospitals, etc.

We consulted subject matter experts in the lung cancer field and through their advice added additional features such as Elixhauser and Charlson comorbidity indices to enrich our existing dataset. By delving deep into the clinical features, we also made sure the chosen variables are pre-procedure information and verified there is no information leakage from post-operative or known future level variables.

The features were then analyzed to check whether they had statistical significance with our selection of predictive models by looking at correlation matrices and feature importance charts.

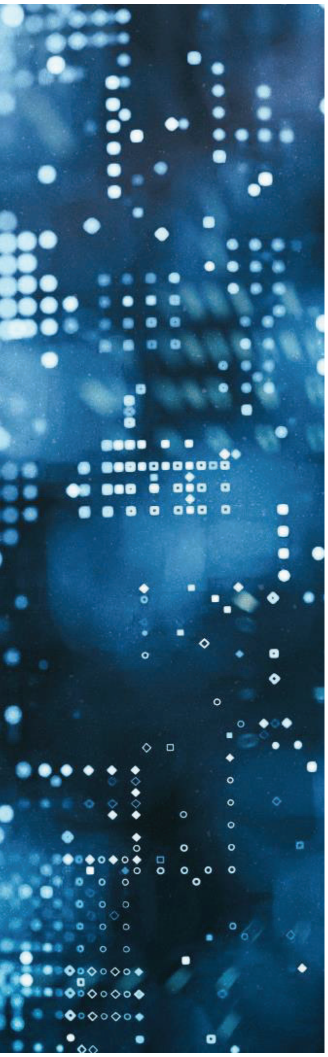
Analyzing the initial data distribution for many of the features required us to remove outliers, transform skewed distributions and scale majority of the features for algorithms that were particularly sensitive to non-normalized variables. Diagnosis codes were grouped into 22 categories to reduce dimensionality and improve interpretation.





Rank	Diagnosis Code	Description
1	Z00-Z99	Factors influencing health status and contact with health services
2	I00-I99	Diseases of the circulatory system
3	E00-E89	Endocrine, nutritional and metabolic diseases
4	J00-J99	Diseases of the respiratory system
5	C00-D49	Neoplasms

Figure 4. Distribution of diagnosis categories for the lobectomy patients



The resulting dataset was highly imbalanced in terms of the readmitted and not readmitted classes, 8% and 92% respectively. Most classification models are extremely sensitive to imbalanced datasets and multiple data balancing techniques such as oversampling the minority class, under sampling the majority class, and Synthetic Minority Oversampling Technique (SMOTE) were used to train our algorithms and compare the outcomes.

Initial machine learning models had both low precision and recall scores. Although this could be due to many different reasons, the Allwyn team focused mainly on additional feature engineering to remove high dimensionality of initial input variables while also comparing different data balancing methods. This was an iterative time-consuming process and required training more than thousand different models on different combinations or groupings of diagnosis codes (shown in Table 2) along with other non-medical factors.

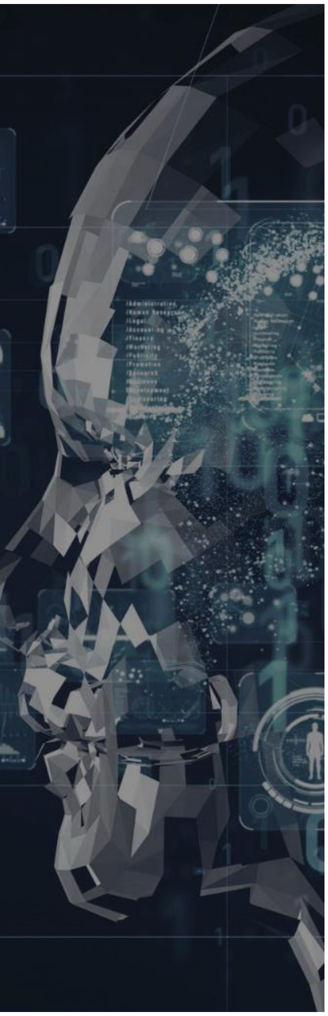
K-fold cross validation was also used during the training and validation to make sure the training results represent the testing. By training models and comparing their validation scores, we weighted the admission and readmission classes to further put importance on classifying the readmitted patients.

We also collaborated with the George Mason University through their DAEN Capstone program. The team led by Dr. James Baldo and several participants from the graduate program, analyzed the underlying data and developed predictive models using various technologies including AWS SageMaker Autopilot. Resulting models and their respective hyperparameters were further analyzed and tuned to achieve high recall.

After choosing the best model, we designed and implemented this workflow in Alteryx Designer to automate our process and put the model into a feedback-reevaluate phase as a Cross-Industry Standard Process for Data Mining (CRISP-DM) to enable our model to evolve and be deployed in production.

Code Group	Description/ICD-10-CM diagnosis
A00-B99	Certain infectious and parasitic diseases
C00-D49	Neoplasms
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving immune mechanism
E00-E89	Endocrine, nutritional, and metabolic diseases
F01-F99	Mental, Behavioral and Neurodevelopmental disorders
G00-G99	Diseases of the nervous system
H00-H59	Diseases of the eye and adnexa
H60-H95	Diseases of the ear and mastoid process
I00-I99	Diseases of the circulatory system
J00-J99	Diseases of the respiratory system
K00-K95	Diseases of the digestive system
L00-L99	Diseases of the skin and subcutaneous tissue
M00-M99	Diseases of the musculoskeletal system and connective tissue
N00-N99	Diseases of the genitourinary system
O00-O9A	Pregnancy, childbirth and the puerperium
P00-P96	Certain conditions originating in the perinatal period
Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S00-T88	Injury, poisoning and certain other consequences of external causes
U00-U85	Codes for special purposes(U07-U85)
V00-Y99	External causes of morbidity
Z00-Z99	Factors influencing health status and contact with health services

Table 2. Diagnosis Code Groups



Machine Learning

More than 10 different classification methods such as Logistic Regression, Random Forest, and Xgboost for different feature combinations were used to compare our target classification metrics and choose an optimum model.

Models that consistently showed the close range of scores in their validation phase were chosen. Through cross-validation and grid search methods, best performing models were further optimized for high recall scores while keeping precision and accuracy in an acceptable range. We chose an XGBoost model with a combination of socioeconomic and medical code groups as the final model due to its 75% recall, ability for interpretation, high efficiency, and fast scoring time.

XGBoost which falls into the gradient boosting framework of machine learning algorithms, has been a consistent highly efficient problem solver and can run in major distributed environments.

Recall is the ability of a model to find all relevant cases within a dataset. In our case, true positives (TP) were the correctly classified readmitted patients and false positives (FP) were the readmitted patients who were incorrectly classified as not readmitted.

We specifically aimed for higher recall scores (TP/TP+FP) since accuracy for an imbalanced dataset would not be a good measure to assess model performance and we had to focus on identifying the readmitted patients to properly target and further analyze their underlying features.

The final model showed that socioeconomic features such as the pay category being Medicare, patient age, gender, wage index, and the population category of patients along with their diagnosis code groups and many other features contribute to classification for readmission.

Threshold	Recall	Precision	Accuracy	AUROC
0.45	0.75	0.09	0.40	0.60

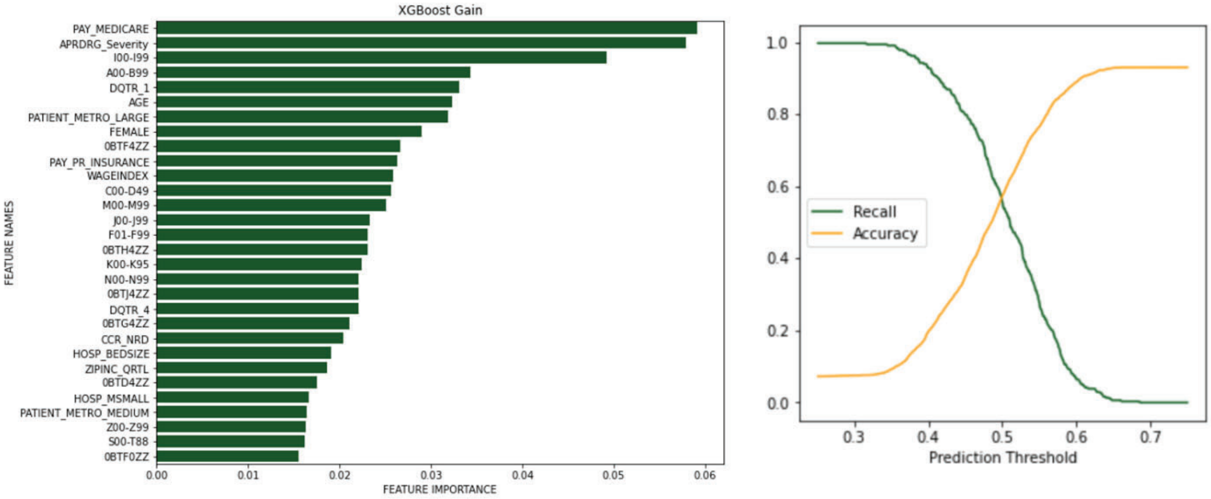


Figure 4. Feature importance of the final XGBoost model and recall/accuracy curve

Conclusion

Our research has enabled us to train models that can target and capture nearly 8 readmitted patients out of every 10. Our final model showed us a combination of demographic and diagnosis related features. These combinations further allowed us to analyze the likelihood of someone being readmitted when going through a lobectomy procedure.

By studying these features, we can further understand which variables have the highest contribution to our model.

Demographic features include a patient's age, gender, healthcare provider, wage index, urban/rural category of the hospital, admission quarter of the year, etc.

Circulatory system diseases (I00-I99), certain infectious and parasitic diseases (A00-B99), neoplasms (C00-D49), musculoskeletal system and connective tissue diseases (M00-M99) were among the top contributing factors to the predictive ability of our model in the medical factors.

By understanding the likelihood of a patient's readmission, pre/post-operative interventions such as weight loss, home monitoring programs, or additional medical procedures can be introduced into a patient's hospital care cycle which would improve their outcome and reduce the relative costs for them, healthcare provider, and the hospital.

Likewise, our approach can be used to target different medical procedures for any dataset that has similar information but not necessarily all the features used in our models.

Limitations

One of the key limitations we faced in our research was the ICD10 data being available only from Q415 to Q417. This limited us to only research the existing data from a 2 year period. Similar research done on readmission cases cover near 10 years' worth of data.

Acquiring more data can enable us further to optimize our models based on the desired target metric and help with class imbalance.

The study is limited to the non-medical factors that are being collected in the NRD and depending on healthcare information providers, final model is subject to change/update.

Next Steps

- To refine the readmission predictive analysis model on a smaller subset of medical and non-medical features and perform more real-world data validation.
- Refine the model by applying to larger data sets from other sources.
- Working with the medical community on possible preventive actions to reduce re-admissions.

About Allwyn Corporation

Allwyn Corporation (www.allwyncorp.com) is a forward thinking, innovative software solutions company, headquartered in the Metropolitan Washington DC area. Allwyn was founded in 2003 with a mission to help organizations address complex technology challenges by providing industry-leading tools, technologies, seasoned professionals, and proven methodologies. We are proud to be certified for ISO 9001 (Quality), ISO 27001 (Security), and ISO 20000 (Service Delivery).
With a team of ~200 professionals, Allwyn delivers high-quality services to a wide range of clients in the public and private sector.

Allwyn has been providing leading-edge IT professional services to various government agencies through the GSA MAS Schedule. We are also on the FAA eFAST, GSA OASIS+, and GSA STARS III contract vehicles.

Allwyn has experience with implementing Artificial Intelligence and Machine Learning solutions and Modernizing Applications using Low Code Technologies. Our relationships with AWS, Appian, ServiceNow, Microsoft, Databricks, Informatica, Salesforce, etc. strengthen our ability to support our customers in their Digital Transformation journey. We are already supporting several of our customers in the public as well as commercial sector with their cloud adoption strategies and Artificial Intelligence and Machine Learning implementations. For additional information on Allwyn's full range of services, please visit our website at www.allwyncorp.com.