Hospital Occupancy Prediction: Association Rule Mining

Justin Ledford
Eric Smith

December, 2018

Abstract

In this paper, we explore frequent itemsets and association rules mined from a dataset consisting of people's insurance claim history, as well as other features such as the number of days they spent in the hospital, number of unique prescriptions and number of lab visits. By searching for association rules that imply our target class (number of days in hospital the next year), we were able to find some sets of medical conditions that could be used as a signal for determining if and for how long a person visits the hospital the next year.

Contents

	Absolute Frequency	**Relative Frequency**
0 days	101,414	88.1%
1 day	6,257	5.4%
2-4 days	5,778	5.0%
5+ days	1,644	1.4%

Table 1: Class Balance

1 Data Preparation

We have flattened the claims into summary statistics, for example age, sex, and a member's maximum Charlson Index. For sex, which is a binary value, we used a 0 for female and 1 for male. We also used the number of lab visits per year, and number of unique prescriptions filled. The number of unique prescriptions, or drug count, is in the range 1-7+. We chose these features as we thought they are most indicative of the number of days spent in the hospital the next year.

For any feature that contained more than one number, like 7+ for drug count, we replaced it with that number, because these values were chosen as they are the 95th percentile for that feature.

We used data from year 1 and year 2 claims, and days in hospital from year 2 and year 3. For members with claims in both years we considered them as separate samples, so our model only considers the claim history of a single year.

1.1 Categorical Feature Counts

In order to create more features and use the categorical features, we have created columns containing the counts of unique values for Primary Condition Group, Specialty and Place of Service. For example, given a member, some of these features would be number of claims with a metastatic cancer, number of claims where the member visited an diagnostic imaging lab, and number of claims where the member visited an internal medicine specialist.

1.2 Discretization

To create transactions from our data, features need to be discretized and itemized. To do this we convert the count features above to logical features, where they are false if the count is 0 and true if 1 or more. For age and Charlson Index, we discretize the values into discrete ranges.

For lab count, we have the ranges [0,1), [1,2), [2,7) and [7,10]. This is a logical grouping, with 0 lab visits, a single lab visit, a small amount and large amount of lab visits. We do the same with prescription counts, with 0, 1, 2-4, and 5+ counts. 2-4 prescriptions are common for illnesses such as respiratory infections, but 5+ prescriptions seem to indicate a chronic condition. Using the logical groupings above better reflect the differences between the ranges, rather than splitting by equal frequency or range length.

For days in hospital, we have the ranges 0 days, 1 day, 2-4 days, and 5+ days. This is another logical grouping, where 1 day could be a non critical overnight stay, 2-4 days could be a longer non critical stay or pregnancy, and 5+ days could be a critical stay.

Because we have a new feature for each of the categorical values, we have about 70 features. It could be beneficial to reduce the number of features by combining the primary condition group features into groups, such as cancer conditions, heart conditions, etc.

1.3 Class Imbalance

After setting the days in hospital classes, there is a very large imbalance which can be seen in Table 1. In order to inspect item sets and association rules for all classes equally, we have downsampled our data so that each class has an equal number of samples. After downsampling, we have 1644 rows in each class. Although this results in an unrealistic distribution of our data, we will be able to see more associations across all classes.

1.4 Condition Groups as Items

We also focused on combinations of conditions groups by treating conditions as items. In this way, a transaction is a grouping of conditions a person had during the year. This will allow us to find interesting sets and rules pertaining to only the conditions more easily, regardless of frequency of other factors such as sex and age.

2 Modeling

2.1 Frequent Itemsets

We created frequent itemsets for both of our datasets, the cumulative dataset, and the conditions dataset. Figure 1 shows the most frequent conditions, while Table 3 shows the top condition combinations. In general, many of the top conditions include "other" and "miscellaneous" groups that act as catch-alls for conditions. However there are a few interesting sets of 2 conditions that are likely common among older members, such as the combination of arthropathies with other conditions experienced in old age such as metabolic, cardiac and respiratory issues.

Primary Condition Group	**Description**
AMI	Acute myocardial infarction
APPCHOL	Appendicitis
ARTHSPIN	Arthropathies
CANCRA	Cancer A
CANCRB	Cancer B
CANCRM	Ovarian and metastatic cancer
CATAST	Catastrophic conditions
CHF	Congestive heart failure
COPD	Chronic obstructive pulmonary disorder
FLaELEC	Fluid and electrolyte
FXDISLC	Fractures and dislocations
GIBLEED	Gastrointestinal bleeding
GIOBSENT	Gastrointestinal, inflammatory bowel disease, and obstruction
GYNEC1	Gynecology
GYNECA	Gynecologic cancers
HEART2	Other cardiac conditions
HEART4	Atherosclerosis and peripheral vascular disease
HEMTOL	Non-malignant hematologic
HIPFX	Hip fracture
INFEC4	All other infections
LIVERDZ	Liver disorders
METAB1	Diabetic ketoacidosis and related metabolic
METAB3	Other metabolic
MISCHRT	Miscellaneous cardiac
MISCL1	Miscellaneous 1
MISCL5	Miscellaneous 3
MSC2a3	Miscellaneous 2
NEUMENT	Other neurological
ODaBNCA	Ingestions and benign tumors
PERINTL	Perinatal period
PERVALV	Pericarditis
PNCRDZ	Pancreatic disorders
PNEUM	Pneumonia
PRGNCY	Pregnancy
RENAL1	Acute renal failure
RENAL2	Chronic renal failure
RENAL3	Other renal
RESPR4	Acute respiratory
ROAMI	Chest pain
SEIZURE	Seizure
SEPSIS	Sepsis
SKNAUT	Skin and autoimmune disorders
STROKE	Stroke
TRAUMA	All other trauma
UTI	Urinary tract infections

Table 2: Primary Condition Groups

Conditions	**Support**	**Count**
{Other metabolic}	0.33	38041
{Arthropathies}	0.29	33526
{Other neurological}	0.26	30235
{Acute respiratory}	0.26	29732
{Miscellaneous cardiac}	0.19	22413
{Skin and autoimmune disorders}	0.17	19847
{All other infections}	0.15	17744
{Gastrointestinal bleeding}	0.14	16436
{Arthropathies,Other metabolic}	0.13	14594
{All other trauma}	0.12	14021
{Other metabolic,Other neurological}	0.11	12806
{Other metabolic,Miscellaneous cardiac}	0.11	12482
{Arthropathies,Other neurological}	0.10	10959
{Miscellaneous 3}	0.09	10624
{Other metabolic,Acute respiratory}	0.08	9177
{Arthropathies,Acute respiratory}	0.08	8966
{Ingestions and benign tumors}	0.08	8930
{Arthropathies,Miscellaneous cardiac}	0.08	8718
{Other neurological,Acute respiratory}	0.07	8432
{Other metabolic,Skin and autoimmune disorders}	0.07	8276
{Chronic obstructive pulmonary disorder}	0.07	8262
{Chest pain}	0.07	7915
{Urinary tract infections}	0.07	7898
{Other renal}	0.07	7871
{Miscellaneous cardiac,Other neurological}	0.07	7860
{Arthropathies,Skin and autoimmune disorders}	0.07	7854
{Gynecology}	0.07	7602
{Arthropathies,All other trauma}	0.06	7328
{Gastrointestinal bleeding,Other metabolic}	0.06	6858
{Other neurological,Skin and autoimmune disorders}	0.06	6717

Table 3: Top 30 Condition Sets

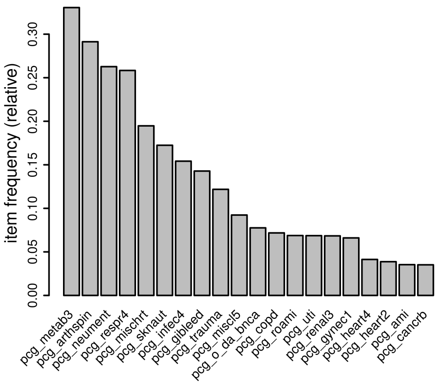


Figure 1: Top 20 Most Frequent Conditions

2.2 Association Rules

After inspecting the frequent itemsets, we mined association rules for both our datasets. Table 4 shows the top 10 condition rules by lift with a minimum support of 0.01. Looking at this table we are able to see some of the condition combinations that can point towards the days in hospital for a person.

For example, having a pregnancy during a year shows that that person will likely be in the hospital the next year for 2-4 days, most likely when they are giving birth. Other combinations of conditions can really lead to 5+ days in the hospital, such as kidney and heart failures. Figure 2 shows a visualization of these rules in the form a graph, to see the relationship among different conditions.

No rules for 0 days in hospital were found with the support at 0.01. This is most likely from the wide variety of condition sets, as this class contains people who were healthy, as well as people who were very unhealthy and may have passed away before the next year.

After lowering the support down to 0.0001, we were able to see other, more rare condition sets with high lift. For example, many condition sets with gastrointestinal issues led to a person spending 5 days in the hospital, and these rules had a confidence of 1.0 and lift of 4.0.

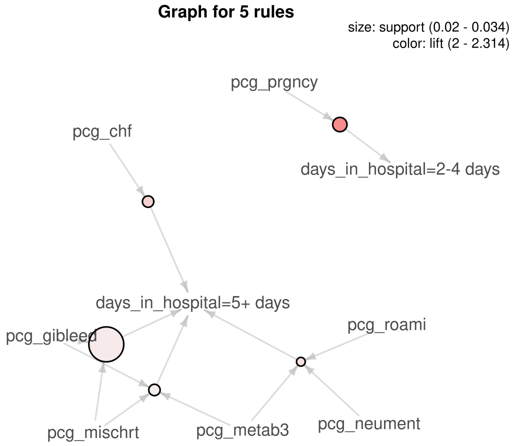
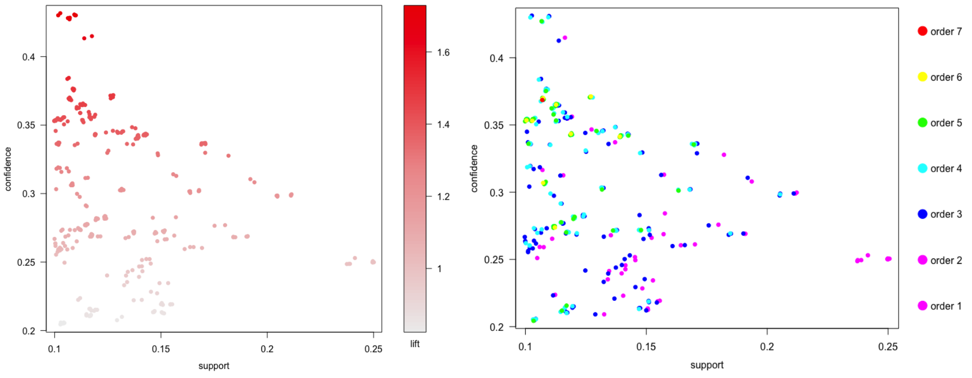


Figure 2: Condition Sets with Support > 0.02

Conditions	Days	Support	Confidence	Lift	Count
{Congestive Heart Failure,Chest Pain}	5+	0.01	0.65	2.59	66
{Pregnancy}	2-4	0.02	0.60	2.42	157
{Congestive Heart Failure,Acute Respiratory}	5+	0.01	0.59	2.38	79
{Gastrointestinal Bleeding,Miscellaneous Cardiac,Chest Pain}	5+	0.01	0.58	2.31	83
{Non-malignant Hematologic,Miscellaneous Cardiac}	5+	0.01	0.57	2.28	86
{Arthropathies,Congestive Heart Failure}	5+	0.01	0.57	2.27	75
{Chronic Renal Failure}	5+	0.01	0.56	2.24	88
{Gastrointestinal Bleeding,Miscellaneous 3,Other Neurological}	5+	0.01	0.56	2.23	72
{Arthropathies,Miscellaneous Cardiac,Other Neurological,Chest Pain}	5+	0.01	0.56	2.23	69
{Arthropathies,Gastrointestinal Bleeding,Miscellaneous 3}	5+	0.01	0.56	2.22	75

Table 4: Top 10 Condition Rules by Lift with Support > 0.01



(a) 315 rules

(b) 315 rules by order

Figure 3: Support, confidence, and lift of Rules with DaysInHospital

Of the 315 rules with *DaysInHospital* on the right-hand side had a confidence below 0.5 (See Figure 3). Meanwhile, most rules had less than 20% lift. When these rules are broken down by order (i.e. how many items are in each rule), we see that lower-order rules are those with the highest support (Figure-3). This is expected given the apriori principle, which states that if an itemset is frequent, than all of its subsets must also be frequent. In other words, this means that a subset of a frequent set is at least as frequent as the set, so subsets of 1 item will have the highest frequency (support).

LHS	**Days in Hospital**	**Support**	**Confidence**	**Lift**
{Sex: Male}	0	0.119	0.286	1.142
{Maximum Charlson Index: 0}	0	0.178	0.329	1.316
{Sex: Female}	0	0.131	0.225	0.899
{General Practice}	0	0.141	0.241	0.962
{Primary Condition: Miscellaneous 2}	0	0.146	0.223	0.890
{Internal Specialist}	0	0.137	0.214	0.856
{Laboratory Specialist}	0	0.156	0.217	0.866
{Independent Lab}	0	0.156	0.216	0.866
{Office}	0	0.237	0.248	0.992
{Sex: Male}	1	0.105	0.253	1.012

Table 5: Top 5 *DaysInHospital* Sets

When we examine the top 10 rules with *DaysInHospital* (Table 5), we find that zero days in hospitals dominates the association rules, even though the dataset was downsampled. Being of male sex has the most support for determining if a patient will spend zero days in the hospital, followed by a Charlson index of zero and being female sex.

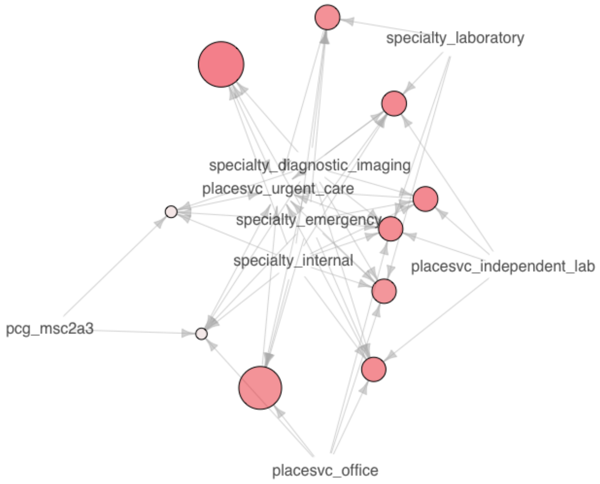


Figure 4: Graph of top 10 rules containing DaysInHospital

3 Evaluation

By looking at the frequent condition sets, as well as rules with high lift, we were able to find some interesting combinations of sets that point towards a person spending 2-4, and 5+ days in the hospital. For example, having pregnancy conditions and no other health issues almost guarantees that a person will be in the hospital the next year. More critical combinations such as those with gastrointenstinal issues and heart failures usually lead to more days in the hospital.

4 Deployment

While the findings above could be useful they are also somewhat obvious to health professionals, and using this model to determine if a person will visit the hospital may be overkill, but could help in automating a basic prediction. One of the biggest issues with the data is determining if the person doesn't visit the hospital the next year because they are either healthy or pass away. It would be beneficial to train another model given the conditions and other factors to determine if their current condition is fatal and their chance of death within the current year.