REPORT

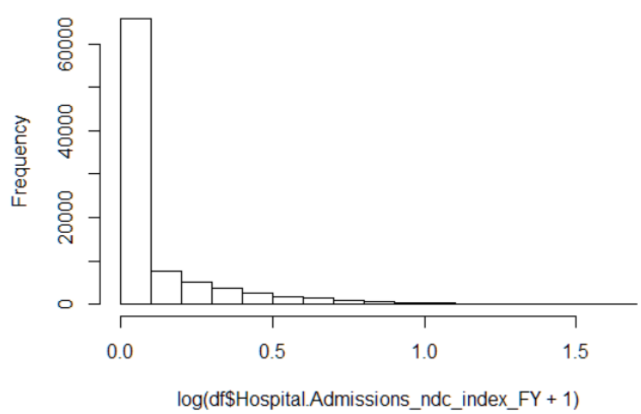
Rule Mining

By Jaideep Reddy

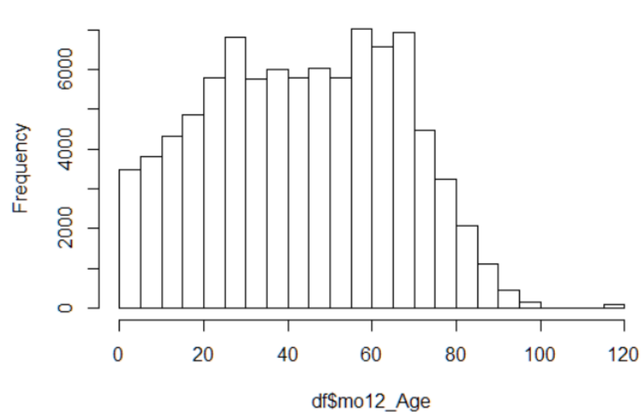
Dataset:90656 rows and 21 columns

Top 20 variables from the feature importance by XGBoost.
The last column is risk probability.
0.2 is the cut-off. Data points with probability greater than 0.2 are classified as high risk and low risk otherwise.

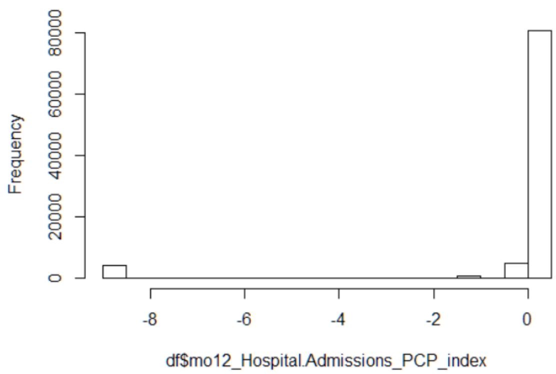
Var 1: `Hospital.Admissions_ndc_index_FY`
Used log(x+1) for binning the variable. Got a good spread and therefore went ahead and binned them accordingly



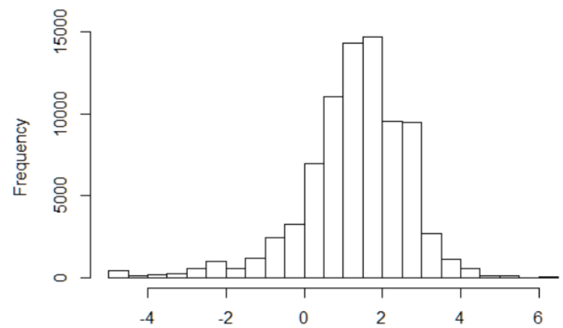
Var 2: `mo12_Age`
Binned age into 6 equal width bins based on histogram. Age had a good spread so didn't require any transformation.



Var 3: `mo12_Hospital.Admissions_PCP_index`
Binned -9, -1 and 0 separately. For the positive values used log(30x) and log(1000x) to compare. Went forward with log(1000x) because it shifts the median just above 1 which means at least 50 percent of the positive values are above 1.
We wanted majority of the values on the positive side of the log(x) vs. x curve therefore went with log(1000x)

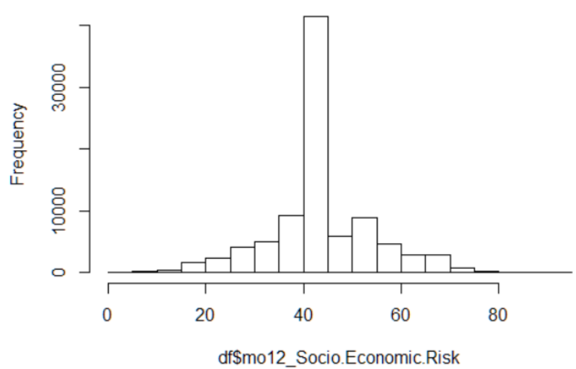


Original histogram with -1 and -9

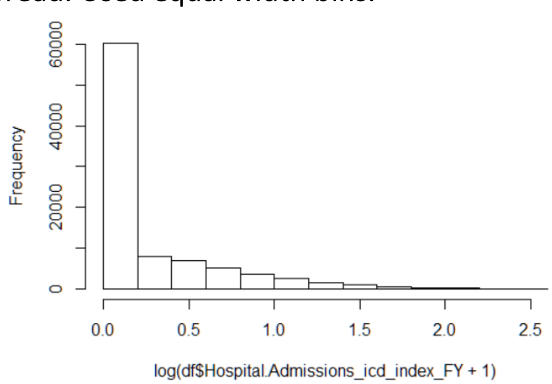


Only +ve values with log(1000x) transform

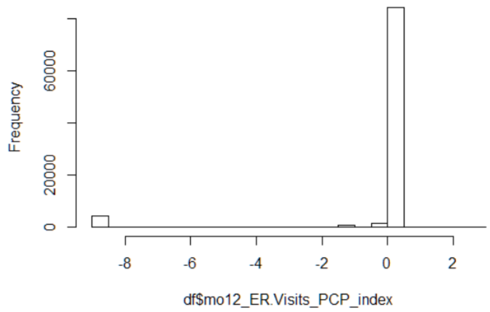
Var 4: `mo12_Socio.Economic.Risk`
Good spread. Didn't need any transformation. Binned based on histogram.



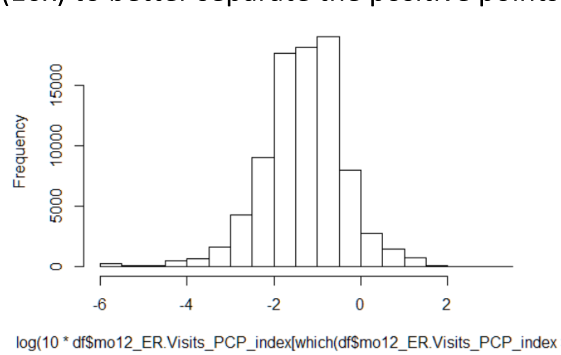
Var 5: `Hospital.Admissions_icd_index_FY`
Used log(x+1) for better spread. Used equal width bins.



Var 6: mo12_ER.Visits_PCP_index
Saw the data and binned it based on that.
Majority of the values were binned into -9,-1. All the positive points are very small numbers between 0 and 1. The median is 0.02. I'm using log(10x) to better separate the positive points

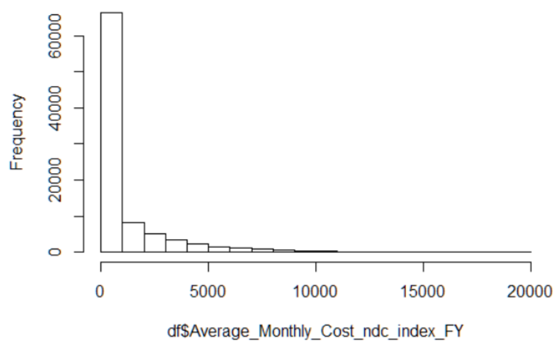


The full histogram

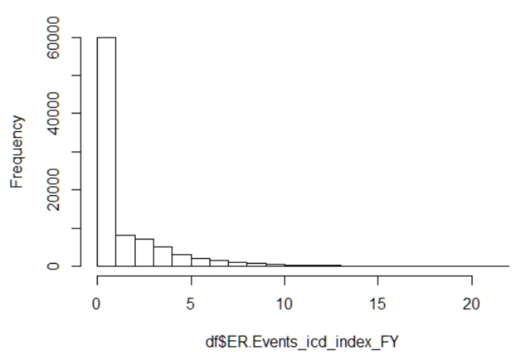


log(10x) histogram of values greater than 0

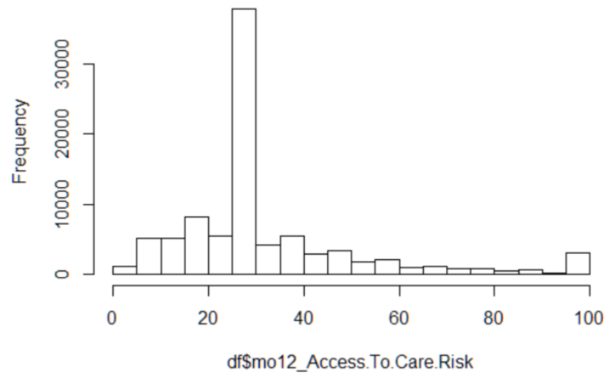
Var 7: Average_Monthly_Cost_ndc_index_FY
The cost already has a good spread across the x axis with values ranging from 0 to 20000



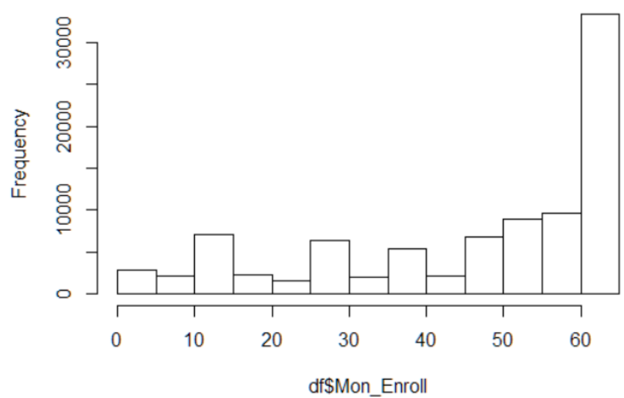
Var 8: ER.Events_icd_index_FY
Similarly as above did not need any transformation. Binned the values based on the data



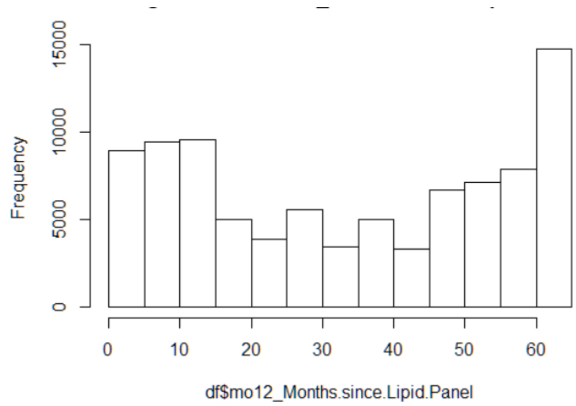
Var 9: mo12_Access.To.Care.Risk
Did not need any transformation. Data was well spread across. Split into equal width bins.



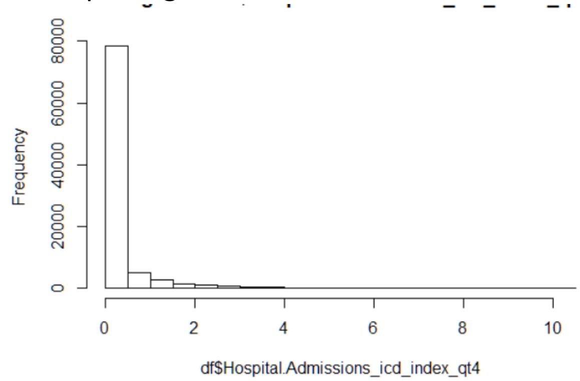
Var 10: Mon_Enroll
Similarly, did not need any transformation. Data was well spread across. Split into equal width bins.



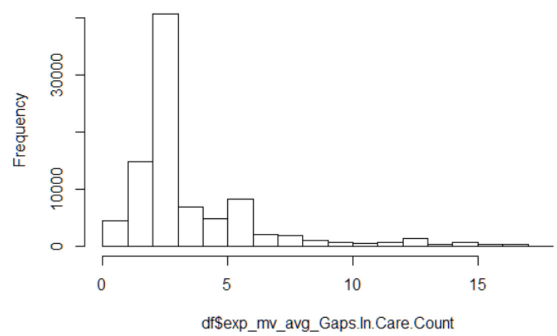
Var 11: mo12_Months.since.Lipid.Panel
Again, did not need any transformation. Data was well spread across. Split into equal width bins.



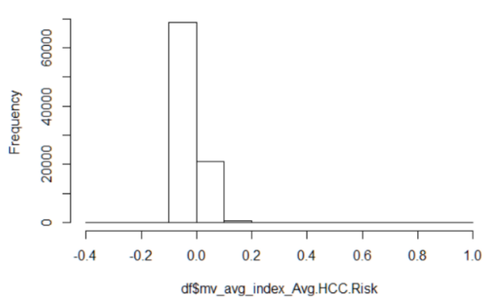
Var 12: Hospital.Admissions_icd_index_qt4
No transformation. Subjective splitting.



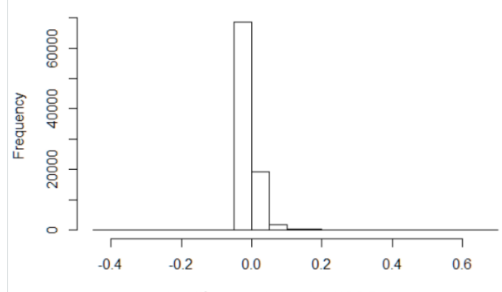
Var 13: exp_mv_avg_Gaps.In.Care.Count
Didn’t need any transformation



Var 14: mv_avg_index_Avg.HCC.Risk
No change at all with log transformation. Therefore, discretizing by looking at the data.

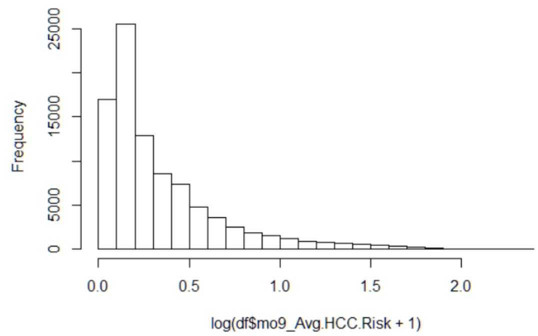


Without transformation

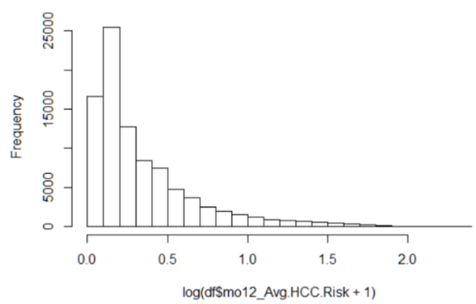


With log transformation

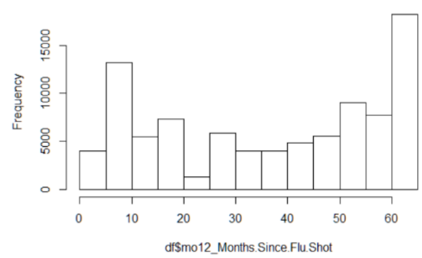
Var 15: mo9_Avg.HCC.Risk
Used log(x+1) transformation. Got a better spread for smaller values.



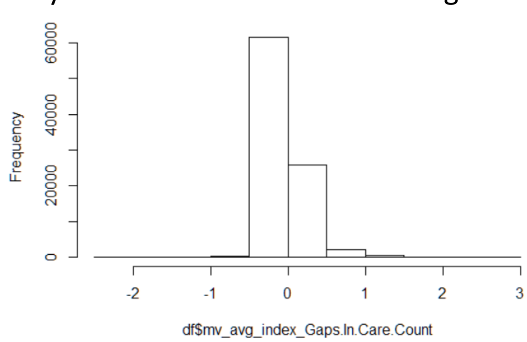
Var 16: mo12_Avg.HCC.Risk
The spread here was more or less the same with or without transformation but using log(x+1) h elped in better spread less frequent data therefore better discretization of minority values.



Var 17 : mo12_Months.Since.Flu.Shot
No transformation required.

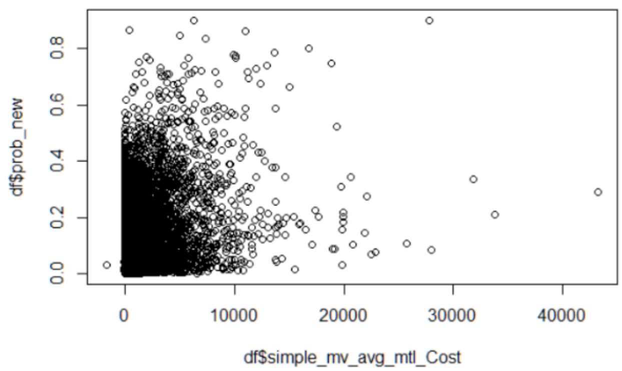


Var 18: mv_avg_index_Gaps.In.Care.Count
Transformation did not give any better results. Binned the original data.



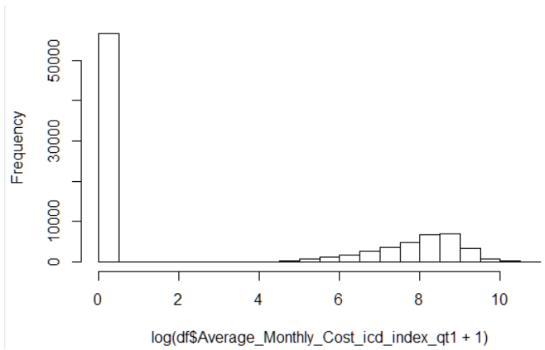
Var 19: simple_mv_avg_mtl_Cost

Just as discussed during the meeting I binned these values into 3 bins based on the scatterplot



Var 20: Average_Monthly_Cost_icd_index_qt1

The log(x+1) transformation helped in binning the minority values better.



Association rule mining:

The top 5 rules sorted by lift in decreasing order.

```
    lhs                                                     rhs               support     confidence  lift     count
[1] {Hospital.Admissions_ndc_index_FY=(1,2],
     ER.Events_icd_index_FY=(10,15],
     simple_mv_avg_mtl_Cost=(1e+03,4.5e+04]}             => {prob_new=risk} 0.002371603  0.9148936 36.20279   215
[2] {Hospital.Admissions_ndc_index_FY=(1,2],
     ER.Events_icd_index_FY=(10,15],
     Average_Monthly_Cost_icd_index_qt1=(8,10],
     simple_mv_avg_mtl_Cost=(1e+03,4.5e+04]}             => {prob_new=risk} 0.002106866  0.9138756 36.16251   191
[3] {Hospital.Admissions_ndc_index_FY=(1,2],
     ER.Events_icd_index_FY=(10,15]}                     => {prob_new=risk} 0.002570155  0.9137255 36.15657   233
[4] {Hospital.Admissions_ndc_index_FY=(1,2],
     ER.Events_icd_index_FY=(10,15],
     Average_Monthly_Cost_icd_index_qt1=(8,10]}          => {prob_new=risk} 0.002294388  0.9122807 36.09940   208
[5] {Average_Monthly_Cost_ndc_index_FY=(1e+04,1.5e+04],
     ER.Events_icd_index_FY=(10,15]}                     => {prob_new=risk} 0.002040681  0.9113300 36.06178   185
```

From the above rules we can say that, if these patterns are present in a record then with almost 90% confidence we can say that the patient has a higher risk to be admitted to ER next year.

```
    lhs                                              rhs                support confidence     lift count
[1] {mo12_ER.Visits_PCP_index=(0.03,1],
     ER.Events_icd_index_FY=(10,15]}              => {prob_new=risk} 0.00559257   0.758982 30.03329   507
[2] {mo12_Hospital.Admissions_PCP_index=(0,0.1],
     mo12_ER.Visits_PCP_index=(0.03,1],
     ER.Events_icd_index_FY=(10,15]}              => {prob_new=risk} 0.00543814   0.756135 29.92063   493
```

For these specific rules there is a much higher support. There are 2291 high risk patients and in almost 50% of them the above patterns are observed. With the above factors alone we can be 75% confident that the patient has a higher chance of getting admitted to ER next year.

Conclusion:

The results from Association rule mining are pretty much in line with the results obtained from XGBoost. Association rules are pretty good at finding frequently occurring patterns which gives us a lot more understanding of the data and helps us figuring out causation or correlation between different features(such as chronic illness or symptoms leading to a health issue) in health data. However, here we did not have the data in such format so could not mine rules with enough quality.