Mini Project

Rahul Dhakecha

15 November 2016

Introduction

This case pertains to the readmission rate in various hospital accross the country. Readmission rate implies poor medication and higher rate of untreated diseases. In this case, we particularly focus on diabetic patients which amount to large proportion of the total people admitted in hospital. Readmission in hospital is a costly affair in terms of economic factor as well as a failure to provide adequate treatment.

Thus, it is of utmost importance to determine the significant factors which lead to readmission of a patient. We focus on readmission rate within 30 days primarily because of following reasons: Firstly, readmission within 30 days signify that disease of patient is not cured to a significant level and therefore he/she is readmitted within short span of time. Secondly, if there is a readmission after 30 days, then there might be several other factors which acts as the causes of disease. For example, patient might develop a different diet plan over the long the span which may result in readmission.

In this report, we analyse data in raw form as well as through couple of sophisticated dala analysis algorithms. We try to related output of these algorithms with intuitive understanding of the data.

Specifically, we run Random Forest to select important predictors among various differnt variables available in dataset. Advantages and limitations of both models are discussed in the coming sections.

We then finally use random forest as our classification algorithm to predict whether a patient will be readmitted or not. Approaches described in this report are limited by the computational capability and the available dataset.

Data Analysis

Complete data set given to us in file "diabetic.data.csv" has many redundant variables. Therefore we work with filtered dataset provided in file "readmission.csv" which contains only 31 variables. It has almost around 100000 data samples of different patients. We cannot work this large dataset because of the computational limit. The maximum number of trees formed in random forest is directly limited by the samples we take in training our model. More the samples, less number of trees can be formed owing to computational limit.

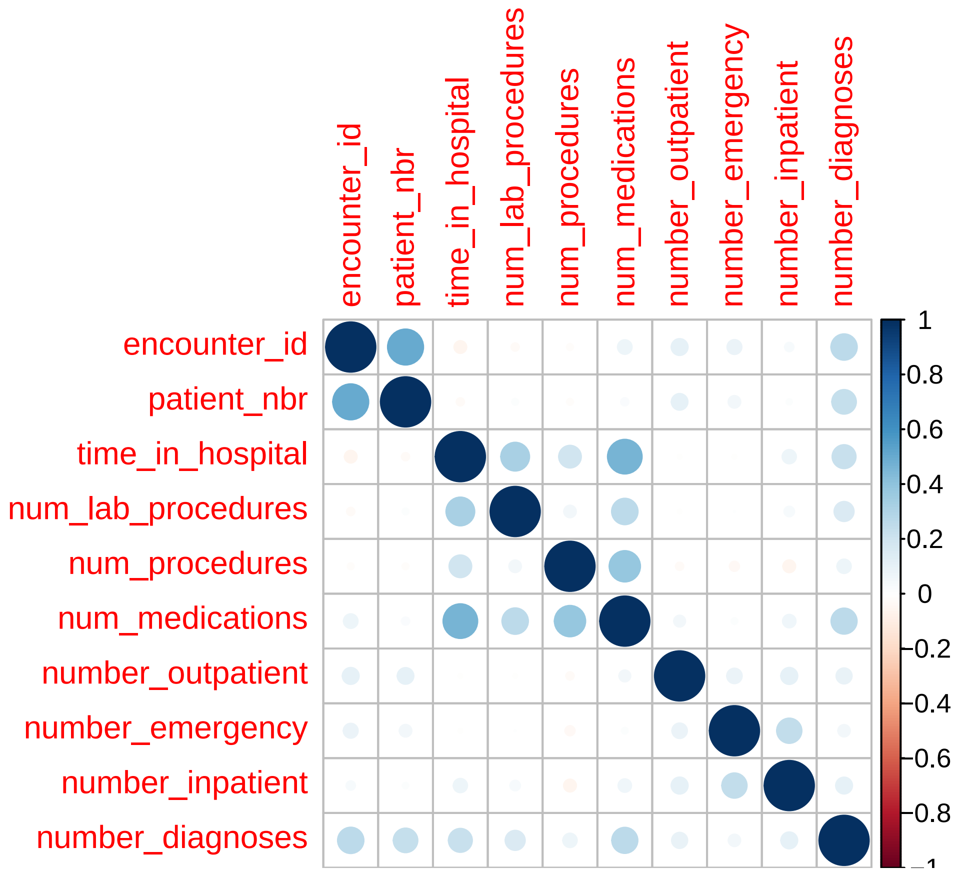
"readmission.csv" contains following data variables as its predictors:

```
##  [1] "encounter_id"       "patient_nbr"        "race"
##  [4] "gender"             "time_in_hospital"   "num_lab_procedures"
##  [7] "num_procedures"     "num_medications"    "number_outpatient"
## [10] "number_emergency"   "number_inpatient"   "number_diagnoses"
## [13] "max_glu_serum"      "A1Cresult"          "metformin"
## [16] "glimepiride"        "glipizide"          "glyburide"
## [19] "pioglitazone"       "rosiglitazone"      "insulin"
## [22] "change"             "diabetesMed"        "disch_disp_modified"
## [25] "adm_src_mod"        "adm_typ_mod"        "age_mod"
## [28] "diag1_mod"          "diag2_mod"          "diag3_mod"
## [31] "readmitted"
```

We split the entire data into two parts, one for training and other for testing. This selection is random and we try to keep around two third samples in training test and remainin ones in testing set. We intend to take a naive look at almost all of the predictors. We primarily exploit the relation between different variables with “readmitted” variable, which is our response variable.

In our data set we have categorical as well as numerical variables. Relation between numerical variables can be directly reflected by correlation values but to understand the interaction between categorical variables, we need to use two way tables.

Covariance between numerical variables



Above plot of covariance gives us an intuitive glance at the relation between different numeric variables. We see that there is no significance relation between any of the two variables so as to consider one of them redundant in our model selection. There is minor relation between total number of medications and time spent in hospital. This aligns with our intuition that a patient who spends more time in hospital would perhaps be given more medication.

Now we relate different categorical tables with readmission rate by two way tables which gives us the conditional probability of readmission given a particualar categorical predictor. Tables are included in the Appendix, and here we try to discuss important variables.

Our analysis shows that there are many variables which are important for the predicting patient’s readmission rate, but there are very few which directly related with readmission rate within 30 days.

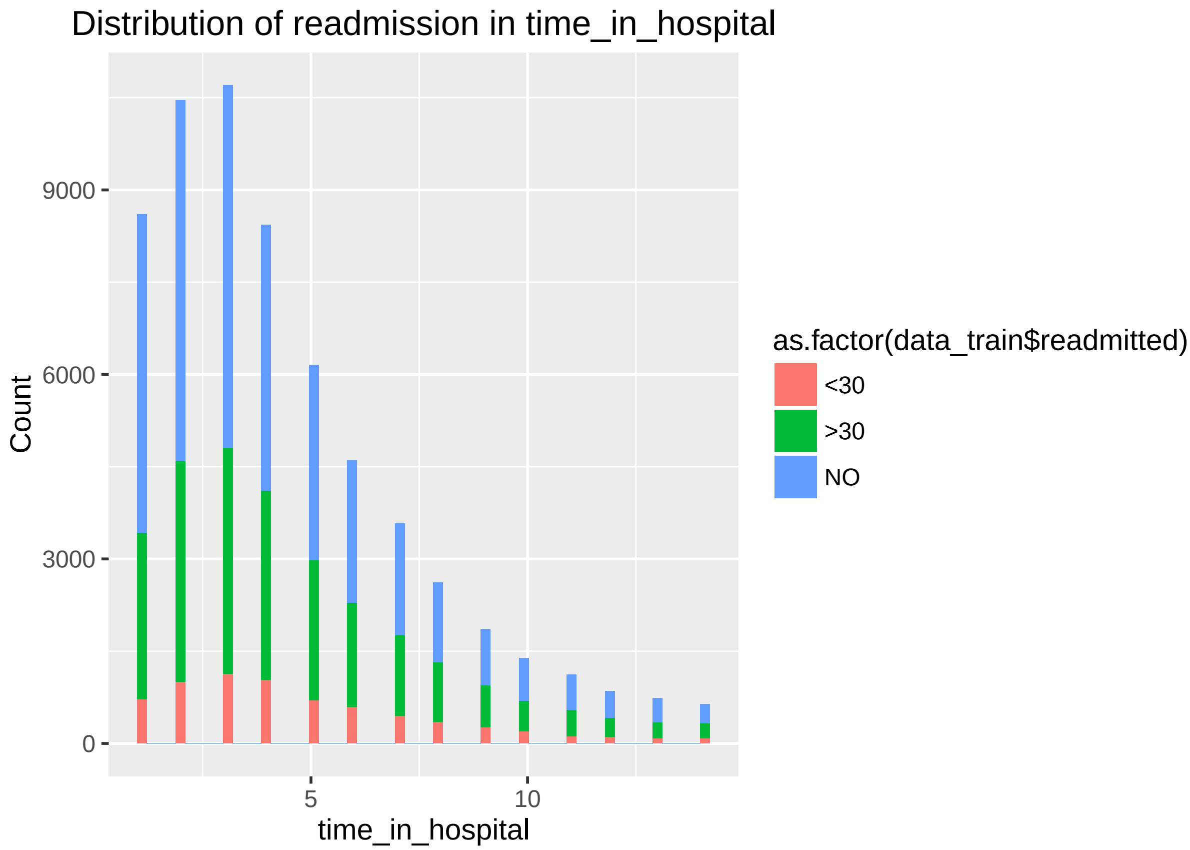
Important predictors for readmission:

- max_glu_serum: glucose serum test gives us the sugar level of patient. This is one of the predictor which directly correlates with the severeness of disease. It can be seen from proportion table, that case with >200 and >300 serum levels have high probability of readmission compared to other cases.

- change: this variable shows us whether there was change in any of the medication of patient. Conditional probability table shows us that there is no significant impact, but patients whose medication are changes have slightly higher probability of readmission.
- diabetesmed: this variable shows that there is higher probability of readmission given that diabetes medication is given to patient. This is one of the main factor which raises doubts on the previous treatment of patient. Also we see that this factor shows considerable difference between percent of patients readmitted within 30 days and after 30 days.
- disch_disp_modified: There are four levels with this predictor:
- discharged to home
- discharged to home with home health service
- discharged/transferred to Skilled Nursing Facility
- other from the table, we see that people discharged to home have less probability of readmission. Patients who are provided with home health service or those who are transfered to SNF are more vulnerable to readmission. This is quite intuitive as patients not cured completely are the ones who need extra care. And eventually they are the ones who have high probability of readmission.
- adm_src_mod: from the table, we see that patients who are admitted to hospital on emergency basis or those who are transferred from home health serive have higher probability of readmission. Emergency case indicates that there is some severe malfunctioning with patient and it needs serious diagnosis. If this emergency is not well treated, then there is high probability of patient being readmitted. If the patient is transfered from home health service, then it is a sign of prolonged treatment, which in turn means that disease might be incurable and patient may be readmitted again. If the patient is admitted on the basis of physician referral, then it is quite possible that he is admitted for the first time and his disease is curable.
- diag1_mod: diag1_mod gives us the ICD9 codes for primary treatment for various diseases. From the conditional probability table we see that this turns out to be one of the significant factor which differes for various different levels. For example, patient with ICD9 code equal to 250.6 has 21.6% probability of readmission compared to patient with ICD9 code equal to 996 which has only 5.17% probability of readmission. Analyzing this variable further, we see that 250.6 code corresponds to diabetes with neurological manifestations. Clearly this shows that patients who have undergone primary treatment for diabetes with neurological manifestations are far more vulnerabel to readmission within 30 days compared to other patients.
- diag2_mod: we see that patient who has undergone secondary treatment of Other cellulitis and abscess(682), has very high probability of readmission within 30 days.
- diag3_mod: we see that patient who has undergone tertiary treatment for Alteration of consciousness(780), has very high probability of readmission within 30 days

From the three level of treatments undergone by a patient we see that that patient with diabetes who has undergone primary treatment is very much likely for readmission within 30 days but this probability goes down considerably after secondary and tertiary treatment. Also analyzing changes in various medications, we find that their variation gives us significant information about the readmission rate.

Next we exploit the relation between “readmitted” variable with various numerical variables.



Graphically we find the fundamental relation between variables but it is more convenient to compute the correlation. We find that readmitted is highly correlated to variables such as time_in_hospital, number_inpatients, number_outpatients, number_diagnosis and number_emergencies.

Predictors important for readmission rate < 30 days

Up till now we saw the predictors important for overall readmission rate. Now we specifically list out variables which are important for readmission rate less than 30 days.

- diabetesmed
- diag1_mod
- diag2_mod
- diag3_mod
- metformin
- glimepiride
- glipizide
- glyburide
- pioglitazone
- rosiglitazone
- time_in_hospital
- num_lab_procedures
- num_procedures
- num_medications
- number_outpatient
- number_inpatient+number_diagnoses

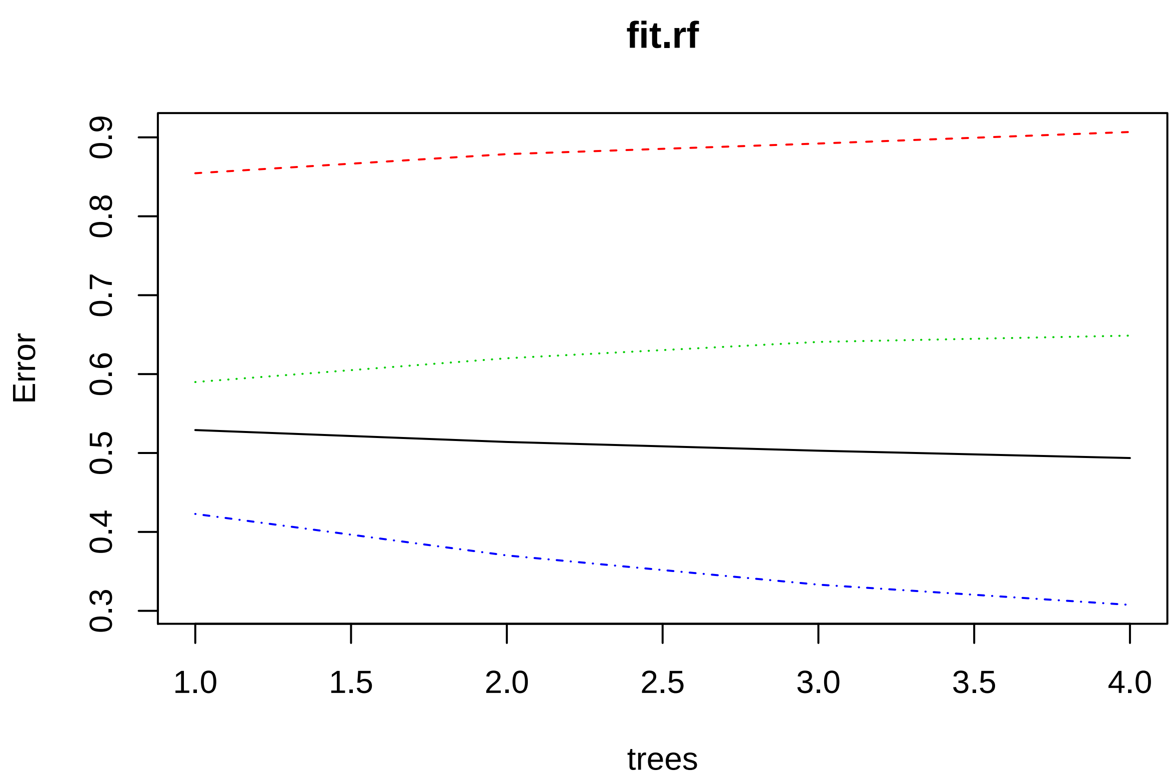
Exceptions

During the raw analysis of data, we found couple of contradictory results.

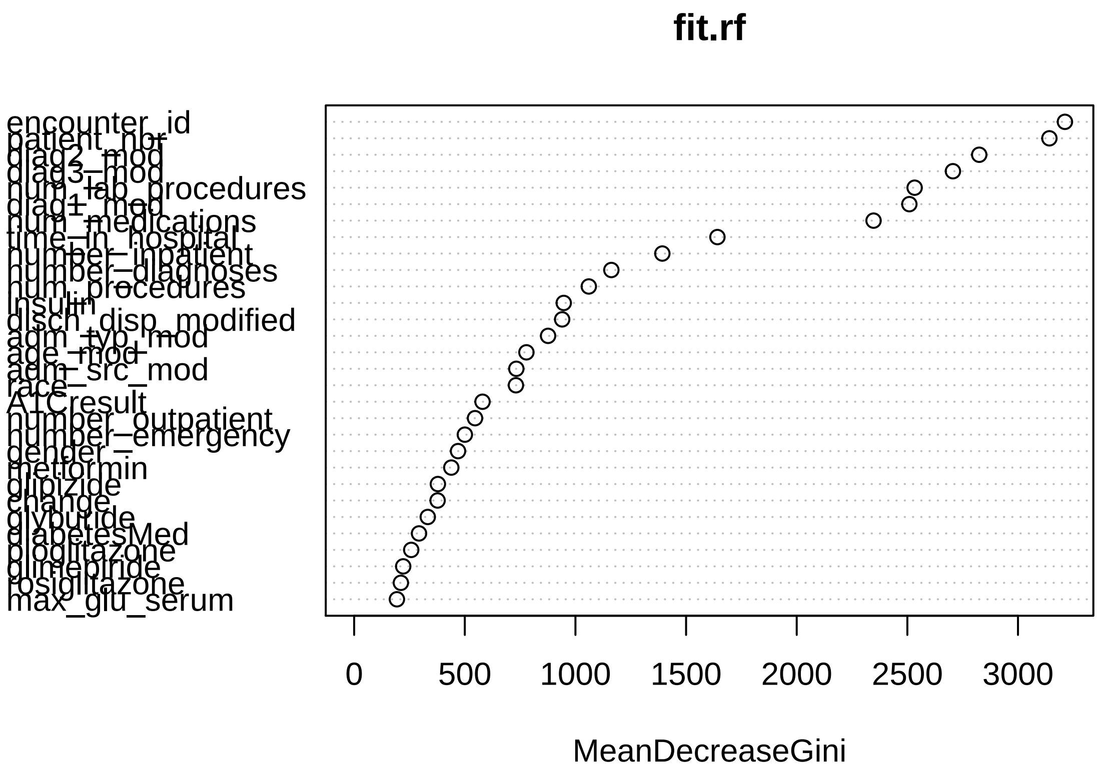
- Insulin: We expect that this variable would have significant impact on the readmission rate but it turns out that there is very bleak relation between this variable and readmission rate. Perhaps this abnormality owes to the fact that all diabetic patients, whether readmitted or not, have insulin levels which are not wihtin the range.
- A1Cresult: This test result gives the average sugar level in a patient's body for the past three months. We expect that patients who are readmitted should have high A1Cresult but on the contrary, this variable provides us no significant information.

Model Selection through Random Forest

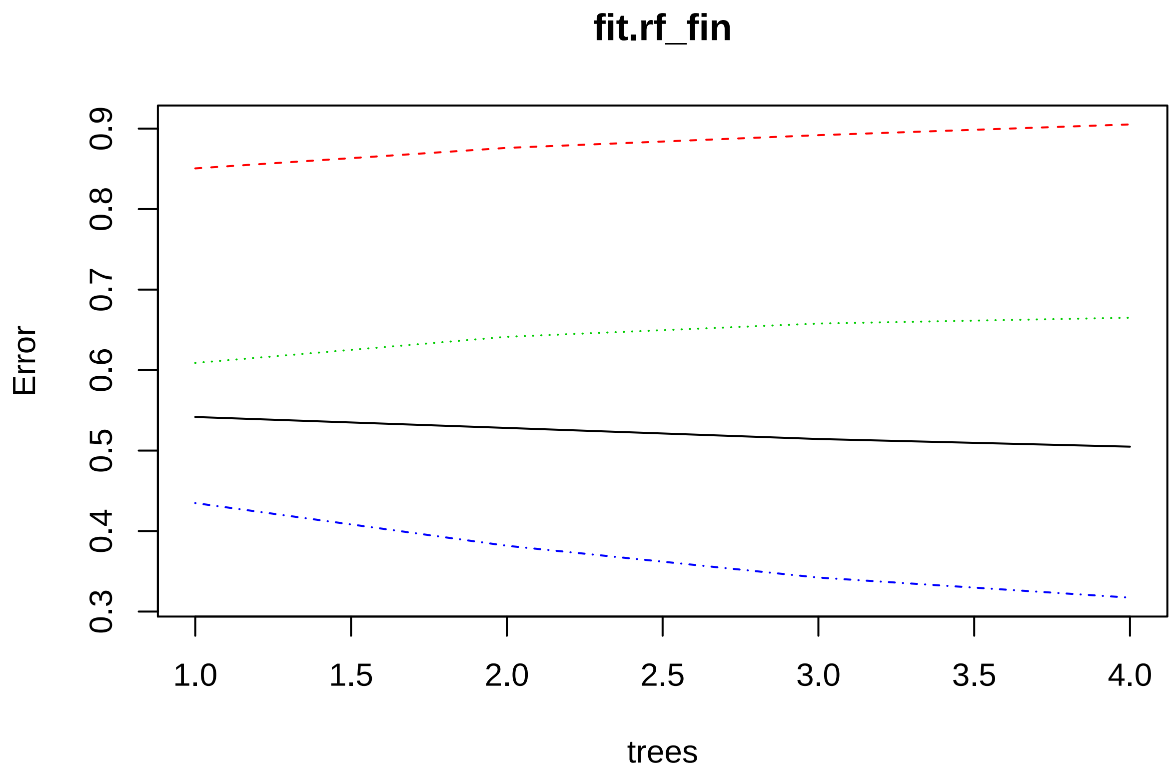
We now run the random forest algorithm on our entire data set.



We get mean classification error of 0.39. Since there are too many predictors and all are not important, therefore we need to select significant predictors.



We use following predictors to build our final model which can be used for prediction.



Using important predictors, we obtain mean classification error of 0.35.

Conclusion

From the analysis of this case, we found out various important predictors. It can be noticed that various important predictors given by Random Forest are also accounted from the raw analysis of data. But for the prediction purpose, we will use model create by Random Forest.

Appendix

1. readmitted vs gender

```
##           gender
## readmitted Female  Male Unknown/Invalid
##        <30   3753  3106               0
##        >30  11944  9741               0
##        NO   17640 15579               3
```

2. readmitted vs race

```
##           race
## readmitted     ? AfricanAmerican Asian Caucasian Hispanic Other
##        <30   112            1309    40      5158      143    97
```

```
##        >30   315            4104    95     16505      386   280
##        NO    963            6279   249     24450      730   551
```

3. readmitted vs max_glu_serum

```
##           max_glu_serum
## readmitted      >200      >300      None      Norm
##        <30 0.1255459 0.1436031 0.1102986 0.1146907
##        >30 0.3515284 0.4190601 0.3502529 0.3485825
##        NO  0.5229258 0.4373368 0.5394485 0.5367268
```

```
##           max_glu_serum
## readmitted      >200      >300      None      Norm       Sum
##        <30 0.1255459 0.1436031 0.1102986 0.1146907 0.4941383
##        >30 0.3515284 0.4190601 0.3502529 0.3485825 1.4694238
##        NO  0.5229258 0.4373368 0.5394485 0.5367268 2.0364379
##        Sum 1.0000000 1.0000000 1.0000000 1.0000000 4.0000000
```

4. readmitted vs A1Cresult

```
##           A1Cresult
## readmitted         >7         >8       None       Norm        Sum
##        <30 0.09727626 0.10367825 0.11317823 0.09730986 0.41144262
##        >30 0.34241245 0.35509297 0.35244883 0.32779807 1.37775232
##        NO  0.56031128 0.54122878 0.53437294 0.57489206 2.21080506
##        Sum 1.00000000 1.00000000 1.00000000 1.00000000 4.00000000
```

5. readmitted vs change

```
##           change
## readmitted        Ch        No       Sum
##        <30 0.1177026 0.1053076 0.2230102
##        >30 0.3703419 0.3344692 0.7048111
##        NO  0.5119555 0.5602232 1.0721787
##        Sum 1.0000000 1.0000000 2.0000000
```

6. readmitted vs diabetesmed

```
##           diabetesMed
## readmitted         No        Yes        Sum
##        <30 0.09346387 0.11629667 0.20976054
##        >30 0.31222707 0.36268079 0.67490787
##        NO  0.59430906 0.52102254 1.11533159
##        Sum 1.00000000 1.00000000 2.00000000
```

7. readmitted vs disch_disp_modified

```
##           disch_disp_modified
## readmitted Discharged to home Discharged to home with Home Health Service
##        <30         0.09248176                                  0.12660317
```

```
##        >30        0.35880408                                   0.41752381
##        NO         0.54871416                                   0.45587302
##        Sum        1.00000000                                   1.00000000
##           disch_disp_modified
## readmitted Discharged/Transferred to SNF      Other        Sum
##        <30                    0.14197237 0.14442947 0.50548677
##        >30                    0.35647928 0.25550314 1.38831031
##        NO                     0.50154836 0.60006739 2.10620292
##        Sum                    1.00000000 1.00000000 4.00000000
```

8. readmitted vs adm_src_mod

```
##           adm_src_mod
## readmitted Emergency Room     Other Physician Referral
##        <30      0.1146807 0.1016119          0.1069004
##        >30      0.3794676 0.2185268          0.3284363
##        NO       0.5058516 0.6798613          0.5646633
##        Sum      1.0000000 1.0000000          1.0000000
##           adm_src_mod
## readmitted Transfer from Home Health       Sum
##        <30                 0.1095958 0.4327889
##        >30                 0.3677163 1.2941471
##        NO                  0.5226878 2.2730640
##        Sum                 1.0000000 4.0000000
```

9. readmitted vs adm_typ_mod

```
##           adm_typ_mod
## readmitted  Elective Emergency     Other    Urgent       Sum
##        <30 0.1039052 0.1132997 0.1095825 0.1126058 0.4393932
##        >30 0.3094491 0.3584669 0.3921569 0.3489532 1.4090261
##        NO  0.5866457 0.5282333 0.4982606 0.5384410 2.1515806
##        Sum 1.0000000 1.0000000 1.0000000 1.0000000 4.0000000
```

10. readmitted vs age_mod

```
##           age_mod
## readmitted       0-19      20-59      60-79        80+        Sum
##        <30 0.05029014 0.10188239 0.11513158 0.11865251 0.38595661
##        >30 0.29013540 0.33786595 0.36062127 0.35197478 1.34059739
##        NO  0.65957447 0.56025166 0.52424715 0.52937272 2.27344600
##        Sum 1.00000000 1.00000000 1.00000000 1.00000000 4.00000000
```

11. readmitted vs diag1_mod

```
##           diag1_mod
## readmitted       250.6       250.8         276          38         410
##        <30  0.18953324  0.09940945  0.14834674  0.11132623  0.09962929
##        >30  0.46393211  0.42027559  0.37890974  0.33301065  0.29147359
##        NO   0.34653465  0.48031496  0.47274352  0.55566312  0.60889713
```

```
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag1_mod
## readmitted         414         427         428         434         435
##        <30  0.09431100  0.09323583  0.13571255  0.15785256  0.08452951
##        >30  0.32161820  0.35588056  0.44355426  0.30608974  0.38277512
##        NO   0.58407080  0.55088361  0.42073319  0.53605769  0.53269537
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag1_mod
## readmitted         486         491         493         518         577
##        <30  0.08666346  0.12005650  0.08520179  0.09495549  0.12261146
##        >30  0.39768897  0.47457627  0.49925262  0.31157270  0.38057325
##        NO   0.51564757  0.40536723  0.41554559  0.59347181  0.49681529
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag1_mod
## readmitted         584         599         682         715         780
##        <30  0.12977099  0.11111111  0.08986928  0.09457364  0.09496284
##        >30  0.34569248  0.37537538  0.36764706  0.26821705  0.37241949
##        NO   0.52453653  0.51351351  0.54248366  0.63720930  0.53261767
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag1_mod
## readmitted         786         820         996       Other         Sum
##        <30  0.07174888  0.17050691  0.13360324  0.11205965  2.73158165
##        >30  0.34121484  0.27803379  0.39757085  0.33098714  8.83834145
##        NO   0.58703628  0.55145929  0.46882591  0.55695320 12.43007690
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000 24.00000000
```

12. readmitted vs diag2_mod

```
##           diag2_mod
## readmitted         250      250.01      250.02         276         285
##        <30  0.07329124  0.12964931  0.11721908  0.11973716  0.08836207
##        >30  0.27944002  0.34643996  0.37105901  0.34728644  0.33405172
##        NO   0.64726873  0.52391073  0.51172191  0.53297639  0.57758621
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag2_mod
## readmitted         401         403         411         413         414
##        <30  0.07449112  0.15531178  0.09786700  0.08598726  0.08931918
##        >30  0.28237332  0.46016166  0.35006274  0.37261146  0.33791380
##        NO   0.64313556  0.38452656  0.55207026  0.54140127  0.57276702
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag2_mod
## readmitted         424         425         427         428         486
##        <30  0.09726444  0.11111111  0.11547269  0.11943967  0.10218140
##        >30  0.35410334  0.40661939  0.35721295  0.41189481  0.35935706
##        NO   0.54863222  0.48226950  0.52731436  0.46866552  0.53846154
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag2_mod
## readmitted         491         496         518         584         585
##        <30  0.12798265  0.10438729  0.09573092  0.10969638  0.15156794
##        >30  0.45336226  0.40847201  0.29883571  0.34280118  0.43815331
##        NO   0.41865510  0.48714070  0.60543338  0.54750245  0.41027875
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag2_mod
```

```
## readmitted         599         682         707         780       Other
##        <30  0.11111111  0.13992762  0.14003436  0.09846827  0.11385748
##        >30  0.36326738  0.40289505  0.42268041  0.34901532  0.33009028
##        NO   0.52562151  0.45717732  0.43728522  0.55251641  0.55605224
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag2_mod
## readmitted         Sum
##        <30  2.76946853
##        >30  9.18016061
##        NO  13.05037086
##        Sum 25.00000000
```

13. readmitted vs diag3_mod

```
##           diag3_mod
## readmitted           ?         250      250.02       250.6         272
##        <30  0.06306306  0.08443045  0.13679245  0.18085106  0.07027942
##        >30  0.24211712  0.31296508  0.38443396  0.40577508  0.27265030
##        NO   0.69481982  0.60260446  0.47877358  0.41337386  0.65707028
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag3_mod
## readmitted         276         285         401         403         414
##        <30  0.11326758  0.10000000  0.08476266  0.16300496  0.08887896
##        >30  0.35475660  0.35540541  0.30295174  0.41318214  0.36176865
##        NO   0.53197582  0.54459459  0.61228560  0.42381290  0.54935239
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag3_mod
## readmitted         424         425         427         428         496
##        <30  0.10869565  0.10869565  0.10836177  0.12818671  0.12820513
##        >30  0.40217391  0.42463768  0.37627986  0.40359066  0.38836773
##        NO   0.48913043  0.46666667  0.51535836  0.46822262  0.48342714
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag3_mod
## readmitted         585         599         707         780       Other
##        <30  0.16065574  0.12984823  0.14002478  0.10929648  0.11721754
##        >30  0.43606557  0.35160202  0.39900867  0.34170854  0.34994368
##        NO   0.40327869  0.51854975  0.46096654  0.54899497  0.53283878
##        Sum  1.00000000  1.00000000  1.00000000  1.00000000  1.00000000
##           diag3_mod
## readmitted         V45         Sum
##        <30  0.09785203  2.42237035
##        >30  0.41288783  7.69227224
##        NO   0.48926014 10.88535741
##        Sum  1.00000000 21.00000000
```