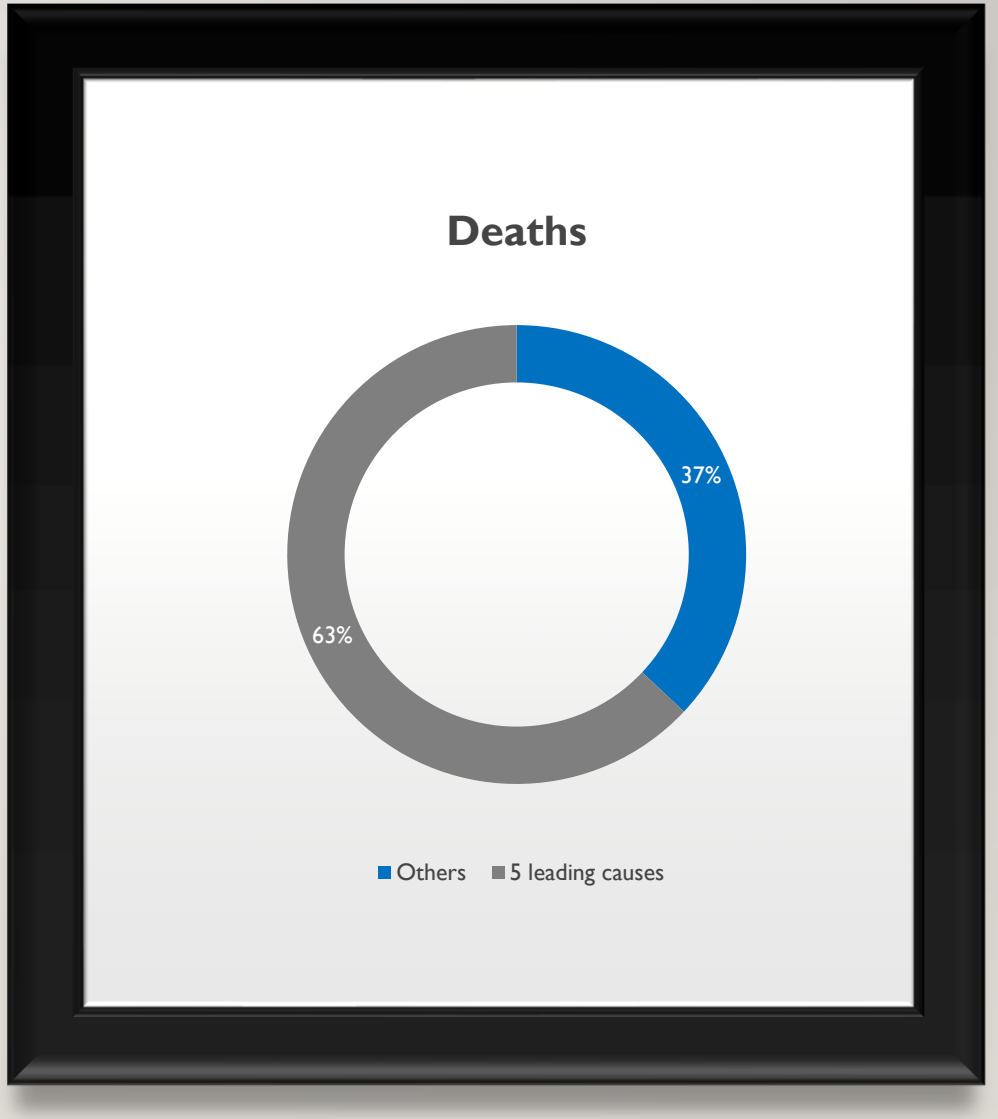
PROPHYLACTIC TREATMENT

KIRAN PRASANNADAS

INTRODUCTION

- Each year, nearly 900,000 Americans die prematurely from the five leading causes of death
- Heart disease, Cancer, Chronic lower respiratory diseases, Stroke, and Unintentional injuries
- Together they accounted for 63% of all U.S. deaths



Courtesy: Centers for Disease Control and Prevention
https://cdc.gov

SAVE LIFE

- 20% – 40% of the 900,000 deaths due to the leading causes are preventable
- Identify who is at risk!
- Proper screening

PROBLEM STATEMENT

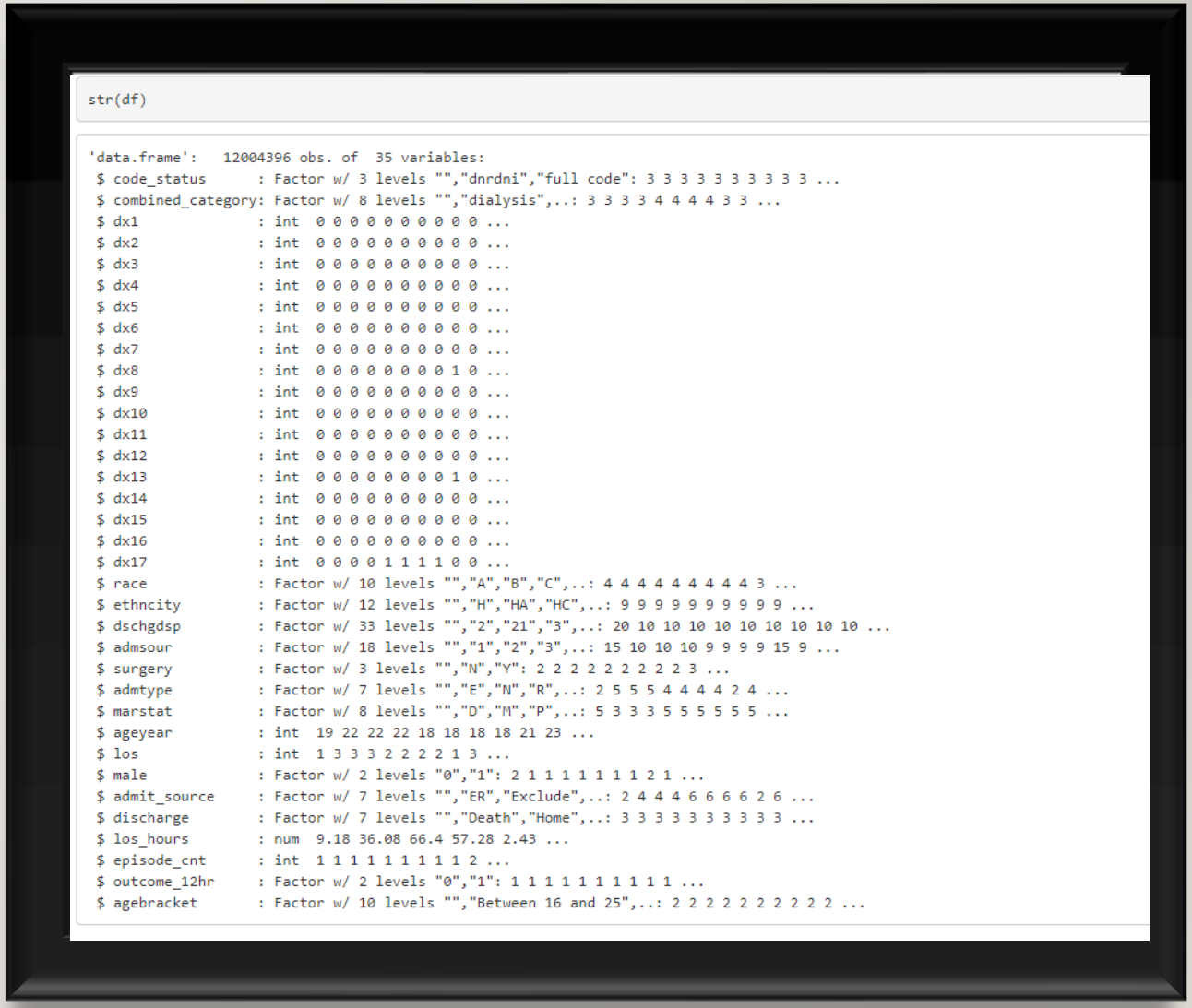
- Prophylactic treatment against a future bad outcome may be possible if some knowledge or intelligence is available to predict the onset.
- A patient once admitted is observed across time.
- We focus on to predict the onset of bad outcome after 12 hours.

DATASET

- 13,178,226 observations, Each row is a time-stamped observation entry
- 146 variables
- Target : Outcome_12hr
- Location details, dates, admit source, discharge disposition, LOS, ethnicity, marital status, lab values, Charlson Comorbidity indexes

FEATURE & OBSERVATION SELECTION

- Removed the features with majority NA values
- Highly correlated features (LOS, LOS hours, marital status, age..)
- Factors with high number of levels (Drug name)
- Removed the observations that are marked as Exclude (Exclude==1)
- Deteriorating Patient Eligible is “in”
- Reduced the dataset to 12,004,396 obs. & 32 variables



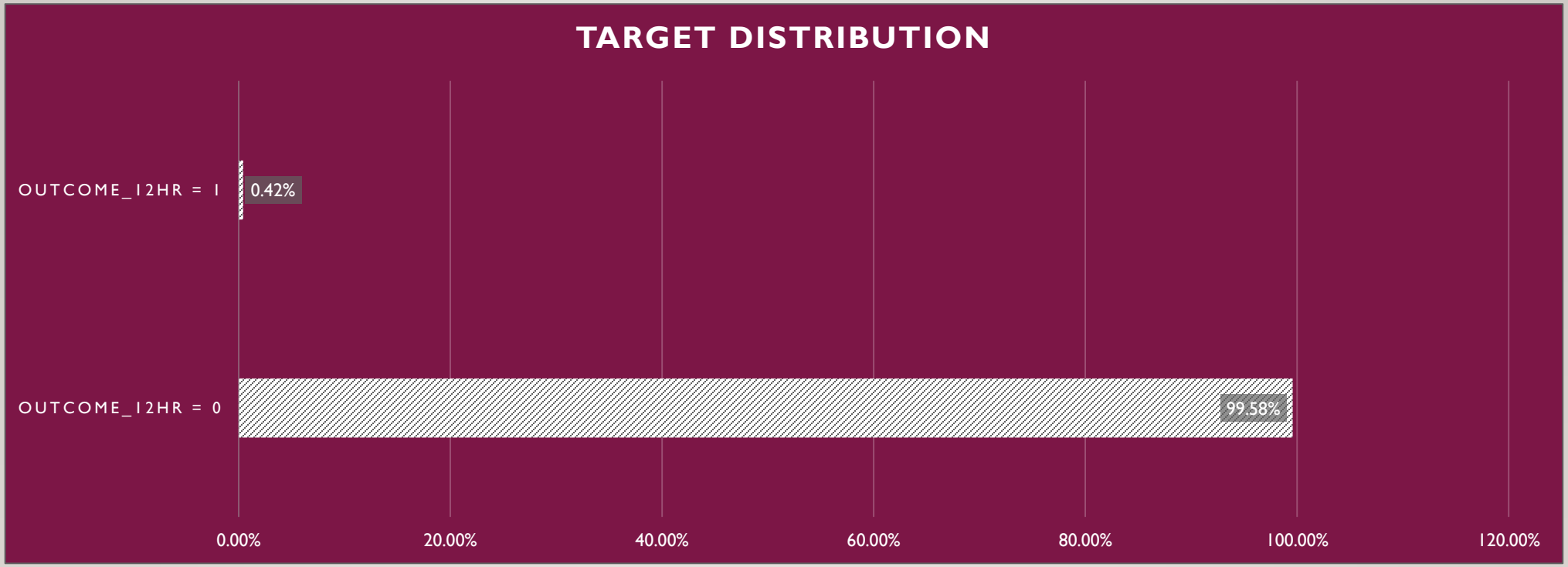
```
str(df)
```

```
'data.frame':   12004396 obs. of  35 variables:
 $ code_status      : Factor w/ 3 levels "","dnrdni","full code": 3 3 3 3 3 3 3 3 3 3 ...
 $ combined_category: Factor w/ 8 levels "","dialysis",..: 3 3 3 3 4 4 4 4 3 3 ...
 $ dx1              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx2              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx3              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx4              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx5              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx6              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx7              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx8              : int  0 0 0 0 0 0 0 0 1 0 ...
 $ dx9              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx10             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx11             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx12             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx13             : int  0 0 0 0 0 0 0 0 1 0 ...
 $ dx14             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx15             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx16             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dx17             : int  0 0 0 0 1 1 1 1 0 0 ...
 $ race             : Factor w/ 10 levels "","A","B","C",..: 4 4 4 4 4 4 4 4 4 3 ...
 $ ethncity         : Factor w/ 12 levels "","H","HA","HC",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ dschgdsp         : Factor w/ 33 levels "","2","21","3",..: 20 10 10 10 10 10 10 10 10 10 ...
 $ admsour          : Factor w/ 18 levels "","1","2","3",..: 15 10 10 10 9 9 9 9 15 9 ...
 $ surgery          : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 3 ...
 $ admtype          : Factor w/ 7 levels "","E","N","R",..: 2 5 5 5 4 4 4 4 2 4 ...
 $ marstat          : Factor w/ 8 levels "","D","M","P",..: 5 3 3 3 5 5 5 5 5 5 ...
 $ ageyear          : int  19 22 22 22 18 18 18 18 21 23 ...
 $ los              : int  1 3 3 3 2 2 2 2 1 3 ...
 $ male             : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ admit_source     : Factor w/ 7 levels "","ER","Exclude",..: 2 4 4 4 6 6 6 6 2 6 ...
 $ discharge        : Factor w/ 7 levels "","Death","Home",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ los_hours        : num  9.18 36.08 66.4 57.28 2.43 ...
 $ episode_cnt      : int  1 1 1 1 1 1 1 1 1 2 ...
 $ outcome_12hr     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ agebracket       : Factor w/ 10 levels "","Between 16 and 25",..: 2 2 2 2 2 2 2 2 2 2 ...
```

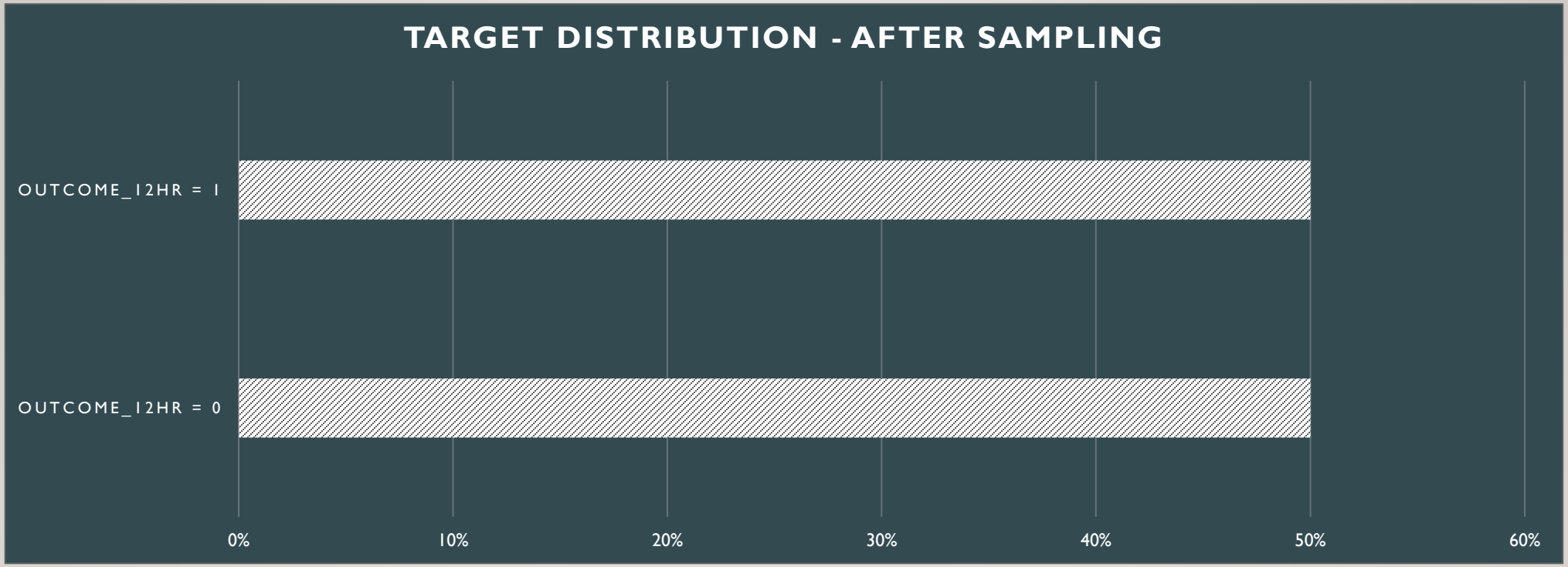
CORRELATION: CHARLSON COMORBIDITY INDEXES

	dx1	dx2	dx3	dx4	dx5	dx6	dx7	dx8	dx9	dx10	dx11	dx12	dx13	dx14	dx15	dx16	dx17
dx1	1.00	0.16	0.11	0.07	0.00	0.05	0.00	-0.02	0.06	0.06	-0.03	0.10	-0.04	-0.02	0.00	0.01	-0.05
dx2	0.16	1.00	0.10	0.07	0.00	0.25	-0.01	0.02	0.09	0.14	-0.05	0.34	-0.01	-0.09	-0.01	0.03	-0.07
dx3	0.11	0.10	1.00	0.15	0.01	0.09	0.02	-0.02	0.03	0.14	-0.01	0.16	-0.03	-0.04	-0.01	0.03	-0.05
dx4	0.07	0.07	0.15	1.00	0.09	0.02	0.02	0.00	0.03	0.06	0.08	0.06	0.00	-0.03	0.00	0.01	-0.03
dx5	0.00	0.00	0.01	0.09	1.00	0.00	-0.01	-0.02	0.01	0.00	0.04	0.02	-0.01	-0.02	0.00	0.00	-0.02
dx6	0.05	0.25	0.09	0.02	0.00	1.00	0.01	0.01	0.06	0.04	-0.03	0.15	0.00	-0.04	-0.01	0.03	-0.04
dx7	0.00	-0.01	0.02	0.02	-0.01	0.01	1.00	0.05	0.01	0.02	-0.01	0.04	0.06	0.00	0.00	0.01	-0.01
dx8	-0.02	0.02	-0.02	0.00	-0.02	0.01	0.05	1.00	0.05	0.02	-0.01	0.04	0.46	-0.02	0.02	-0.02	0.02
dx9	0.06	0.09	0.03	0.03	0.01	0.06	0.01	0.05	1.00	-0.09	0.01	0.09	0.04	-0.02	-0.01	-0.01	-0.01
dx10	0.06	0.14	0.14	0.06	0.00	0.04	0.02	0.02	-0.09	1.00	-0.02	0.28	0.02	-0.05	0.00	-0.02	-0.05
dx11	-0.03	-0.05	-0.01	0.08	0.04	-0.03	-0.01	-0.01	0.01	-0.02	1.00	-0.03	-0.01	-0.01	0.00	-0.02	-0.01
dx12	0.10	0.34	0.16	0.06	0.02	0.15	0.04	0.04	0.09	0.28	-0.03	1.00	0.04	-0.08	0.00	0.01	-0.04
dx13	-0.04	-0.01	-0.03	0.00	-0.01	0.00	0.06	0.46	0.04	0.02	-0.01	0.04	1.00	-0.02	0.01	-0.02	0.01
dx14	-0.02	-0.09	-0.04	-0.03	-0.02	-0.04	0.00	-0.02	-0.02	-0.05	-0.01	-0.08	-0.02	1.00	0.00	-0.02	0.19
dx15	0.00	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.02	-0.01	0.00	0.00	0.00	0.01	0.00	1.00	0.00	0.00
dx16	0.01	0.03	0.03	0.01	0.00	0.03	0.01	-0.02	-0.01	-0.02	-0.02	0.01	-0.02	-0.02	0.00	1.00	-0.02
dx17	-0.05	-0.07	-0.05	-0.03	-0.02	-0.04	-0.01	0.02	-0.01	-0.05	-0.01	-0.04	0.01	0.19	0.00	-0.02	1.00

dx1	Myocardial Infarct
dx2	Congestive Heart Failure
dx3	Peripheral Vascular Disease
dx4	Cerebrovascular Disease
dx5	Dementia
dx6	Chronic Pulmonary Disease
dx7	Ulcer
dx8	Mild Liver Disease
dx9	Diabetes
dx10	Diabetes with Organ Damage
dx11	Hemiplegia
dx12	Moderate/Severe Renal Disease
dx13	Moderate/Severe Liver Disease
dx14	Metastatic Solid Tumor
dx15	Aids
dx16	Rheumatologic Disease
dx17	Other Cancer



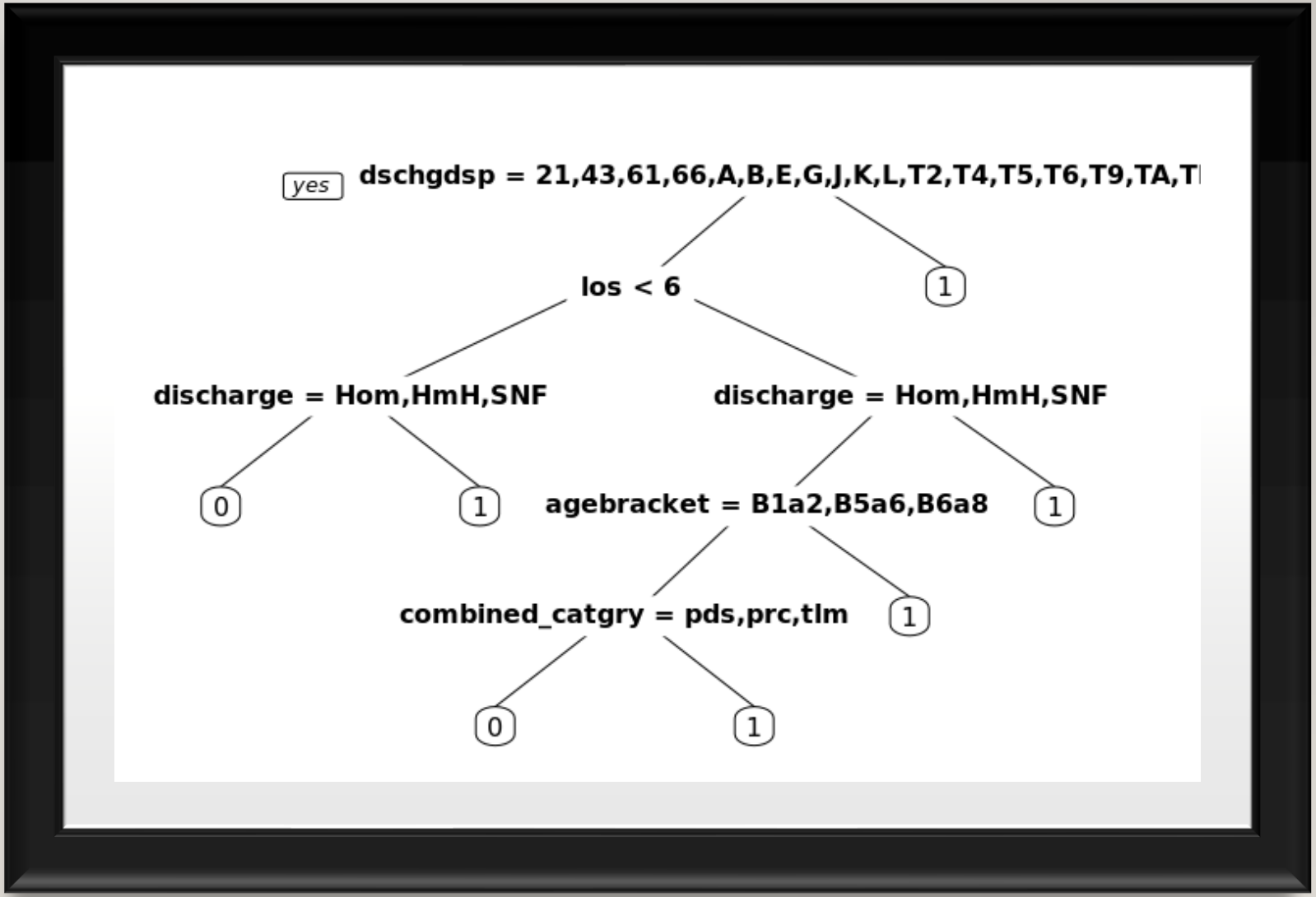
HIGHLY IMBALANCED TARGET DISTRIBUTION



SYNTHETIC MINORITY OVER SAMPLING TECHNIQUE – SMOTE (DMWR)

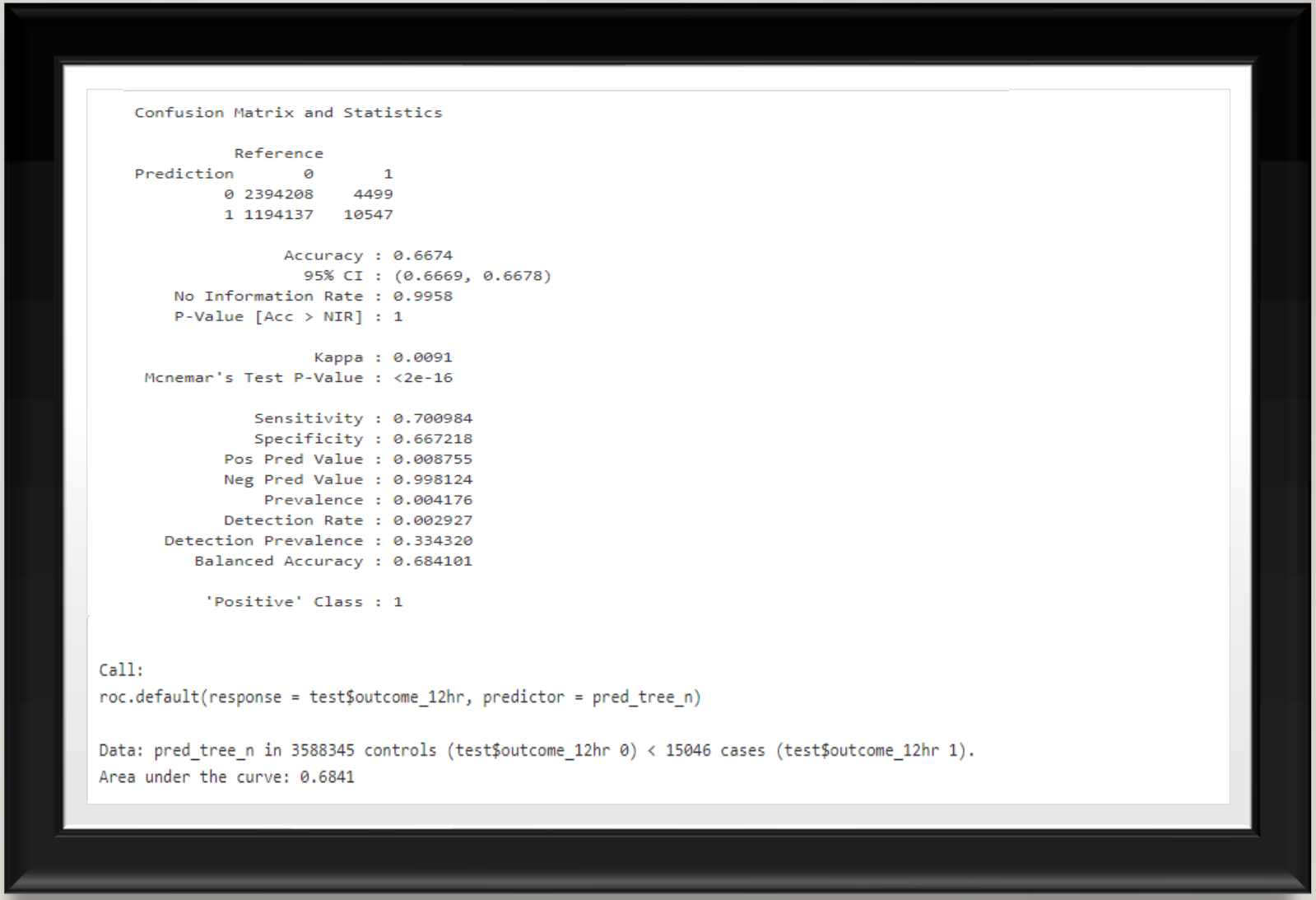
MODEL – DECISION TREE

DISCHARGE DISPOSITION, LOS, DISCHARGE, AGEBRACKET, COMBINED CATEGORY



PREDICTION ACCURACY : DT

Accuracy: 66.74%
Sensitivity : 70.09%
AUC: 68.41%



```
Confusion Matrix and Statistics

          Reference
Prediction       0       1
         0 2394208    4499
         1 1194137   10547

               Accuracy : 0.6674
                 95% CI : (0.6669, 0.6678)
    No Information Rate : 0.9958
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0091
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.700984
            Specificity : 0.667218
         Pos Pred Value : 0.008755
         Neg Pred Value : 0.998124
             Prevalence : 0.004176
         Detection Rate : 0.002927
   Detection Prevalence : 0.334320
      Balanced Accuracy : 0.684101

       'Positive' Class : 1


Call:
roc.default(response = test$outcome_12hr, predictor = pred_tree_n)

Data: pred_tree_n in 3588345 controls (test$outcome_12hr 0) < 15046 cases (test$outcome_12hr 1).
Area under the curve: 0.6841
```

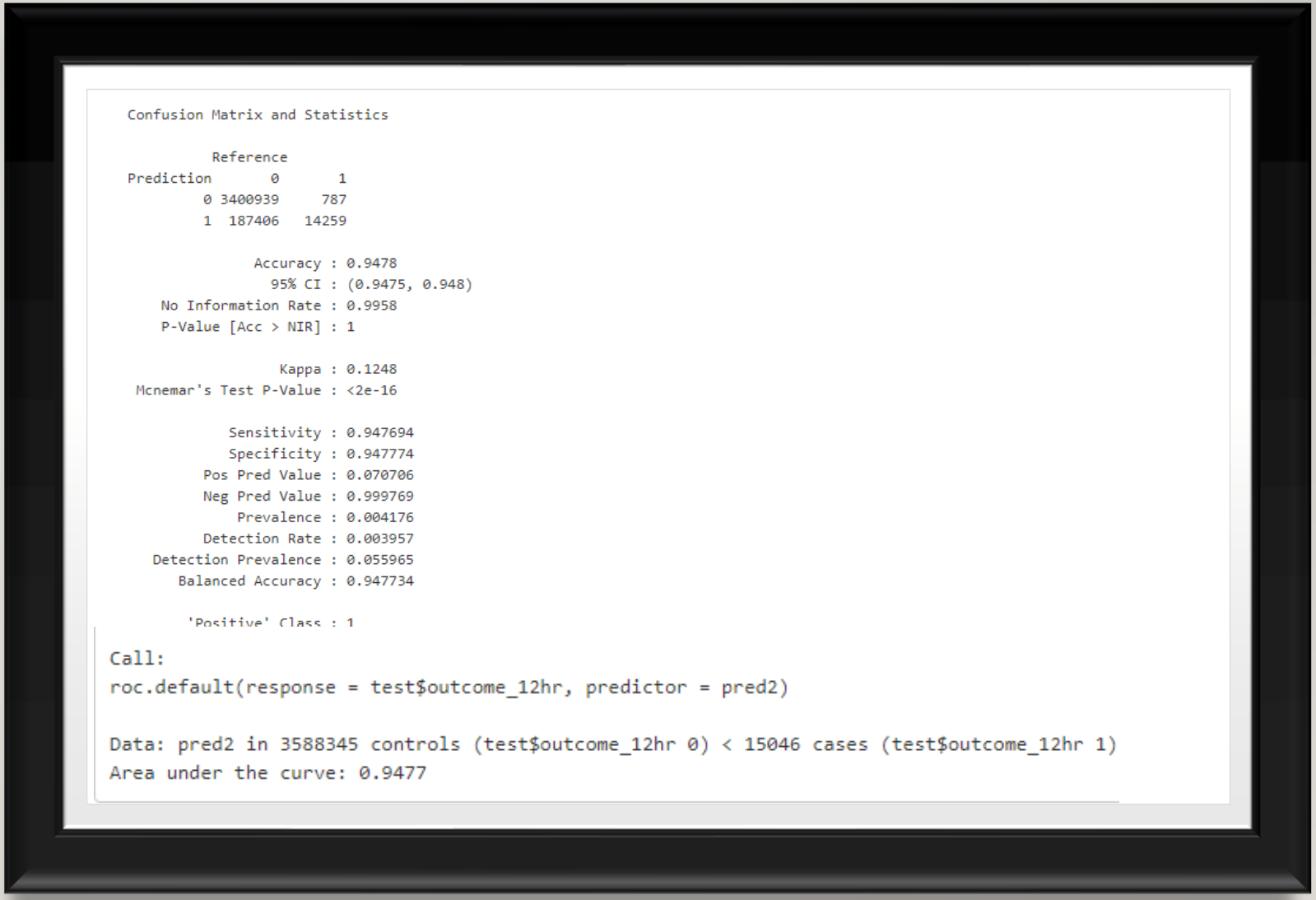
MODEL – RANDOM FOREST

LOS, DISCHDSP, DISCHARGE

	MeanDecreaseGini
code_status	1066.132677
combined_category	1656.745808
dx1	741.727784
dx2	1164.601875
dx3	847.006348
dx4	695.222459
dx5	414.680225
dx6	1228.845688
dx7	349.042218
dx8	656.757751
dx9	985.631283
dx10	635.357300
dx11	457.086957
dx12	1222.739901
dx13	355.991920
dx14	566.996492
dx15	9.459095
dx16	448.184911
dx17	926.741839
race	1736.575532
ethncity	1143.709682
dschgdsp	7806.416722
admsour	3563.825564
surgery	1116.412195
admtype	2361.641137
marstat	2172.283218
los	9321.935275
male	1038.631758
admit_source	2660.919641
discharge	5401.737412
episode_cnt	3428.047263
agebracket	3004.960724

PREDICTION ACCURACY : RF

Accuracy: 94.78%
Sensitivity: 94.77%
AUC: 94.77%



```
Confusion Matrix and Statistics

          Reference
Prediction       0       1
         0 3400939     787
         1  187406   14259

               Accuracy : 0.9478
                 95% CI : (0.9475, 0.948)
    No Information Rate : 0.9958
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1248
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.947694
            Specificity : 0.947774
         Pos Pred Value : 0.070706
         Neg Pred Value : 0.999769
             Prevalence : 0.004176
         Detection Rate : 0.003957
   Detection Prevalence : 0.055965
      Balanced Accuracy : 0.947734

       'Positive' Class : 1
```

```
Call:
roc.default(response = test$outcome_12hr, predictor = pred2)

Data: pred2 in 3588345 controls (test$outcome_12hr 0) < 15046 cases (test$outcome_12hr 1)
Area under the curve: 0.9477
```

CONCLUSION

- Developed a model that could predict the onset of a bad outcome within 12 hours.
- Random forest with 70:30 split gave good accuracy.
- The model prediction of the onset of bad outcome can save lives by giving urgent attention and taking preventive measures.