Predicting the length of stay and readmission probability of a patient in a hospital in Barcelona

Max van Esso Castellet

26 de junio de 2016

Introduction

Health service and hospitals are not topics that take so much time or importance on our day to day thoughts for most of us, but they are indeed crucial in guaranteeing our living standards. Public health care and public hospitals function on a complex balance between efficiency, budget restraints and the fulfill of moral and etic duties. The intricacy of how all this is achieved year after year is only known by the specialists; doctors, nurses, managers etc.

A common issue they have to deal with is the budget restrictions or limitations (a condition imposed by the government in order to conceed public funds) based on the number of readmissions that the health center has. On one hand a deeper study and prediction of the readmissions could be both interesting and key in the future development of hospitals, not only to reduce the budget sanctions but to improve the quality of the atention and health after a patient leaves the hospital.

On the other hand, the efficient allocation and use of their resources (human assets, lab tests and procedures. . .) can contribute to the profit of health services and hospitals. The availability of beds is an indicator of this efficient allocation of resources. Therefore, predicting the patient's length of stay in the hospital can be beneficial for the doctors and managers to act and react efficiently to each patient and circumstance.

All these challenges are particularly interesting for data science projects since their identification, model and prediction are not only important but they can also be rather complicated. Understanding this data and its challenges can lead to a powerful breakthough, identifying waste in the system, which can allow the healthcare service to work more efficiently. The increase in efficiency could terefore potientially improve the allocation of resources and lead to an improved level of care for patients.

Data description

For this project we were provided with data from a hospital in Barcelona. The dataset provides us with the following variables:

-*Age*, *Gender*, *City* (where the patient lives), *Region* (where the patient lives), *Country* (where the patient lives), *Zip Code*, *Basic Health Area*, *Country of Birth* (where the patient was borned), *Insurance Company*, *Primary Diagnostic*, *Secondary diganostic/s*, *Number of comorbidities*, *Especial clinical condition/s*, *Mental illness* (binary whether the patient has one or not), *Alcohol and drugs* (binary whether the patient took some or not), *Neoplasm* (binary whether the patient has it or not), *Mentally handicapped* (binary whether the patient has it or not), *Respiratory disease* (binary whether the patient has one or not), *Diabetes* (binary whether the patient has it or not), *Heart failure* (binary whether the patient had it or not), *Loss of wieght* (binary whether the patient had it or not), *Depression* (binary whether the patient has one or not), *Anemia* (binary whether the patient has it or not), *PCC* (binary whether the patient is a chronic complex patient or not), *MACA* (binary whether the patient has a complex chronic disease or not), *Number of drugs prescrived at discharge*, *Date of admission*, *Time of admission*, *Day of the week of admission*, *Year of admission*, *Month of admission*, *Date of discharge*, *Time of discharge*, *Day of the week of discharge*, *Year of discharge*, *Month of discharge*, *Lenght of stay* (LoS), *Medical specialty that handled that patient*, *Oncological area* (binary whether the patient was treated by the oncological area or not), *Admission origin*, *Number of procedures*, *Number of lab tests*, *Number of yearly admissions*, *Number of yearly emergencies*, *Number of yearly visits*, *Days until* *readmission*, *Readmission of type 1* (whether a patient was admited within the next 30 days after he was discharged and by a diagnostic that belongs to the same diagnostic group the original diagnostic he was admited for belonged to), *Readmission of type 2* (same as readmission of type 1 but within the next 60 days after the patient was discharged instead of 30), *Readmission of type 3* (any patient that came back to the hospital within the next 15 days for any reason).

The response variables used in this project's analysis will be *LoS*, *Readmission of type 1*, *Readmission of type 2* and *Readmission on type 3*.

The number of observations of the original dataset are 18892.

```
##   EdatEnAlta Sexe  Poblacio Provincia    Pais CodiPostal AreaBasicaSalut
## 1         67    D BARCELONA BARCELONA ESPANYA      08036   BARCELONA 2-E
## 2         67    D BARCELONA BARCELONA ESPANYA      08017   BARCELONA 5-C
## 3         55    D BARCELONA BARCELONA ESPANYA      08003   BARCELONA 1-A
## 4         83    D BARCELONA BARCELONA ESPANYA      08022   BARCELONA 5-E
## 5         85    D BARCELONA BARCELONA ESPANYA      08032   BARCELONA 7-D
##          PaisNaixement Cobertura GrupDiag NumComorbiditats CondClinEspec
## 1              ESPANYA       SCS      07I                9             Y
## 2              ESPANYA       SCS      02B                8             N
## 3 REPUBLICA DOMINICANA       SCS      02B                8             Y
## 4              ESPANYA       COL       VG                7             N
## 5              ESPANYA       COL       VG                1             N
##   MalaltiaMental AlcoholDrogues Neoplasia RetardMental
## 1              N              N         N            N
## 2              N              N         Y            N
## 3              N              N         Y            N
## 4              N              N         N            N
## 5              N              N         N            N
##   MalaltiaRespitatoria Diabetis IC PerduaPes Depressio Anemia PCC MACA
## 1                    N        Y  N         N         N      N   N    N
## 2                    N        N  N         N         N      N   N    N
## 3                    N        Y  N         N         N      N   N    N
## 4                    N        N  N         N         N      N   N    N
## 5                    N        N  N         N         N      N   N    N
##   NumMedAlta              DataIngres HoraIngres HoraIngres2
## 1          0 2012/11/22 00:00:00.000       1070   1200-1800
## 2          0 2012/11/26 00:00:00.000        523   0600-1200
## 3          0 2012/11/28 00:00:00.000        598   0600-1200
## 4          0 2012/11/30 00:00:00.000        801   1200-1800
## 5          0 2012/12/05 00:00:00.000       1041   1200-1800
##   DiaSetmanaIngres PeriodeIngres MesIngres                DataAlta
## 1                3          2012        10 2013/02/25 00:00:00.000
## 2                0          2012        10 2013/01/04 00:00:00.000
## 3                2          2012        10 2013/01/28 00:00:00.000
## 4                4          2012        10 2013/01/04 00:00:00.000
## 5                2          2012        11 2013/01/03 00:00:00.000
##   HoraAlta HoraAlta2 DiaSetmanaAlta PeriodeAlta MesAlta DuradaEstada
## 1      573 0600-1200              0        2013       1           95
## 2      360 0000-0600              4        2013       0           39
## 3      225 0000-0600              0        2013       0           61
## 4      876 1200-1800              4        2013       0           35
## 5      922 1200-1800              3        2013       0           29
##   Servei ServeiOnco              OrigenAdmissio NumProc NumLab
## 1   DERM          N                      Normal       0      0
## 2   CGEN          N                      Normal       0      0
## 3   ONCO          S                      Normal       0      0
## 4     MI          N OTROS CENTROS HOSPITALARIOS       0      0
## 5     MI          N OTROS CENTROS HOSPITALARIOS       0      0
##   NumAdmissionsAny NumUrgenciesAny NumVisitesAny DiesReadmissio
## 1                0               0             0              0
## 2                0               0             0              0
## 3                0               0             0              0
## 4                0               0             0              0
## 5                0               0             0              0
##   ReadmissioN1 ReadmissioN2 ReadmissioN3
## 1            N            N            N
## 2            N            N            N
## 3            N            N            N
## 4            N            N            N
## 5            N            N            N
```

Literature review and software

Literature review

A literature review revealed that little has been published with regard to this topic within Europe and that the vast majority of publications are from the USA [1, 3, 4, 5]. However, the main focus of this literature review will be on the following two papers due to their approach to the challenge of prediciting hospital readmissions which could later be replicated by our study.

- Development and Implementation of a Real-Time 30-Day Readmission Predictive Model

Cronin et al., (2014) used the data of Massachuesetts General Hospital (MGH) to explain the reasons behind the pressure on the hospitals to reduce readmission rates of patients. Cronin et al., (2014) explained that reliable predicitions for patients rehospitalisation would allow for tailored interventions to be introduced for patients most at risk. This would require "the creation of a functional predictive model specifically designed to support real-time clinical operations" which comes with some challenges to solve and options to test.

- A comparison of models for predicting early hospital readmissions

Differently to the previous paper, Futuoma et al.[2], (2015) uses a dataset called "the New Zealand National Minimum Dataset, obtained from the New Zealand Ministry of Health". The authors take into account that "there are a number of published risk models predicting 30 day readmissions for particular patient populations, however they often exhibit poor predictive performance and would be unsuitable for use in a clinical setting" and they "describe and compare several predictive models, some of which have never been applied to this task and which outperform the regression methods that are typically applied in the healthcare literature".

Software

The software used in this thesis to clean, process, analyze, model and predict has been the open source statistical program R as well as the data mining with open source machine learning software in java Weka.

In the case of R, several packages have been used in order to maximize the efficiency and the quality of the outcome. Such packages are:

- For the data vizualization work: corrplot, ggbiplot, ggplot2, reshape2, caret and splines.

- For the analysis, modeling and predicting part: caret, e1071, plyr, parallel, gbm, ranger, caTools, glmnet and mlr.

In the case of Weka, the analysis conducted has used the costMatrix function (manually tunned) with a random forest of 100 trees and a data splitting of 75% training set and 25% test set.

Further explanation on how exactly the algorithms have been trained and tuned along with the challenges overcame and results of the process will be explained in subsequent sections of the project.

Case study

Aim

The goal is to predict the lenght of stay of a patient and the probability of readmission. This predictions (if made accurately) will both trigger the efficiency and quality of the hospital in treating their patients and reduce the sanctions by the government. This would therefore augment the budget of the hospital, increasing the capacity of the hospital to invest in R&D, new equipment, personal training etc.

Methodology

Models will be trained and tested to predict both the length of stay and the probability of readmissions. We will then compare the models and evaluate which model performs better, according to the specific criteria that this problem requires.

To train the models we will use almost all the variables listed in the data description section. The choose of the final variables along with the variable transformation will be explained in more detail in the analysis section.

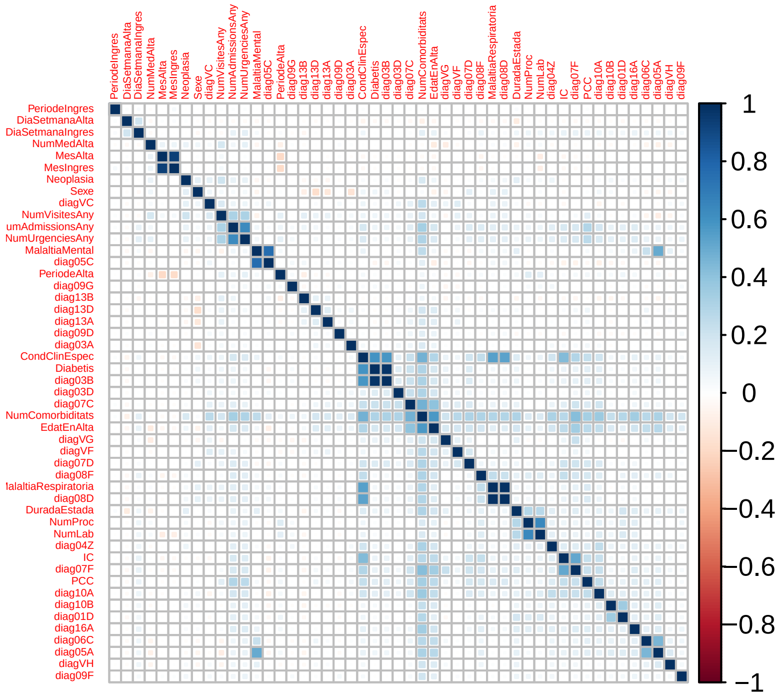
Visualization

To grasp a sense of what the dataset looks like, how the variables are related to each other and how the patients characteristics are distributed we used some visualisation techniques that will present right after.

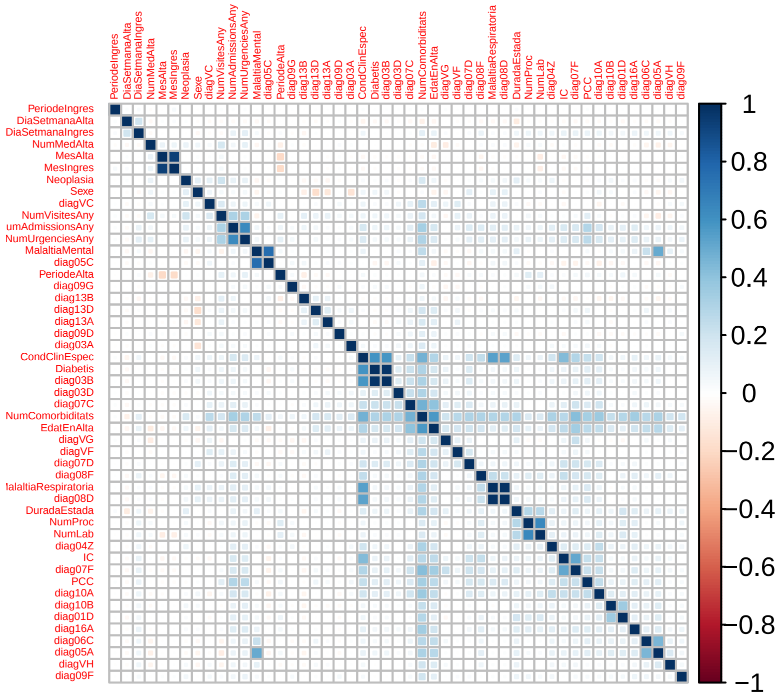
We first started by plotting some correlograms to see the possible correlation between the non-categoric variables (those that are numeric or binary and also doesn't have 0 or near zero variance).

There are three correlograms, the first one considers all the observations. The following two are subject to the readmission response variables, therefore one correlogram represents the correlation of the variables whose observations have a negative value in the response readmission variables. The other correlogram is considering those observations that have a positive response value in at least one of them.

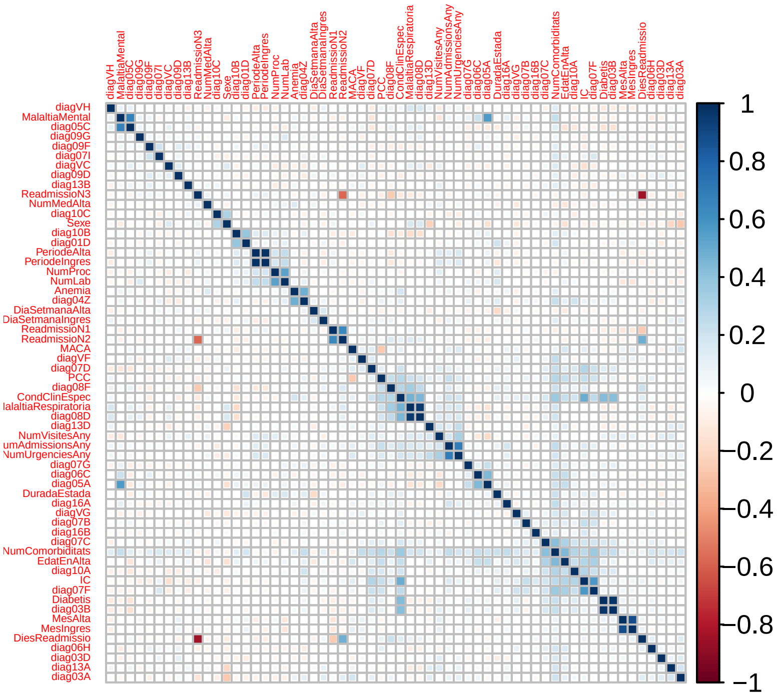
Correlogram with all observations



Correlogram with negative readmited observations



Correlogram with positive readmited observations



We can appreciate how we find some strong correlations like the correlation between *Month of admission* and *Month of discharge*. This is because the average length of stay in hospital is less than 30 days, therefore the month of admission and month of discharge are usually within the same month. We also see a strong correlation between some diagnostic codes and their strictly related diseases such as respiratory diseases, mental illness or diabetes. Finally the number of comorbidities has a strong correlation with some diagnostic codes (representing the most common secondary diagnostics).

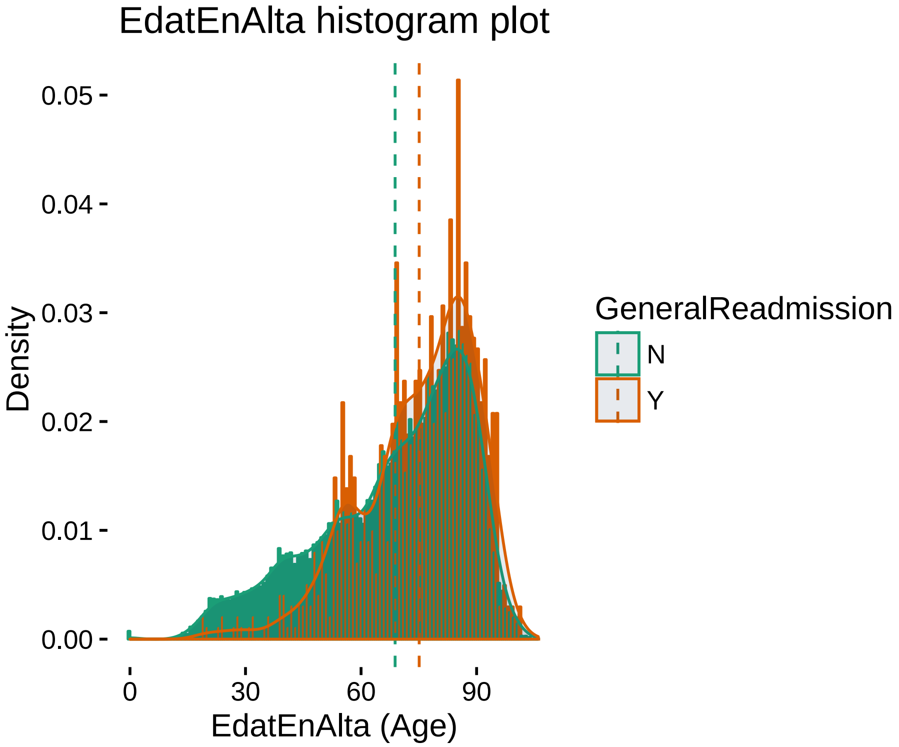
Most of the correlations we observe are positive (shown in blue while negative correlations will have a red tonality) and the pattern of correlation replicates through all three correlograms. One noteable difference is in the one including only the observations with any kind of readmission where some strong negative correlations appear.

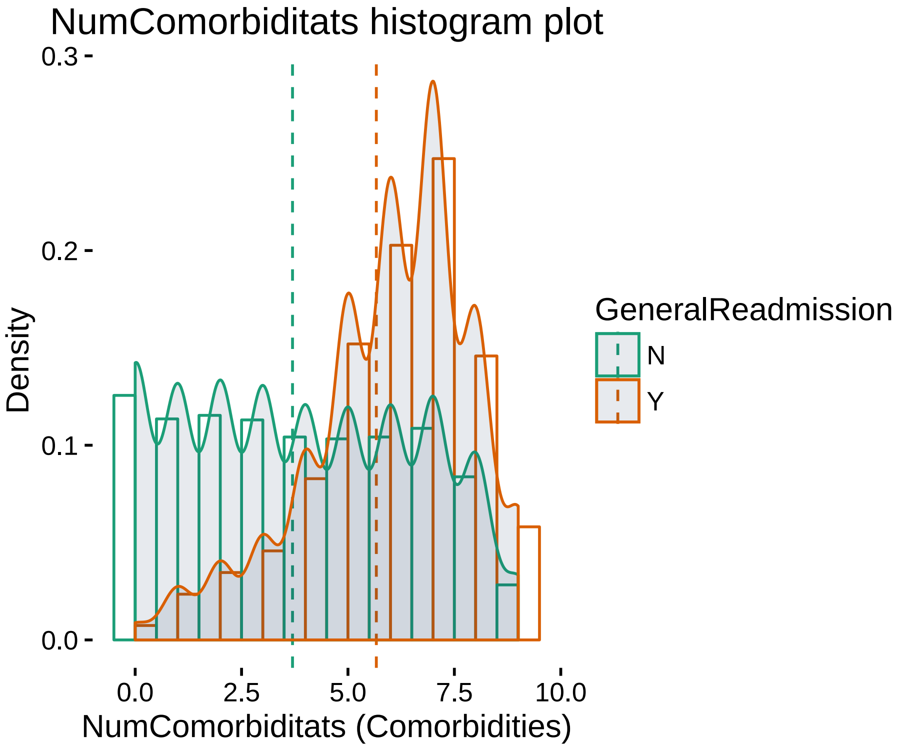
The strongest negative correlation appears between the *Days until readmission* variable and the *Readmission type 3* (R3) variable. This is due to the *R3* variable being defined as every readmission that happens for any reason within 15 days after discharge. As a result the positive responses of *R3* match exactly with the lower values of the *Days until readmission* variable.

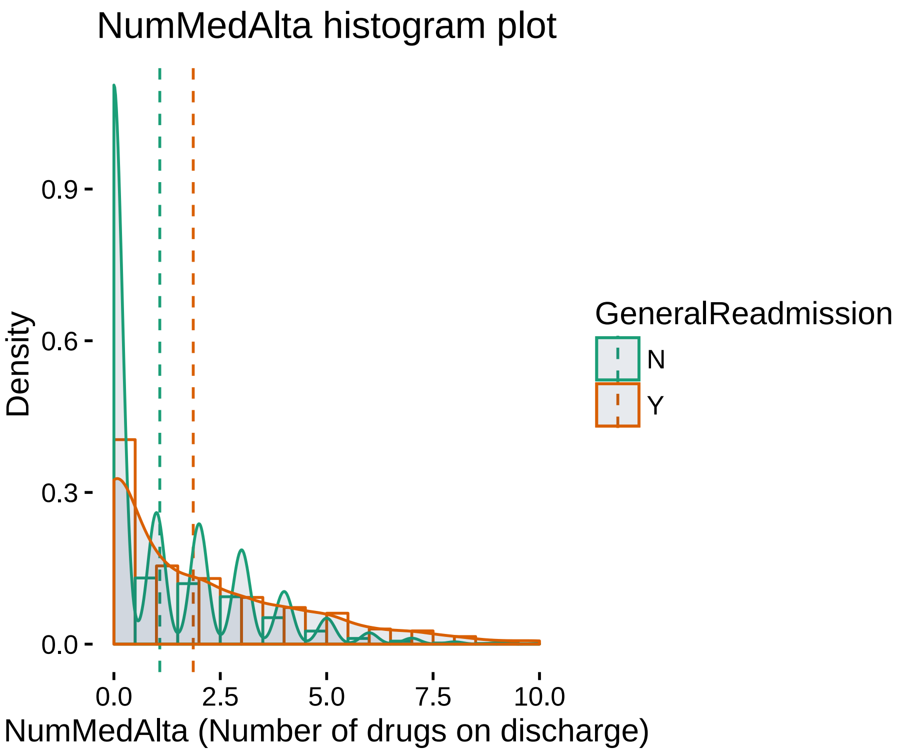
The other negative correlation appears between *R3* and the *Readmission type 2* (R2) variable. One reason for this could be because the R2 accepts readmissions up till 60 days after discharge and will capture those readmissions with the higher *Days until readmission* values (as opposed to *R3*). *R2* also have some medical restrictions: just returning to the hospital is enough for *R3* but not enough for *R1* and *R2*, as we explained above in the definition of the variables, so no all *R3* values will be at the same time *R2* or *R1*.

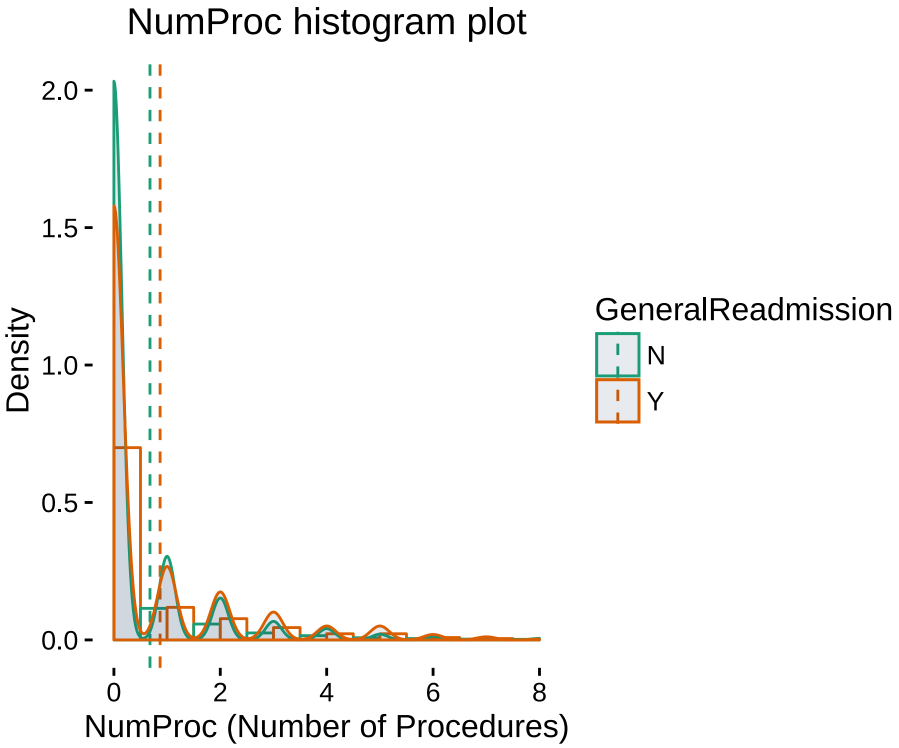
For some of the numeric variables it is also interesting to see how different the distributions are between the individuals with a positive response in, at least, one of the readmission variables and those with a negative response in al *R1*, *R2* and *R3*.

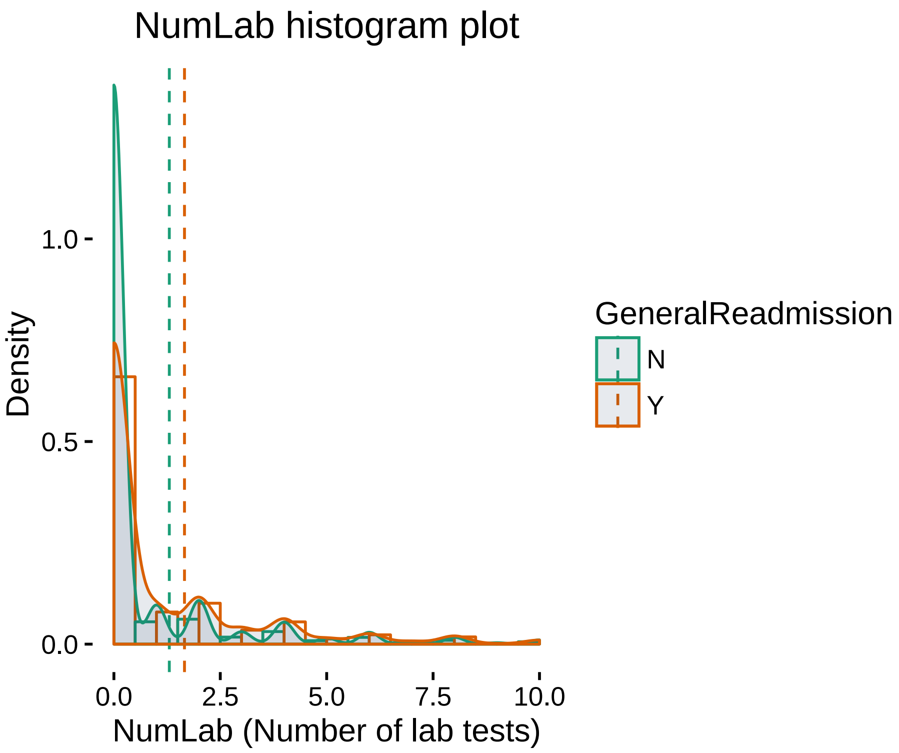
We built histograms to better display the information, plotting in different colors the different distributions and adding a dashed line that represents the mean of each group.


EdatEnAlta histogram plot
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
30
60
90
EdatEnAlta (Age)
GeneralReadmission
N
Y

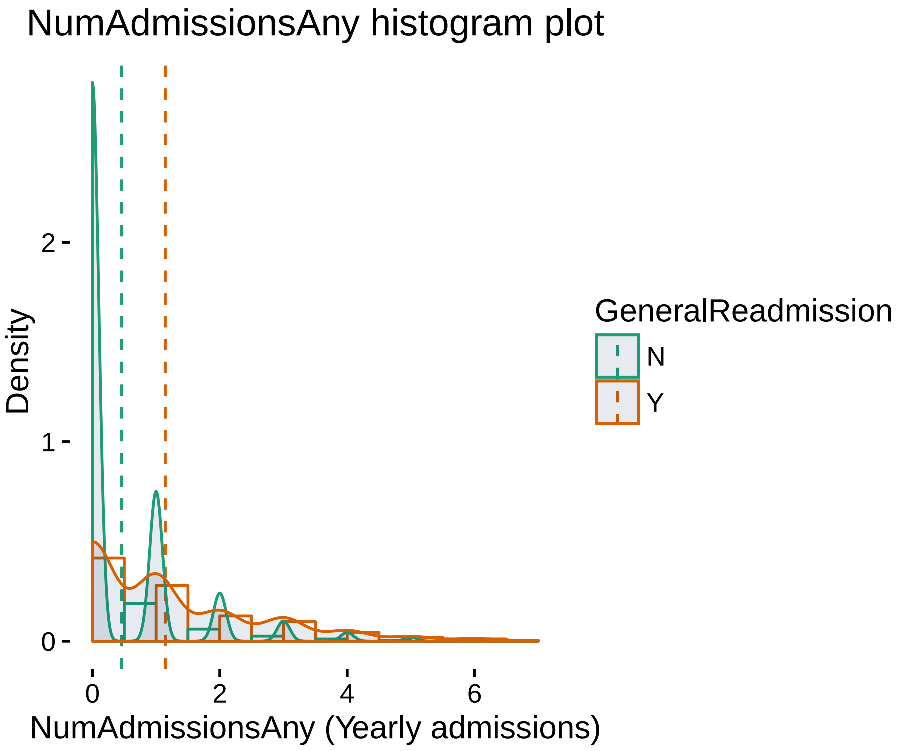

NumComorbiditats histogram plot
Density
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
NumComorbiditats (Comorbidities)
GeneralReadmission
N
Y


NumMedAlta histogram plot
Density
0.9
0.6
0.3
0.0
0.0
2.5
5.0
7.5
10.0
NumMedAlta (Number of drugs on discharge)
GeneralReadmission
N
Y

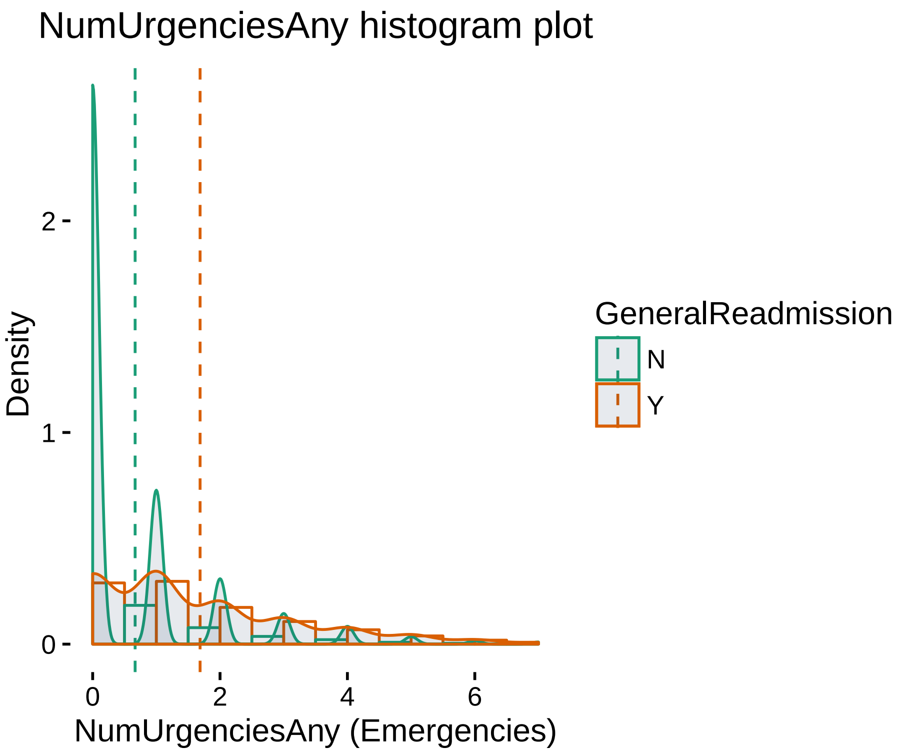

NumProc histogram plot
Density
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
NumProc (Number of Procedures)
GeneralReadmission
N
Y

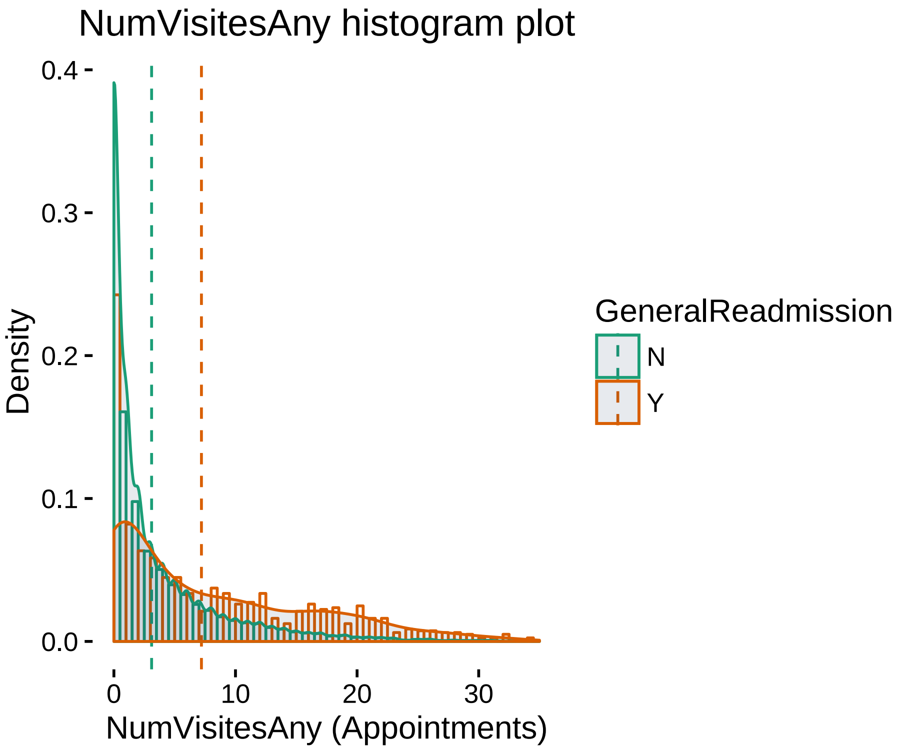


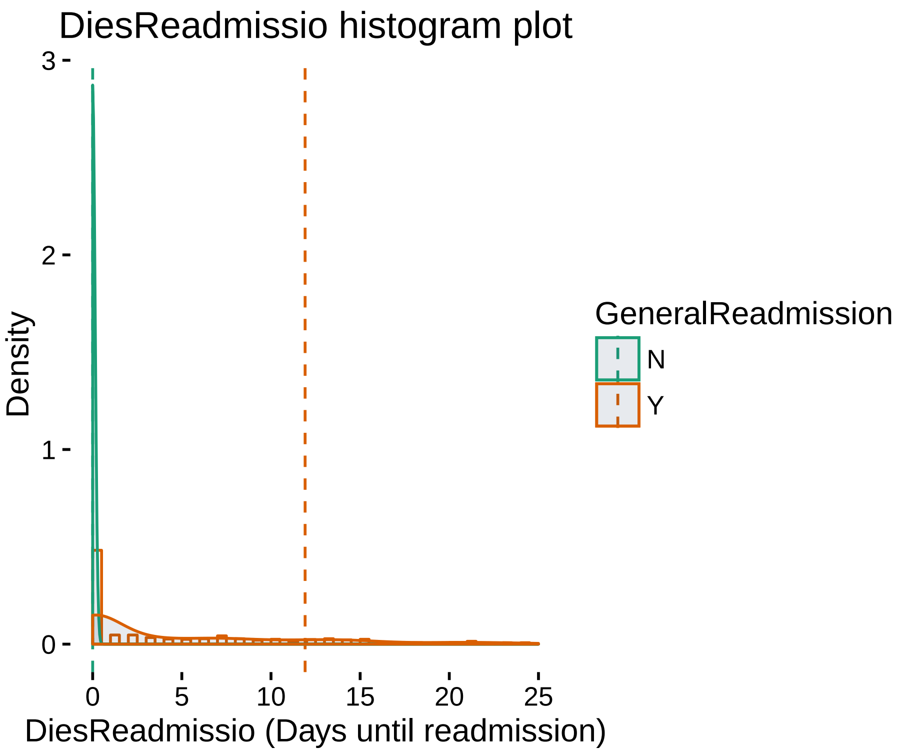
NumLab histogram plot



NumAdmissionsAny histogram plot

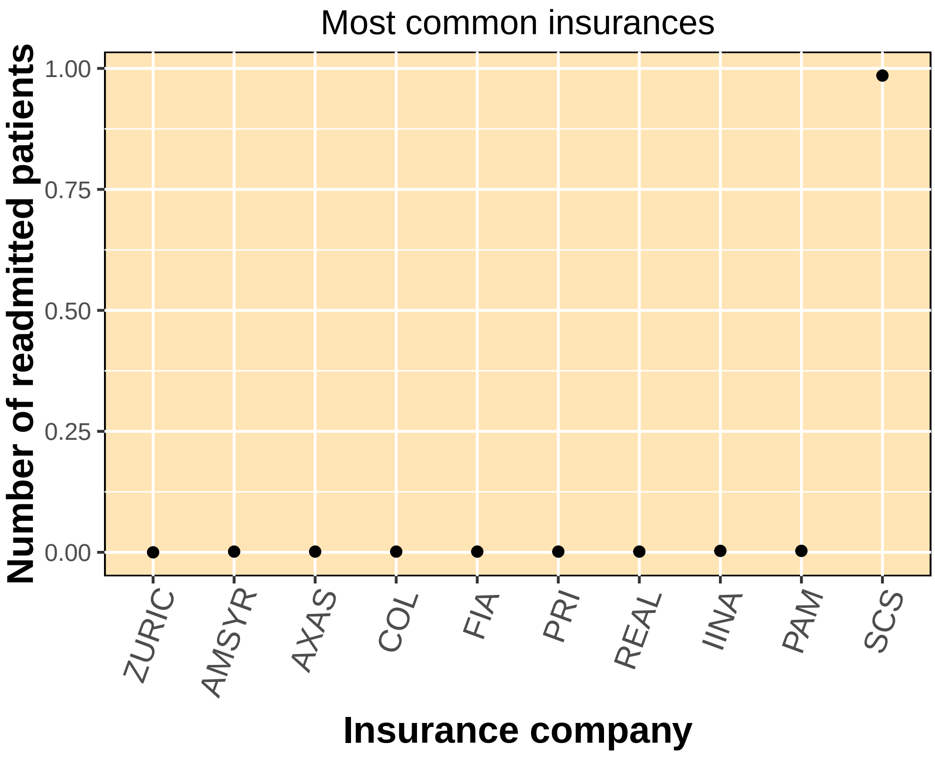

NumUrgenciesAny histogram plot

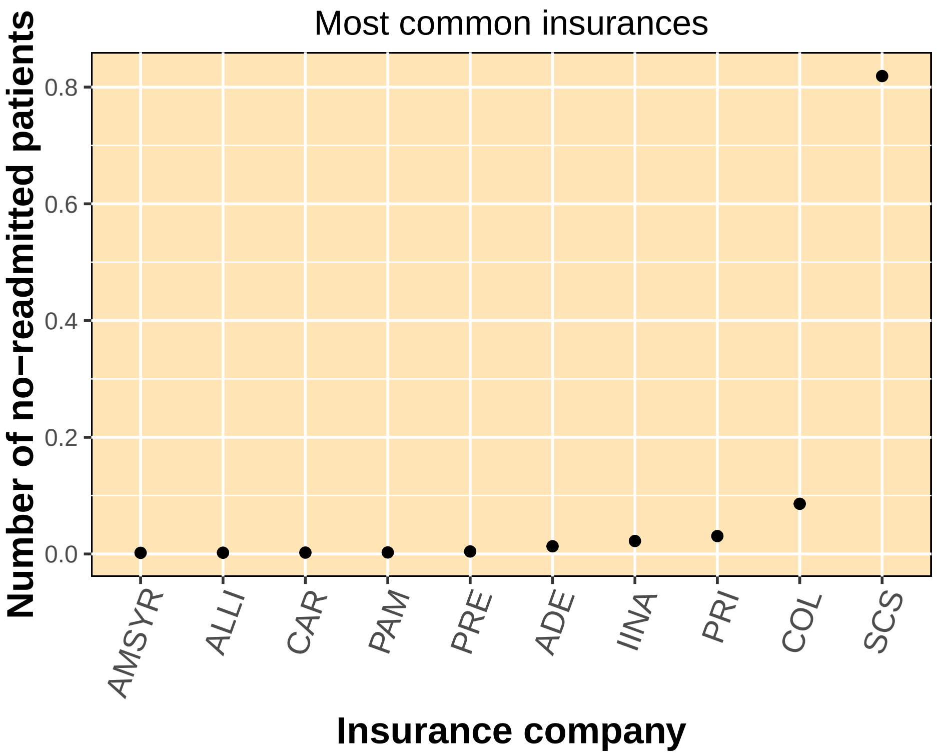

NumVisitesAny histogram plot

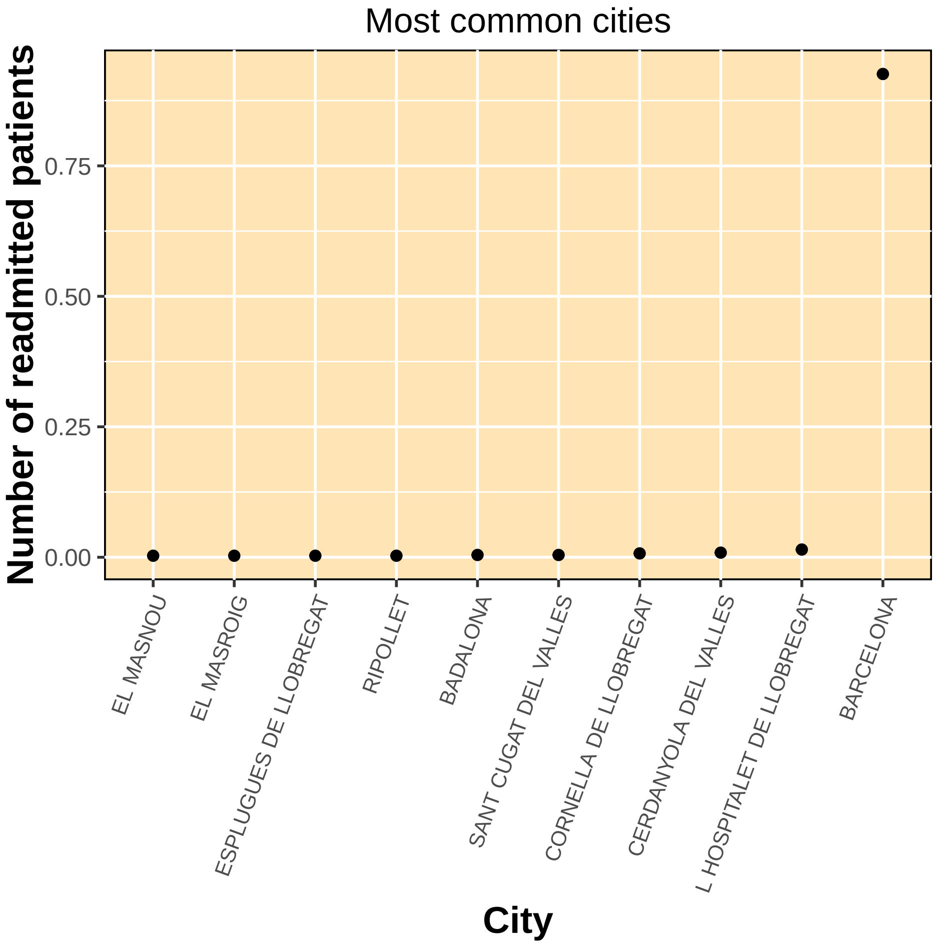


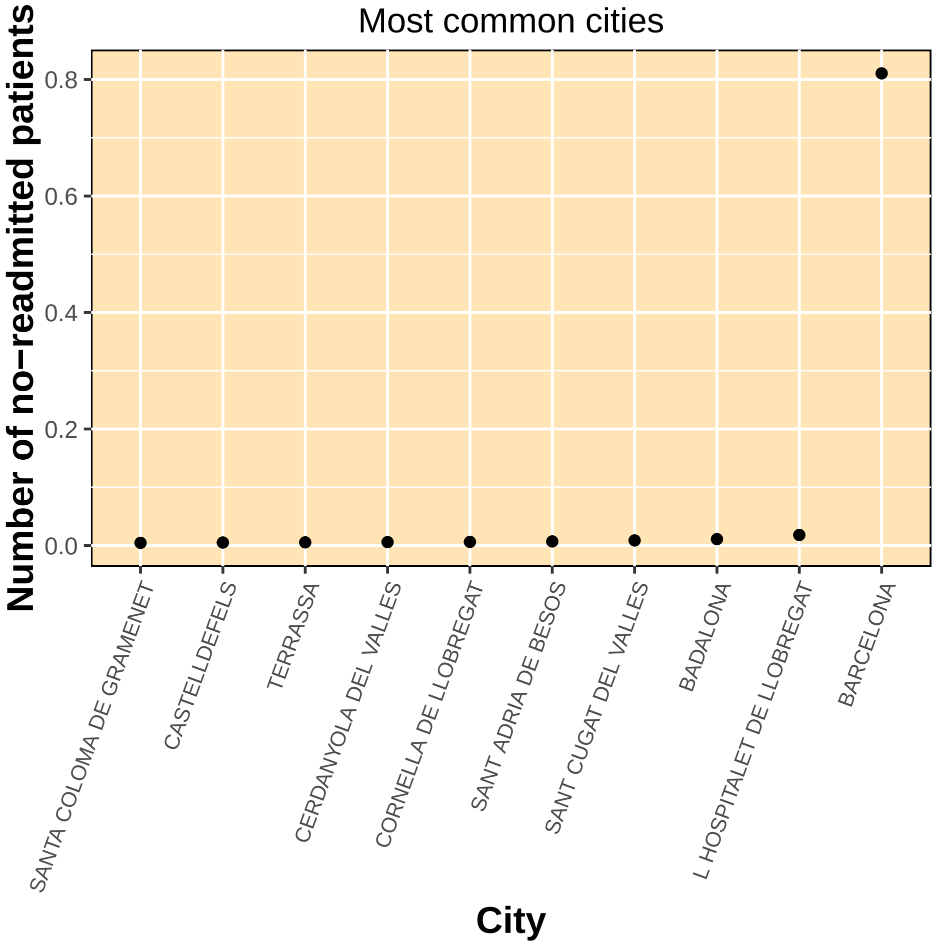
We can appreciate how the means are higher for the group with positive response in the readmission variables in all the histograms, this indicates that the patients that readmit are older, with more health problems, that required a longer and more complicated treatment and that attend the hospital more times along the year.

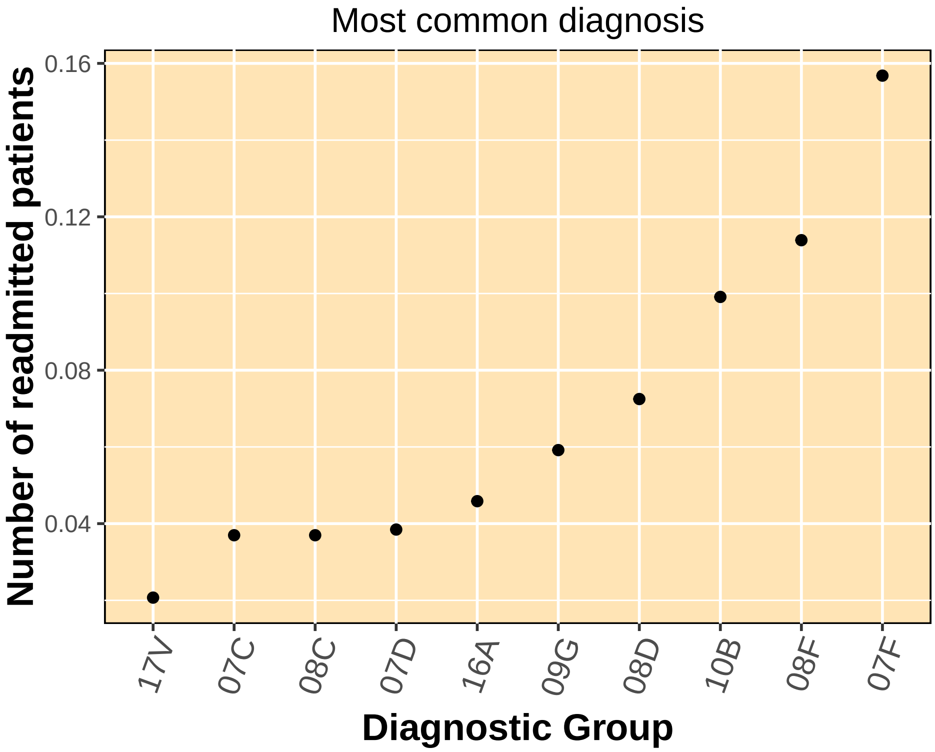
To visualize the most important categorical variables we made some plots where we can see the top 10 categories of each variable and the percentage of the number of observations belonging to each category.

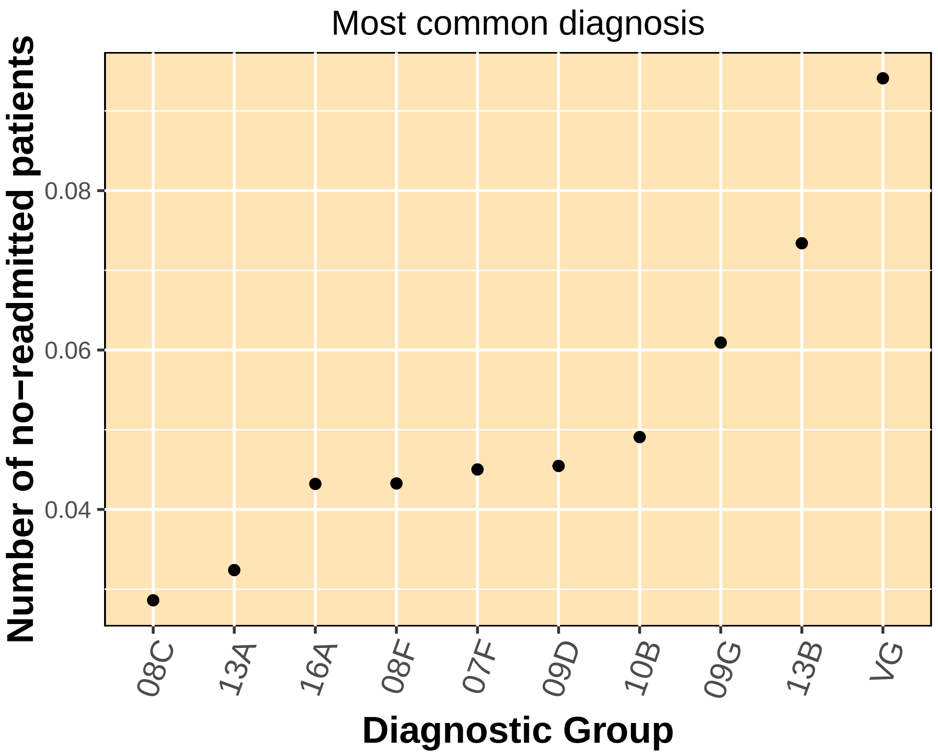


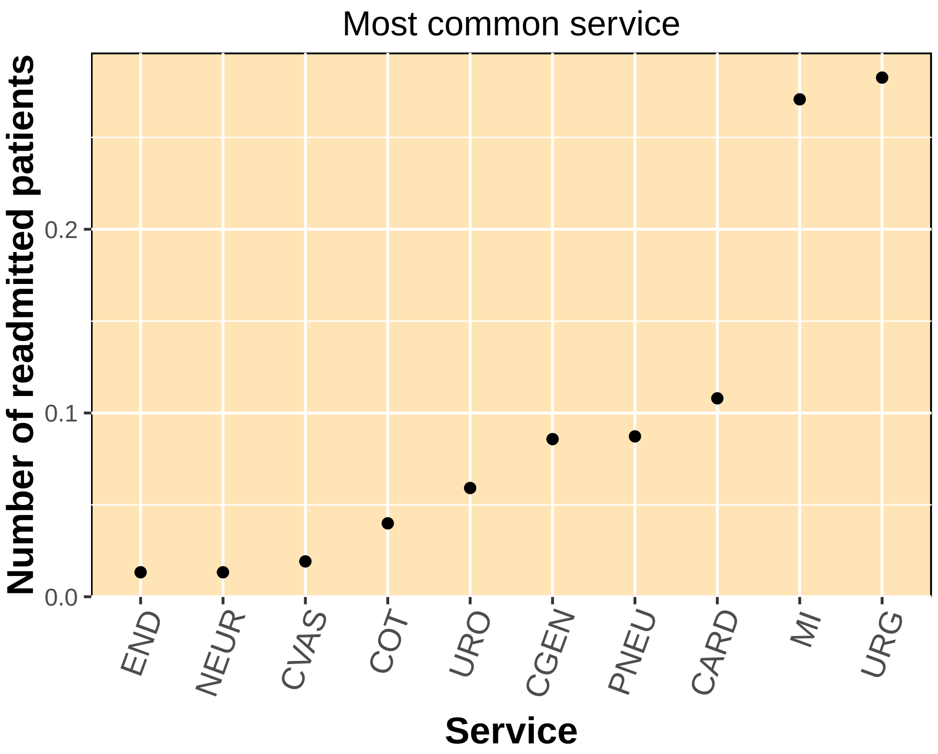

Most common insurances
Number of no−readmitted patients
0.8
0.6
0.4
0.2
0.0
AMSYR
ALLI
CAR
PAM
PRE
ADE
IINA
PRI
COL
SCS
Insurance company

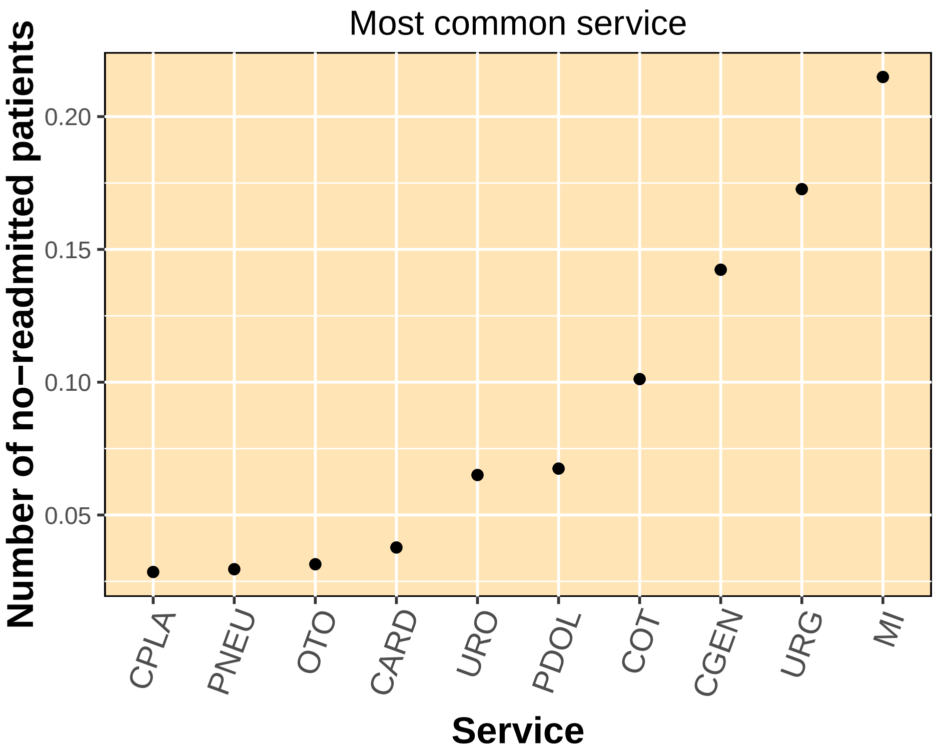

Most common cities
Number of readmitted patients
0.75
0.50
0.25
0.00
EL MASNOU
EL MASROIG
ESPLUGUES DE LLOBREGAT
RIPOLLET
BADALONA
SANT CUGAT DEL VALLES
CORNELLA DE LLOBREGAT
CERDANYOLA DEL VALLES
L HOSPITALET DE LLOBREGAT
BARCELONA
City

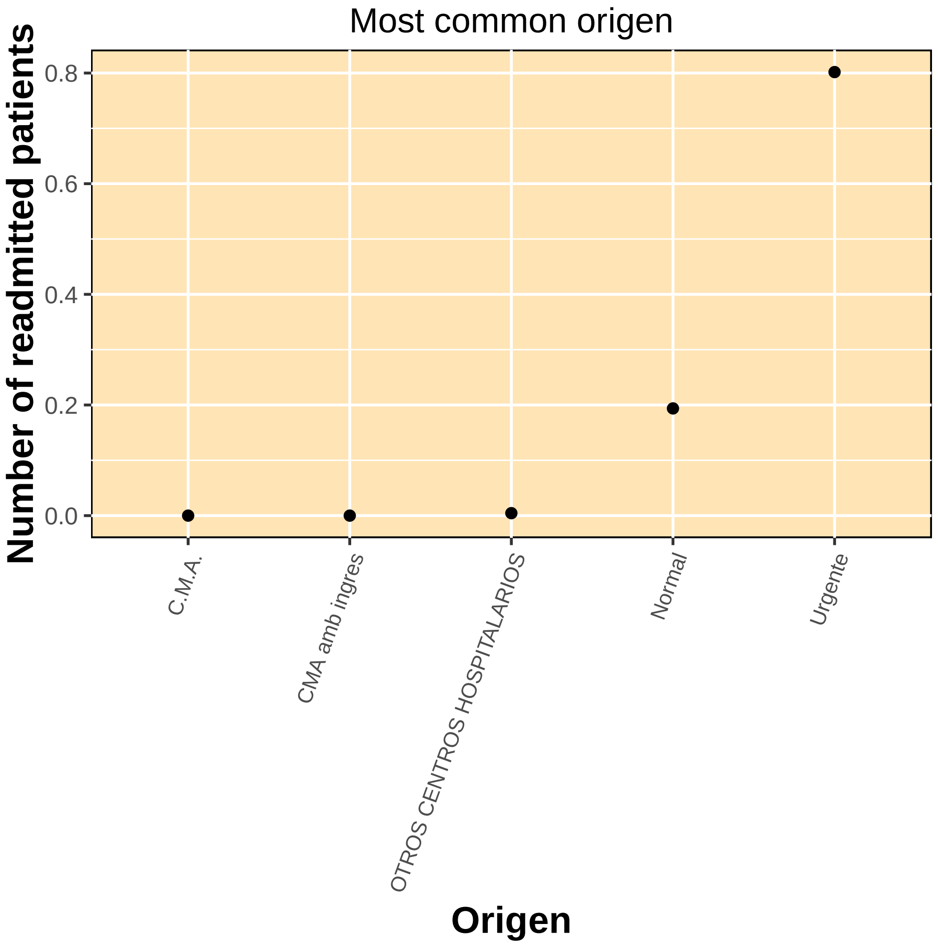

Most common cities
Number of no−readmitted patients
0.0
0.2
0.4
0.6
0.8
SANTA COLOMA DE GRAMENET
CASTELLDEFELS
TERRASSA
CERDANYOLA DEL VALLES
CORNELLA DE LLOBREGAT
SANT ADRIA DE BESOS
SANT CUGAT DEL VALLES
BADALONA
L HOSPITALET DE LLOBREGAT
BARCELONA
City

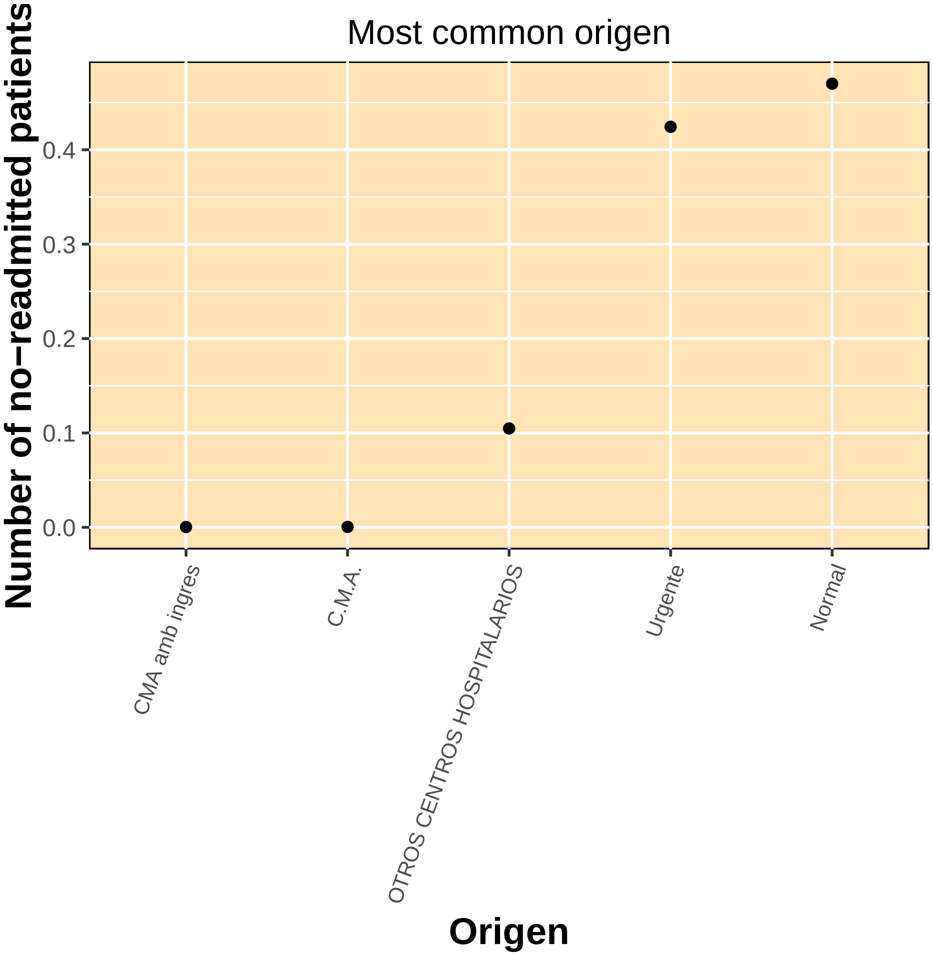

Most common diagnosis
Number of readmitted patients
0.04
0.08
0.12
0.16
17V
07C
08C
07D
16A
09G
08D
10B
08F
07F
Diagnostic Group


Most common diagnosis
Number of no−readmitted patients
0.04
0.06
0.08
08C
13A
16A
08F
07F
09D
10B
09G
13B
VG
Diagnostic Group


Most common service
Number of readmitted patients
0.0
0.1
0.2
END
NEUR
CVAS
COT
URO
CGEN
PNEU
CARD
MI
URG
Service


Most common service
Number of no−readmitted patients
0.20
0.15
0.10
0.05
CPLA
PNEU
OTO
CARD
URO
PDOL
COT
CGEN
URG
MI
Service


Most common origen
Number of readmitted patients
0.8
0.6
0.4
0.2
0.0
C.M.A.
CMA amb ingres
OTROS CENTROS HOSPITALARIOS
Normal
Urgente
Origen



We can see that in almost all variables the vast amount of observations are concentrated in one category. The data is highly clustered in this sense because all of the data cames from the same hospital and most of the patients there share a common profile.

There are interesting differences to be noted between the two *Most common diagnostic* plots and *Most common services* plots. They are significantly different between the no-readmitted patients and the readmitted ones. Both plots are correlated to each other because different diagnostics were probably taken care in different health services. Taking into account this differences we can start assuming that maybe the diagnostic has something to do with the readmission.

Analysis

To analyze our dataset and predict our response variables we droped some variables and needed to make a few changes in some other variables.

We droped the *Region*, *Country* (both where the patient was born and lives), *Zip Code* and *Basic Health Area*, *Date of Admission*, *Days until readmission* and *Date of Discharge* for efficient porpuses.

We generated the following variables:

-Population 2: this new variable breaks the *Population* variable into two categories: Barcelona and others. We did that because, as we saw in the visualization part we had the majority of observations grouped under the category Barcelona and therefore grouping all the other cities into one category gave us more predictive power and made more sense in general since the hospital is indeed based in Barcelona so the difference would be if the patient lives in Barcelona or not.

- Insurance company 2: this new variable breaks the *Insurance company* variable in two categories: social security and others. We did this for simillar reasons as the ones explained above in the *Population 2* variable.

- Primary diagnostic: what we did in this case was to group the primary diagnostics in groups that encapsule similar pathologies.

- Secondary diagnostic: what we did to this categorical variable was to transform each secondary diagnostic code into a binary variable and give it a 1 to all the observations that had that code in the old secondary diagnostic variable and a 0 otherwise. With this process we created 157 new binary variables.

- Time of admission: we transformed this variable and defined into the following blocks: from 00:00 to 06:00, from 6:00 to 12:00, from 12:00 to 18:00 and from 18:00 to 00:00, the intervals are with the first time included and the second one excluded. We justify this again for efficiency and predictive reasons, by grouping the times in these intervals we have more observations represented in each one than by considering the exact time for each observation.

- Time of discharge: we applied exactly the same procedure as we applied to the *Time of admission* variable.

Predictions

After this we moved on to the predictions. We first tried to predict the *LoS*. For that we tried gbm, glmnet and random forest (RF) methods and we realized that due to the distribution of the data (most of the values for LoS are 2) it was extremely hard to predict with accuracy the exact number of days. After some considerations on the actual value of predicting an exact number of days with a high probability of predicting them wrong we considered that for the information that the doctors are actually going to get from this prediction and for prediction accuracy purposes we should transform this numeric variable into a categorical one. The new *LoS* categorical variable had the following categories: less than 2 days, from 3 to 7 days and more than 7 days.

With this new variable we got the following results:

```
+-------------------+------------------+-----------------+------------------+
| GBM               | Less than 2 days |    3 to 7 days  | More than 7 days |
+===================+==================+=================+==================+
| Less than 2 days  |      93.59%      |      4.31%      |       6.1%       |
+-------------------+------------------+-----------------+------------------+
| 3 to 7 days       |       2.64%      |      80.7%      |     25.33%       |
+-------------------+------------------+-----------------+------------------+
| More than 7 days  |       3.76%      |       15%       |     68.57%       |
+-------------------+------------------+-----------------+------------------+
```

```
+-------------------+------------------+-----------------+------------------+
| GLMNET            | Less than 2 days |    3 to 7 days  | More than 7 days |
+===================+==================+=================+==================+
| Less than 2 days  |      90.41%      |     17.31%      |      4.69%       |
+-------------------+------------------+-----------------+------------------+
| 3 to 7 days       |       7.94%      |     62.71%      |     27.57%       |
+-------------------+------------------+-----------------+------------------+
| More than 7 days  |       1.64%      |     18.97%      |     67.73%       |
+-------------------+------------------+-----------------+------------------+
+------------------+-----------------+----------------+-----------------+
| RF               | Less than 2 days |    3 to 7 days  | More than 7 days |
+==================+=================+================+=================+
| Less than 2 days |      91.02%     |     8.53%      |      2.72%      |
+------------------+-----------------+----------------+-----------------+
| 3 to 7 days      |       6.82%     |    72.58%      |     25.33%      |
+------------------+-----------------+----------------+-----------------+
| More than 7 days |       2.15%     |    18.89%      |     71.95%      |
+------------------+-----------------+----------------+-----------------+
```

There are 2418 observations of the class *Less than 2 days* and the best prediction is made by GBM, 1207 of the class *3 to 7 days* and the best prediction is made by GBM and 1066 of the class *More than 7 days* and the best prediction is made by RF. Two out three of these predictions are made using GBM. Considering all this and the difference in the other predictions between models, if we had to pick a model as the model that performs better in predicting the response variable *LoS*, that would be GBM.

For the predictions of *R1*, *R2* and *R3* we went with the logistic regression and the glmnet and random forest.

The biggest challenge we had to face was the severe class imbalance that the dataset had. The positive response for *R1* represented only 1.54% of the observations, for *R2* represented 2.31% and for *R3* represented 2.22%. The approaches we took to solve this problem were different depending on the model.

For logistic regression and glmnet we optimized the threshold in order to maximize the F1 score, which is the harmonic mean of the precision and recall.

The following are the results for each model on each response variable:

- Logistic Regression (tuning the threshold and maximizing the F1 score)

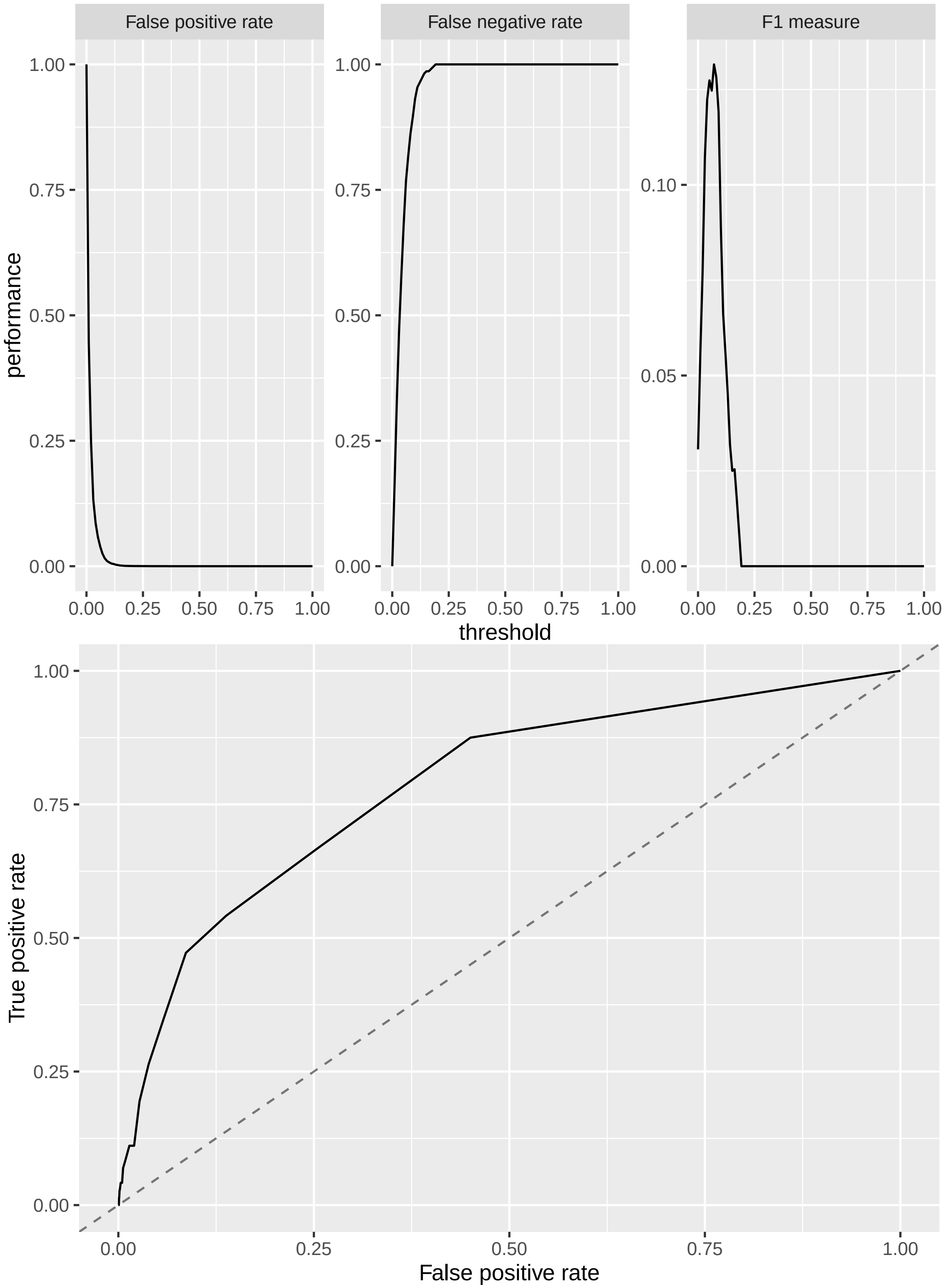
Readmission Type 1 results:

```
$train.metrics
       acc        auc        ppv      brier
0.98444492 0.79704448        NaN 0.01489564

$test.metrics
       acc        auc        ppv      brier
0.96099744 0.80295364 0.10071942 0.01467203

$train.matrix
       predicted
true        0 1 -SUM-
  0     13860 0     0
  1       219 0   219
  -SUM-   219 0   219

$test.matrix
       predicted
true       0   1 -SUM-
  0     4495 125   125
  1       58  14    58
  -SUM-   58 125   183
```


False positive rate
False negative rate
F1 measure
performance
1.00
0.75
0.50
0.25
0.00
0.10
0.05
0.00
0.25
0.50
0.75
1.00
threshold
True positive rate
False positive rate

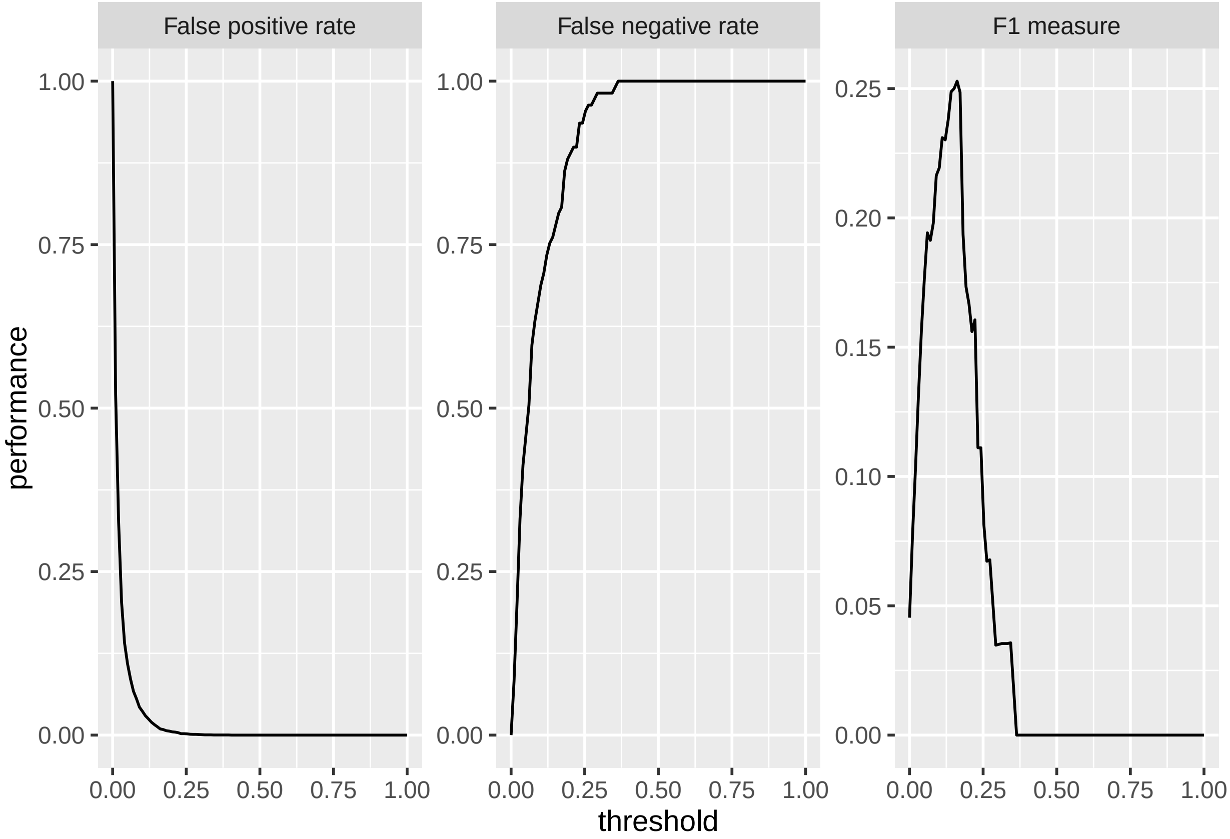
Readmission Type 2 results:

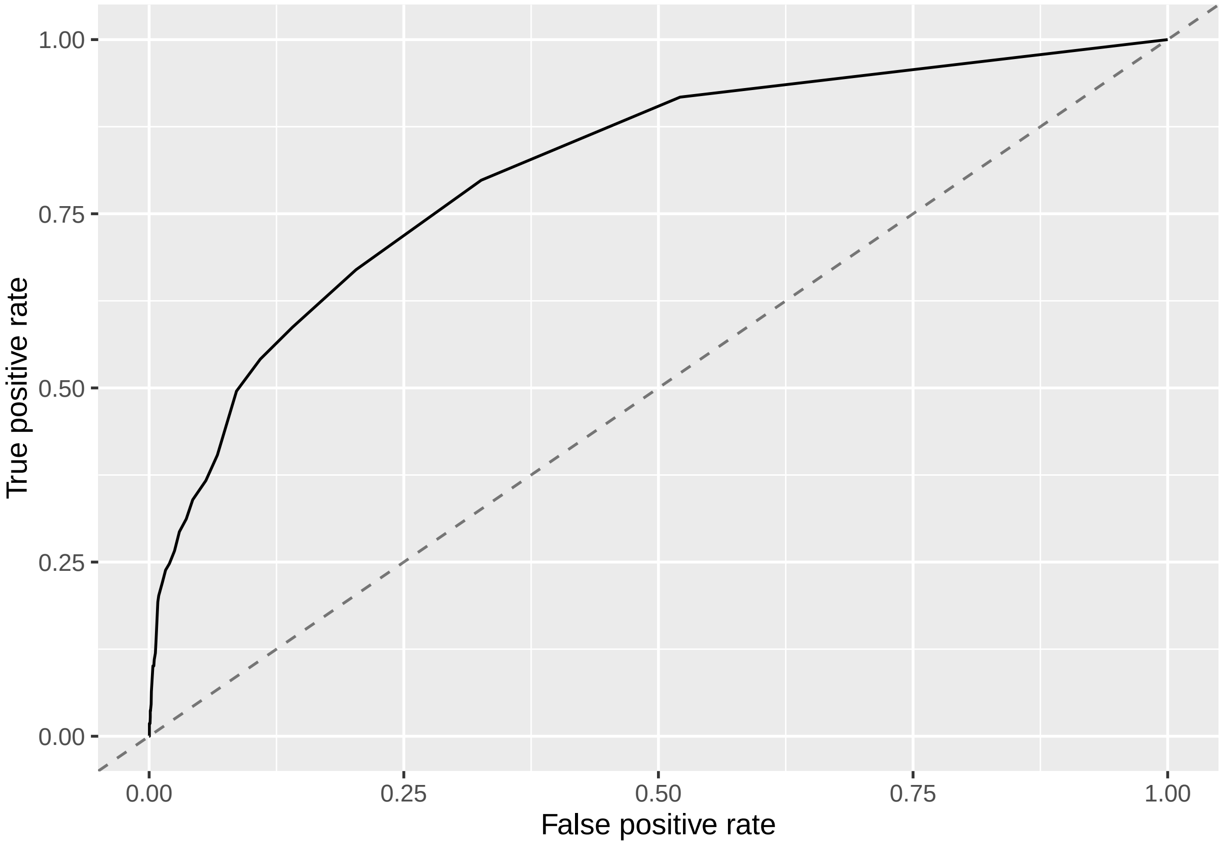
```
$train.metrics
       acc        auc        ppv      brier
0.97677392 0.82348734        NaN 0.02147668

$test.metrics
       acc        auc        ppv      brier
0.94842285 0.81646672 0.16915423 0.02108385

$train.matrix
       predicted
true        0 1 -SUM-
  0     13752 0     0
  1       327 0   327
  -SUM-   327 0   327

$test.matrix
       predicted
true       0   1 -SUM-
  0     4416 167   167
  1       75  34    75
  -SUM-   75 167   242
```





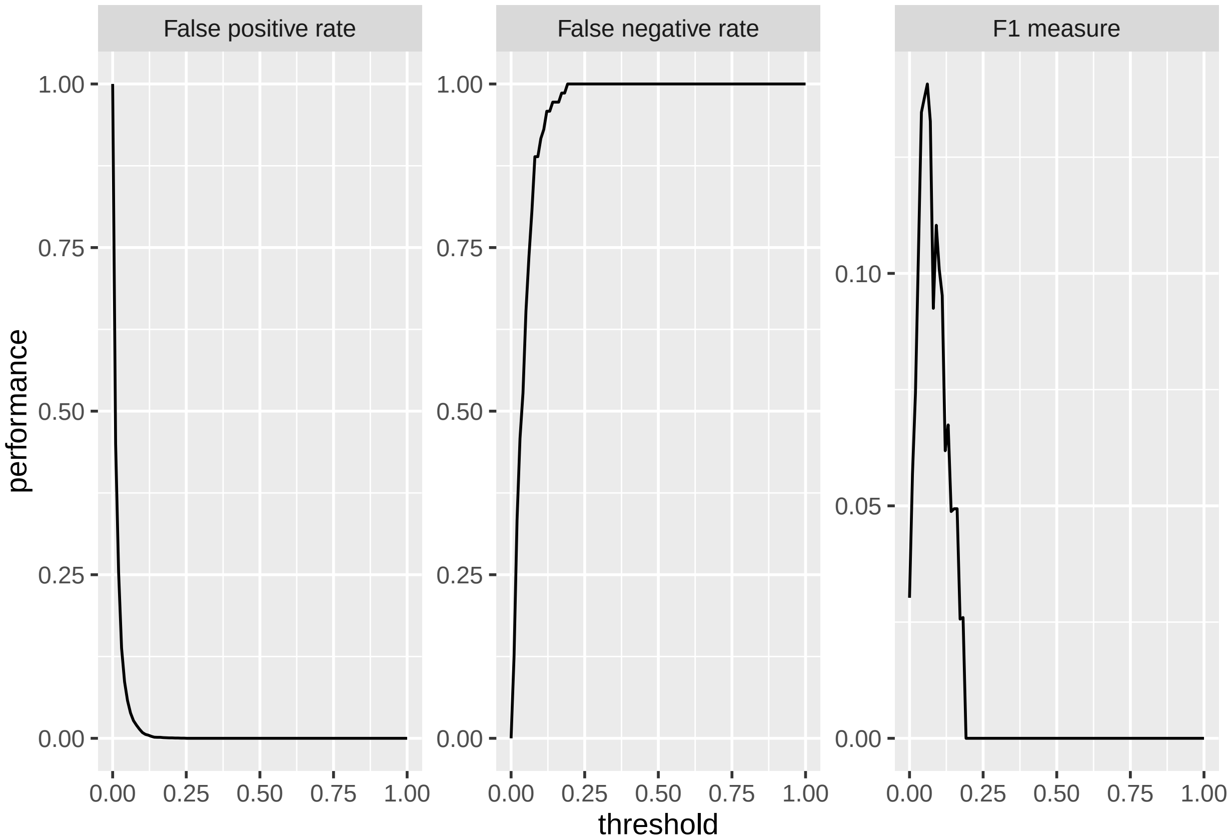
Readmission Type 3 results:

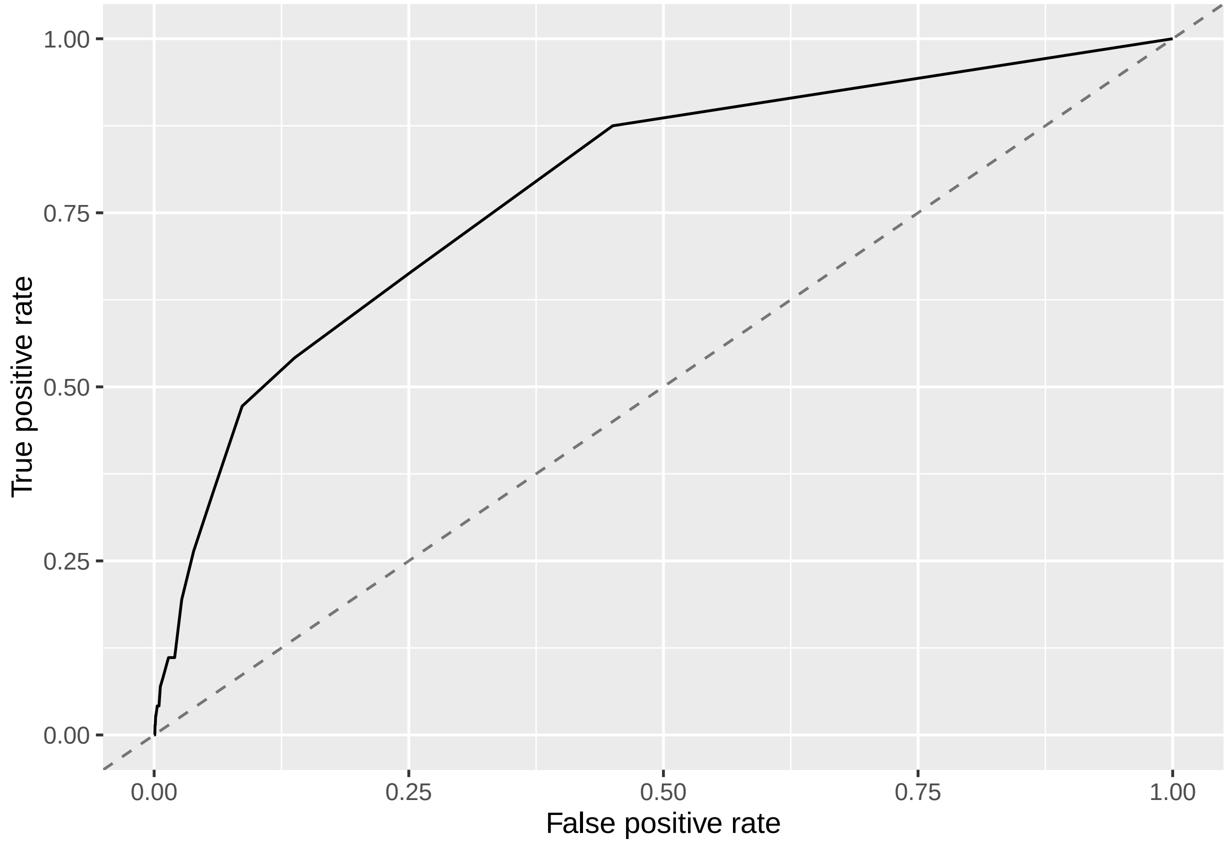
```
$train.metrics
       acc        auc        ppv      brier
0.97755522 0.78373454 0.00000000 0.02131409

$test.metrics
       acc        auc        ppv      brier
0.90920716 0.79600631 0.08740360 0.02143638

$train.matrix
       predicted
true        0 1 -SUM-
  0     13763 1     1
  1       315 0   315
  -SUM-   315 1   316

$test.matrix
       predicted
true       0   1 -SUM-
  0     4232 355   355
  1       71  34    71
  -SUM-   71 355   426
```


False positive rate
False negative rate
F1 measure
performance
1.00
0.75
0.50
0.25
0.00
0.10
0.05
0.00
0.00 0.25 0.50 0.75 1.00
threshold


True positive rate
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
False positive rate

- Glmnet (with 10 folds of cross validation and grid tuning the alpha parameter)

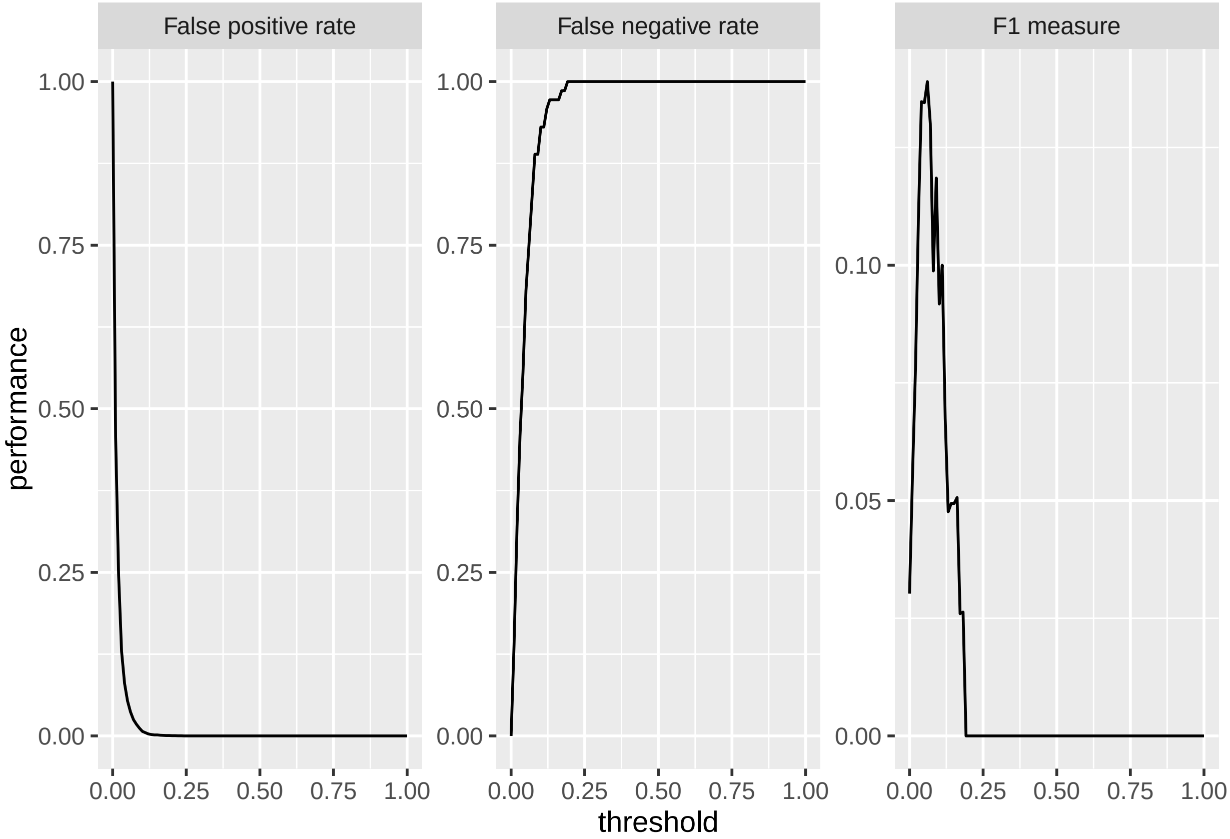
Readmission Type 1 results:

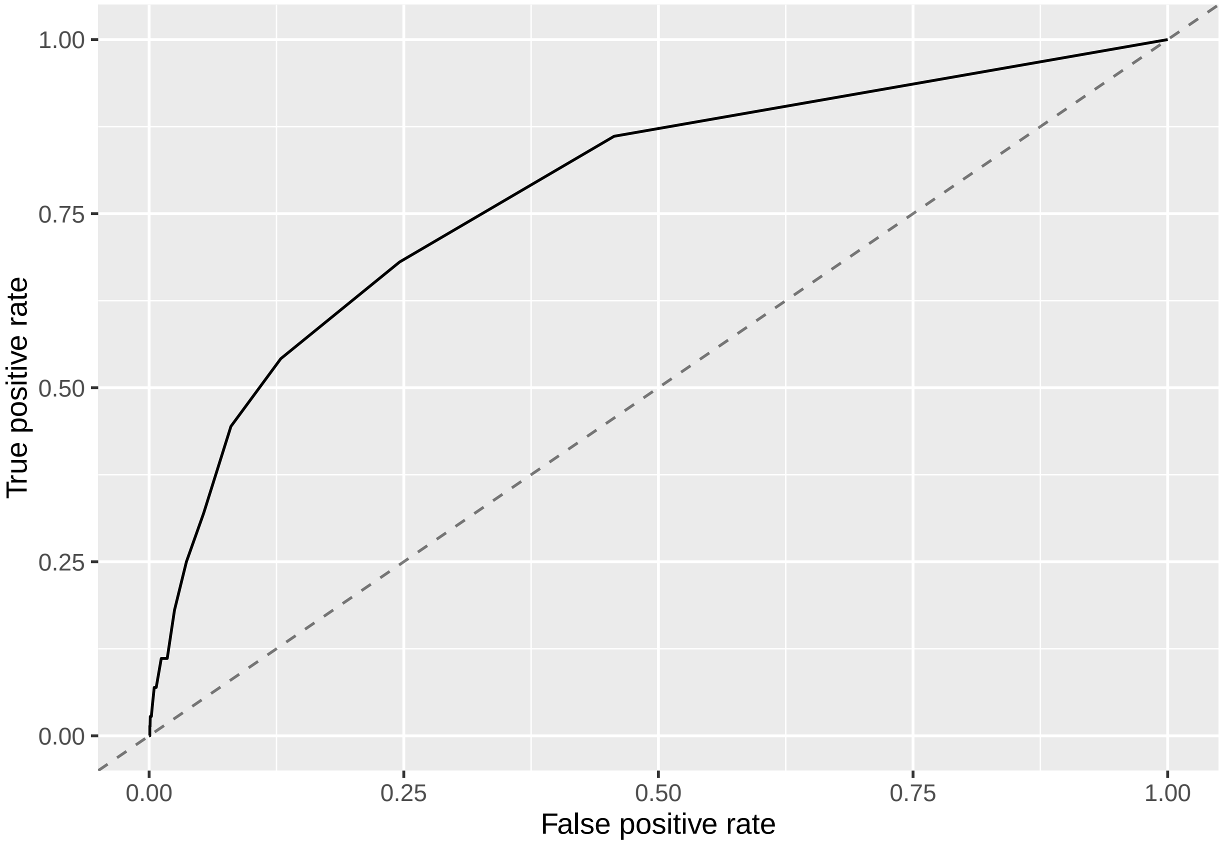
```
$train.metrics
       acc        auc        ppv      brier
0.98444492 0.79560823        NaN 0.01489673

$test.metrics
       acc        auc        ppv      brier
0.96291560 0.80520984 0.10156250 0.01466212

$train.matrix
       predicted
true        0 1 -SUM-
  0     13860 0     0
  1       219 0   219
  -SUM-   219 0   219

$test.matrix
       predicted
true       0   1 -SUM-
  0     4505 115   115
  1       59  13    59
  -SUM-   59 115   174
```





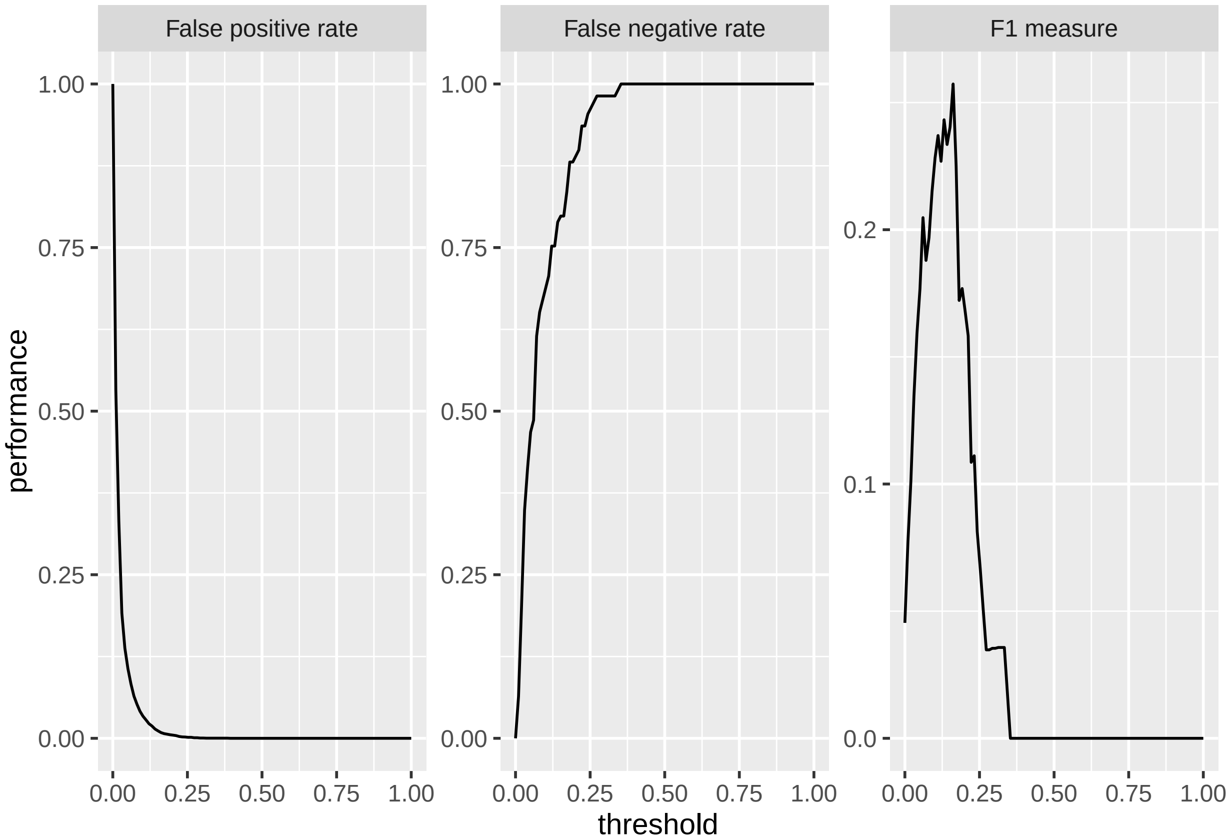
Readmission Type 2 results:

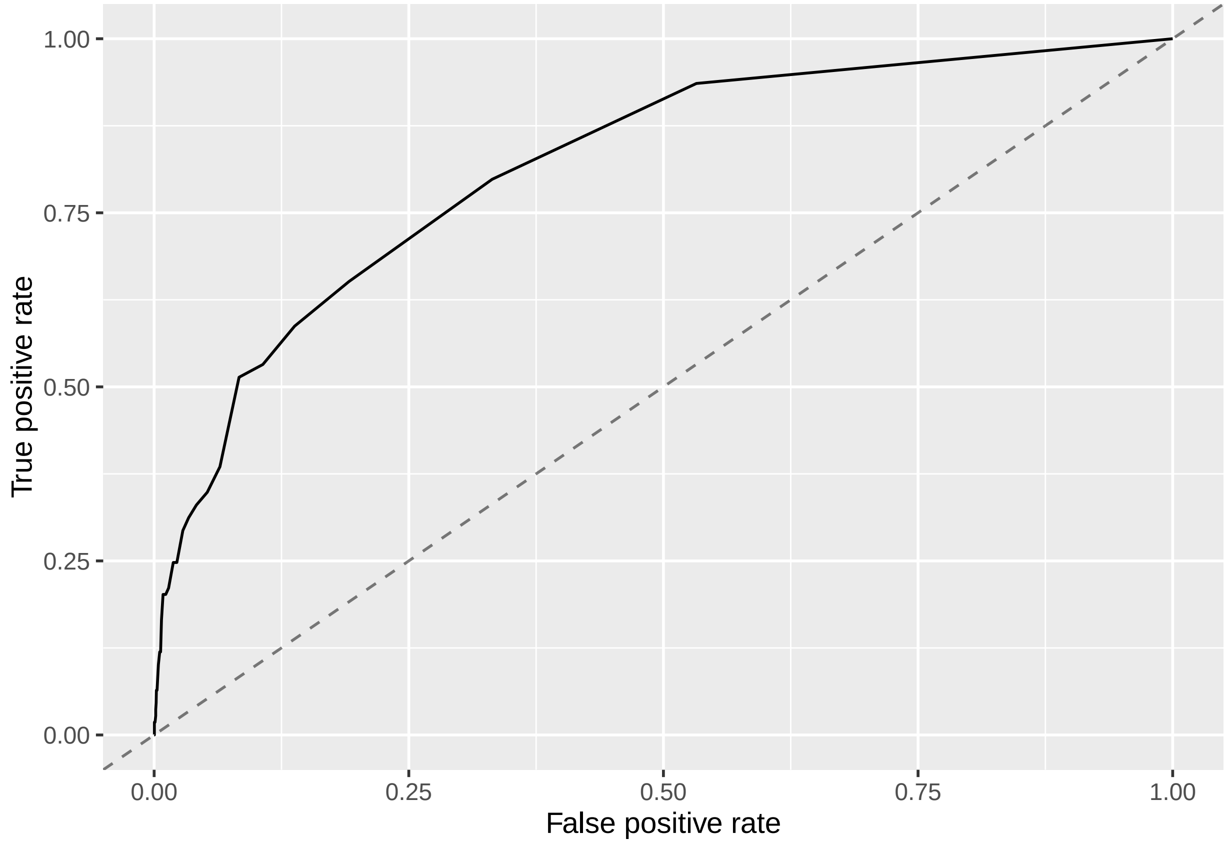
```
$train.metrics
       acc        auc        ppv      brier
0.97677392 0.82371616        NaN 0.02148113

$test.metrics
       acc        auc        ppv      brier
0.94394714 0.82219591 0.15929204 0.02109058

$train.matrix
       predicted
true        0 1 -SUM-
  0     13752 0     0
  1       327 0   327
  -SUM-   327 0   327

$test.matrix
       predicted
true       0   1 -SUM-
  0     4393 190   190
  1       73  36    73
  -SUM-   73 190   263
```


False positive rate
False negative rate
F1 measure
performance
1.00
0.75
0.50
0.25
0.00
0.2
0.1
0.0
0.00 0.25 0.50 0.75 1.00
threshold


True positive rate
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
False positive rate

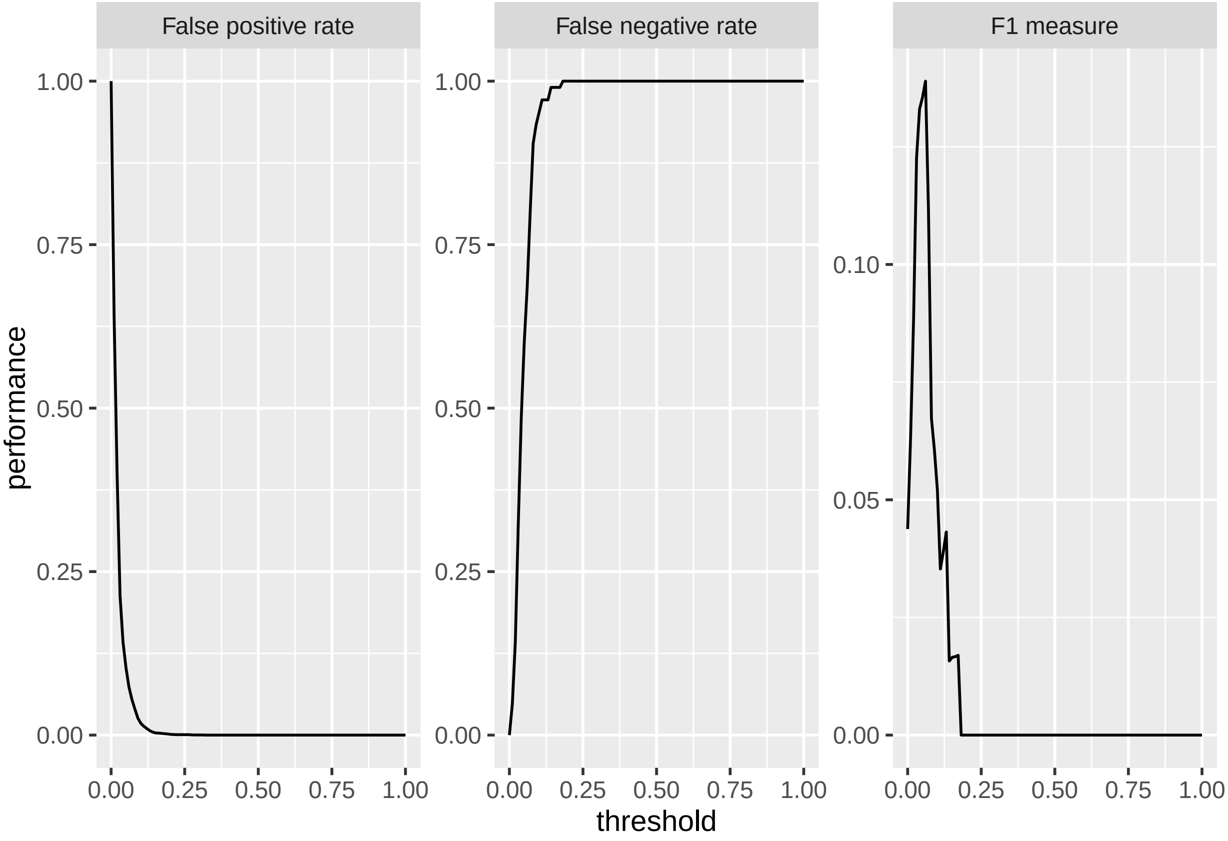
Readmission Type 3 results:

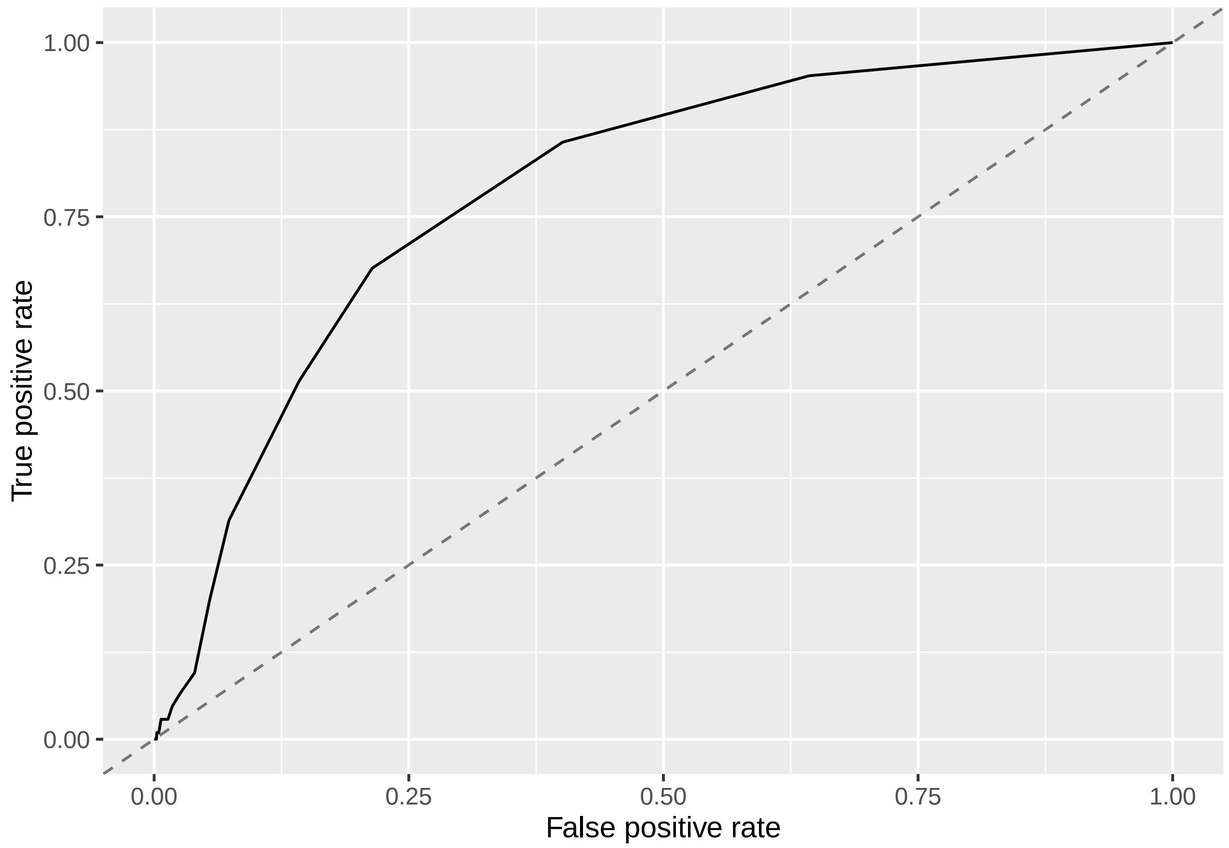
```
$train.metrics
       acc        auc        ppv      brier 
0.97755522 0.78381400 0.00000000 0.02131442 

$test.metrics
       acc        auc        ppv      brier 
0.91283035 0.79830889 0.08918919 0.02141667 

$train.matrix
       predicted
true        0 1 -SUM-
  0     13763 1     1
  1       315 0   315
  -SUM-   315 1   316

$test.matrix
       predicted
true       0   1 -SUM-
  0     4250 337   337
  1       72  33    72
  -SUM-   72 337   409
```





For the random forest we introduced different cost matrices to try to optimize our results. Of course in our case of study the false negatives (FN) are more important than the false positives (FP) so that is why the cost matrix penalizes the FN more than the FP. This also gave us a way to compare the results with the previous models.

- Weka results of random forest:

Readmission Type 1 results:

```
Readmissio N1, cost matrix 20-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4642               98.285  %
Incorrectly Classified Instances        81                1.715  %
Kappa statistic                          0.0445
Mean absolute error                      0.0421
Root mean squared error                  0.1307
Relative absolute error                135.4414 %
Root relative squared error            102.5708 %
Total Number of Instances             4723
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.974      0.984     0.999     0.991      0.76     N
                 0.026     0.001      0.286     0.026     0.047      0.76     Y
Weighted Avg.    0.983     0.958      0.972     0.983     0.976      0.76

=== Confusion Matrix ===

    a    b   <-- classified as
 4640    5 |    a = N
   76    2 |    b = Y


------------------------------------------------
Readmissio N1, cost matrix 100-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        3802               80.4997 %
Incorrectly Classified Instances       921               19.5003 %
Kappa statistic                          0.0622
Mean absolute error                      0.2459
Root mean squared error                  0.3576
Relative absolute error                790.2177 %
Root relative squared error            280.5712 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.809     0.41       0.992     0.809     0.891      0.807    N
                 0.59      0.191      0.049     0.59      0.091      0.807    Y
Weighted Avg.    0.805     0.407      0.976     0.805     0.878      0.807

=== Confusion Matrix ===


    a    b   <-- classified as
 3756  889 |    a = N
   32   46 |    b = Y


---------------------------------------------------------
Readmissio N1, cost matrix 60-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4328               91.6367 %
Incorrectly Classified Instances       395                8.3633 %
Kappa statistic                          0.1074
```

```
Mean absolute error                      0.1482
Root mean squared error                  0.2486
Relative absolute error                476.3702 %
Root relative squared error            195.0567 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.925     0.615      0.989     0.925     0.956      0.813    N
                 0.385     0.075      0.08      0.385     0.132      0.813    Y
Weighted Avg.    0.916     0.606      0.974     0.916     0.942      0.813

=== Confusion Matrix ===

    a    b   <-- classified as
 4298  347 |    a = N
   48   30 |    b = Y



---------------------------------------------------------------
Readmissio N1, cost matrix 50-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4435               93.9022 %
Incorrectly Classified Instances       288                6.0978 %
Kappa statistic                          0.1257
Mean absolute error                      0.1211
Root mean squared error                  0.2176
Relative absolute error                389.1505 %
Root relative squared error            170.7035 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.949     0.679      0.988     0.949     0.968      0.794    N
                 0.321     0.051      0.096     0.321     0.148      0.792    Y
Weighted Avg.    0.939     0.669      0.973     0.939     0.955      0.794

=== Confusion Matrix ===

    a    b   <-- classified as
 4410  235 |    a = N
   53   25 |    b = Y
```

Readmission Type 2 results:

```
-------------------------------------------------------------------
Readmissio N2, cost matrix 20-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4561               96.57   %
Incorrectly Classified Instances       162                3.43   %
Kappa statistic                          0.1727
Mean absolute error                      0.0777
Root mean squared error                  0.1749
Relative absolute error                171.4773 %
Root relative squared error            115.973  %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.985     0.827      0.98      0.985     0.982      0.792    N
                 0.173     0.015      0.211     0.173     0.19       0.792    Y
Weighted Avg.    0.966     0.808      0.962     0.966     0.964      0.792

=== Confusion Matrix ===

    a    b   <-- classified as
 4542   71 |    a = N
   91   19 |    b = Y



------------------------------------------------------------------
Readmissio N2, cost matrix 100-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        3179               67.3089 %
Incorrectly Classified Instances      1544               32.6911 %
Kappa statistic                          0.0603
Mean absolute error                      0.3429
Root mean squared error                  0.4671
Relative absolute error                757.0491 %
Root relative squared error            309.6936 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.67      0.209      0.993     0.67      0.8        0.818    N
                 0.791     0.33       0.054     0.791     0.101      0.818    Y
Weighted Avg.    0.673     0.212      0.971     0.673     0.784      0.818
```

```
=== Confusion Matrix ===

    a    b   <-- classified as
 3092 1521 |    a = N
   23   87 |    b = Y


-------------------------------------------------------------------
Readmissio N2, cost matrix 60-1, random forest, split 75%


=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        3823               80.9443 %
Incorrectly Classified Instances       900               19.0557 %
Kappa statistic                          0.0865
Mean absolute error                      0.237 
Root mean squared error                  0.3588
Relative absolute error                523.2731 %
Root relative squared error            237.8628 %
Total Number of Instances             4723     

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.815     0.418      0.988     0.815     0.893      0.814    N
                 0.582     0.185      0.07      0.582     0.125      0.814    Y
Weighted Avg.    0.809     0.413      0.967     0.809     0.875      0.814

=== Confusion Matrix ===

    a    b   <-- classified as
 3759  854 |    a = N
   46   64 |    b = Y


-------------------------------------------------------------------
Readmissio N2, cost matrix 50-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4028               85.2848 %
Incorrectly Classified Instances       695               14.7152 %
Kappa statistic                          0.1049
Mean absolute error                      0.203 
Root mean squared error                  0.3208
Relative absolute error                448.1796 %
Root relative squared error            212.6706 %
Total Number of Instances             4723     
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.861     0.482      0.987     0.861     0.92       0.815    N
                 0.518     0.139      0.082     0.518     0.141      0.815    Y
Weighted Avg.    0.853     0.474      0.966     0.853     0.901      0.815

=== Confusion Matrix ===

    a    b   <-- classified as
 3971  642 |    a = N
   53   57 |    b = Y
```

Readmission Type 3 results:

```
-------------------------------------------------------------------
Readmissio N3, cost matrix 20-1, random forest, split 75%

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4606               97.5228 %
Incorrectly Classified Instances       117                2.4772 %
Kappa statistic                          0.0133
Mean absolute error                      0.0658
Root mean squared error                  0.1596
Relative absolute error                148.9269 %
Root relative squared error            105.811  %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.991      0.977     0.998     0.987      0.762    N
                 0.009     0.002      0.111     0.009     0.017      0.762    Y
Weighted Avg.    0.975     0.968      0.957     0.975     0.965      0.762

=== Confusion Matrix ===

    a    b   <-- classified as
 4605    8 |    a = N
  109    1 |    b = Y
```

```
-------------------------------------------------------------------
Readmissio N3, cost matrix 100-1, random forest, split 75%


=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        2990               63.3072 %
Incorrectly Classified Instances      1733               36.6928 %
```

```
Kappa statistic                          0.0495
Mean absolute error                      0.3669
Root mean squared error                  0.4878
Relative absolute error                829.8135 %
Root relative squared error            323.4365 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.629     0.209      0.992     0.629     0.77       0.782    N
                 0.791     0.371      0.048     0.791     0.091      0.782    Y
Weighted Avg.    0.633     0.213      0.97      0.633     0.754      0.782

=== Confusion Matrix ===

    a    b   <-- classified as
 2903 1710 |    a = N
   23   87 |    b = Y



-------------------------------------------------------------------
Readmissio N3, cost matrix 60-1, random forest, split 75%


=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        3744               79.2716 %
Incorrectly Classified Instances       979               20.7284 %
Kappa statistic                          0.0702
Mean absolute error                      0.2512
Root mean squared error                  0.3645
Relative absolute error                568.0387 %
Root relative squared error            241.6528 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.799     0.455      0.987     0.799     0.883      0.774    N
                 0.545     0.201      0.061     0.545     0.109      0.774    Y
Weighted Avg.    0.793     0.449      0.965     0.793     0.865      0.774

=== Confusion Matrix ===

    a    b   <-- classified as
 3684  929 |    a = N
   50   60 |    b = Y
```

```
-------------------------------------------------------------------
Readmissio N3, cost matrix 50-1, random forest, split 75%


=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        4013               84.9672 %
Incorrectly Classified Instances       710               15.0328 %
Kappa statistic                          0.0799
Mean absolute error                      0.2107
Root mean squared error                  0.3179
Relative absolute error                476.4429 %
Root relative squared error            210.7556 %
Total Number of Instances             4723

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.86      0.573      0.984     0.86      0.918      0.775    N
                 0.427     0.14       0.068     0.427     0.117      0.776    Y
Weighted Avg.    0.85      0.563      0.963     0.85      0.899      0.775

=== Confusion Matrix ===

    a    b   <-- classified as
 3966  647 |    a = N
   63   47 |    b = Y
```

After considering our analysis, we have established that in order to achieve the perfect balance between the FP and the FN and optimize the results for each model, careful consideration should be given to the real cost for the hospital to have a FN than a FP. In the case of the logistic regression and the glmnet we should therefore design a measure that could capture accurately the cost we are trying to minimize. In the case of the random forest with cost matrix we should also optimize the cost matrix subject to the real cost of allowing more FP to reduce the FN. That is an important, complicated and precise challenge that falls out of the scope of this thesis but that we will clearly encourage anyone working on the subject to spend time on it.

To increase the reliability of our analysis an outlier detection analysis was performed in order to remove the outliers from our data. This analysis did not lead to an increase in the quality of our results. This could be because those “outliers” are extreme values that appear precisely in those “readmited” observations and most likely are part of the reason those patients were readmited. Therefore all extreme values were included in our final analysis.

Conclusion

A readmission predictive model can be successfully implemented in a academic hospital using existing data.

However it is not something easy to predict, developers of those predictive models need to consider more than the statistics when developing models. An implementable model balances clinical priorities, statistical requirements, availability of data and technical requirements.

The following are the conclusions we can get from the models predicting each response variable:

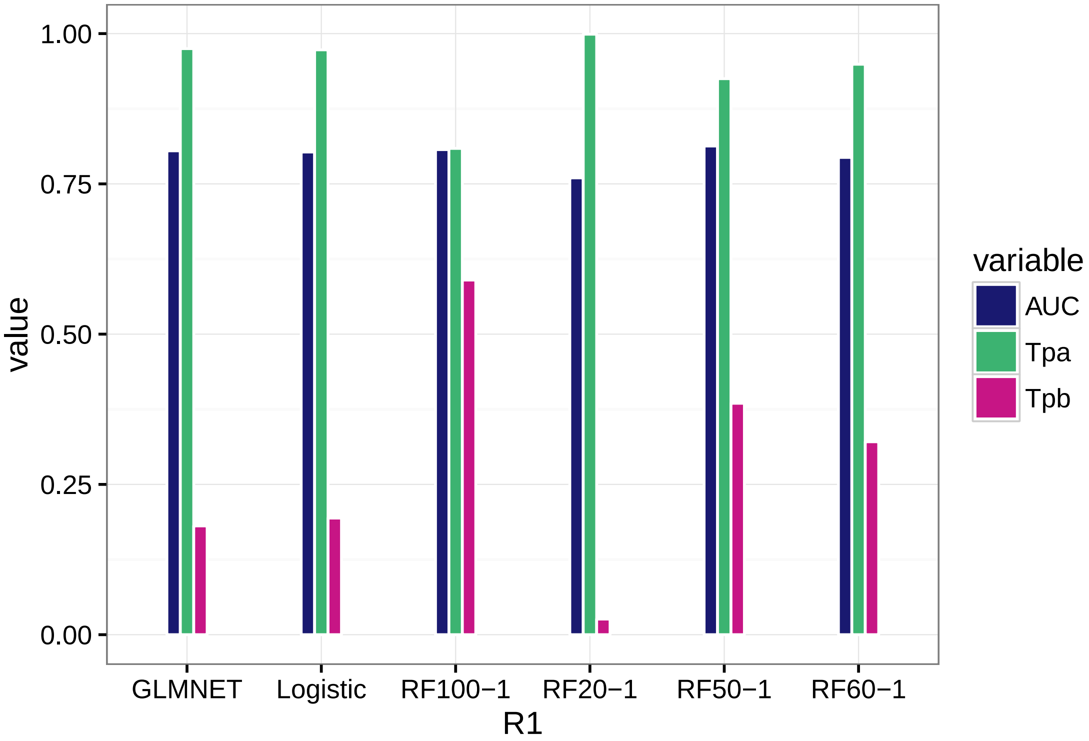
- Length of stay

The best model for predicting the *LoS* in our case study is the GBM for the *Less than 2 days* and *3 to 7 days* categories and the Random Forest for the *More than 7 days* category. If we had to pick just one model we would choose GBM because the difference between the GBM and the RF in the first two categories is significantly larger than the difference between the RF and the GBM in the third category.

- Readmissions

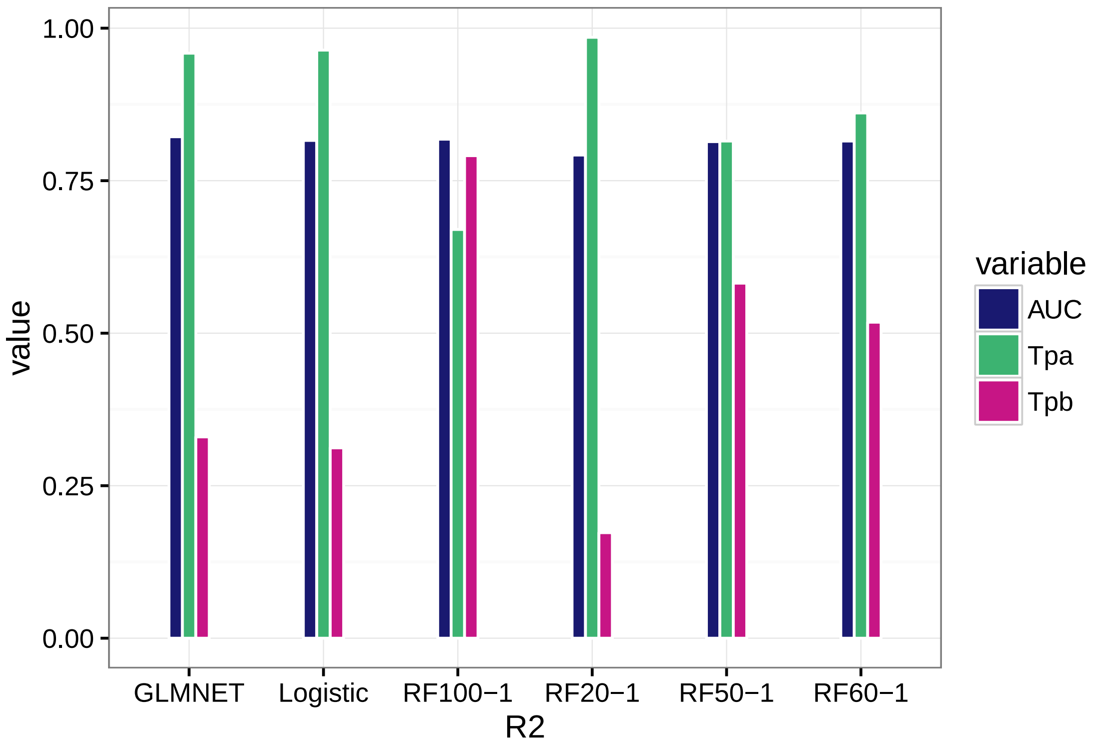
It is not that apparent through model comparison which one is the best model in the readmissions. We plotted all the models and 3 different measures to try to pick one. The three different measures are the AUC (Area under the curve), the Tpa (true positive values for class a, the not readmited patients) and the Tpb (true positive values for class b, the readmited patients). The three of this measures should be maximized but sometimes from one model to another there is a tradeoff where one measure increases and another decreases. In this case study we picked the higher AUC in order to decide on those cases.

- *Readmission type 1*



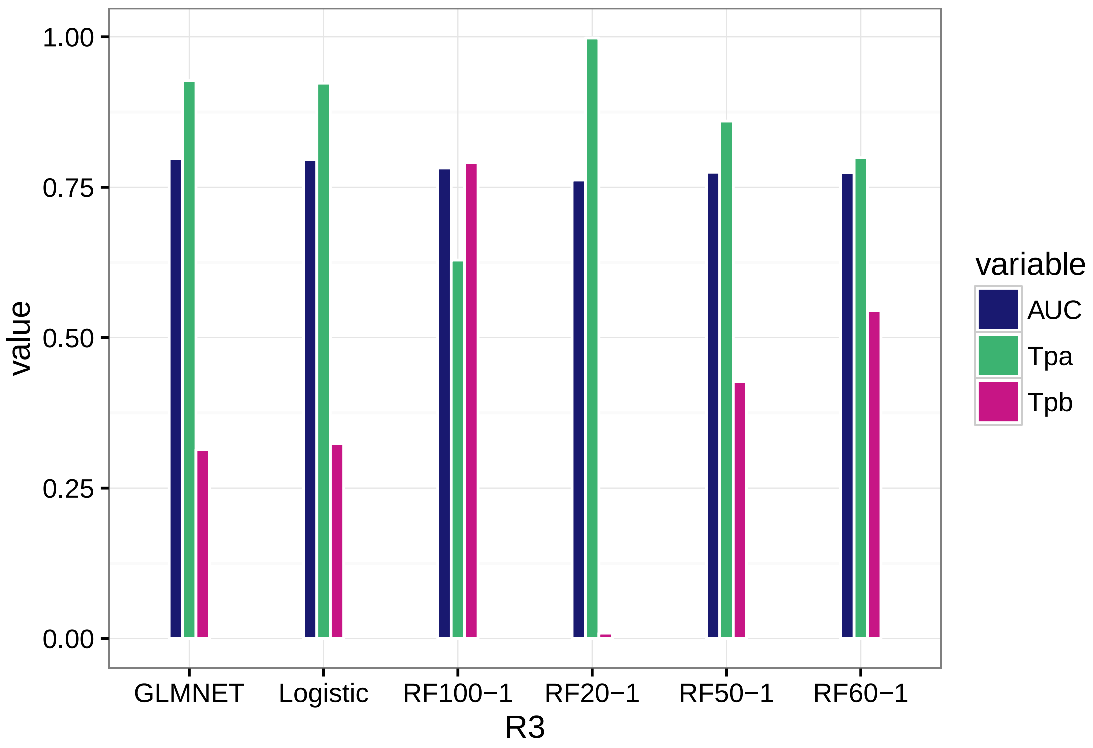
In this case the best model is the RF100-1.

- *Readmission type 2*



In this case the best model is the GLMNET.

- *Readmission type 3*



In this case the best model is the GLMNET.

At a first look the RF100-1 might appear to be the best model in predicting *R2* and *R3* because it predicts better the class b, but when checking the AUC we pick GLMNET. It is true that the AUC of the RF100-1 is close to GLMNET in both cases. It is also true that we said above in the report that we care more about predicting the class b better even at expenses of predicting the class a worse. But as we stated at the very end of the *Predictions* section, to correctly determine which is the exact tradeoff between the increase in predictions of class *b* and the decrease in predictions of class *a* that we can assume, a better cost measure should be designed. For that we would need more information on the economic value that the hospital gives to each wrongly predicted observation.

Bibliography

[1] Patrick R. Cronin, Jeffrey L. Greenwald, Gwen C. Crevensten, Henry C. Chueh, and Adrian H. Zai, 2014. "Development and Implementation of a Real-Time 30-Day Readmission Predictive Model". National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419988/

[2] Joseph Futomaa, Jonathan Morrisb, Joseph Lucas, 2015. "A comparison of models for predicting early hospital readmissions". Journal of Biomedical Informatics. http://www.sciencedirect.com/science/article/pii/S1532046415000969

[3] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, Brian Muckian, 2013. "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients". IEEE International Conference on Big Data. https://cwds.uw.edu/sites/default/files/publications/Big%20Data%20Solutions%20for%20Predicting%20Risk-of-Readmission%20for%20Congestive%20Heart%20Failure.pdf

[4] Phillips Healthcare Transformation Services. "Reducing avoidable readmissions using predictive analytics" URL: http://www.philips.fi/b-dam/b2bhc/us/hts/population-health/Reducing_avoidable_readmissions_using_predictive_analytics.pdf (visited in 15/04/2016)

[5] Issac Shams, Saeede Ajorlou, Kai Yang, 2014. "A predictive analytics approach to reducing avoidable hospital readmission".
https://www.researchgate.net/publication/260366976_A_predictive_analytics_approach_to_reducing_avoidable_hospital_readmission

[6] Christopher A Bain, Peter G Taylor, Geoff McDonnell and Andrew Georgiou, 2010. "Myths of ideal hospital occupancy"- https://www.mja.com.au/system/files/issues/192_01_040110/bai10628_fm.pdf

[7] British Medical Journal (BMJ), 2011. "Hospital safety and complexity". BMJ. http://www.bmj.com/rapid-response/2011/11/03/hospital-bed-occupancy