Exploratory Data Analysis Report

JIE HAN

December 19, 2018

Introduction

1.1 Information of Dataset

The dataset that generated the EDA Report is an 'data.frame' object, which contains 101,766 observations and 50 variables representing patient information.

1.2 Information of Variables

The dataset contains:

patient demographics(race, gender, age and weight),

admission ,discharge and payer details,

3 diagnoses,

laboratory and medicine data,

readmission and time staying in hospital.

We investigate which features relate to time in hospital and readmission of patients in less than 30 days.

Univariate Analysis

2.1 Cleaning data

The original dataset was not suitable for direct analysis.

We need to do some preparations to clean data:

- *Remove*
  - ~ remove the repeated observations according to “patient_nbr”,
  - ~ remove “weight" because of too many unreliable values,
  - ~ remove meaningless variables (“encounter_id","patient_nbr"),
  - ~ remove useless or repetitive meaning variables("max_glu_serum", 25:41 43:47 variables are medicines)
- *Ordinal variables*
  - ~ Change the levels of “age”, “A1Cresult”, “Insulin”, “readmitted” variables to ordered levels
- *Wrong types of variables*
  - ~ Change “admission_type_id", “discharge_disposition_id", “admission_source_id” to factor variables.
- *Collapse levels*
  - ~ “gender”,"admission_type_id","discharge_disposition_id","admission_source_id","medical_specialty","diag_1","diag_2","diag_3" , collapse their levels to 3~5 levels.

2.2 Descriptive Statistics

diab_test : 23 Variables 71518 Observations

---

race

n	missing	distinct
69570	1948	5

Value	AfricanAmerican	Asian	Caucasian	Hispanic	Other
Frequency	12887	497	53491	1517	1178
Proportion	0.185	0.007	0.769	0.022	0.017

---

gender

n	missing	distinct
71515	3	2

Value	Female	Male
Frequency	38025	33490
Proportion	0.532	0.468

---

time_in_hospital

n	missing	distinct	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95
71518	0	14	4.289	8.698	2.949	1	1	2	3	6	9	11

---

num_lab_procedures

n	missing	distinct	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95
71518	0	116	22.55	398.096	19.952	4	13	31	44	57	68	74

---

num_procedures

n	missing	distinct	Mean	Var	Sd
71518	0	7	1.431	3.097	1.760

---

num_medications

n	missing	distinct	Info	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95
71518	0	75	0.998	15.71	69.075	8.311	5	7	10	14	20	26	31

---

number_outpatient

n	missing	distinct	Info	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95
71518	0	33	0.341	0.2801	1.142	1.068	0	0	0	0	0	1	2

---

number_emergency

n	missing	distinct	Info	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95
71518	0	18	0.203	0.1035	0.259	0.509	0	0	0	0	0	0	1

---

number_inpatient

n	missing	distinct	Info	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95
71518	0	13	0.313	0.1778	0.365	0.604	0	0	0	0	0	1	1

---

number_diagnoses

n	missing	distinct	Info	Mean	Var	Sd	.05	.10	.25	.50	.75	.90	.95

71518	0	16	0.907	7.246	3.979	1.995	4	4	6	8	9	9	9

A1Cresult

n	missing.	distinct
71518	0	4

Value	None	Norm	>7	>8
Frequency	58532	3791	2891	6304
Proportion	0.818	0.053	0.040	0.088

---

insulin

n	missing	distinct
71518	0	4

Value	Down	No	Steady	Up
Frequency	7505	34921	22129	6963
Proportion	0.105	0.488	0.309	0.097

---

change

n	missing	distinct
71518	0	2

Value	Ch	No
Frequency	32024	39494
Proportion	0.448	0.552

---

diabetesMed

n	missing	distinct
71518	0	2

Value	No	Yes
Frequency	17199	54319
Proportion	0.24	0.76

---

readmitted

n	missing	distinct
71518	0	3

Value	NO	<30	>30

Frequency	42985	6293	22240
Proportion	0.601	0.088	0.311

---

new_age

n	missing	distinct
71518	0	3

Value	[0-30)	[30~70)	[70~100)
Frequency	1816	38003	31699
Proportion	0.025	0.531	0.443

---

new_adtype_id

n	missing.	distinct
71518	0	6

Value	Emergency	Urgent	Elective	Newborn.	Trauma Center	Not Mapped
Frequency	36490	13028	13917	9	21	8053
Proportion	0.510	0.182	0.195	0.000	0.000	0.113

---

new_discharge_id

n	missing	distinct
71518	0	3

Value	hospice	expired	others
Frequency	461	1084	69973
Proportion	0.006	0.015	0.978

---

new_adsource_id

n	missing	distinct
71518	0	4

Value	referral	emergency	transfer	others
Frequency	23071	38290	4942	5215
Proportion	0.323	0.535	0.069	0.073

---

new_med_spec

n	missing	distinct
37041	34477	4

Value	Cardiology	Emergency/Trauma	InternalMedicine	others
Frequency	4266	4465	10919	17391
Proportion	0.115	0.121	0.295	0.470

---

new_diag1

n	missing	distinct
70579	939	8

Value	Cir	Dia	Res	Dig	Inj	Gen	Can	Others
Frequency	21894	5805	9776	6570	4779	3514	6402	11839
Proportion	0.310	0.082	0.139	0.093	0.068	0.050	0.091	0.168

---

new_diag2

n	missing	distinct
69424	2094	8

Value	Cir	Dia	Res	Dig	Inj	Gen	Can	Others
Frequency	22534	9759	7242	2907	1858	5468	9185	10471
Proportion	0.325	0.141	0.104	0.042	0.027	0.079	0.132	0.151

---

new_diag3

n	missing	distinct
66752	4766	8

Value	Cir	Dia	Res	Dig	Inj	Gen	Can	Others
Frequency	21313	12660	4873	2746	1443	4199	9489	10029
Proportion	0.319	0.190	0.073	0.041	0.022	0.063	0.142	0.150

2.3 Visualization of important variables

The types of variables:

Continous: time_in_hospital, num_lab_procedures, num_procedures, num_medications, numeber_outpatient, number_emergency, number_inpatient, number_diagnoses

Binary: gender, change, diabetesMed

Nominal: race, new_diag1, new_diag2, new_diag3, new_adtype_id, new_discharge_id, new_adsource_id, new_med_spec

Ordinal: new_age, A1Cresult, insulin, readmitted

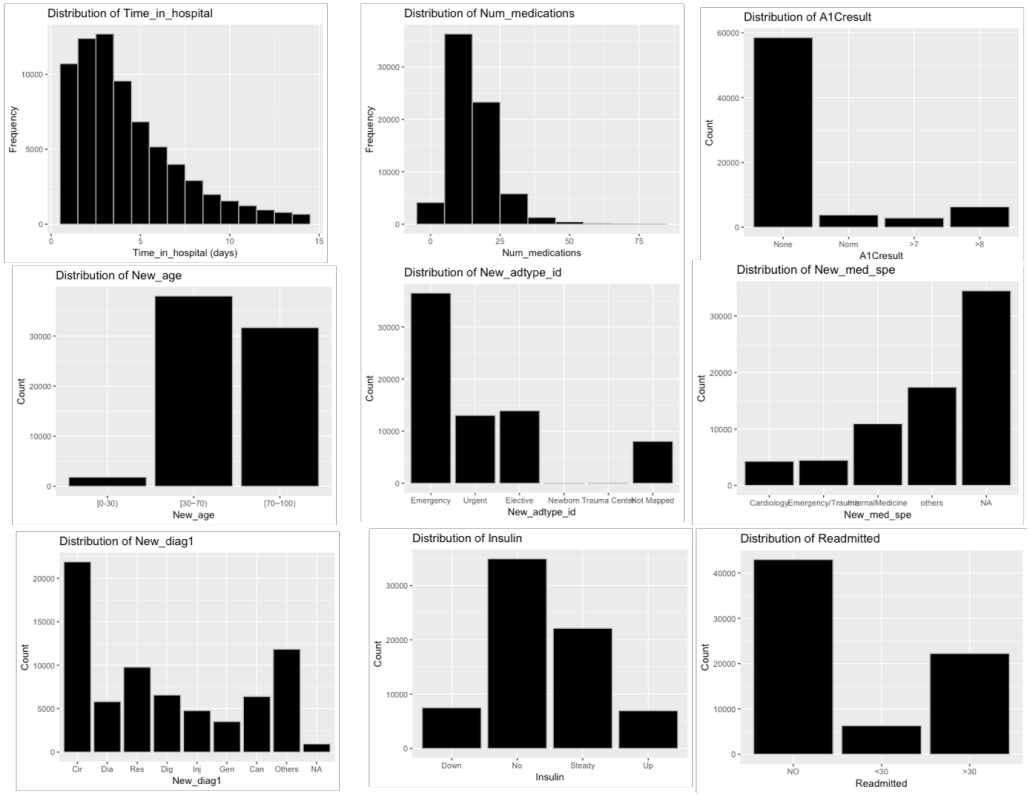


Figure 1. Distribution of variables :time_in_hospital, num_medications, A1Cresult, insulin, diabetesMed, readmitted, new_age, new_adtype_id, new_med_spec, new_diag1

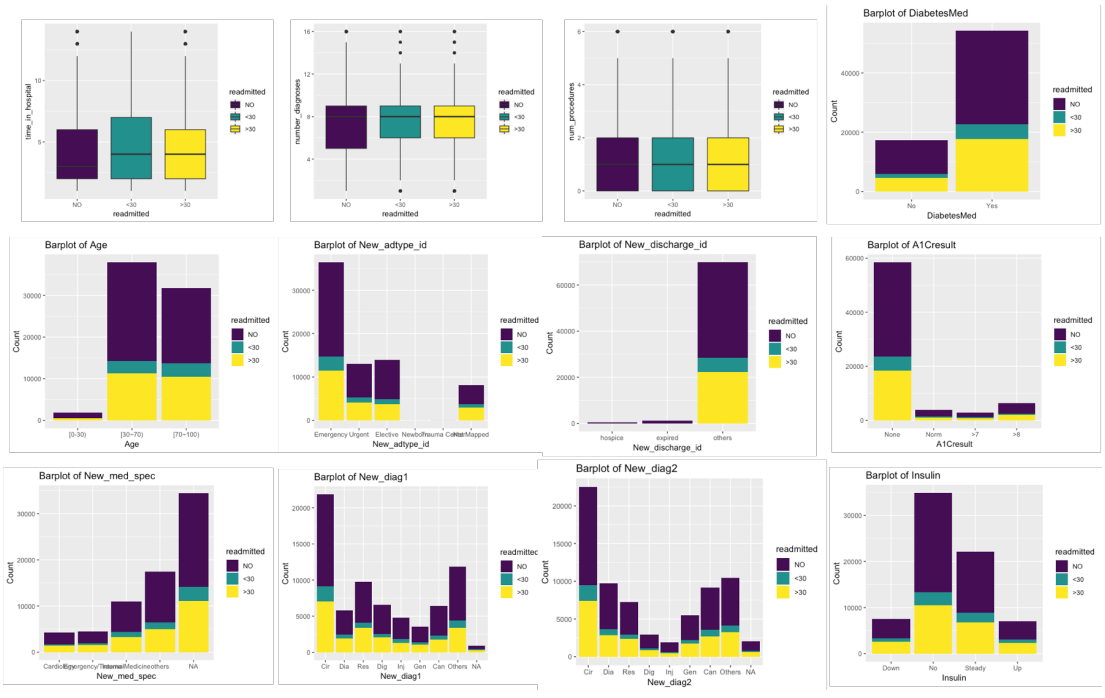


Figure 2. Distribution of variables grouped by readmitted: new_discharge_id, number_diagnoses, new_age , time_in_hospital, diabetesMed, new_adtype_id, new_diag1, num_procedures, A1Cresult, insulin, new_diag2, new_adsource_id, new_med_spec

Bivariate data EDA

3.1 correlation Coefficient

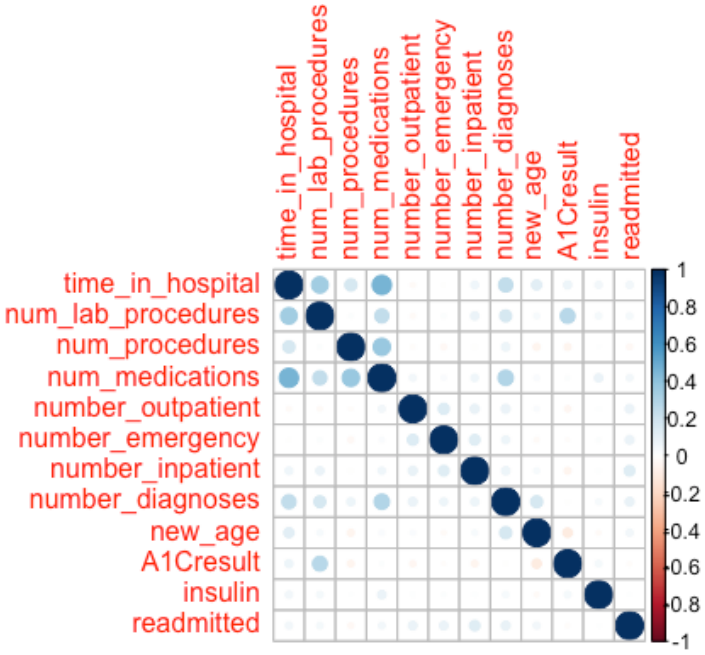


Figure 3. Correlation Plot of Numerical and Ordinal Variables

"Time_in_hospital" has positive relationship with num_lab_procedures, num_procedures, num_medications, num_diagnoses, num_procedures, new_age.

"Num_meditions" has positive relationship with num_lab_procedures, num_diagnoses, num_procedures.

There is no correlation coefficient > 0.5, indicates that variables have no strong relationship between them.

3.2 Statistics analysis

3.2.1 "Time_in_hospital" as dependent variable

The association was analyzed using Welch Two Sample t-test for binary variables, one-way ANOVA for norminal variables.

Significant level = 0.05

- Two Sample t-test results: all p-value of tests are less than 0.05, we have sufficient evidence to reject HO that there is no difference between the true means of two variables.

  Example:

  ```
  Welch Two Sample t-test

  data:  time_in_hospital by gender
  t = 6.5421, df = 70215, p-value = 6.109e-11
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   0.1012951 0.1879538
  sample estimates:
  mean in group Female   mean in group Male
          4.356897            4.212272
  ```

  For time_in_hospital and diabetesMed: although p-value is statistical significant, the 95% CI interval is very tiny, may not have actual effect in reality.

- One-way ANOVA results: all p-value of tests are less than 0.05, we have sufficient evidence to suggest that the multi means grouped by categorical variables are not equivalent.

  Example:

  ```
  One-way analysis of means (not assuming equal variances)

  data:  time_in_hospital and new_adtype_id
  F = 21.427, num df = 5.000, denom df = 66.646, p-value = 1.026e-12
  ```

  For time_in_hospital and new_adtype_id: we have enough evidence to suggest that the true means of time in hospital grouped by types of admission id are not equal.

3.2.2 “Readmitted” as dependent variable

We use chi-square tests for categorical variables.

- Chi-square results: all p-value of tests are less than 0.05. There is evidence to suggest that some kind of dependency exists between those two categorical variables.

4 Regression Analysis

4.1 Linear regression analysis for “time_in_hospital"

• Final interpretable model:

time_in_hospital = 2.68 + 0.17 * num_medications -
0.37 * raceAsian -0.17 * raceCaucasian +
0.33 * A1Cresult.L - 0.23*A1Cresult.Q + 0.13 * A1Cresult.C +
0.26 * insulin.L + 0.32 * insulin.Q - 0.10 * insulin.C -
0.30 * diabetesMedYes +
0.16 * readmitted.L - 0.28 * readmitted.Q +
0.66 * new_age.L + 0.20 * new_age.Q +
0.28 * new_adtype_idUrgent - 0.67 * new_adtype_idElective + 0.18 * new_adtype_idNot Mapped -
1.82 * new_discharge_idexpired - 1.28 * new_discharge_idothers +
0.28 * new_adsource_idemergency + 0.62 * new_adsource_idtransfer - 0.46 * new_adsource_idothers +
0.38 * new_med_specEmergency/Trauma + 0.68 * new_med_specInternalMedicine + 0.52 * new_med_specothers +
0.21 * new_diag1Dig + 0.14 * new_diag1Inj + 0.44 * new_diag1Can+0.44 * new_diag1Others -
0.21 * new_diag2Dia + 0.46 * new_diag2Res + 0.54 * new_diag2Dig + 0.65 * new_diag2Inj + 0.37 * new_diag2Gen + 0.27 * new_diag2Can + 0.59 * new_diag2Others -
0.24 * new_diag3Dia + 0.43 * new_diag3Res + 0.20 * new_diag3Dig + 0.60 * new_diag3Inj + 0.46 * new_diag3Gen + 0.26 * new_diag3Can + 0.43 * new_diag3Others

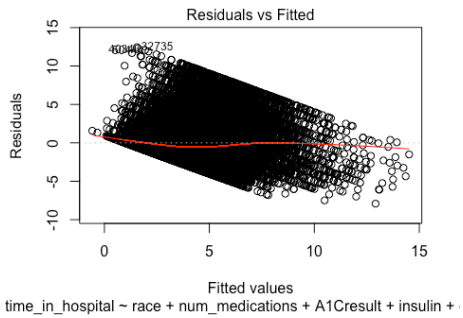
• Evaluation

~ Adjusted R-squared

0.2962, this model explains 29.6% variability of the time_in_hospital data around its mean.

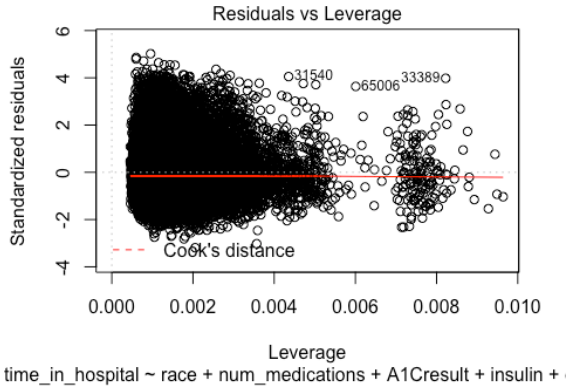
~ Residuals vs Fitted

In this model, residuals are not randomly distributed, there’s some pattern we should work on to find a better model.



~ Residuals vs Leverage

No point is out of cook's distance, which indicates that the multi predictors have linear relationship with response variable " time_in_hospital".



~ *Numeric variable fitted coeficients*

For every 1 medication increase, the time stay in hospital increases by 0.17 day on average, when all other variables in the model are held constant.

~ *Categorical variable*

If dianose2_level == "injure", then time stay in hospital increases by 0.65 day on average as compared with time_in_hospital of the reference group when all other variables in the model are held constant.

~ *Inclusion*

According to this model, there exists some useful predictors for length of stay in hospital, including race, age, A1Cresult, insulin, admission with emergency, and diagnoses. Diabetes Patients at risk for longer stay in hospital are likely to be Asian, elderly, admitted from emergency, not taking A1C test and using diabetes medicines.

4. 2 Logistic regression analysis for "readmitted<30"

Excluding "readmitted > 30"

```
fit_final<- glm(readmitted ~ number_diagnoses + new_age +
        time_in_hospital + diabetesMed + new_adtype_id + new_diag1 +
        num_procedures + A1Cresult + insulin+ new_diag2 +
        new_med_spec , family=binomial(), data = diab_glm2)
```

• *Evaluation*

~ Pseudo R^2 for logistic regression

Hosmer and Lemeshow R^2	0.03
Cox and Snell R^2	0.023
Nagelkerke R^2	0.042
McFadden	3.017670e-02
r2CU	4.249370e-02

The values of McFadden and r2CU are less than 1, which mean the regression model is valid although it is not well.

~ *Odds Ratios*

A. For predictor variable admission type id, when its level == "Urgent", average value of odds of "readmitted < 30" is multiplied by 1.06, holding all other predictors constant.
B. For number_diagnoses, one increase in gestation, average value of odds of "readmitted <30" is multiplied by 1.11, holing all other predictors constant.

~ *Inclusion*

Many predictor variables are statistically significant to the readmission less than 30 days- number_diagnoses, new_age, time_in_hospital, diabetesMed, new_adtype_id, new_diag1, num_procedures, A1Cresult, insulin, new_diag2, new_med_spec, the P-value < 0.05.

Including "readmitted > 30"

```
fit_final2<- glm(formula = readmitted ~ number_inpatient +
  time_in_hospital + new_age + diabetesMed + number_diagnoses +
  new_diag1 + new_med_spec + num_procedures + number_emergency +
  new_diag2 + A1Cresult, family = binomial(), data = diab_glm3)
```

• *Evaluation*

~ Pseudo R^2 for logistic regression

Hosmer and Lemeshow R^2	0.028
Cox and Snell R^2	0.016
Nagelkerke R^2	0.036

All R^2 values are less than that of above model, shows that this model is worse than the model excluding “readmitted > 30”.

- AIC

    fit_final1: 16509

    fit_final2: 18648

It indicates that the first model is better fitted than the second one.