Peng Shen

1.

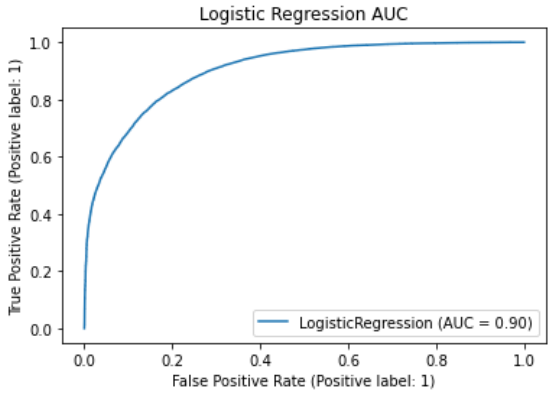
There are a total of 185 features selected. Some of them are 'dep_name', 'esi', 'age', 'gender', 'ethnicity', 'race', 'lang','religion', 'maritalstatus', 'employstatus', ... 'meds_vitamins', 'n_surgeries', 'cc_abdominalpain', 'cc_alcoholintoxication', 'cc_alteredmentalstatus', 'cc_chestpain','cc_neurologicproblem', 'cc_other', 'cc_psychiatricevaluation','cc_shortnessofbreath'.

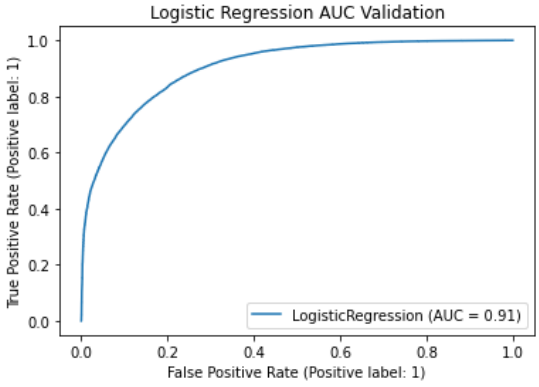
2.

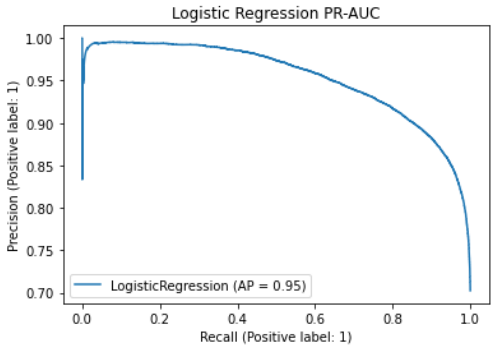
The AUC, PR-AUC for logistic regression in test data are 0.90 and 0.95. The AUC, PR-AUC for logistic regression in validation data are 0.91 and 0.95. The Calibration curve is shown below.

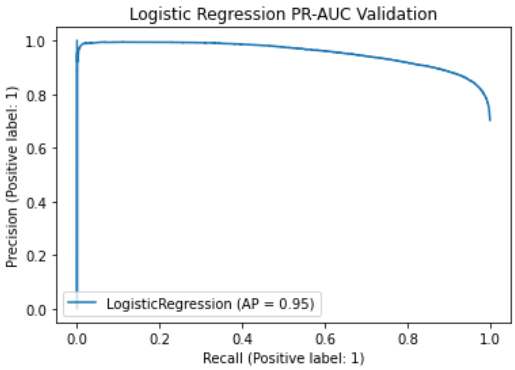
The AUC, PR-AUC for XGBoost in test data are 0.92 and 0.96. The AUC, PR-AUC for logistic regression in validation data are 0.93 and 0.97. The Calibration curve is shown below.

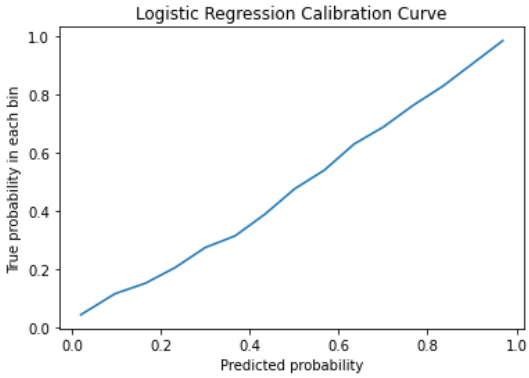
The XGBoost performance is better than logistic regression. The performance of xgboost model for validation dataset is better than in test set, which indicates over-fitting.

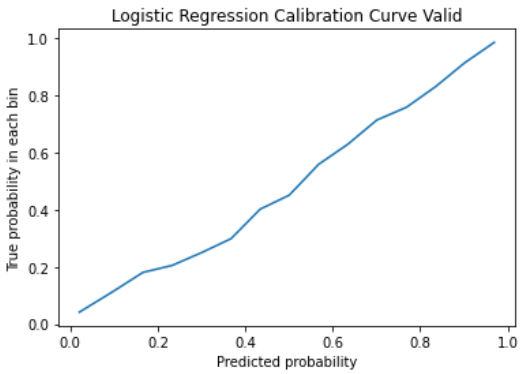


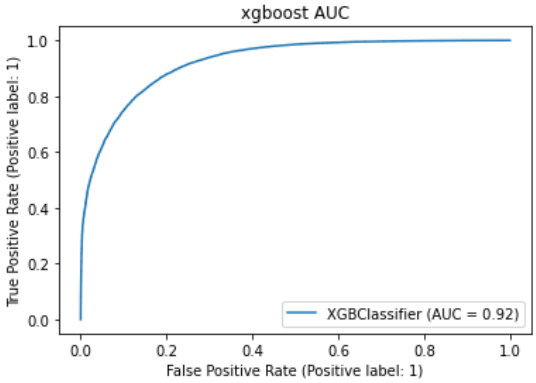


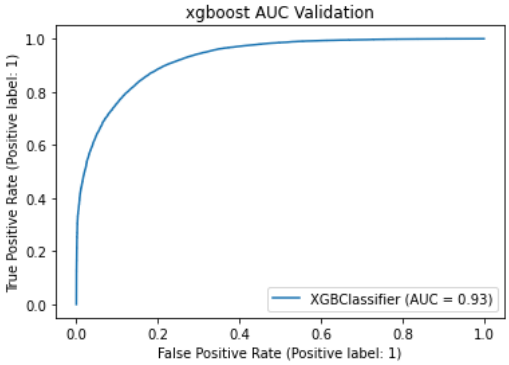


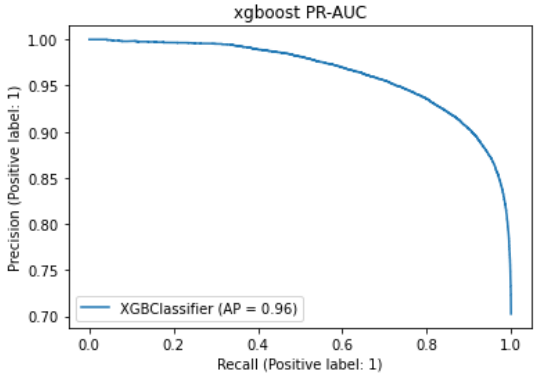


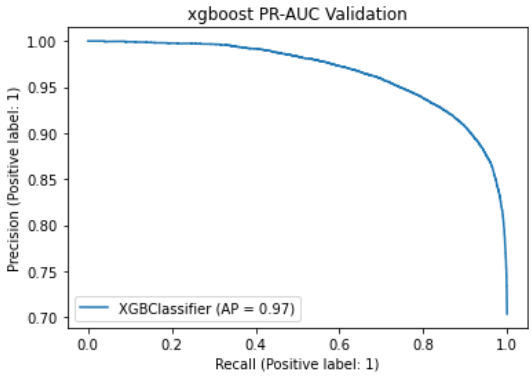

Logistic Regression Calibration Curve
True probability in each bin
Predicted probability

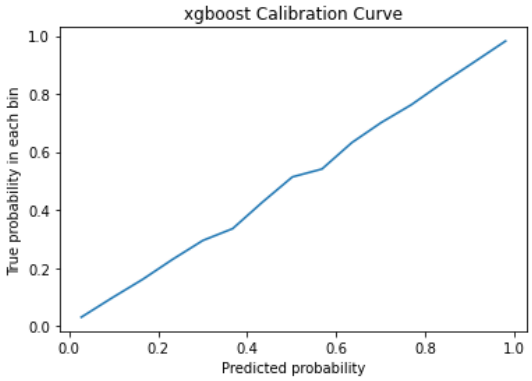

Logistic Regression Calibration Curve Valid
True probability in each bin
Predicted probability

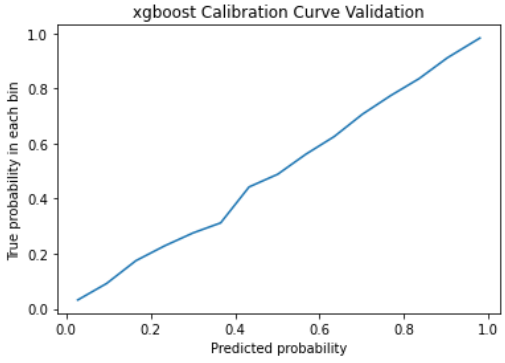

xgboost AUC
True Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.92)
False Positive Rate (Positive label: 1)


xgboost AUC Validation
True Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.93)
False Positive Rate (Positive label: 1)


xgboost PR-AUC
Precision (Positive label: 1)
XGBClassifier (AP = 0.96)
Recall (Positive label: 1)


xgboost PR-AUC Validation
Precision (Positive label: 1)
XGBClassifier (AP = 0.97)
Recall (Positive label: 1)


xgboost Calibration Curve
True probability in each bin
Predicted probability


xgboost Calibration Curve Validation
True probability in each bin
Predicted probability

3.

The odds ratio of top 10 variables of importance in logistic regression is shown below.

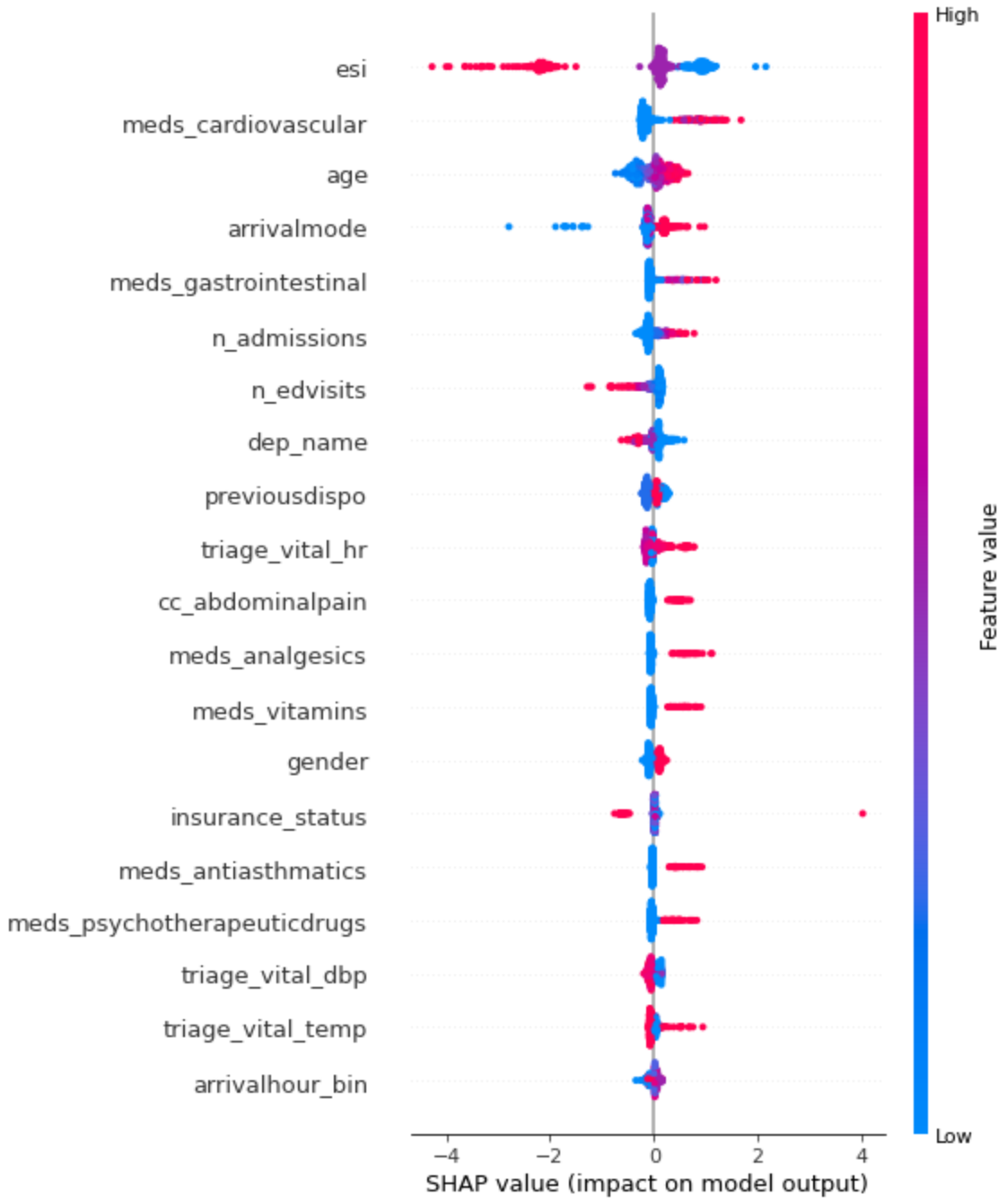
	feature	odds_ratio
0	esi	0.312485
1	meds_analgesics	1.862151
2	cc_abdominalpain	1.748657
3	meds_cardiovascular	1.523392
4	meds_antiasthmatics	1.474066
5	meds_psychotherapeuticdrugs	1.472075
6	meds_gastrointestinal	1.469729
7	meds_vitamins	1.448149
8	meds_antiarthritics	1.430254
9	cc_alcoholintoxication	0.706095

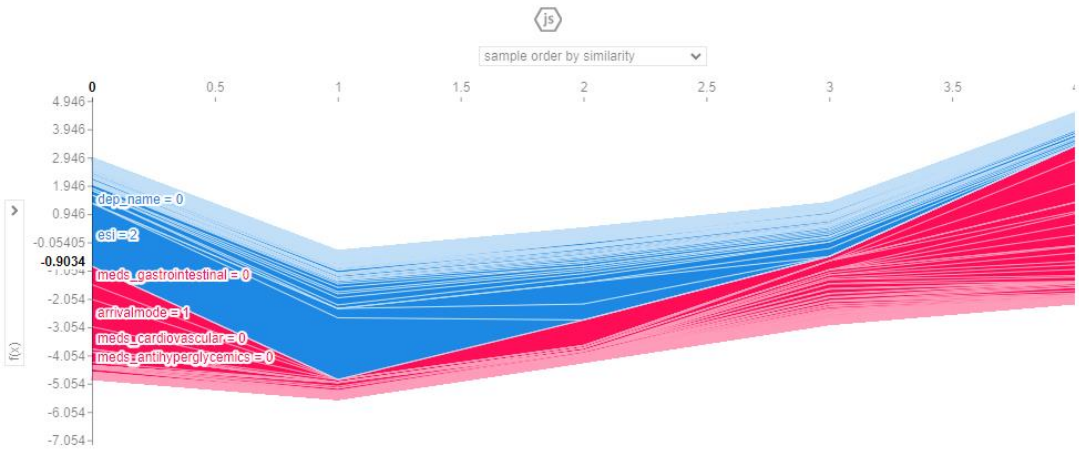
The SHAP summary plot of the top 20 variables in XGBoost is shown, along with the force plot and the dependence plot. Since the dataset is very large, only 400 randomly selected rows are included in the plots.

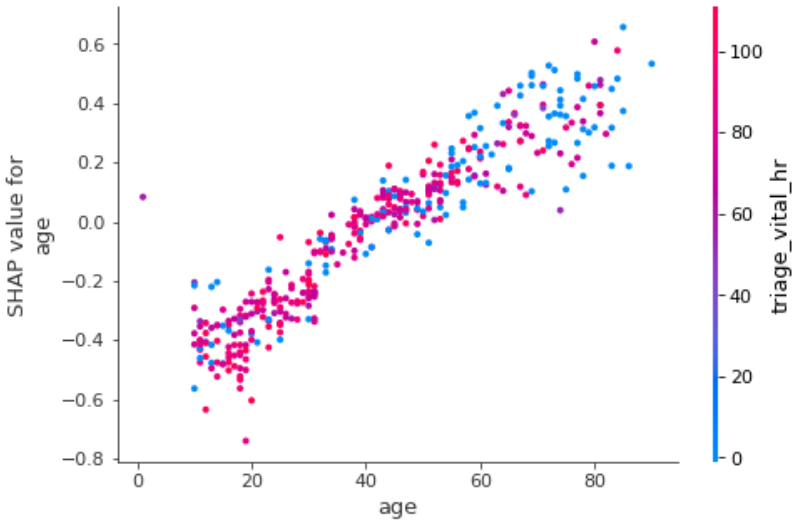
In summary plot, patients with higher esi, or lower meds_cardiovascular or lower age is having a higher probability of discharge. Vice versa, lower esi, or higher meds_cardiovascular or higher age corresponds to higher probability of admission.

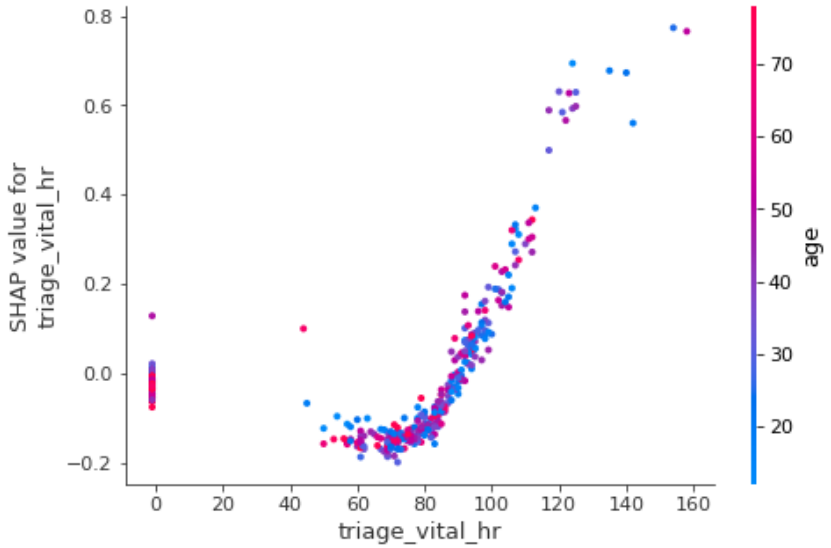
The force plot of the 5 visits is an interactive plot, thus it would be better to interpret using software.

Looking at the dependence plot of age and systolic blood pressure, besides the information already provided by the summary plot, age and systolic blood pressure are negatively correlated for the patients. Higher age often corresponds to a lower systolic blood pressure.


esi
meds_cardiovascular
age
arrivalmode
meds_gastrointestinal
n_admissions
n_edvisits
dep_name
previousdispo
triage_vital_hr
cc_abdominalpain
meds_analgesics
meds_vitamins
gender
insurance_status
meds_antiasthmatics
meds_psychotherapeuticdrugs
triage_vital_dbp
triage_vital_temp
arrivalhour_bin
−4
−2
0
2
4
SHAP value (impact on model output)
High
Low
Feature value


sample order by similarity
dep_name = 0
esi = 2
meds_gastrointestinal = 0
arrivalmode = 1
meds_cardiovascular = 0
meds_antihyperglycemics = 0
f(x)


SHAP value for age
age
triage_vital_hr


SHAP value for triage_vital_hr
triage_vital_hr
age