Lab Report

Ximu Wang

Dataset Pretreatment

Firstly, preview the dataset, we can see there are some observations have too many NA values, I think we should discard these observations because they have too less information and are worthless in prediction. So, I discard the observations whose proportion of NA are greater than 20%. Then I delete the rows which have NA value to get the clean dataset.

Feature Selection

Firstly, I apply all features to create a logistic regression model, and exclude the features that not associated with the target (P > 0.1, same as the lecture). Then I separate the features into 6 categories:

1. "EGFR_CLOSEST" and "FOLLOW_UP_EGFR_VALUE"
2. "AGE_ON_CONTACT_DATE" and "FEMALE", "RACE_F"
3. "BMI"
4. "ALT_CLOSEST_F", "AST_CLOSEST_F" and "CA_CLOSEST_F"
5. "OSTEO_HST_F", "PSORIATIC_ARTHRITIS_HST_F" and "OBS_SLEEPAPNEA_HST_F", "ANXIETY_HST_F"
6. "ARB" and "SGLT2_INHIBITOR"

Model Development

I apply these feature categories into 7 logistic regression models:

Model 1: ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F")

Model 2: ("EGFR_CLOSEST", "FOLLOW_UP_EGFR_VALUE"), ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F")

Model 3: ("EGFR_CLOSEST", "FOLLOW_UP_EGFR_VALUE"), ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F"), ("BMI")

Model 4: ("EGFR_CLOSEST", "FOLLOW_UP_EGFR_VALUE"), ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F"), ("BMI"), ("ALT_CLOSEST_F", "AST_CLOSEST_F", "CA_CLOSEST_F")

Model 5: ("EGFR_CLOSEST", "FOLLOW_UP_EGFR_VALUE"), ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F"), ("BMI") ("OSTEO_HST_F", "PSORIATIC_ARTHRITIS_HST_F", "OBS_SLEEPAPNEA_HST_F", "ANXIETY_HST_F")

Model 6: ("EGFR_CLOSEST", "FOLLOW_UP_EGFR_VALUE"), ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F"), ("BMI"), ("ARB", "SGLT2_INHIBITOR")

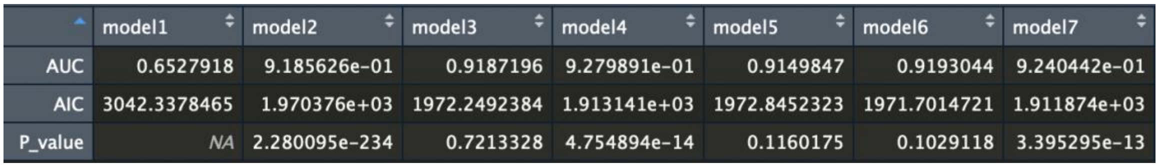
Model 7: ("EGFR_CLOSEST", "FOLLOW_UP_EGFR_VALUE"), ("AGE_ON_CONTACT_DATE", "FEMALE", "RACE_F"), ("BMI"), ("ALT_CLOSEST_F", "AST_CLOSEST_F", "CA_CLOSEST_F"), ("OSTEO_HST_F", "PSORIATIC_ARTHRITIS_HST_F", "OBS_SLEEPAPNEA_HST_F", "ANXIETY_HST_F"), ("ARB", "SGLT2_INHIBITOR")

- ## Model Validation

First validation:

I calculate AUC, AIC and P-value for each model. P-value is each model compared with prior model, except model 5, 6 and 7, which are compared with model 3.

Results:



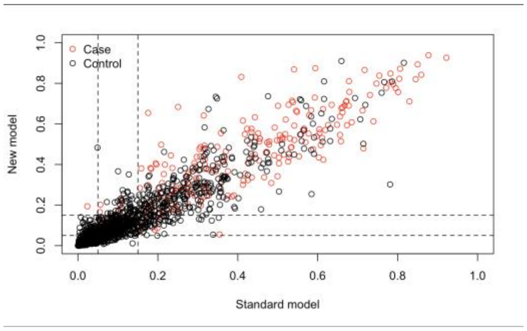
	model1	model2	model3	model4	model5	model6	model7
AUC	0.6527918	9.185626e-01	0.9187196	9.279891e-01	0.9149847	0.9193044	9.240442e-01
AIC	3042.3378465	1.970376e+03	1972.2492384	1.913141e+03	1972.8452323	1971.7014721	1.911874e+03
P_value	NA	2.280095e-234	0.7213328	4.754894e-14	0.1160175	0.1029118	3.395295e-13

We can see model 2, model 4 and model 7 have a good performance. Their AUC, AIC and P-value are good among all models.

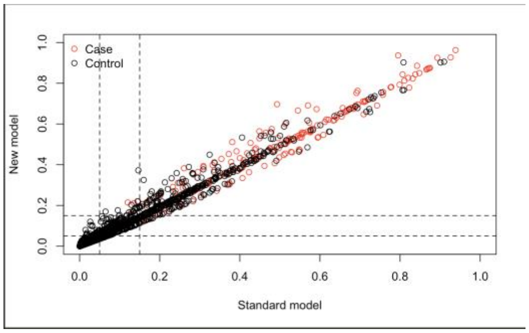
Second validation:

I use NIR to compare the performance improvement between model 2, model 4 and model 7. For CKD stage 3, model 4 outperformed model 2 and model 7 with an NRI of 0.574% and 0.945%. For CKD stage 4, model 4 outperformed model 2 and model 7 with an NRI of 1.089% and 0.122%.

Stage 3

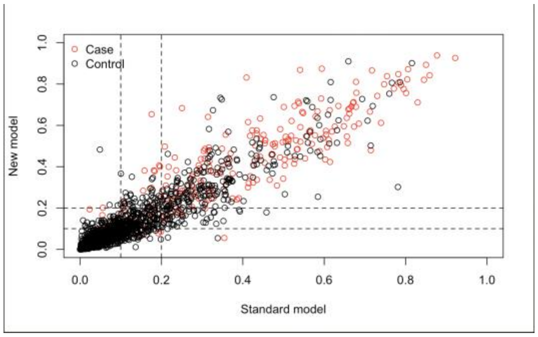


Model 4 vs Model 2

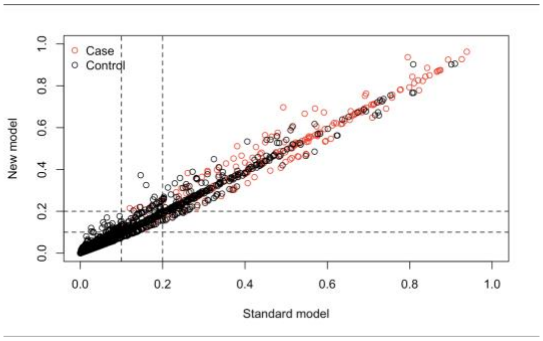


Model 7 vs Model 4

Stage 4



Model 4 vs Model 2



Model 7 vs Model 4

So, we select model 4 as the best model.

- ## Conclusion

I identify the risk level according to the reading lecture: the risk category for CDK stage 3 are 0% to 4.9%, 5% to 14.9% and 15% to more; the risk category for CDK stage 4 are 0% to 9.9%, 10% to 19.9% and 20% to more.

But I would not deploy this model. Because I think the AUC of this model is a little high, it may lead an overfitting in future prediction. I think this is because the raw data has many NA values and I ignore them, if I can have more information, I can get a more reasonable model.