week_8

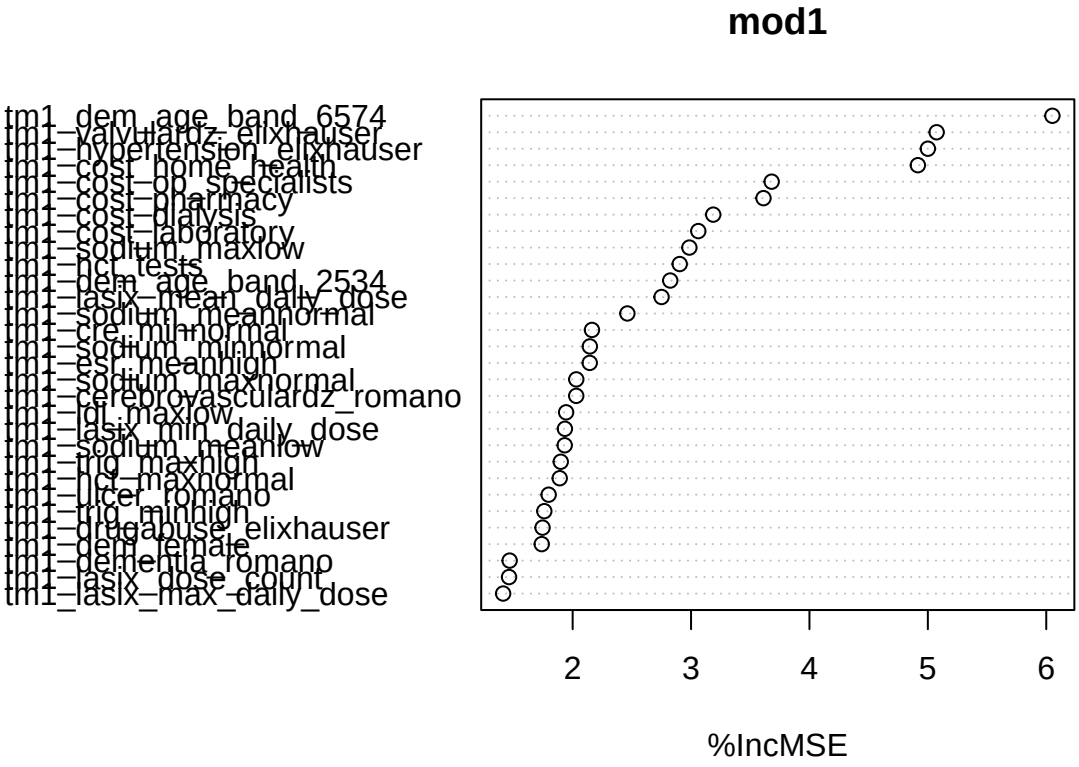
Lindsey Greenhill

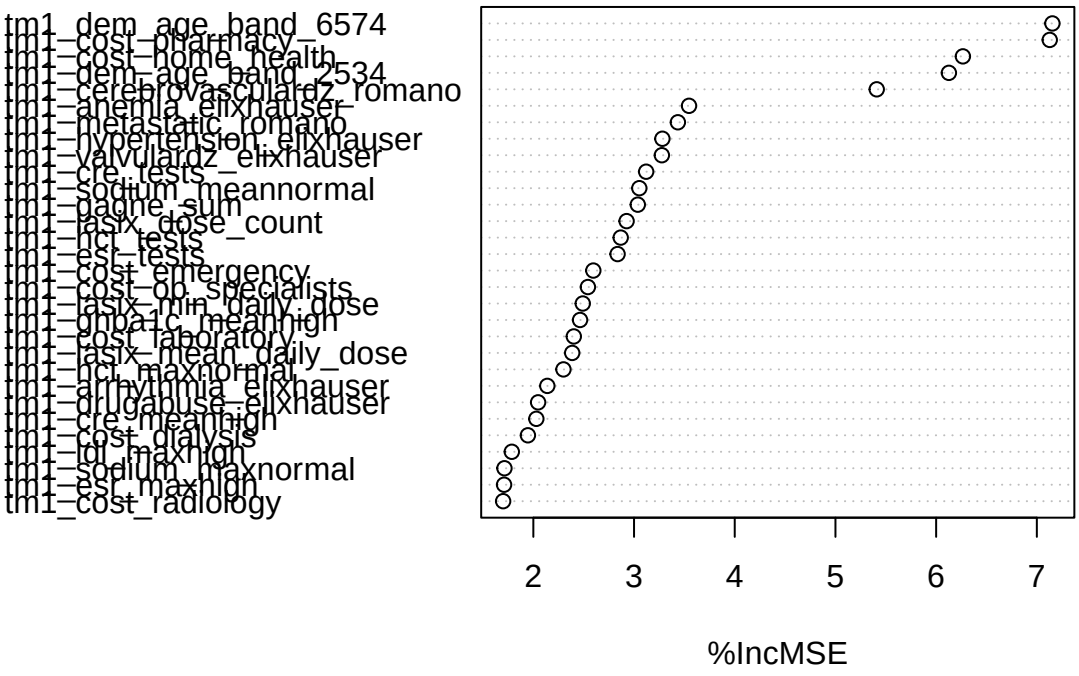
4/14/2022

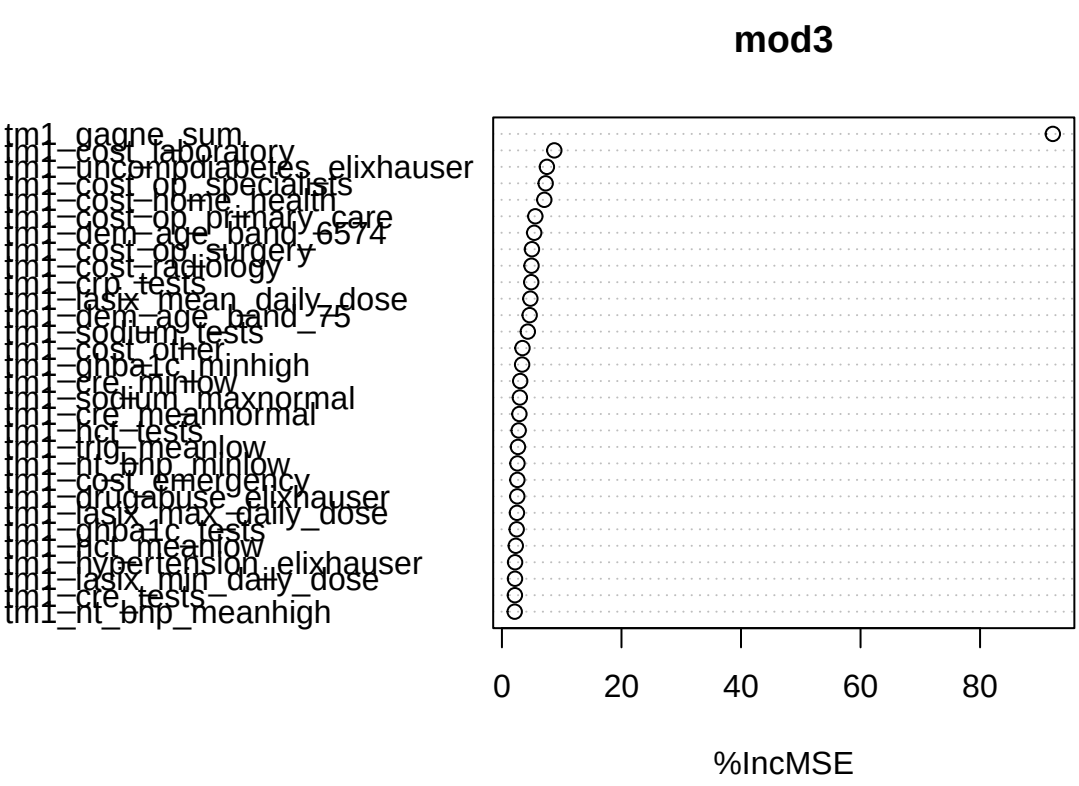
Question 1

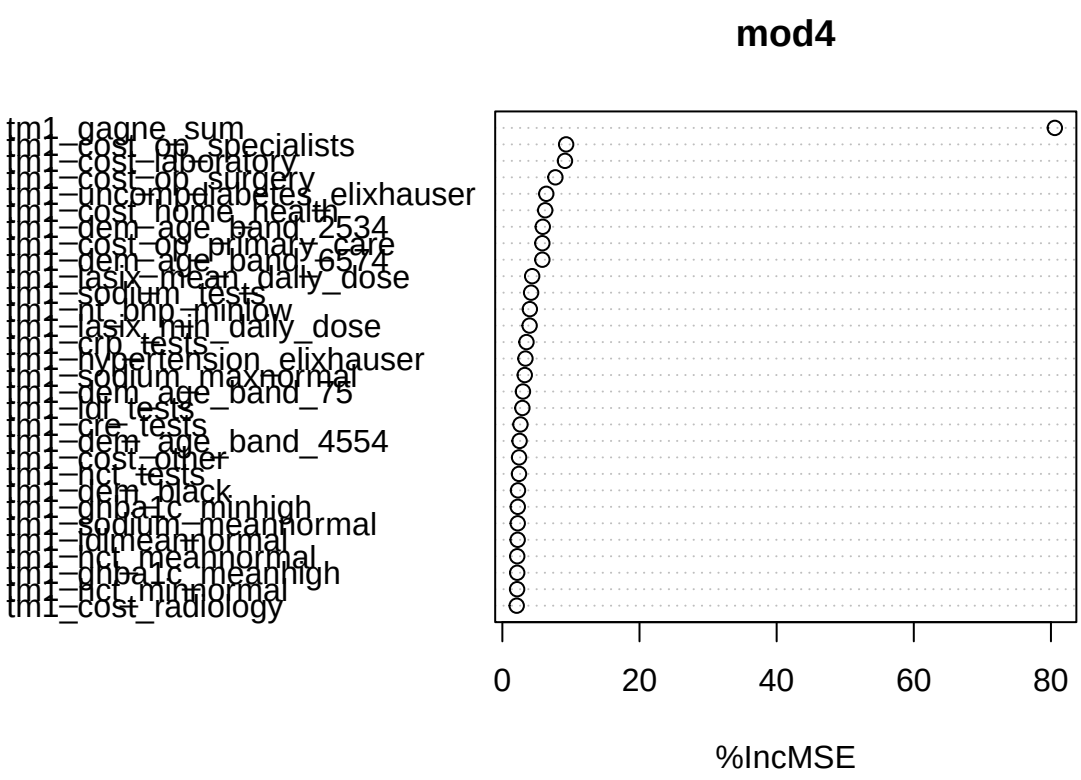
I split the data into a training and test group. There are 4791 observations in the training group and 43993 in the test group.

Question 2


mod1
tm1_dem_age_band_6574
tm1_valvulardz_elixhauser
tm1_hypertension_elixhauser
tm1_cost_home_health
tm1_cost_op_specialists
tm1_cost_pharmacy
tm1_cost_dialysis
tm1_cost_laboratory
tm1_sodium_maxlow
tm1_hct_tests
tm1_dem_age_band_2534
tm1_lasix_mean_daily_dose
tm1_sodium_meannormal
tm1_cre_minnormal
tm1_sodium_minnormal
tm1_esr_meanhigh
tm1_sodium_maxnormal
tm1_cerebrovasculardz_romano
tm1_ldl_maxlow
tm1_lasix_min_daily_dose
tm1_sodium_meanlow
tm1_trig_maxhigh
tm1_hct_maxnormal
tm1_ulcer_romano
tm1_trig_minhigh
tm1_drugabuse_elixhauser
tm1_dem_female
tm1_dementia_romano
tm1_lasix_dose_count
tm1_lasix_max_daily_dose
2
3
4
5
6
%IncMSE


mod2
tm1_dem_age_band_6574
tm1_cost_pharmacy_
tm1_cost_home_health_
tm1_dem_age_band_2534
tm1_cerebrovasculardz_romano
tm1_anemia_elixhauser_
tm1_metastatic_romano
tm1_hypertension_elixhauser
tm1_valvulardz_elixhauser
tm1_cre_tests_
tm1_sodium_meannormal
tm1_pagne_sum
tm1_lasix_dose_count
tm1_hct_tests
tm1_esr_tests
tm1_cost_emergency
tm1_cost_op_specialists
tm1_lasix_min_daily_dose
tm1_ghba1c_meanhigh
tm1_cost_laboratory
tm1_lasix_mean_daily_dose
tm1_hct_maxnormal
tm1_arrhythmia_elixhauser
tm1_drugabuse_elixhauser
tm1_cre_meanhigh
tm1_cost_dialysis
tm1_ldl_maxhigh
tm1_sodium_maxnormal
tm1_esr_maxhigh
tm1_cost_radiology
2
3
4
5
6
7
%IncMSE


mod3
tm1_gagne_sum
tm1_cost_laboratory
tm1_uncomplicdiabetes_elixhauser
tm1_cost_op_specialists
tm1_cost_home_health
tm1_cost_op_primary_care
tm1_dem_age_band_6574
tm1_cost_op_surgery
tm1_cost_radiology
tm1_crp_tests
tm1_lasix_mean_daily_dose
tm1_dem_age_band_75
tm1_sodium_tests
tm1_cost_other
tm1_ghba1c_minhigh
tm1_cre_minlow
tm1_sodium_maxnormal
tm1_cre_meannormal
tm1_hct_tests
tm1_trig_meanlow
tm1_ht_bnp_minlow
tm1_cost_emergency
tm1_drugabuse_elixhauser
tm1_lasix_max_daily_dose
tm1_ghba1c_tests
tm1_hct_meanlow
tm1_hypertension_elixhauser
tm1_lasix_min_daily_dose
tm1_cre_tests
tm1_ht_bnp_meanhigh
0
20
40
60
80
%IncMSE



Question 3 (Training Data)

- The RMSPE is higher for the model that excludes race for both the cost models and the health models

```
##                        algorithm        RMSPE
## 1 Model 1 - Costs (excl. race)  6858.8229010
## 2 Model 2 - Costs (incl. race)  7018.0751853
## 3 Model 3 - Health (excl. race)    0.4554441
## 4 Model 4 - Health (incl. race)    0.4564577
```

Question 4 (Test Data)

- The RMSPE_OOS is higher for the model that includes race for the cost models and is higher for the model that excludes race for the health model

```
##                        algorithm    RMSPE_OOS
## 1 Model 1 - Costs (excl. race)  18836.925703
## 2 Model 2 - Costs (incl. race)  18690.828483
## 3 Model 3 - Health (excl. race)     1.058870
## 4 Model 4 - Health (incl. race)     1.056515
```

Question 5

Question 6

I calculated percentile ranks for each of the 4 model predictions

Question 7

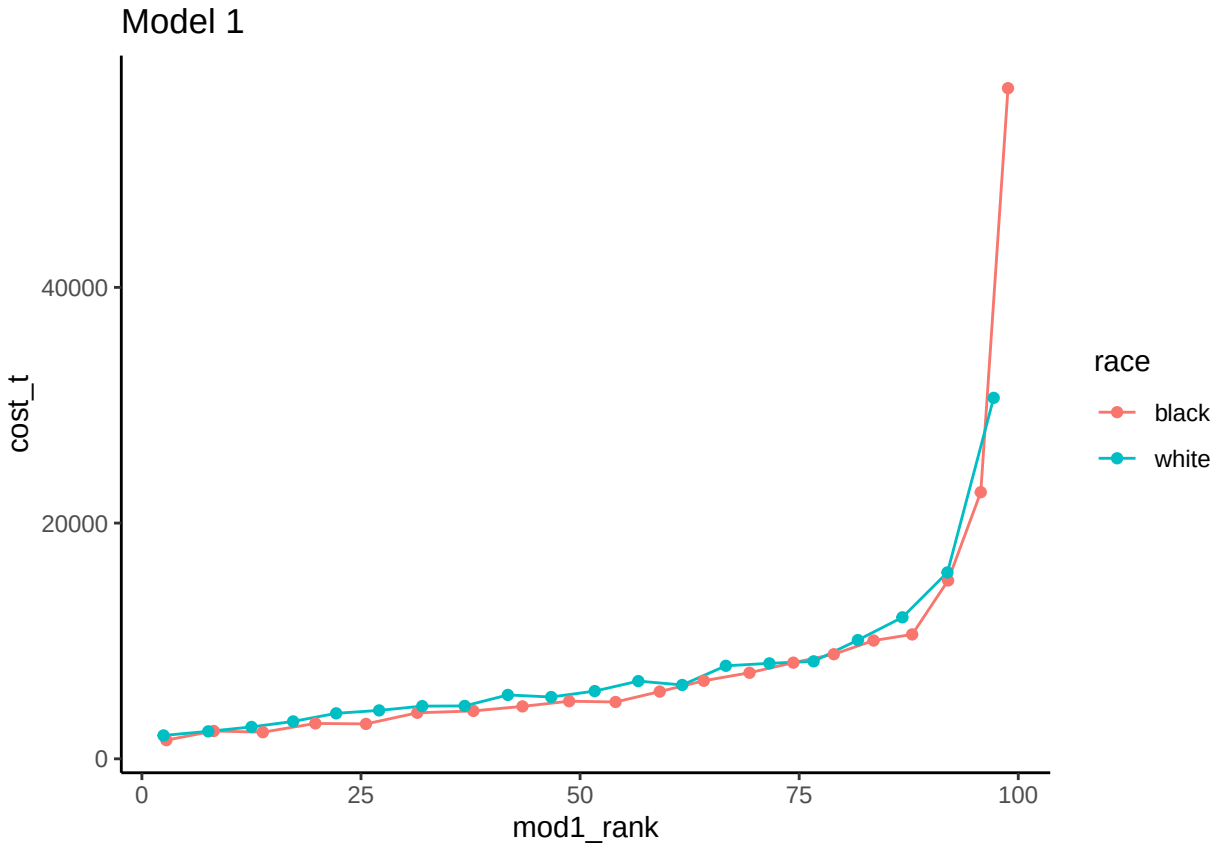
Part a

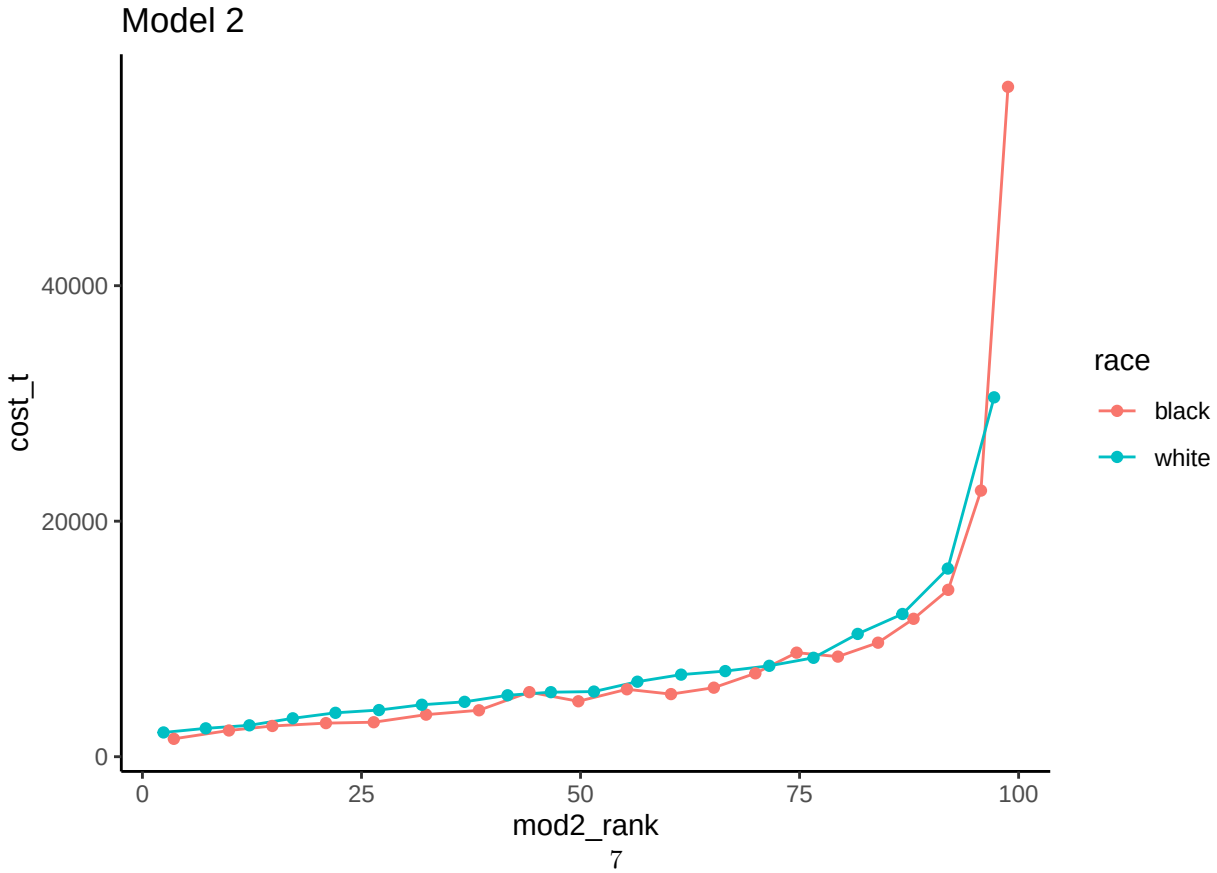
model	fraction_qualify
mod1	0.5153557
mod2	0.5270458
mod3	0.5870814
mod4	0.6166039

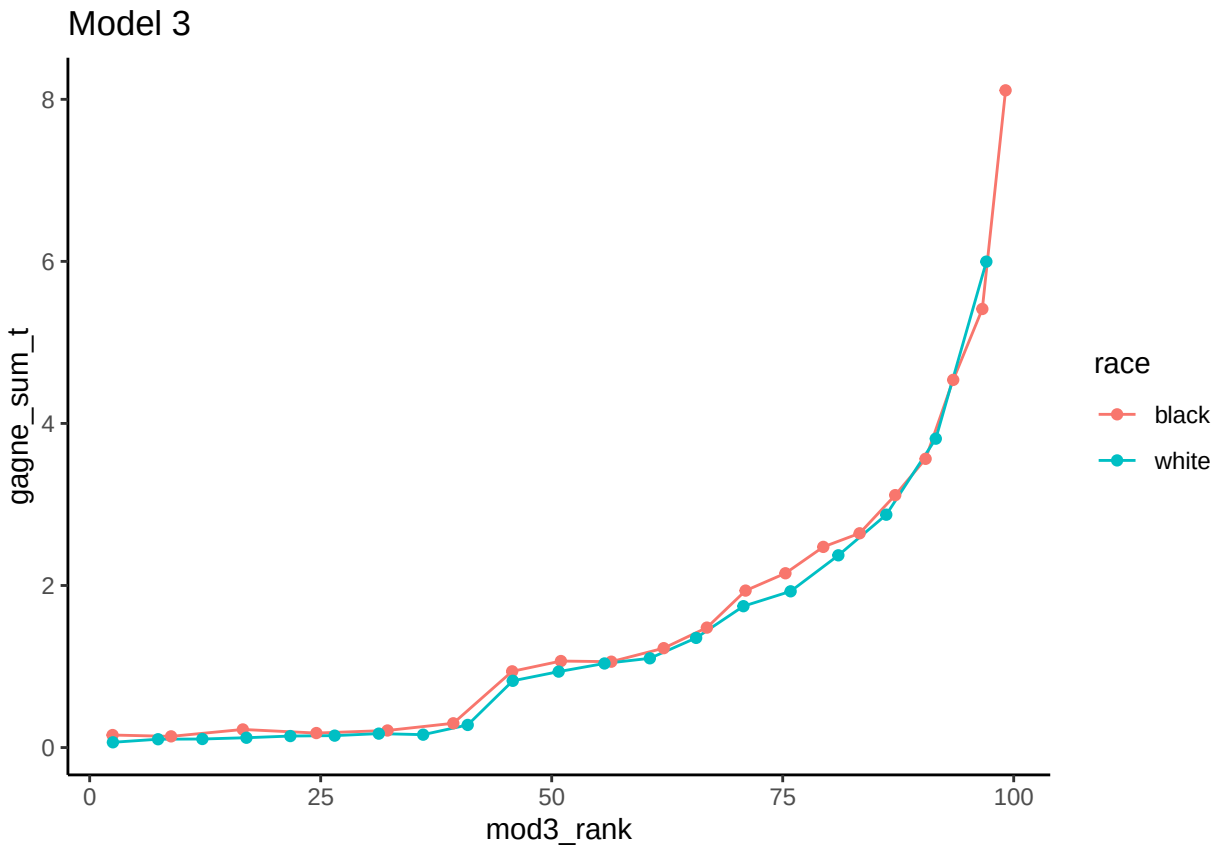
Part b

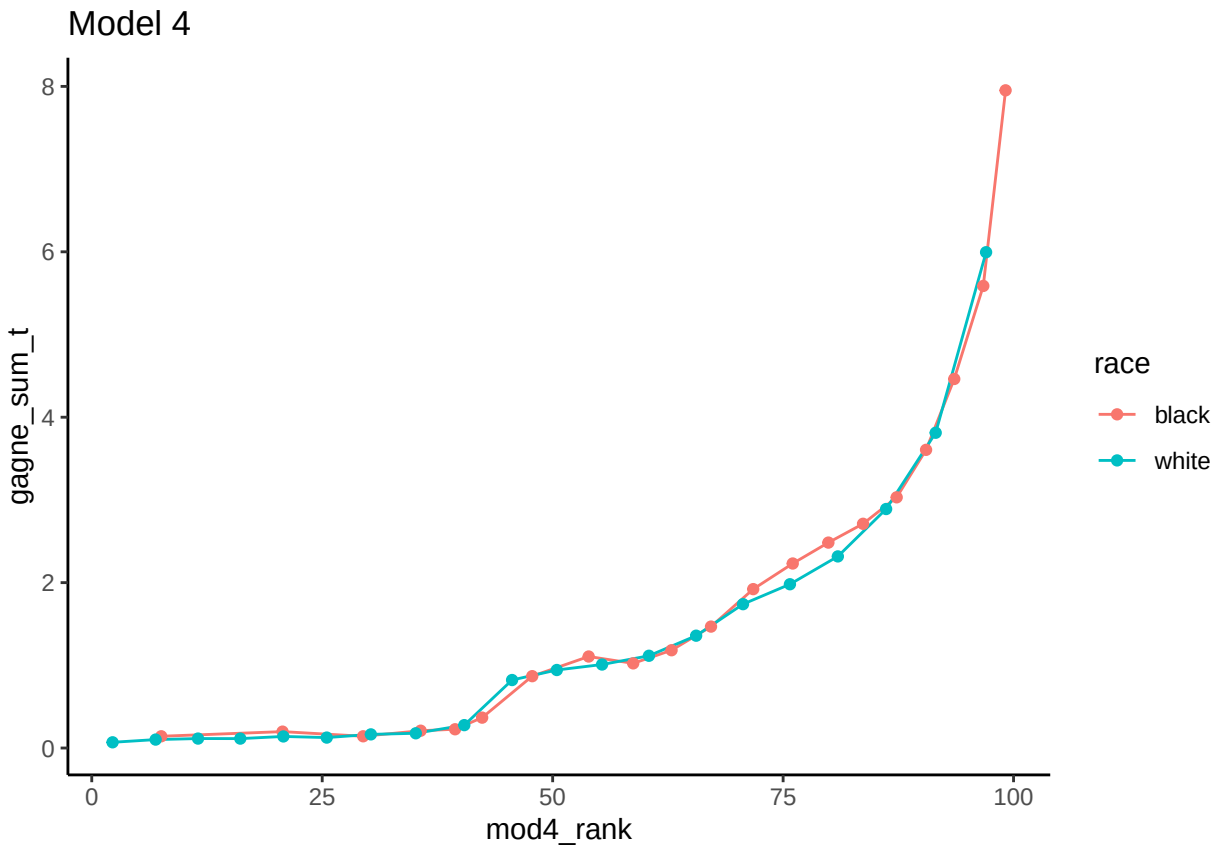
model	fraction_black
mod1	0.1313835
mod2	0.1343638
mod3	0.1496465
mod4	0.1572035

Question 8


Model 1
cost_t
40000
20000
0
0
25
50
75
100
mod1_rank
race
black
white


Model 2
cost_t
40000
20000
0
0
25
50
75
100
mod2_rank
race
black
white


Model 3
gagne_sum_t
0
2
4
6
8
0
25
50
75
100
mod3_rank
race
black
white



Question 9

Professor Obermeyer was referring to the fact that this algorithm was trained to predict health cost as the dependent variable, or "label." The algorithm designers used cost as a proxy for health needs for patients. However, it turns out that using cost as a proxy for need leads to biased results. As we can see in the graphs above, the models that predict gagne_sum_t (total number of active chronic illnesses) is biased against black patients, meaning that it predicts a lower risk score for black patients with the same health as white patients. This happens because, as just discussed, the model was trained to predict cost. As we can see, the model pretty accurately predicts cost for both black and white patients. However, for a variety of reasons, the costs for black patients are often lower than the cost for white patients. As such,the model underestimates the needs of black patients relative to white patients.