ECON1000 - Using Big Data to Solve Economic and Social Problems
Economics and Watson
Brown University
Prof. John Friedman

BROWN

ECON 1000

Empirical Exercise #4 Submission Template

Due TUESDAY 10/31/2023, by 10am EDT on Gradescope

Name: Sonya Rashkovan

Group members with whom you worked[1]:

1. Maurice Silvera
2.
3.
4.

[1] In this class we encourage working in groups because you will learn a great deal from your peers. At the end of the day, however, it is important that you write up your own analysis. Concretely, this means that you should be finding your own interesting tree models throughout this entire assignment. Please then list all group members with whom you worked. If you have any questions, please ask.

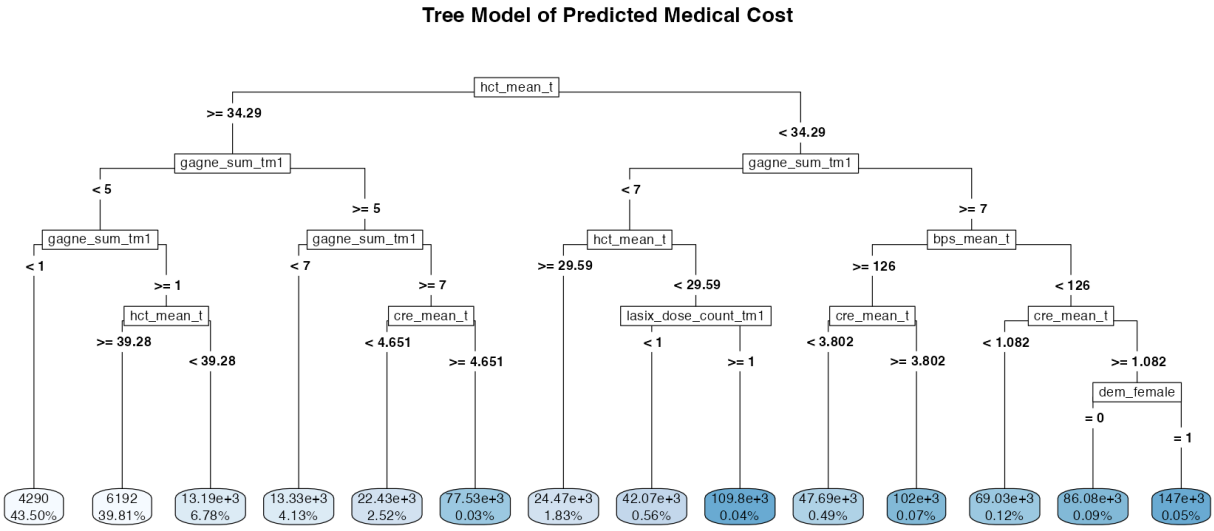
1. Use multivariate regression to predict how much a patient spends on health expenses, cost_t.

```
Call:
lm(formula = cost_t ~ gagne_sum_tm1 + dem_female + alcohol_elixhauser_tm1 +
    bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 +
    ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t, data = training_data)

Residuals:
   Min     1Q Median     3Q    Max
-48862  -6767  -3022   1750 452891

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               55045.745   3027.989  18.179  < 2e-16 ***
gagne_sum_tm1              1285.224    107.466  11.959  < 2e-16 ***
dem_female                 -940.226    461.669  -2.037   0.0417 *
alcohol_elixhauser_tm1    -3737.888   1804.961  -2.071   0.0384 *
bps_mean_t                  -32.178     12.890  -2.496   0.0126 *
cre_mean_t                 5047.642    414.449  12.179  < 2e-16 ***
drugabuse_elixhauser_tm1   1904.836   2193.930   0.868   0.3853
obesity_elixhauser_tm1     -141.871    655.004  -0.217   0.8285
ldl_mean_t                  -10.645      6.233  -1.708   0.0877 .
lasix_dose_count_tm1       3806.182    839.522   4.534 5.88e-06 ***
hct_mean_t                -1147.757     56.132 -20.447  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17340 on 7748 degrees of freedom
  (23405 observations deleted due to missingness)
Multiple R-squared:  0.1543,    Adjusted R-squared:  0.1532
F-statistic: 141.3 on 10 and 7748 DF,  p-value: < 2.2e-16
```

RMSE value: 18064.05

2. Make a decision tree to predict medical cost.

tree_model <- rpart(cost_t ~ gagne_sum_tm1 + dem_female + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t, data = training_data, control = rpart.control(cp = 0.003))

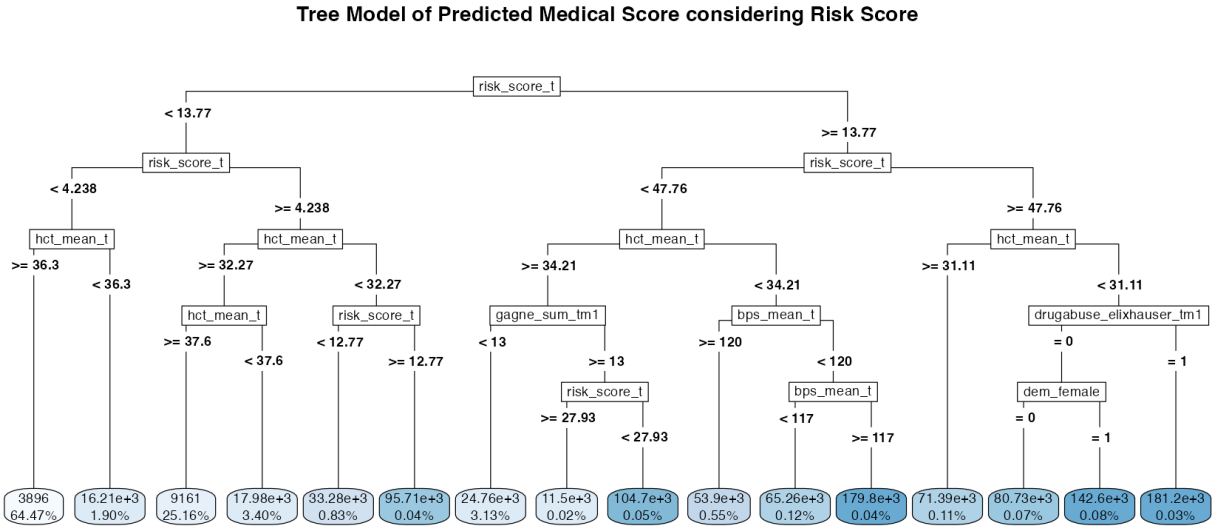


RMSE value: 17157.71

My tree is splitting on the number of hematocrit tests at the threshold of 34.29. Then at total number of active chronic illnesses, mean systolic blood pressure in a year, Mean creatinine in year , number of Lasix doses, and indicator for female gender.

3. Recode your decision tree from **two** but including risk score as a potential split factor.

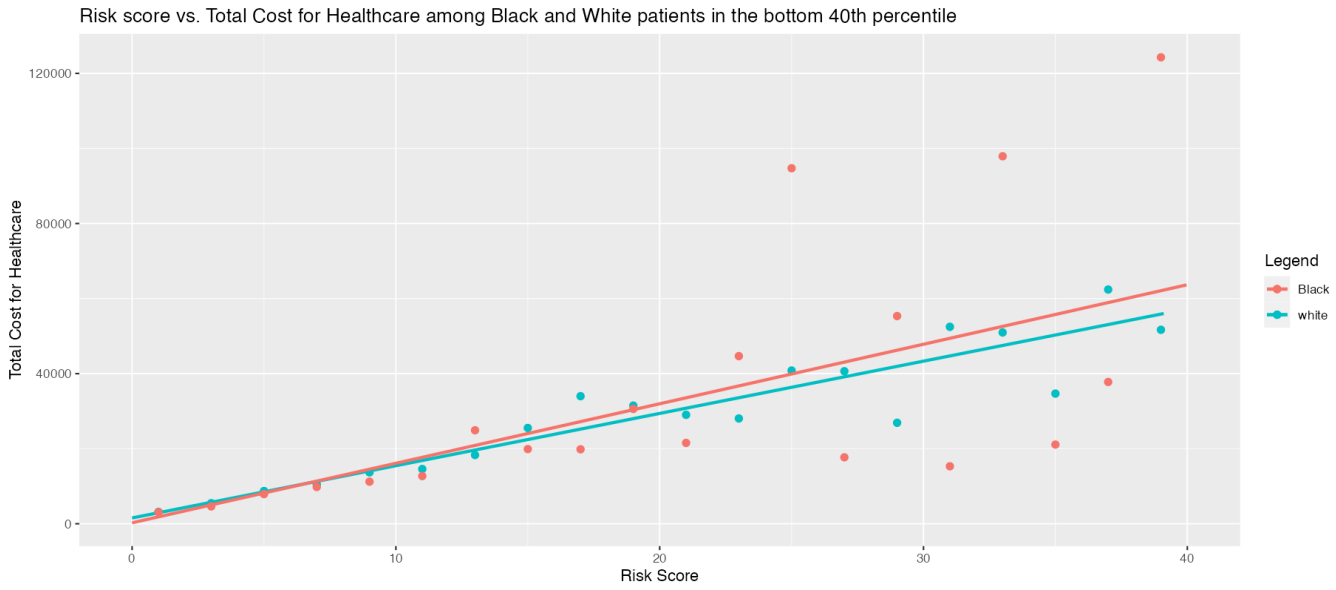
tree_model <- rpart(cost_t ~ gagne_sum_tm1 + dem_female + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t + risk_score_t, data = training_data, control = rpart.control(cp = 0.003))



RMSE value: 16916.77

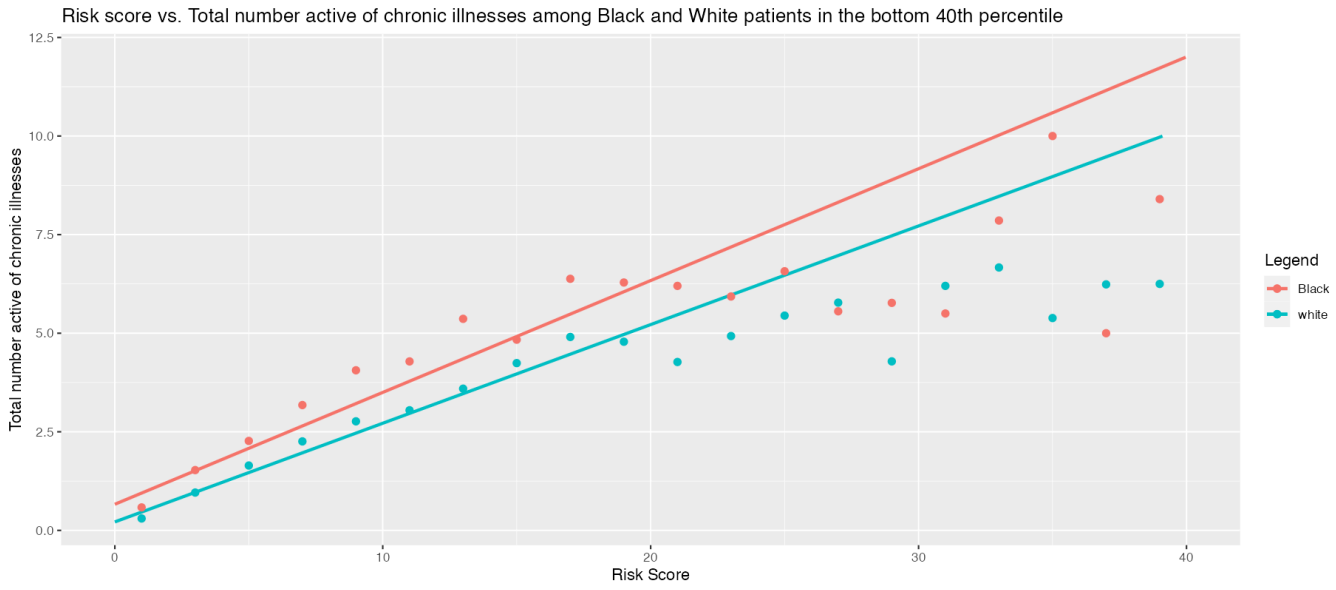
The tree changed drastically: the root node changed from number of hematocrit tests to risk score at the threshold of above/below 13.77. Also, Indicator for drug abuse was added to the tree. Number of Lasix doses and number of creatinine tests don't appear on the tree anymore. The tree might have changed because the cost is mostly correlated with how sick the person and that's determined by the risk factor—thus the risk score becomes the main root node.

4. Examine potential racial bias in cost prediction by creating a binscatter of risk score vs total expenditure.



The binscatter shows a positive correlation between risk scores and total healthcare cost for patients of both races. There is not that much variation in the data between black and white patients. It indicated that there is not that much bias in cost prediction for black and white patients at the same risk factor. This means that black and white patients at the same risk factor probably pay the same amount for their health costs.

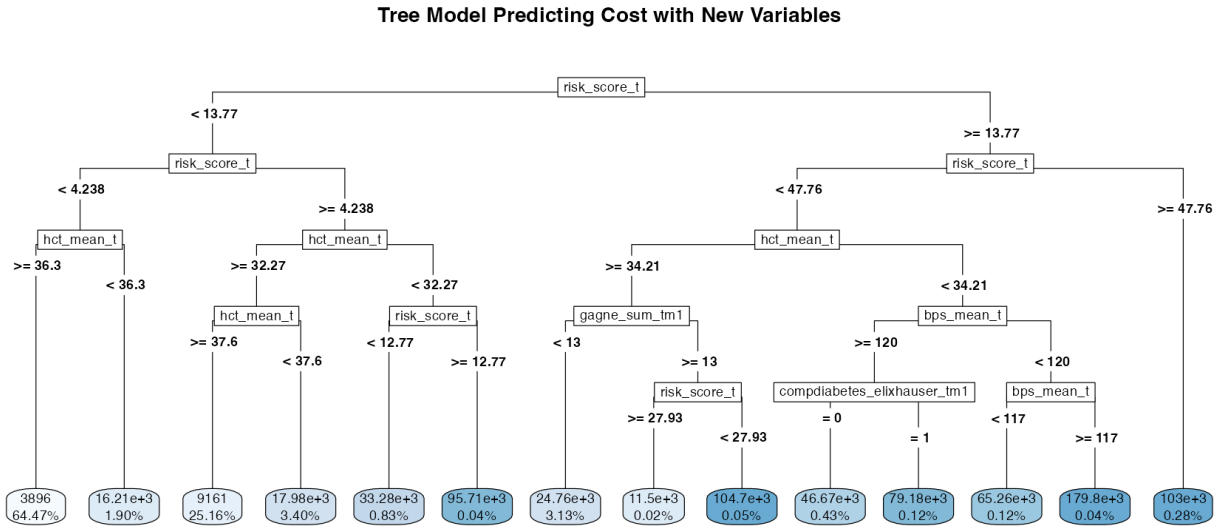
5. Examine potential racial bias in health prediction by creating a binscatter of risk score vs comorbidity.



In this binscatter we observe the relationship between the risk score and the total number of active chronic illnesses in which we observe a positive correlation of as the risk score and total number of chronic illnesses both moving into the same direction. The comobidity score means that the average number of chronic illnesses in all people within one bucket of risk score. It does make sense that for people in higher risk score categories, they would have more chronic illnesses becasue chronic illnesses because the more sick a person is, the more illnesses they would have. The binscatter does show a large data variation in the binscatters and trend lines for white and Black patients which indicates a racial bias. For patients with the same risk score, Black patients would have more average active chronic illnesses. This suggests that using number of chronic active illnesses rather than cost would not be more effective to eliminate racial bias but increase it.

6. Create the best model! **Remember to submit this code in the additional Gradescope assignment.**

```
tree_model <- rpart(cost_t ~ gagne_sum_tm1 + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t + risk_score_t + compdiabetes_elixhauser_tm1 + race, data = training_data,  control = rpart.control(cp = 0.003))
```



RMSE value: 16910.57

For the competition decision tree, I decided to predict cost considering risk factor, total number of active illnesses, Indicator for alcohol abuse, mean systolic blood pressure in year, Mean creatinine in year, indicator for drug abuse, indicator for obesity, number of LDL tests, number of Lasix doses, number of hematocrit tests, indicator for diabetes, and race. The variables that changed from the initial "Tree Model Predicting Cost" are: I added risk factor, Indicator for alcohol abuse, indicator for drug abuse, indicator for obesity, number of LDL tests, number of hematocrit tests, indicator for diabetes, and race while removing the indicator for female gender.

The root node remained risk factor, as in the Tree Model Predicting Cost considering Race, which suggests that it is the critical determinant of the end cost for patients. The major changes in the are happening on the right end of the tree model towards patients paying more. We see that the indicator for diabetes creates a new split where patients who have diabetes are paying more (79.18e+3) than patients without diabetes (46.67e+3). What is interesting is that while I included the drug abuse in the code, it doesn't show up on the tree model. And then, of course, there is not split for gender because I didn't include that on the code. What doesn't change is the lowest price but in this model the highest price is lower than in second tree model from 181.2e+3 to 179.8e+3.

The RMSE does drop from the first Tree Model from 17157.71 to 16910.57 and doesn't change that significantly from the second tree model from 16916.77 to 16910.57 which makes sense because the risk factor remains the most considerable variable.

Something fascinating and at the same time very logical is that the more health issues I included in the tree, the more splits there were on the right side of the model where we observe patients paying more which suggests that the more sick a person is, the more they would pay for health expenditures – it was fun to observe that.

Please include your code report here. The simplest way of doing this is to open the html file (generated from the Download Code button at the end of the HW - see below) and copy all of that text here.

	values
1	Sofiia Rashkovan
2	
3	set.seed(26) sample <- sample(c(TRUE,FALSE), nrow(health), replace = TRUE, prob=c(.8,.2)) training_data <- health[sample,] test_data <- health[!sample,]
4	multregr_health <- lm(formula = cost_t ~ gagne_sum_tm1 + dem_female + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t, data = training_data) summary(multregr_health)
5	test_data[c('predicted_cost')] <- predict(object = multregr_health, newdata = test_data) test_drop_na <- test_data %>% drop_na(predicted_cost) %>% drop_na(cost_t) rmse(actual = test_drop_na$cost_t, predicted = test_drop_na$predicted_cost)

6	# create decision tree
	tree_model <- rpart(cost_t ~ gagne_sum_tm1 + dem_female + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t, data = training_data, control = rpart.control(cp = 0.003))

7	rpart.plot(tree_model, main = "Tree Model of Predicted Medical Cost", type = 5, digits = 4)
8	test_data[c('predicted_cost')] <- predict(object = tree_model, newdata = test_data) test_drop_na <- test_data %>% drop_na(predicted_cost) %>% drop_na(cost_t) rmse(actual = test_drop_na$cost_t, predicted = test_drop_na$predicted_cost)

9	# create decision tree
	tree_model <- rpart(cost_t ~ gagne_sum_tm1 + dem_female + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 + obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t + risk_score_t, data = training_data, control = rpart.control(cp = 0.003))

10	rpart.plot(tree_model, main = "Tree Model of Predicted Medical Score considering Risk Score", type = 5, digits = 4)
11	test_data[c('predicted_cost')] <- predict(object = tree_model, newdata = test_data) test_drop_na <- test_data %>% drop_na(predicted_cost) %>% drop_na(cost_t) rmse(actual = test_drop_na$cost_t, predicted = test_drop_na$predicted_cost)
12	white <- health %>% filter(race == 'white') black <- health %>% filter(race == 'black') white %>% ggplot(aes(risk_score_t, cost_t, color = "white")) + geom_smooth(method = "lm", se = FALSE, formula = "y ~ x", na.rm = TRUE) + xlim (0,40) + stat_summary_bin(fun='mean', bins=20, size=2, geom='point', na.rm = TRUE) + geom_smooth(data = black, aes(risk_score_t, cost_t, color = "Black"), method = "lm", se = FALSE, formula = "y ~ x", na.rm = TRUE) + stat_summary_bin(data = black, aes(risk_score_t, cost_t, color = "Black"), fun='mean', bins=20, size=2, geom='point', na.rm = TRUE) + labs(x = "Risk Score", y = "Total Cost for Healthcare", title = "Risk score vs. Total Cost for Healthcare among Black and White patients in the bottom 40th percentile", color = "Legend")

13	white <- health %>% filter(race == 'white')
	black <- health %>% filter(race == 'black')
	white %>%
	  ggplot(aes(risk_score_t, gagne_sum_t, color = "white")) +
	  geom_smooth(method = "lm", se = FALSE, formula = "y ~ x", na.rm = TRUE) +
	  xlim (0,40) +
	  stat_summary_bin(fun='mean', bins=20, size=2, geom='point', na.rm = TRUE) +
	  geom_smooth(data = black, aes(risk_score_t, gagne_sum_t, color = "Black"), method = "lm", se = FALSE, formula = "y ~ x", na.rm = TRUE)
	+
	  stat_summary_bin(data = black, aes(risk_score_t, gagne_sum_t, color = "Black"), fun='mean', bins=20, size=2, geom='point', na.rm = TRUE)
	+
	  labs(x = "Risk Score",
	  y = "Total number active of chronic illnesses",
	  title = "Risk score vs. Total number active of chronic illnesses among Black and White patients in the bottom 40th percentile",
	  color = "Legend")

14	# create decision tree
	tree_model <- rpart(cost_t ~ gagne_sum_tm1 + alcohol_elixhauser_tm1 + bps_mean_t + cre_mean_t + drugabuse_elixhauser_tm1 +
	obesity_elixhauser_tm1 + ldl_mean_t + lasix_dose_count_tm1 + hct_mean_t + risk_score_t + compdiabetes_elixhauser_tm1 + race, data =
	training_data, control = rpart.control(cp = 0.003))

15	rpart.plot(tree_model, main = "Tree Model Predicting Cost with New Variables", type = 5, digits = 4)
16	test_data[c('predicted_cost')] <- predict(object = tree_model, newdata = test_data) test_drop_na <- test_data %>% drop_na(predicted_cost) %>% drop_na(cost_t) rmse(actual = test_drop_na$cost_t, predicted = test_drop_na$predicted_cost)

Please submit this entire document to Gradescope.