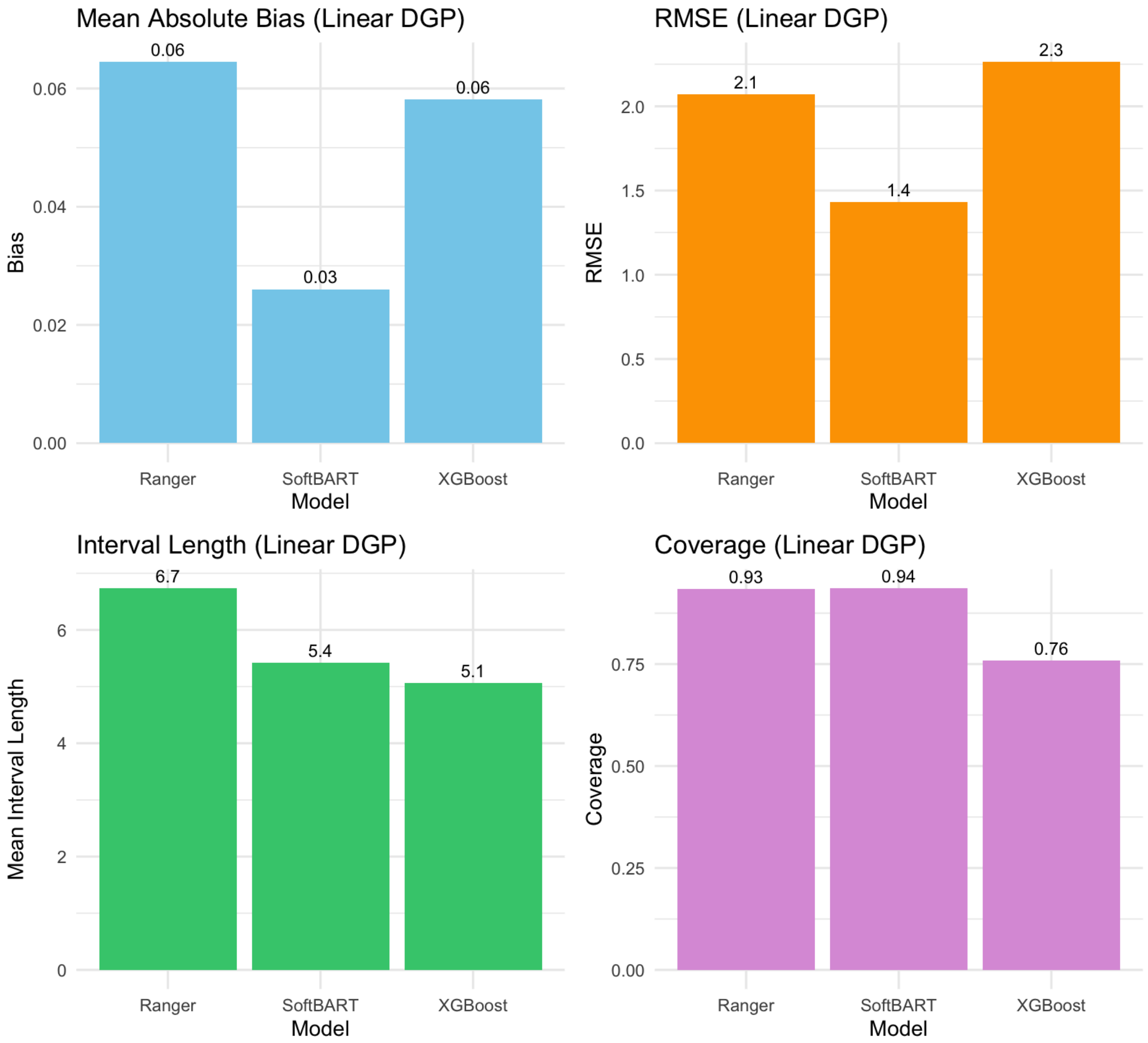
STAT5400 Computing Project Proposal

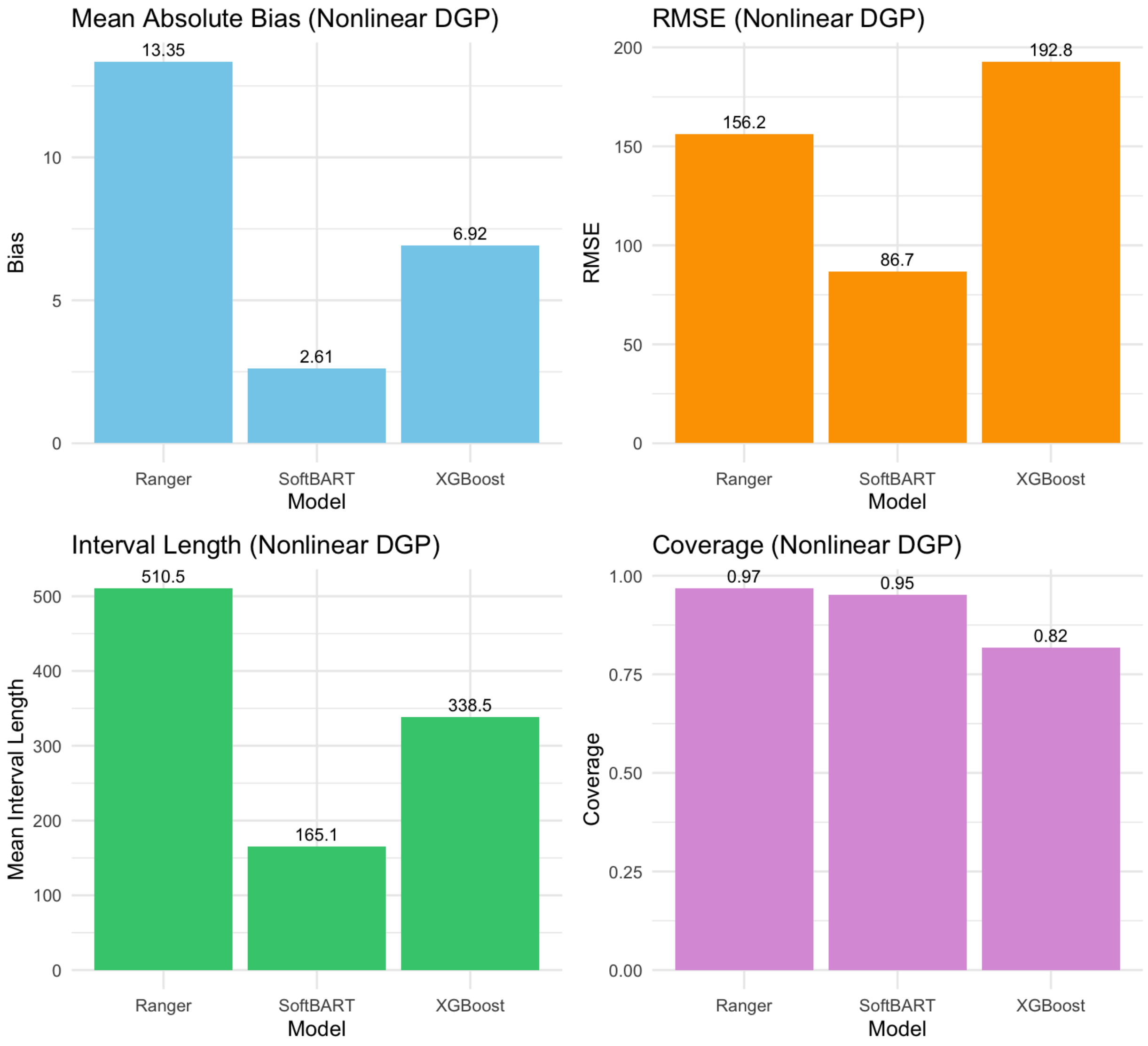
Authors: Tyler Schmidt, Dylan Day, Camden Foster, Ashwin Dervesh

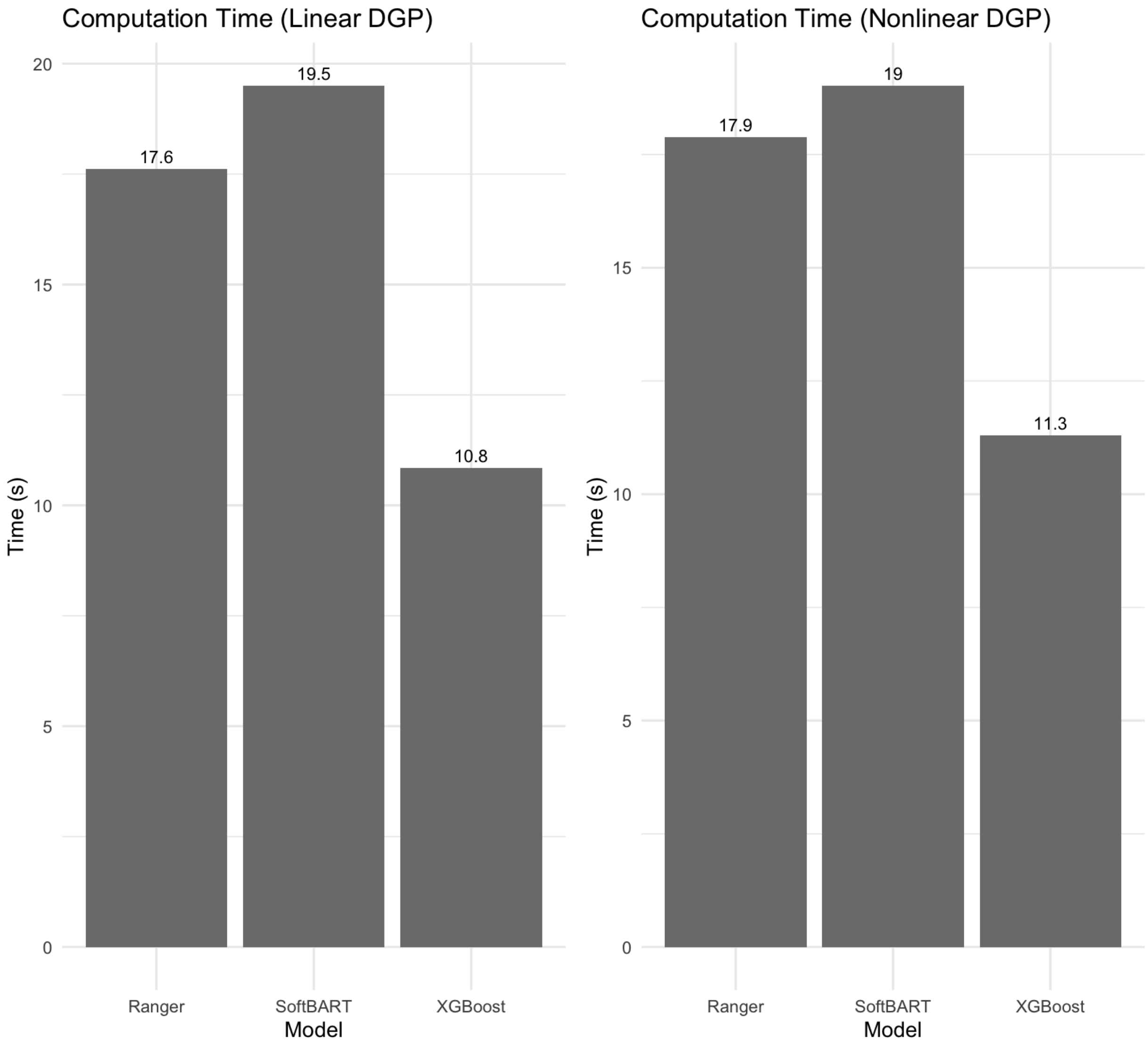
Our Project will be titled: Automatic versus Bootstrapped Uncertainty Quantification. That is, we plan to compare SoftBART (A newer version of Bayesian Additive Regression Trees) with xgBoost and random forests for uncertainty quantification of the prediction interval. SoftBART gives automatic uncertainty quantification due to its Bayesian formalism, while for xgBoost and random forests, we will need to bootstrap them. After initial simulations, we hypothesize that SoftBART will have better coverage, particularly for more complicated data-generating processes, at the expense of greater computational complexity compared to the other two models. On the next several pages, we include preliminary results for the iid case with n = 100, B =1000, and J (# simulations) = 5.

To evaluate this hypothesis, we plan to perform multiple simulation studies, first, on a simple model where covariates will enter the model linearly. Second, we plan to extend this to the case where covariates enter the model nonlinearly with higher-order interactions. Finally, we will extend to the case of time-series data. In all cases, p<<n due to computational limitations, and we will explore multiple sample sizes. The exact number of simulations and sample sizes will be determined later based on the computational requirements of the models. Furthermore, we plan to investigate the performance of these methods on an iid and a time-series real-world data set. As far as software is concerned, we will use R to write all the code, and simulations may have to be sent to the Argon cluster to run in parallel. Also, we plan to use the R packages: SoftBart, ranger, and xgboost to fit their respective models and ggplot2 to help communicate the results.

As far as the division of labor is concerned, Tyler has already written all the code for the iid case to obtain the preliminary results, and he will extend the simulation framework and SoftBART model to time-series data. Dylan and Camden will extend the xgBoost and random forest model and their bootstrap to the time-series case. Ashwin will investigate real-world data sets and make the presentation for this computing project.


Mean Absolute Bias (Linear DGP)
Bias
0.00
0.02
0.04
0.06
0.06
0.03
0.06
Ranger
SoftBART
XGBoost
Model
RMSE (Linear DGP)
RMSE
0.0
0.5
1.0
1.5
2.0
2.1
1.4
2.3
Ranger
SoftBART
XGBoost
Model
Interval Length (Linear DGP)
Mean Interval Length
0
2
4
6
6.7
5.4
5.1
Ranger
SoftBART
XGBoost
Model
Coverage (Linear DGP)
Coverage
0.00
0.25
0.50
0.75
0.93
0.94
0.76
Ranger
SoftBART
XGBoost
Model


Mean Absolute Bias (Nonlinear DGP)
Bias
0
5
10
13.35
2.61
6.92
Ranger
SoftBART
XGBoost
Model
RMSE (Nonlinear DGP)
RMSE
0
50
100
150
200
156.2
86.7
192.8
Ranger
SoftBART
XGBoost
Model
Interval Length (Nonlinear DGP)
Mean Interval Length
0
100
200
300
400
500
510.5
165.1
338.5
Ranger
SoftBART
XGBoost
Model
Coverage (Nonlinear DGP)
Coverage
0.00
0.25
0.50
0.75
1.00
0.97
0.95
0.82
Ranger
SoftBART
XGBoost
Model


Computation Time (Linear DGP)
Time (s)
20
15
10
5
0
17.6
19.5
10.8
Ranger
SoftBART
XGBoost
Model
Computation Time (Nonlinear DGP)
Time (s)
15
10
5
0
17.9
19
11.3
Ranger
SoftBART
XGBoost
Model