Machine Learning Applied to Hepatocellular Carcinoma

Sangil Lee, Yujing Lu, Josh Tomiyama

2022-04-27

Section 1

Background HCC

HepatoCellular Carcinoma (HCC)

- HepatoCellular Carcinoma (HCC) 6th most frequently diagnosed cancer.

HepatoCellular Carcinoma (HCC)

- HepatoCellular Carcinoma (HCC) 6th most frequently diagnosed cancer.
- Data mining approach to tailor evaluation and treatment for HCC are limited in the literature.

HepatoCellular Carcinoma (HCC)

- HepatoCellular Carcinoma (HCC) 6th most frequently diagnosed cancer.
- Data mining approach to tailor evaluation and treatment for HCC are limited in the literature.
- Using the HCC dataset, we undertook the data mining approach to evaluate the patient level factors to identify those who are at risk of one year mortality.

Section 2

Data Summary

Data Summary

- Clinical data of 165 pts with HCC (demographic, risk factors, lab data, and survival features)
- 49 features from HCC clinical practice guidelines (Table 1)
- About 80
- Missing data represents 10.22
- The target variable is the survival at 1 year, coded as 0 (dies) and 1 (lives).

Section 3

Data Summary

Data Summary

Table 1: Table 1. HCC Data Summary

	Overall (N=165)
gender	
female	32 (19.4%)
male	133 (80.6%)
symptom	
no	53 (36.1%)
yes	94 (63.9%)
N-Miss	18
alc	
no	43 (26.1%)
yes	122 (73.9%)

Section 4

Random Forest Model

Description

- Tree model

Description

- Tree model
- 10 fold CV

Description

- Tree model
- 10 fold CV
- hyperparameters tuned by Bayesian Optimization

Description

- Tree model
- 10 fold CV
- hyperparameters tuned by Bayesian Optimization
- KNN imputation

Estimated Performance

Calibration Plots

Variable Importance

Partial Dependence

Section 5

XGBoost Model

Description I

Input: training set $\{(x_i, y_i)\}_{i=1}^N$, a differentiable loss function $L(y, F(x))$, a number of weak learners M and a learning rate α.
Algorithm:

1. Initialize model with a constant value: $\hat{f}_{(0)}(x) = argmin_\theta \sum_{i=1}^N L(y_i, \theta)$.
2. For $m = 1, ..., M$:
   a. Compute gradients and hessians:
   $$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{m-1}(x)}$$
   $$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{m-1}(x)}.$$
   b. Fit a base leanrer using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^N$ by solving the optimization problem below:
   $$\hat{\phi}_m = argmin_{\phi \in \Phi} \sum_{i=1}^N \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2.$$

Description II

- Tuning parameters: number of boosting iterations M, shrinkage of variable weights at each iteration to prevent overfitting η, maximum tree depth.

Description II

- Tuning parameters: number of boosting iterations M, shrinkage of variable weights at each iteration to prevent overfitting η, maximum tree depth.
- put some advantages and disadvantages here

Estimated Performance I

Metric	KNN_none	KNN_corr	KNN_pca	MeanMode_none
Brier	0.2212310	0.2109347	0.2385479	0.2061546
Accuracy	0.7011029	0.7015114	0.6848039	0.6965686
Kappa	0.3566100	0.3636247	0.3306148	0.3405815
ROC AUC	0.7524242	0.7661255	0.7249567	0.7835931
Sensitivity	0.7736364	0.7736364	0.7427273	0.7827273
Specificity	0.5833333	0.5880952	0.5928571	0.5595238

- All models are not doing well for specificity.

Estimated Performance I

Metric	KNN_none	KNN_corr	KNN_pca	MeanMode_none
Brier	0.2212310	0.2109347	0.2385479	0.2061546
Accuracy	0.7011029	0.7015114	0.6848039	0.6965686
Kappa	0.3566100	0.3636247	0.3306148	0.3405815
ROC AUC	0.7524242	0.7661255	0.7249567	0.7835931
Sensitivity	0.7736364	0.7736364	0.7427273	0.7827273
Specificity	0.5833333	0.5880952	0.5928571	0.5595238

- All models are not doing well for specificity.
- KNN_corr appears to do better among these models for its highest accuracy, highest kappa, reasonable brier, roc auc, sensitivity and specificity.

Estimated Performance I

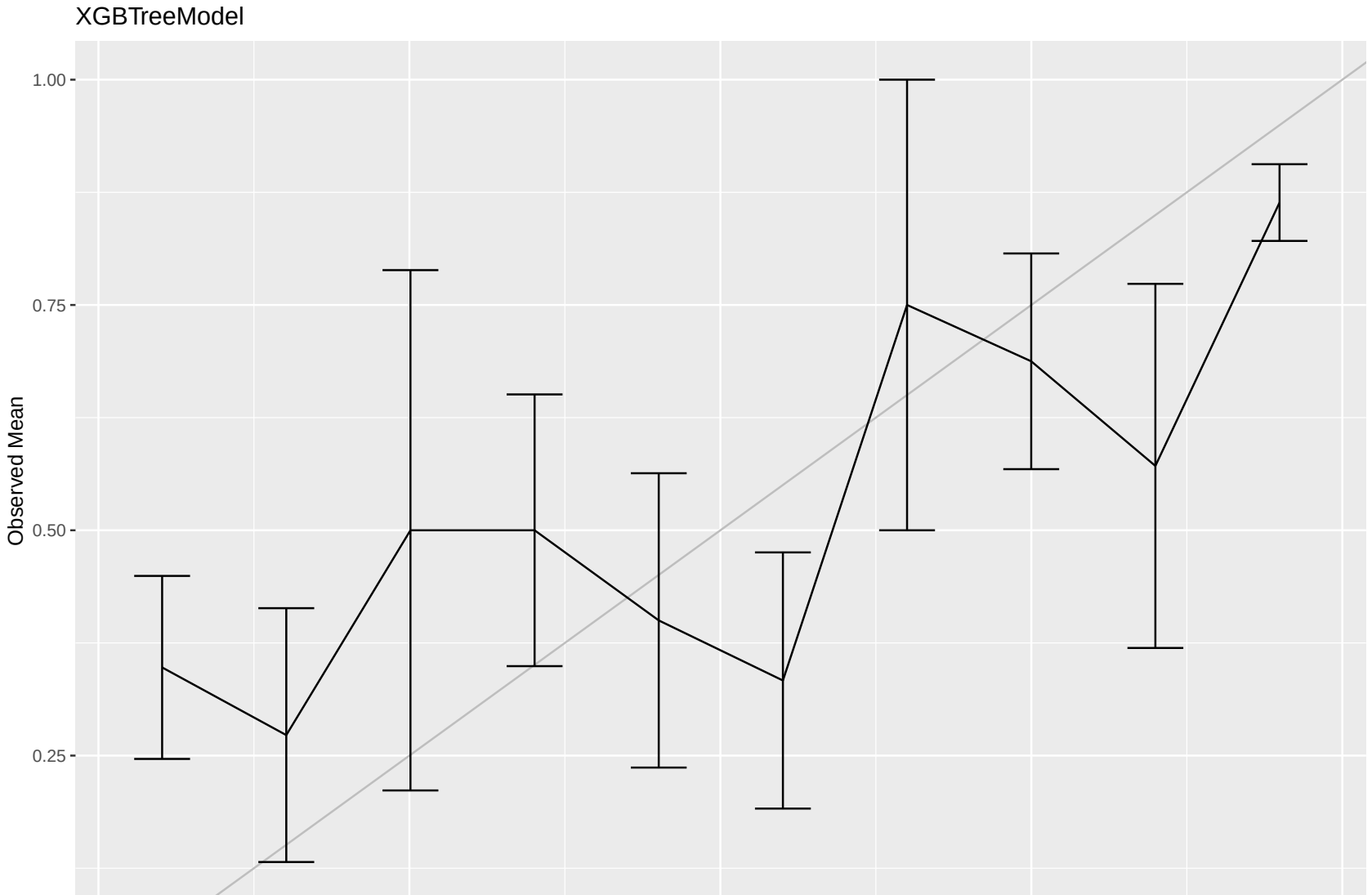
Metric	KNN_none	KNN_corr	KNN_pca	MeanMode_none
Brier	0.2212310	0.2109347	0.2385479	0.2061546
Accuracy	0.7011029	0.7015114	0.6848039	0.6965686
Kappa	0.3566100	0.3636247	0.3306148	0.3405815
ROC AUC	0.7524242	0.7661255	0.7249567	0.7835931
Sensitivity	0.7736364	0.7736364	0.7427273	0.7827273
Specificity	0.5833333	0.5880952	0.5928571	0.5595238

- All models are not doing well for specificity.
- KNN_corr appears to do better among these models for its highest accuracy, highest kappa, reasonable brier, roc auc, sensitivity and specificity.
- p-values are not significant.

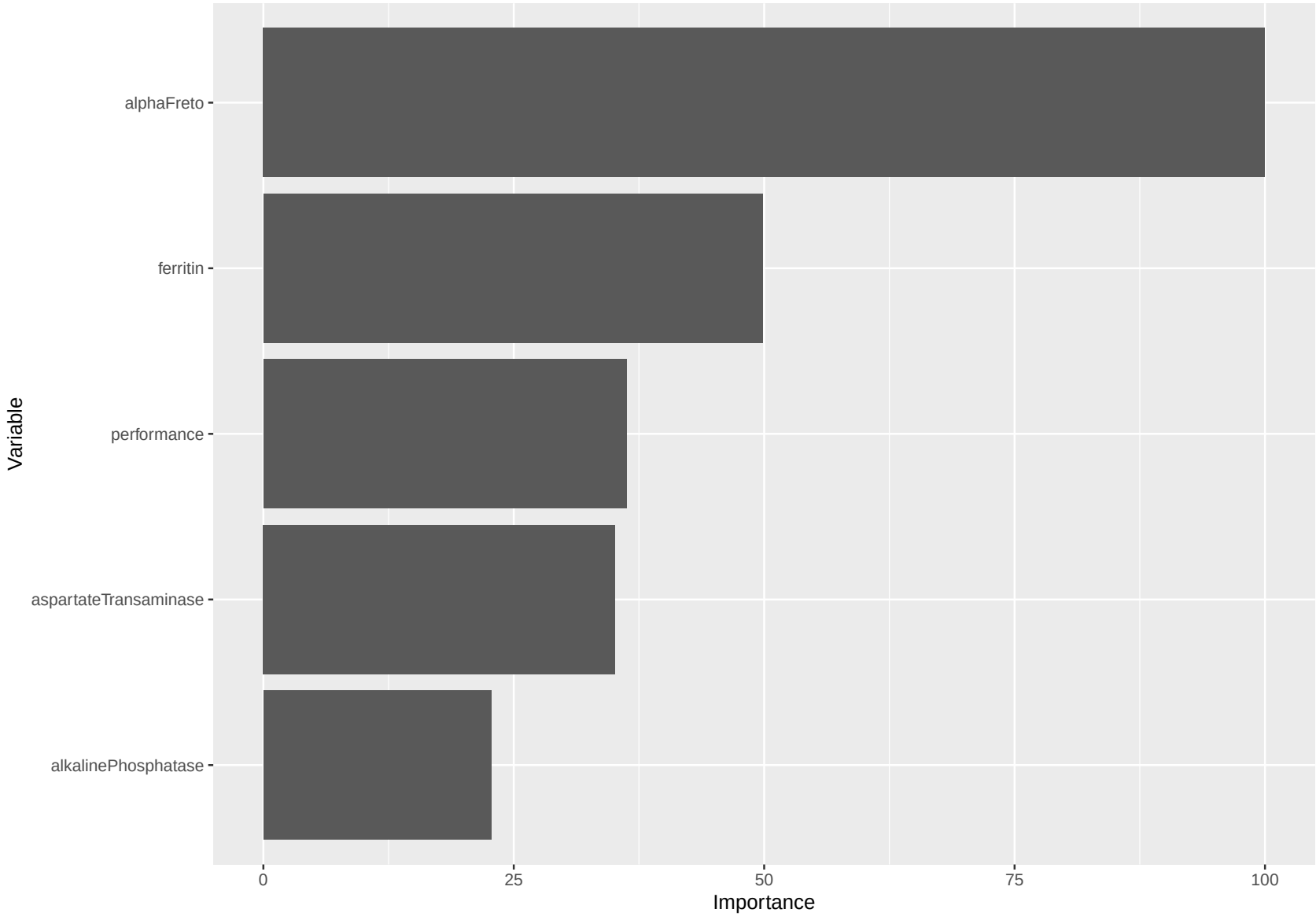
Estimated Performance II

Calibration Plots

```
## $XGBTreeModel
```

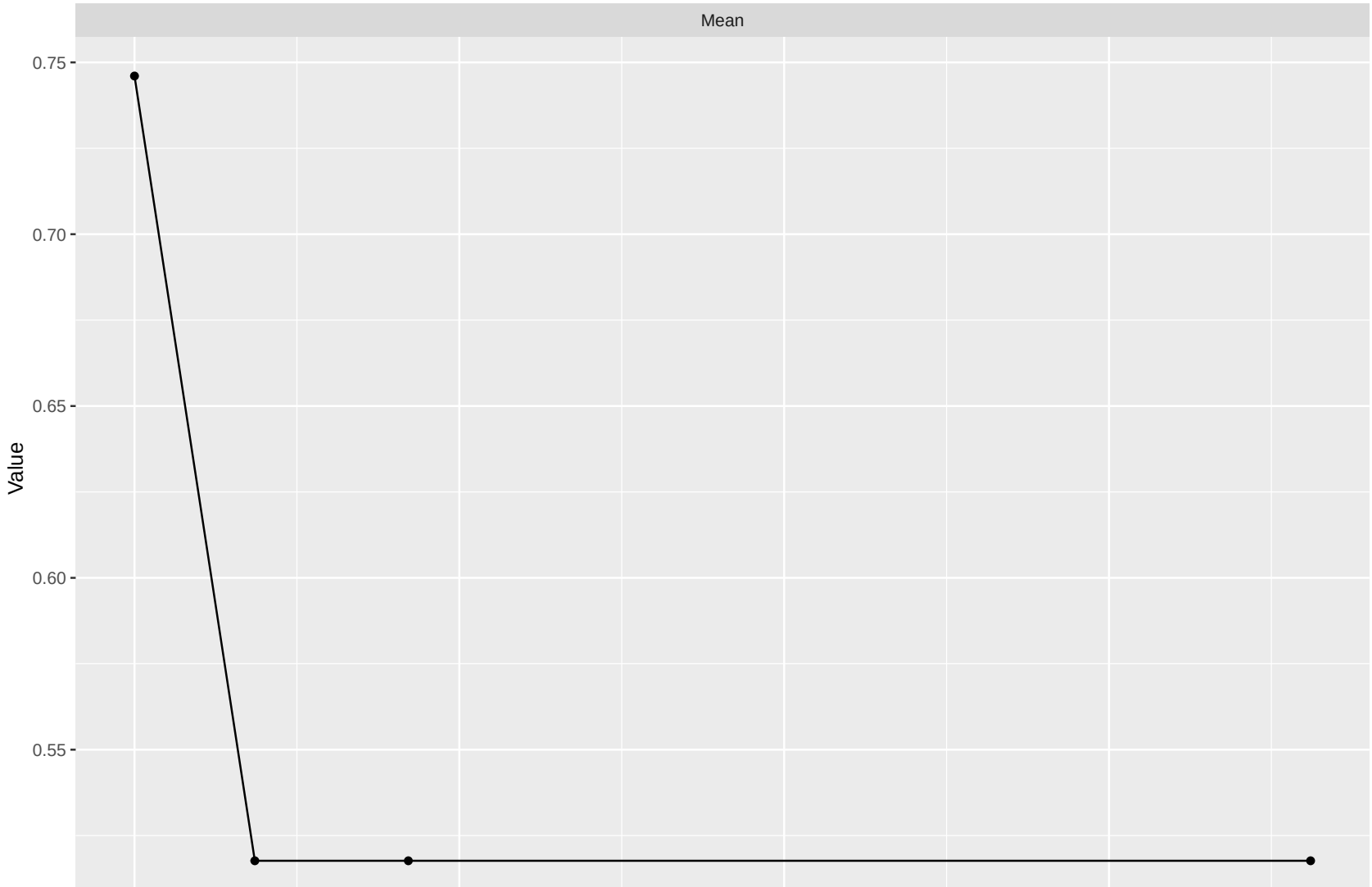


Variable Importance



Partial Dependence

```
## $alphaFreto
```



Section 6

Support Vector Machine Model

Description

- A support vector machine attempts to draw a non-linear boundary to classify the outcome by transforming the predictors using basis functions

Description

- A support vector machine attempts to draw a non-linear boundary to classify the outcome by transforming the predictors using basis functions
- In this transformed space, find linear boundaries using support vector classifier.

Description

- A support vector machine attempts to draw a non-linear boundary to classify the outcome by transforming the predictors using basis functions
- In this transformed space, find linear boundaries using support vector classifier.
- We just need to know the kernel function on the basis functions, don't need to know basis function itself.

Description

- A support vector machine attempts to draw a non-linear boundary to classify the outcome by transforming the predictors using basis functions
- In this transformed space, find linear boundaries using support vector classifier.
- We just need to know the kernel function on the basis functions, don't need to know basis function itself.
- Radial Basis kernel function: $\exp(-\sigma^2\|x - x'\|^2)$; $\sigma = 0.0123873$

Description

- A support vector machine attempts to draw a non-linear boundary to classify the outcome by transforming the predictors using basis functions
- In this transformed space, find linear boundaries using support vector classifier.
- We just need to know the kernel function on the basis functions, don't need to know basis function itself.
- Radial Basis kernel function: $\exp(-\sigma^2\|x - x'\|^2)$; $\sigma =$ 0.0123873
- Cost of misclassification C =0.8956493

Description

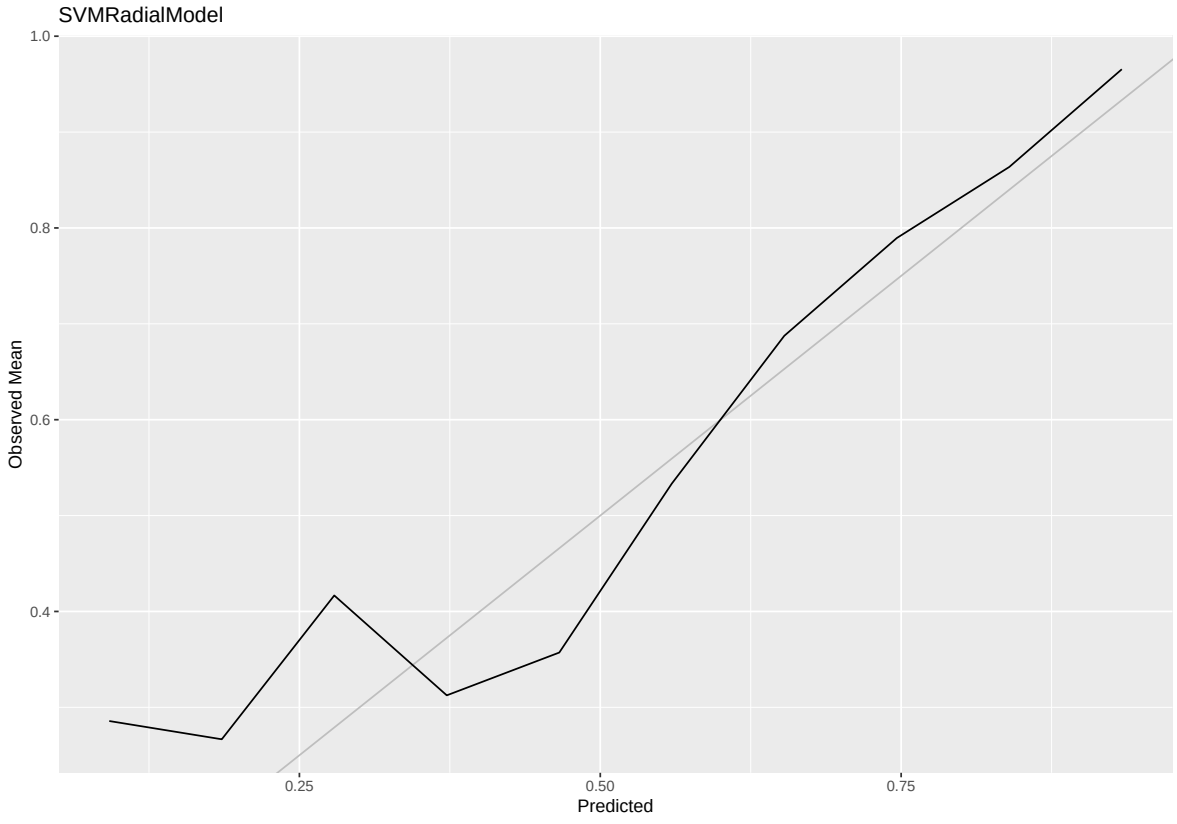
- A support vector machine attempts to draw a non-linear boundary to classify the outcome by transforming the predictors using basis functions
- In this transformed space, find linear boundaries using support vector classifier.
- We just need to know the kernel function on the basis functions, don't need to know basis function itself.
- Radial Basis kernel function: $\exp(-\sigma^2\|x - x'\|^2)$; $\sigma =$ 0.0123873
- Cost of misclassification C =0.8956493
- imputation using knn with neighbors = 4

Estimated Performance

Table 2: SVMRad results with knn imputation

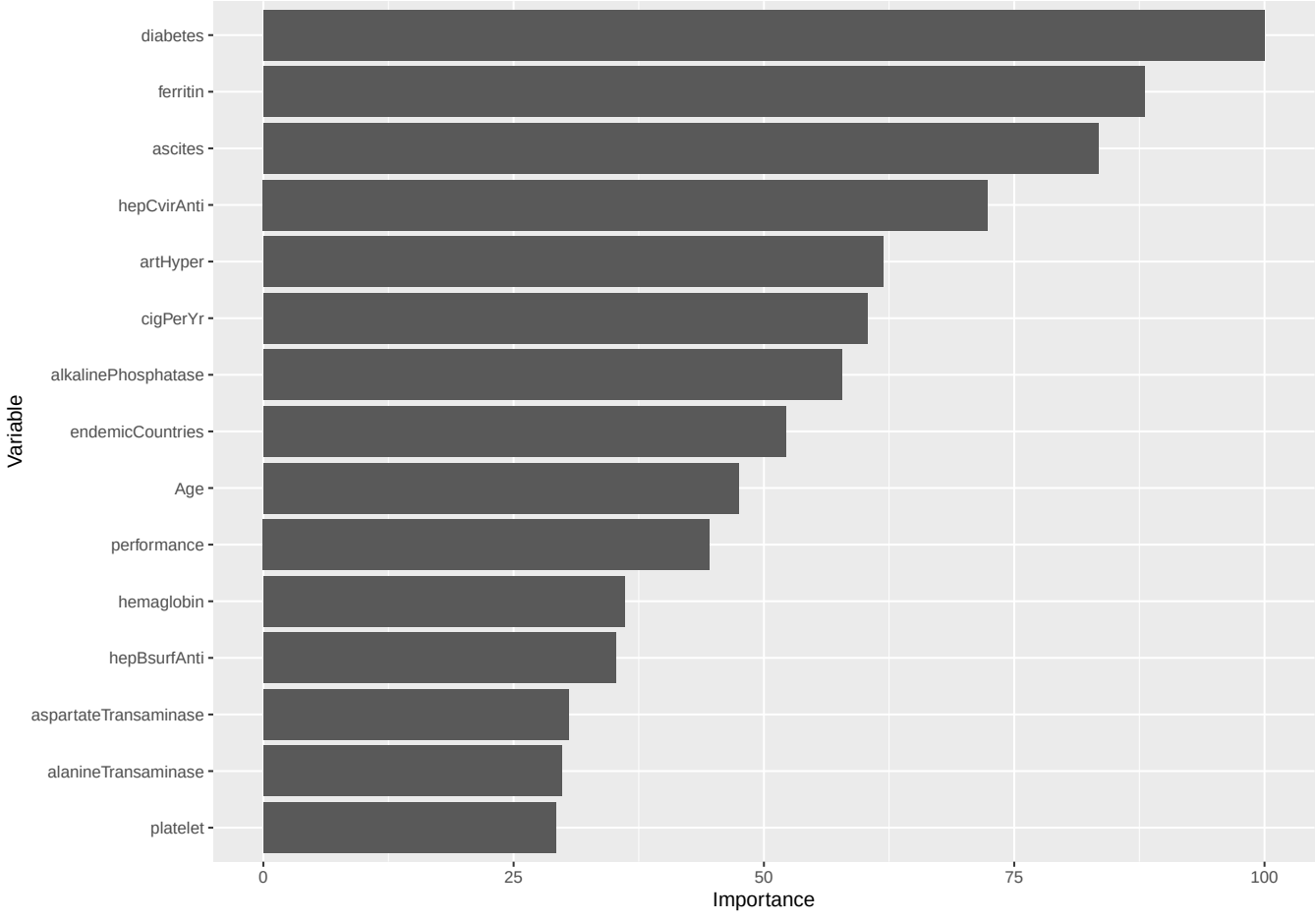
Metric	Mean	Median	SD	Min	Max	NA
Brier	0.176	0.166	0.057	0.101	0.287	0
Accuracy	0.756	0.764	0.119	0.562	0.938	0
Kappa	0.482	0.503	0.254	0.097	0.871	0
ROC AUC	0.818	0.829	0.101	0.600	0.967	0
Sensitivity	0.803	0.809	0.107	0.600	0.909	0
Specificity	0.681	0.643	0.190	0.500	1.000	0

Calibration Plots



- Decently calibrated. Low probabilities have many false negatives.

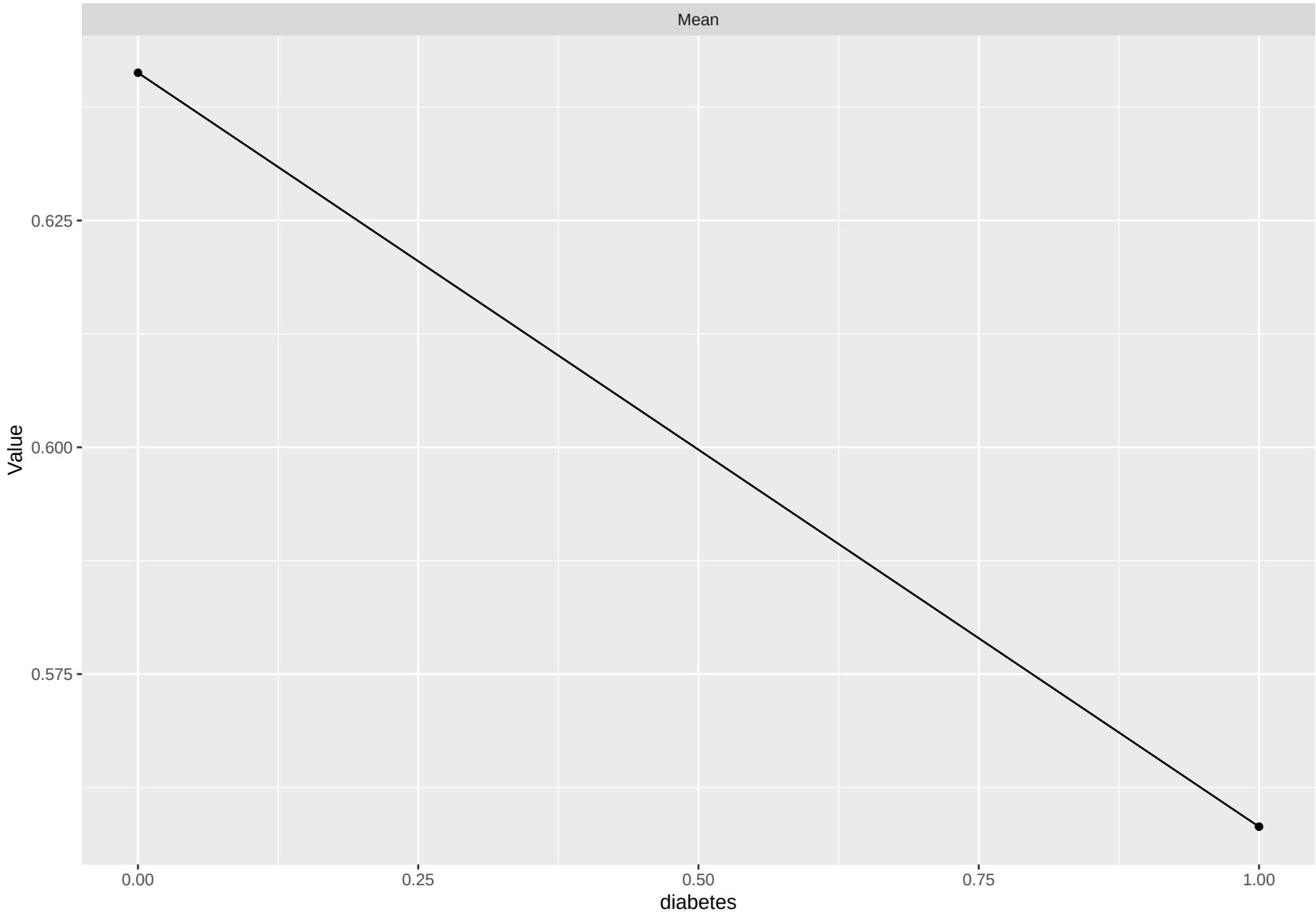
Variable Importance



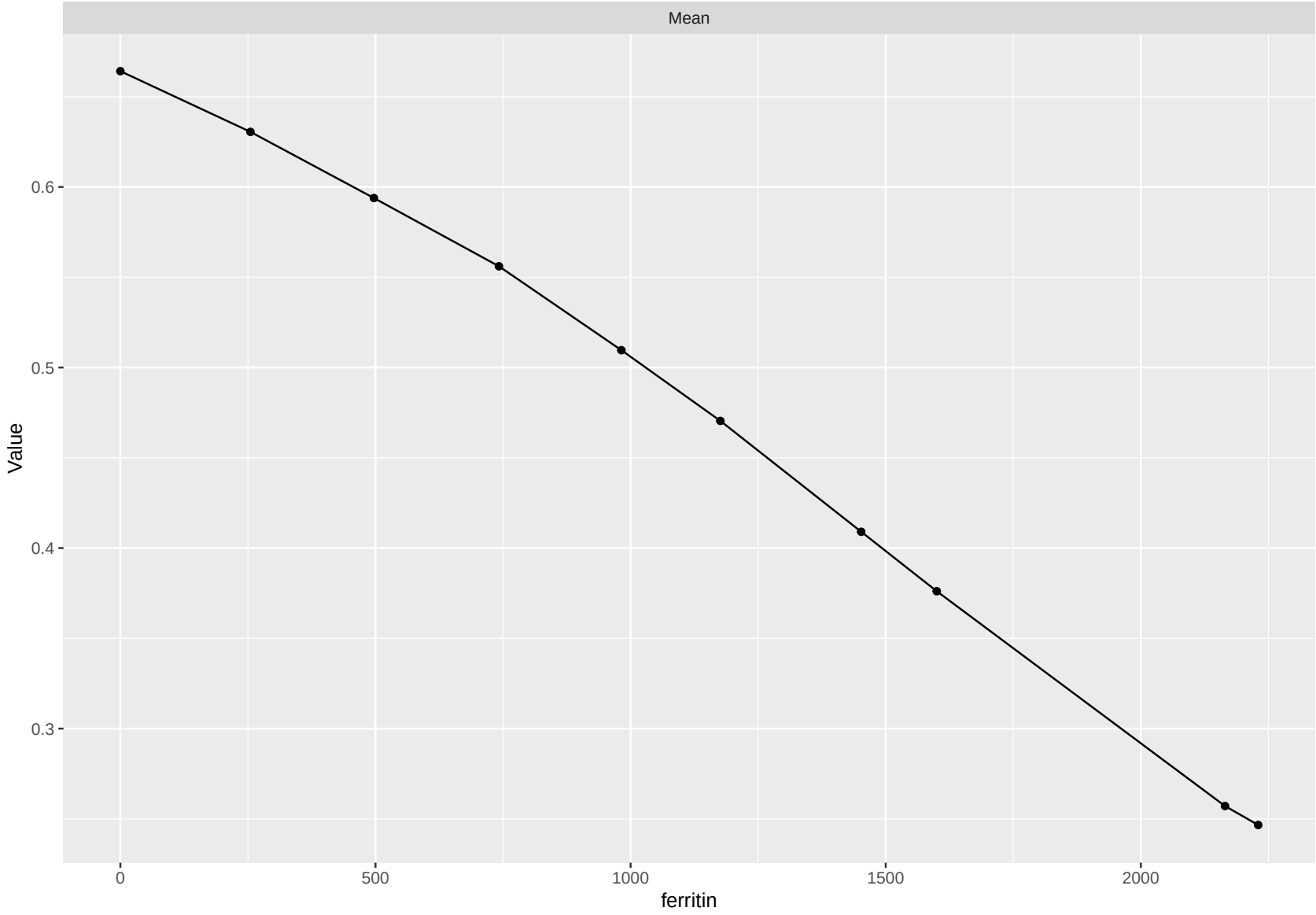
Variable Importance

- Symptoms (performance)
- Indicators of liver injury/disease/infection (hepBsurfAnti, hepCvirAnti,aspartateTransaminase, alkalinePhosphatase, ferritin, totalProteins, ascites, hemoglobin)
- Biological Characteristics (age, portalVeinThromb)
- Risk factors (diabetes, artHyper)
- behavioral/demographic (endemicCountries, cigPerYr)

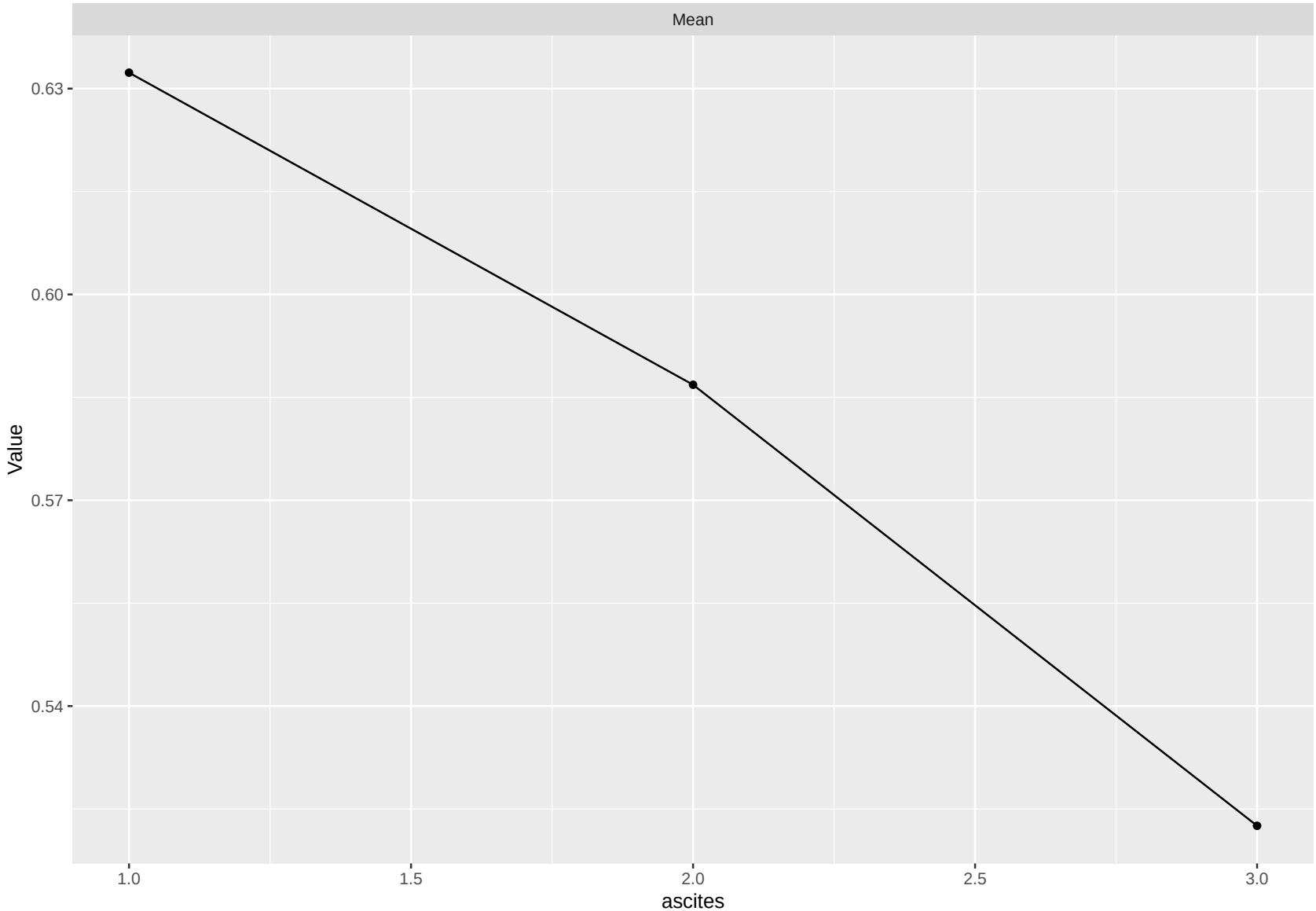
Partial Dependence: diabetes



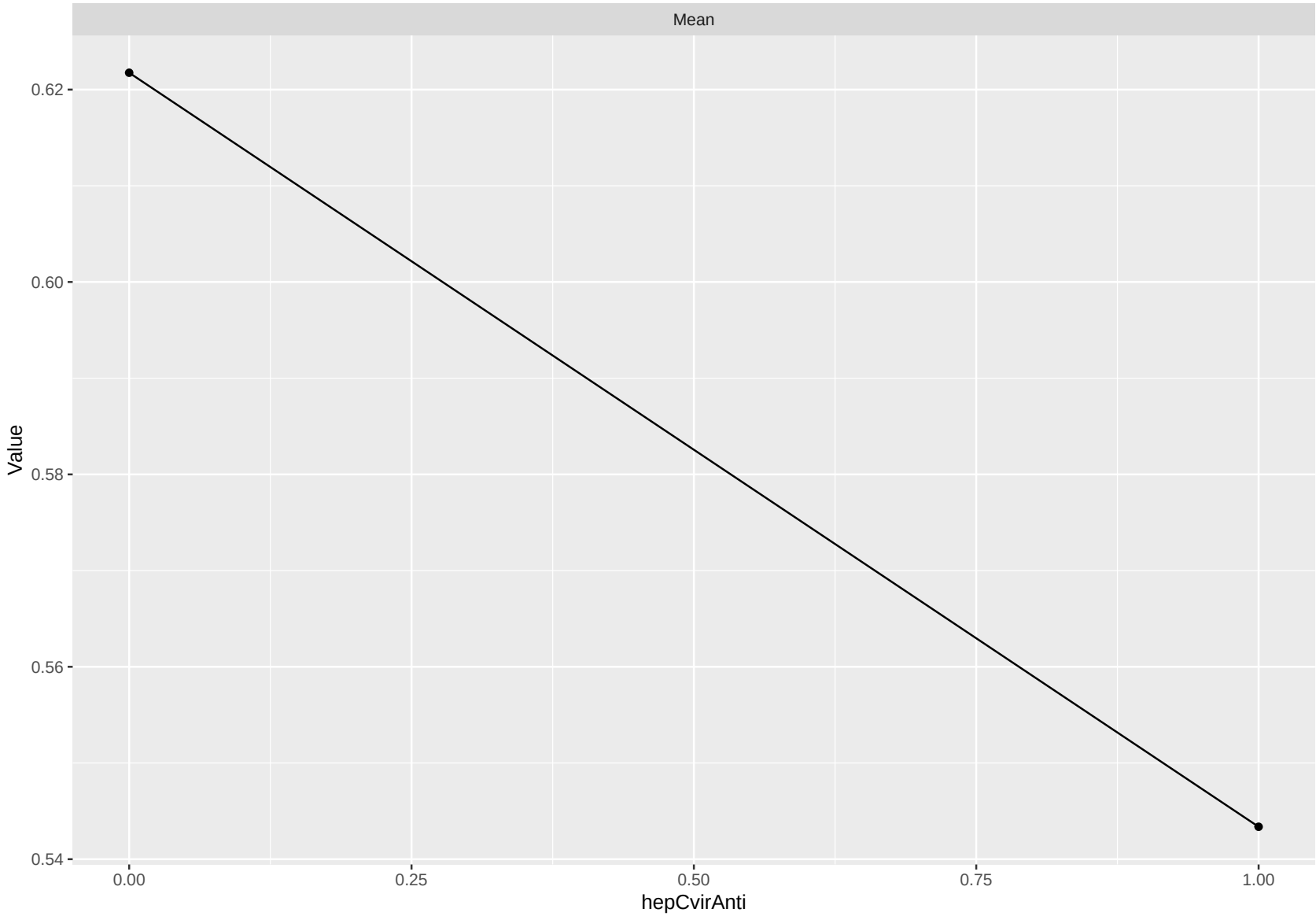
Partial Dependence: ferritin



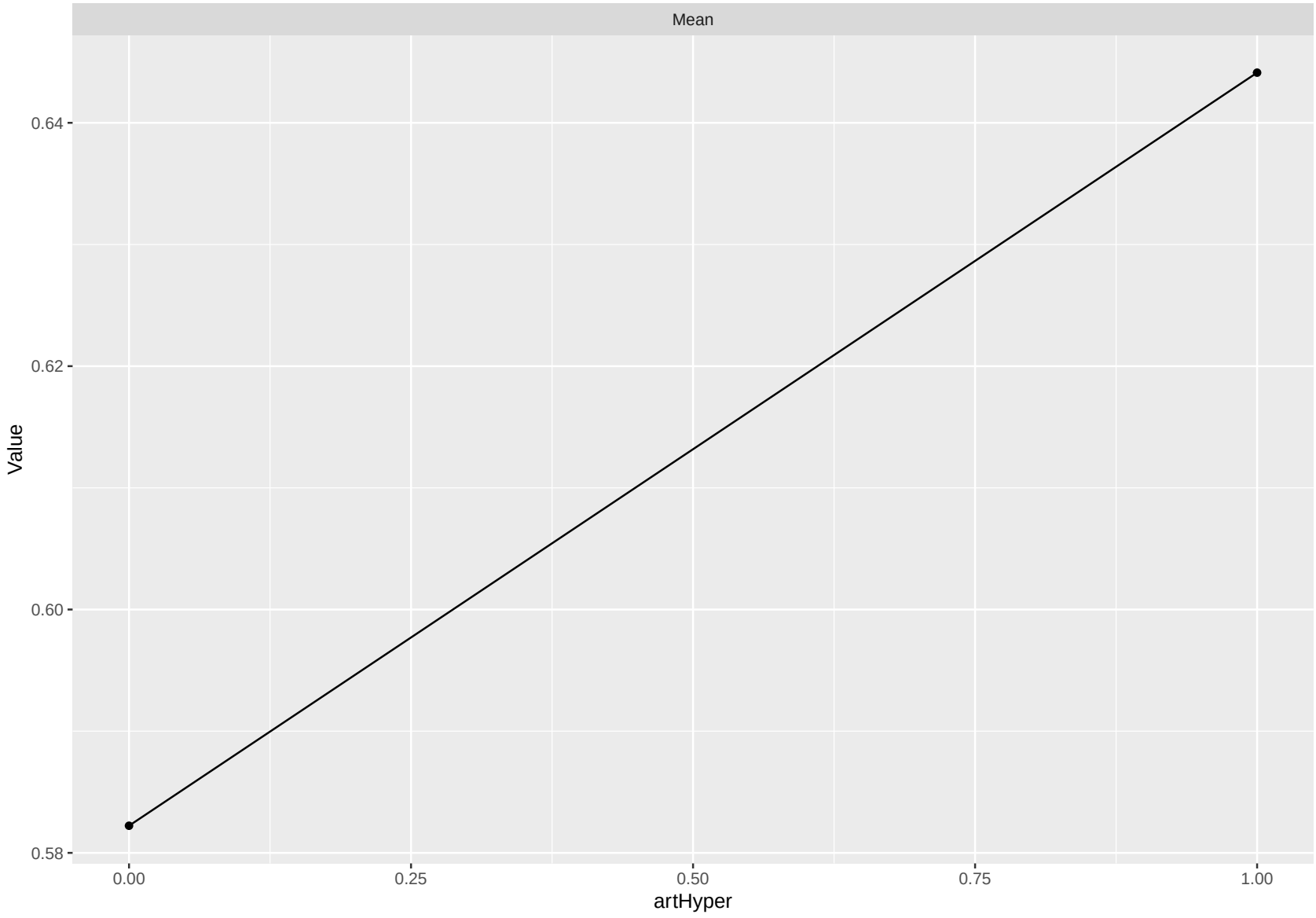
Partial Dependence ascites



Partial Dependence: artHyper



Partial Dependence: symptom



Section 7

Final Model

Final Model

- Final model was chosen by the model with the highest Sensitivity - 2SD

Section 8

References

References I

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bache, Stefan Milton, and Hadley Wickham. 2020. *Magrittr: A Forward-Pipe Operator for r*. https://CRAN.R-project.org/package=magrittr.

Corporation, Microsoft, and Stephen Weston. 2022a. *doSNOW: Foreach Parallel Adaptor for the 'Snow' Package*. https://CRAN.R-project.org/package=doSNOW.

Corporation, Microsoft, and Steve Weston. 2022b. *doParallel: Foreach Parallel Adaptor for the 'Parallel' Package*. https://CRAN.R-project.org/package=doParallel.

References II

Heinzen, Ethan, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson, and Gregory Dougherty. 2021. *Arsenal: An Arsenal of 'r' Functions for Large-Scale Statistical Summaries*. https://CRAN.R-project.org/package=arsenal.

Kuhn, Max, and Hadley Wickham. 2022. *Recipes: Preprocessing and Feature Engineering Steps for Modeling*. https://CRAN.R-project.org/package=recipes.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

References III

Santos, Miriam Seoane, Pedro Henriques Abreu, Pedro J. García-Laencina, Adélia Simão, and Armando Carvalho. 2015. "A New Cluster-Based Oversampling Method for Improving Survival Prediction of Hepatocellular Carcinoma Patients." *Journal of Biomedical Informatics* 58: 49–59. https://doi.org/https://doi.org/10.1016/j.jbi.2015.09.012.

Smith, Brian J. 2021. *MachineShop: Machine Learning Models and Tools*. https://cran.r-project.org/package=MachineShop.

Therneau, Terry M. 2021. *A Package for Survival Analysis in r*. https://CRAN.R-project.org/package=survival.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

References IV

Wickham, Hadley, and Maximilian Girlich. 2022. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.