Predicting Presence of Heart Disease

Sanjeev Singh

6/10/2020

Abstract

Heart diseases or cardiovascular disease (CVD) are widely known, and many people suffer from it. CVD includes Angina causing chest pain, Heart Attack, Congenital Heart Disease, and several others. A few of the underlying factors that relate to heart diseases are lack of exercise, obesity, high blood cholesterol, high blood pressure, and others. Knowing that it is the leading cause of death across the world, we took this study to explore what significant factors we can relate to the presence of heart disease. Getting early signals of the presence of CVD can help save lives.

In our study, we investigated several factors associated with CVD and narrowed down to critical ones that helped detect the presence of the disease. To eliminate features, we used Bayesian methodology and used Laplace Prior that makes non-essential factor effect to zero. Our final model shows that out of 18 attributes, Sex, High blood pressure in women, asymptomatic chest pain (Angina), and Irreversible Thalassemia were significant predictors of CVD. Being a simpler model with less variable, our generalized model performed better than the baseline model on the test data set with an AUC score of 86%.

Keywords: CVD, Logistic Regression, Bayesian, Laplace Prior, Jags, Variable Selection

Introduction

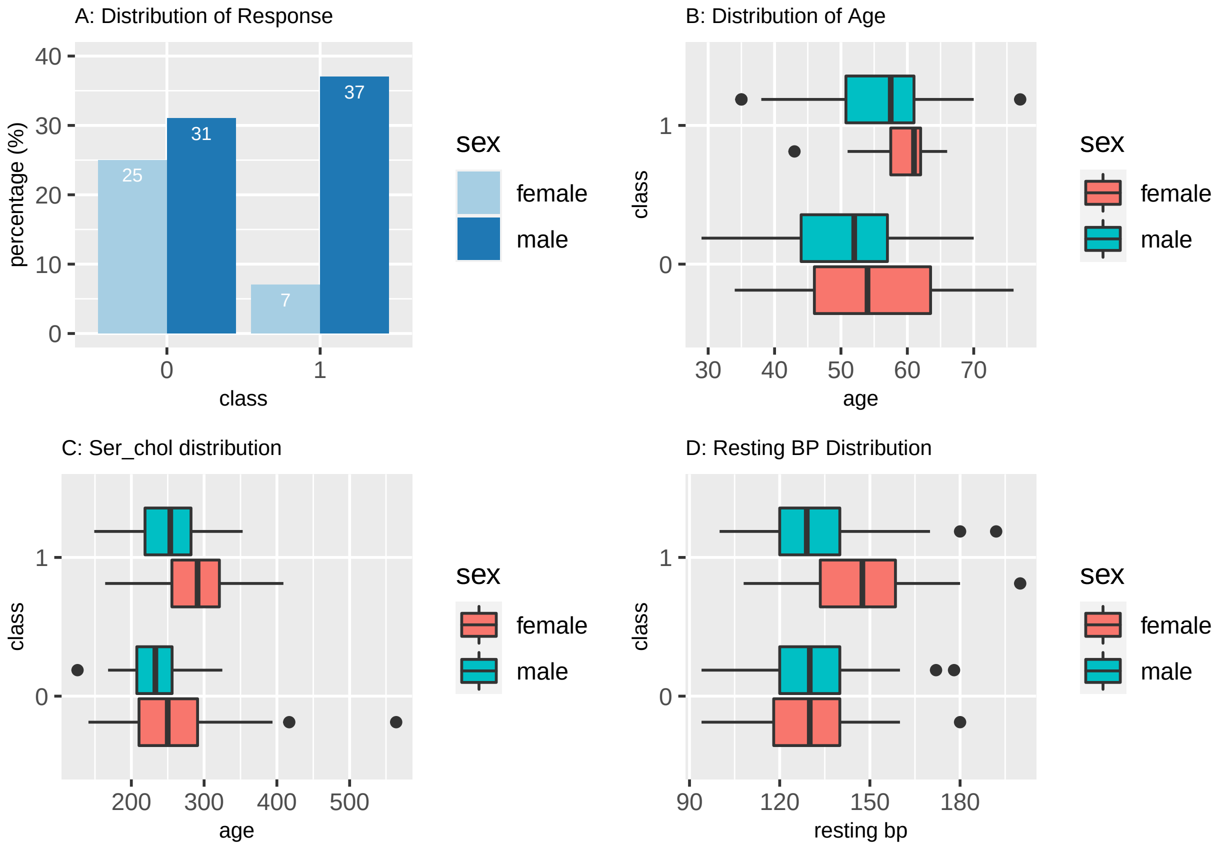
In the present problem, we are trying to narrow down on underlying factors that lead to heart disease. Since CVD constitutes a significant concern across all the countries, this project would help us know better the major factors that can cause this disease. To analyze this problem, we collected a data set available through the UCI ML repository. A major part of this project is related to variable selection; variable selection can be time-consuming if we fit many models carrying the subset of original variables. To help solve this case, we relied on Bayesian statistics, where we can use Laplace-prior to steer non-significant parameters to zero. The present study has the following components, Data: here, we discuss the details of the acquired data and explain the EDA. Next, in the Model section, we discuss data preprocessing and variable selection. In the Result and conclusion section, we discuss the model's performance and conclude on the selected variables strongly related to the response variable.

Table 1: Table carrying attributes used to predict presence of CVD.

Attribute	Definition
age	Age of the person
resting_bp	resting blood pressure (in mm Hg on admission to the hospital)
ser_chol	serum cholestoral in mg/dl
max_hr	maximum heart rate achieved
vessels	number of major vessels (0-3) colored by flourosopy
oldpeak	ST depression induced by exercise relative to rest
sex	Sex (1=Male; 0=Female)

Attribute	Definition
angina	Chest pain type(1=typical angina; 2=atypical angina; 3=non_anginal pain; 4=asymptomatic)
BSgt120	fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
exang	exercise induced angina (1 = yes; 0 = no)
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
slope	the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
y	Absence (1) or presence (2) of heart disease

Data

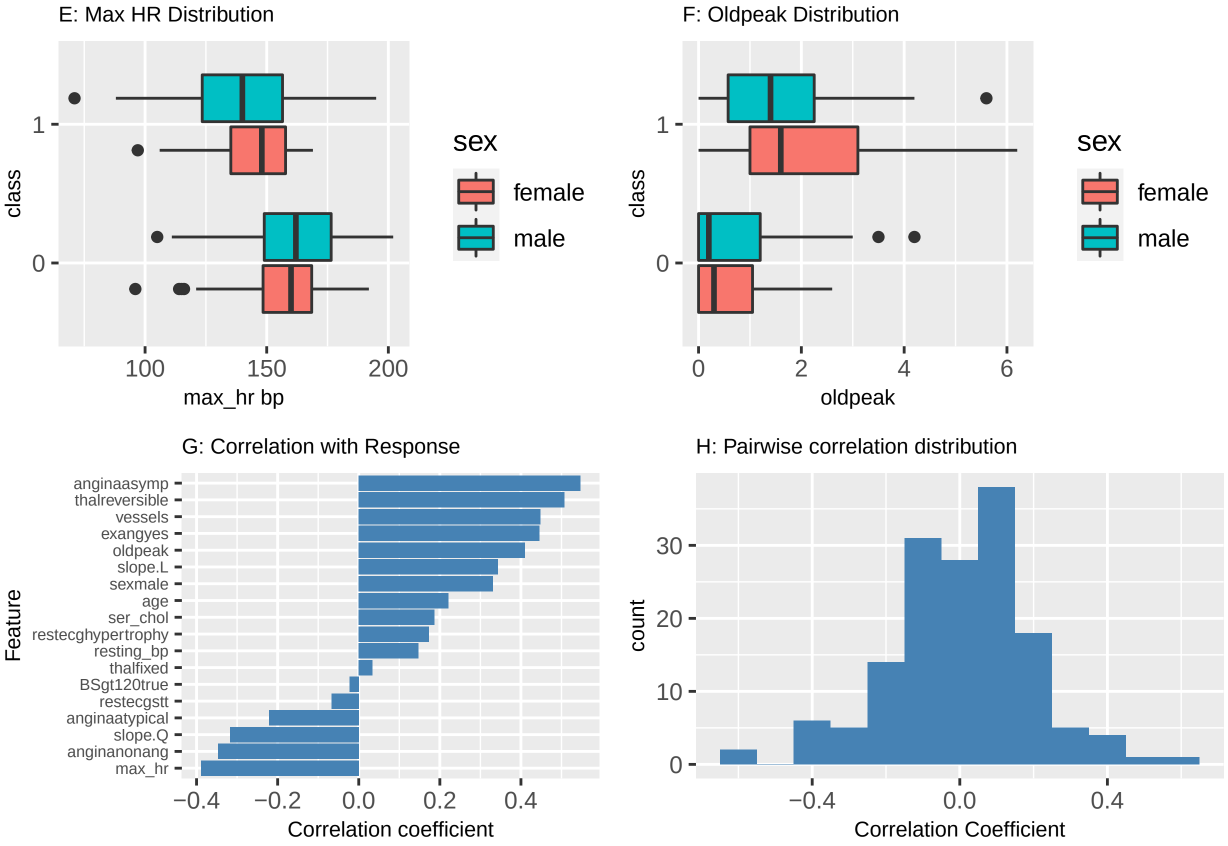


For the study, we collected data from the UCI ML Repository under the name "Statlog (Heart) Data Set." There were no missing values in the data set, and it has 270 instances with 13 attributes that were ordinal, nominal, and real. We can find more details on the attributes in the above table. On encoding the nominal and ordinal variables, we landed with 18 attributes. From the above figure's section A, we found that 44% of data had the presence of CVD: indicating no major skewness in the classes, thus we can use the AUC score for model selection.

On the EDA front, we plotted several boxplots for those real variables where we can see dissimilarities w.r.t. Sex. From above figure B, females show the presence of CVD at a higher avg. Age when compared to males. In figure D, not factoring the Sex, the average resting BP was the same for both the classes. However, on including Sex, females showed a higher average resting BP in the presence of CVD: this shows that there might be an interaction effect between these variables.

In the next chart G and H, we analyzed the correlation between predictors and the response variable. Asymptomatic angina and reversible thalassemia showed a positive correlation, whereas max heart rate and non-angina chest pain were negatively correlated with response. Later, we would see that our final selected variables showed a high correlation with the response.

Next, the pairwise correlation did not show a major correlation between two features. Most of the correlations were near zero, as shown in the picture H. A few with slightly higher correlation were mainly due to factors substituting each other. Thus, we did not see a significant problem with collinearity, which can affect the model inference.



Model

Since the problem in hand is a binary classification task, we selected logistic regression. To start with, we split the data in the train and test set, with 75% of it in the training set. Next, we centered the training data and used its mean and variance to center the test data set. Our model selection criteria were to select one based on the AUC score on the test data set (acting as a validation set) and choosing a simpler model.

To select the significant variables: First, we begin with a baseline "GLM" model in R. The baseline model was fed with all the covariates (As shown in the figure G) to compute a baseline AUC score on the test data set. Second, we implemented logistic regression with Laplace-prior for variable selection. After creating a JAGS model with three chains, followed by a burn-in of 1e3 samples, we saved 10e3 samples per chain for further analysis. After ensuring the convergence of the respective MCMC chain, the variables were selected after observing the MCMC densplot to see if their distribution is not centered near zero. We selected the following nine variables out of the initial 18:

```
## [1] "sexmale"        "anginaasymp"    "resting_bp"     "ser_chol"
## [5] "exangyes"       "oldpeak"        "slope.Q"        "vessels"
## [9] "thalreversible"
```

Third, we fitted a Logistic Regression with the selected nine covariates and used a non-informative normal-prior. On analyzing the densplots, most of the variables had a posterior probability of greater than 95% for being > or < than zero. Except for one, "exangyes." We dropped it, refit the model, and calculated the respective AUC score and other metrics.

Following is the ordered set of parameters w.r.t. their magnitude:

```
##            int    anginaasymp        sexmale thalreversible        vessels
##     -3.5819260      2.6588080      2.1460118      1.3066241      1.0601635
##        slope.Q       ser_chol        oldpeak     resting_bp
##     -0.9956682      0.8299754      0.6539814      0.2830211
```

At the fourth step, we experimented with several interaction terms and settled on Sex and "resting_bp" that gave good AUC scores on the test data set. Following is the ordered set of parameters w.r.t. their magnitude. Note that the interaction term has a negative coefficient, meaning that a higher resting blood pressure in females makes them more prone to heart disease: aligned with what we observed in the EDA.

```
##                int            sexmale        anginaasymp     thalreversible
##         -4.0867410          2.6441232          2.6253547          1.3344767
##         resting_bp            vessels sexmale_resting_bp            slope.Q
##          1.0796092          1.0487476         -1.0028427         -1.0004802
##           ser_chol            oldpeak
##          0.8298902          0.6986281
```

Results & Conclusion

On analyzing the metrics associated with each model, LR with Laplace prior had the highest AUC on the test data set, but it came at the cost of a complex model with 18 variables. The baseline model performed well on the training set, but it suffered from overfitting (by considering all the 18 variables) as it scored poorly on the test data set. The LR model with the interaction term had a slightly better performance compared to the LR with non-inf. prior; however, the marginal gain with the addition of new interaction term was not significant that could have forced its selection. Thus we decided to select the LR with non-inf by the principle of parsimony.

Table 2: Table carrying attributes used to predict presence of CVD.

Model	AUC.Train	AUC.Test	PRC.Train	PRC.Test	DIC/Penalty	#Covariates
GLM (Baseline)	0.946	0.798	0.940	0.713		18
Logistic Regression with Laplace Prior	0.945	0.874	0.937	0.807	150.9/14.2	18
Logistic Regression with non-informative prior	0.940	0.856	0.930	0.753	144.7/8.78	8
Logistic Regression with non-inf. prior and interaction	0.943	0.860	0.934	0.765	144.5/10.06	9

We settled on a probability threshold of 0.25 that gives an overall accuracy of 77.6% at a conservative False-Negative rate of 8.5%. Below confusion matrix provides more details on the same. See the jittered plot in the appendix for more details.

Confusion Matrix:

```
##        Response
##          0  1
##   FALSE 32  3
##   TRUE  12 20
```

Most of the beta coefficients were positive, indicating how a rise in an associated variable can increase CVD's chances. At average values and in the absence of considered factors, being Male carries a probability of 0.2 for having a heart disease.

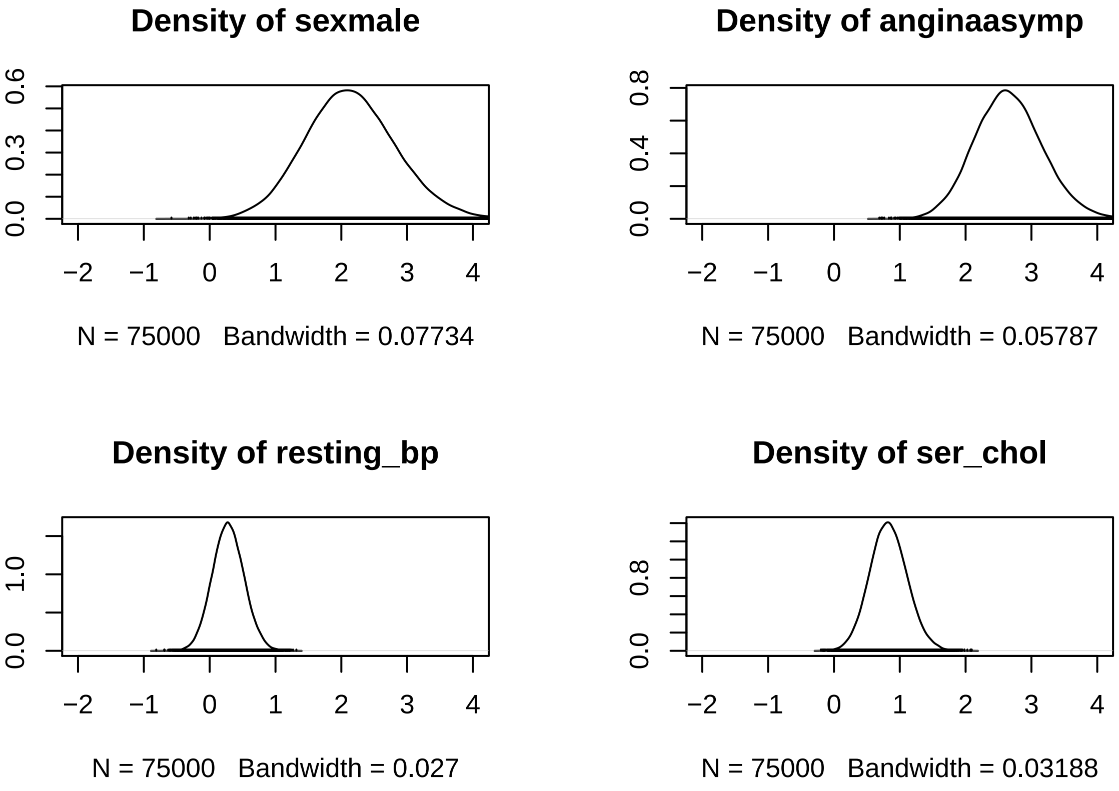
Asymptomatic Angina had the highest importance among all other variables. Considering it in Males, at average values of real variables, it can increase CVD chances by a probability of +0.58(0.77) and by +0.26(0.29) in Females. Thus Asymptomatic Angina alone moved the Males deeper into the CVD territory; Layering in Reversible Thalassemia, in addition to Asymptomatic Angina, for Males, it runs the probability to 0.92, for Females it goes up to 0.59. After we add "vessels," with more than one colored vessel, the Female probability reaches to 0.81.

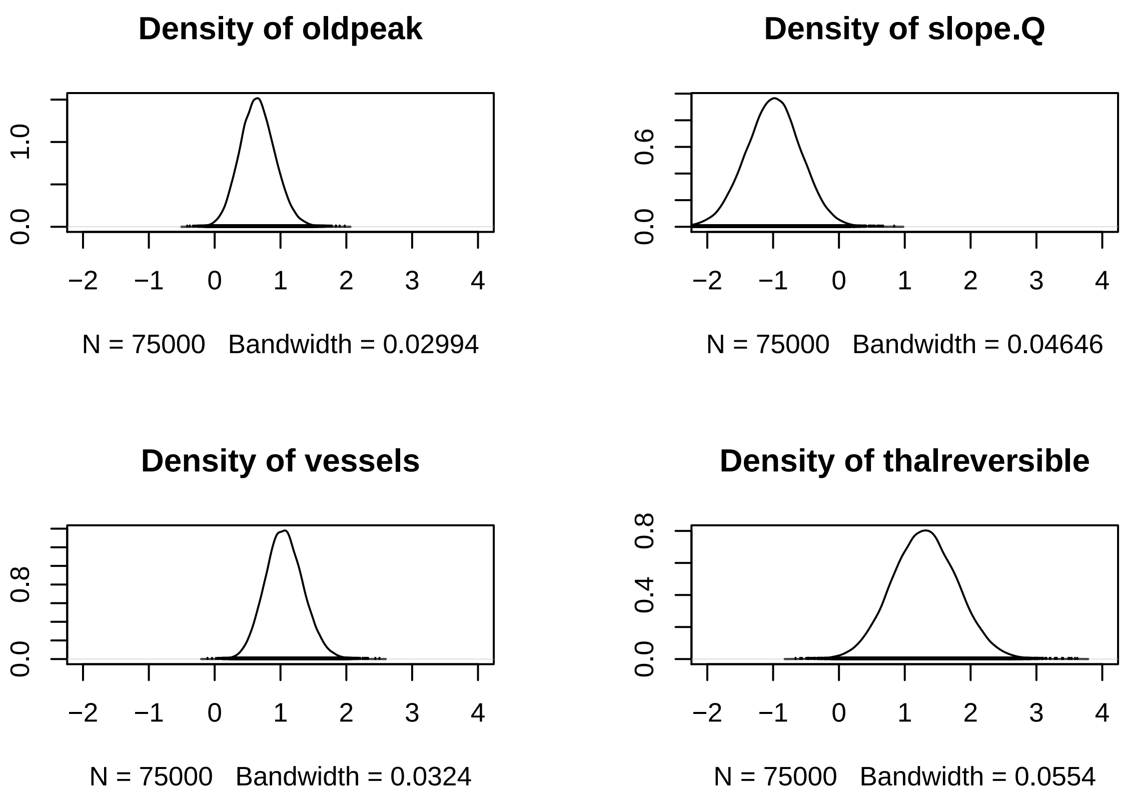
Thus in Male, at average values, Asymptomatic Angina alone can indicate the presence of CVD with relatively fewer chances of False Positives. And in Females, it'll move them far into CVD zone, if we also observe Reversible Thalassemia and colored vessels. The above is one way of interpreting the variables; there can be several other combinations that can lead to CVD's presence.

In the end, we used our test data as a validation set. A standard approach could have been using CV for model selection and selecting threshold probability. But given the scope of this project work, we didn't include it. For future work, we can use different prior other than the non-informative one, transform variables to see if we can get better results. Also, we could fit sophisticated models like XGBoost (boosted tree-based algorithm) and compare our results.

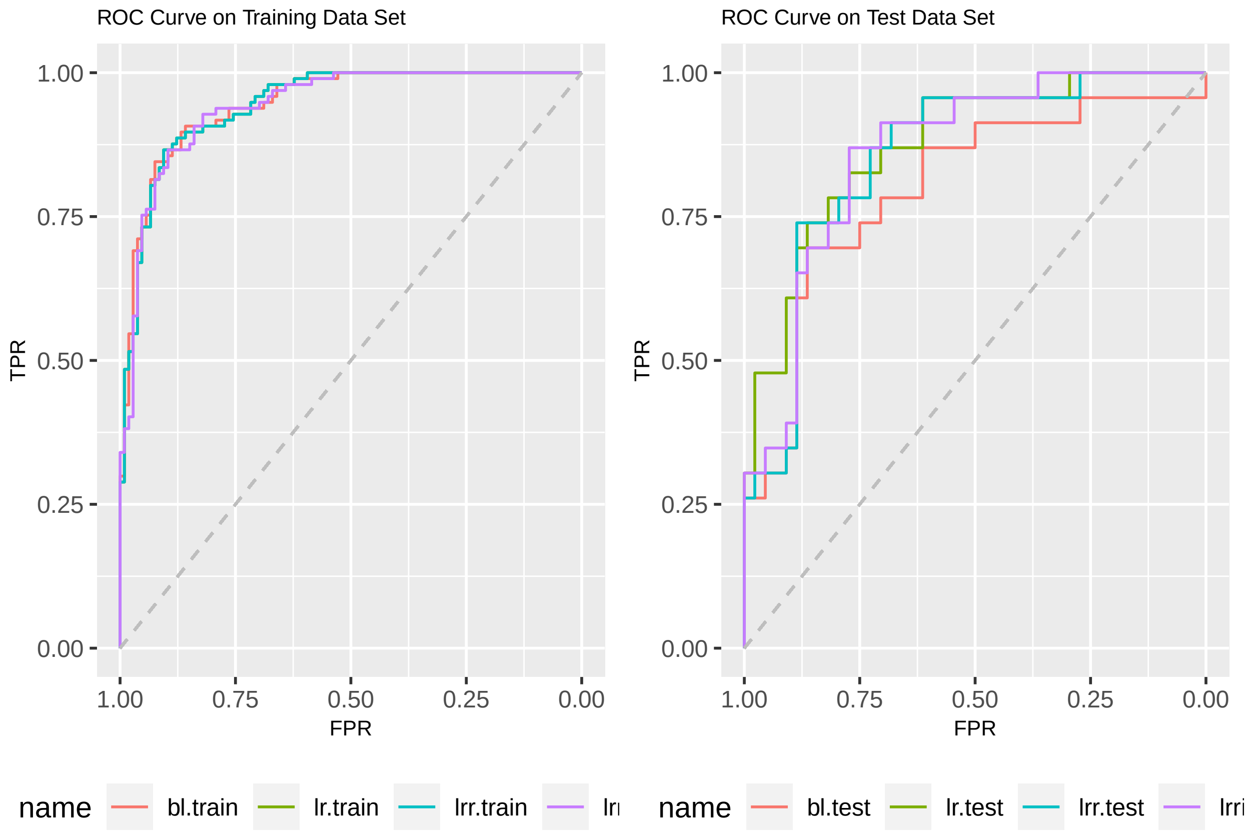
Appendix

1. Densplot of beta coefficient for the Logistic Regression model without an Interaction term.

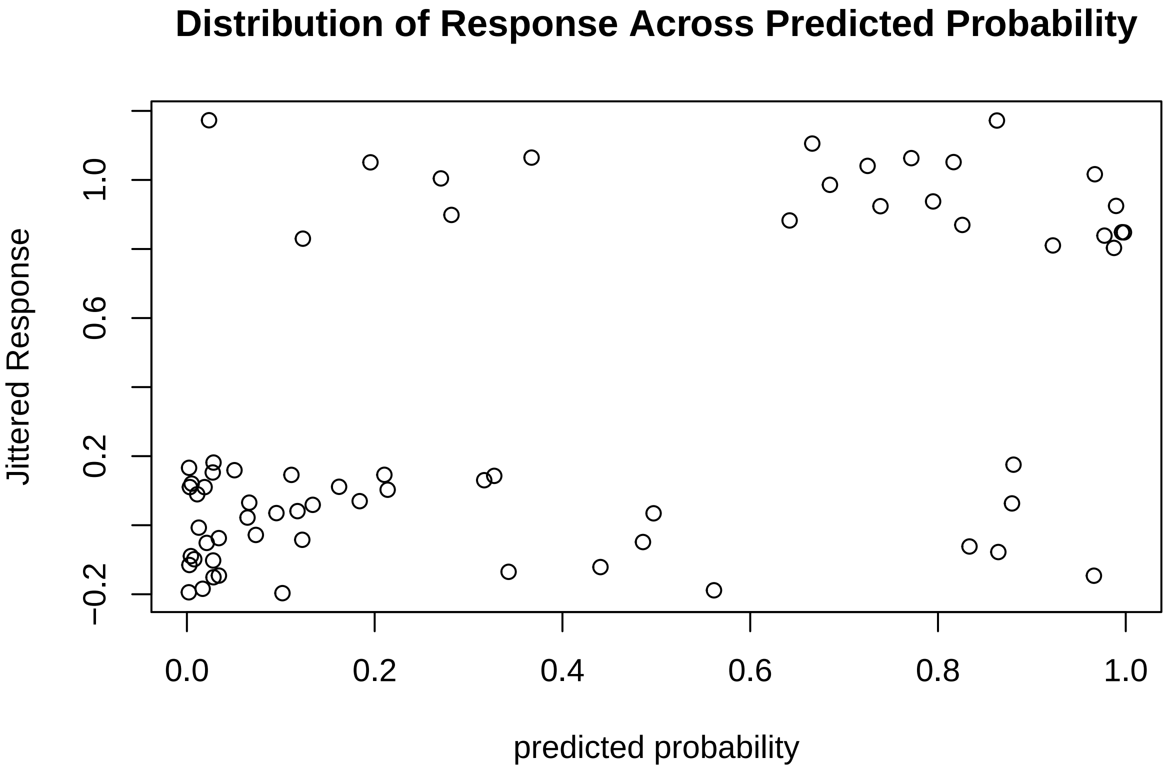




2. AUC computed on training and test data from the four models



3. Response w.r.t. predicted probabilities.


Distribution of Response Across Predicted Probability
Jittered Response
predicted probability