# Comparing Bayesian vs Frequentist Predictive Performance

*Ben Buzzee*

*December 7, 2020*

## Abstract

Supervised learning is a branch of machine learning where an algorithm is trained on a dataset containing known outcomes then used to predict future events where the predictors are known but the outcomes are not. A classic dataset used to practice supervised learning is the Titanic dataset, provided by Kaggle. This dataset provides the outcome survival (0 or 1) and nine potential predictor variables for each passenger aboard the Titanic. The whole dataset is spit into two subsets, one for training the model and one for testing the model. In this project we will compare a bayesian logistic regression that performs model selection via DIC to a standard frequentist logistic regression that performs model selection via AIC. The measure of interest will be the accuracy of the final model (proportion of correct survival predictions on the test dataset).

## Data

The source of the data for this project are the records of the passengers of the famous Titanic voyage. Of the 2224 Passengers that embarked on the journey, only 722 survived. Let's first read in our data and take a look at what we've got:

```
test <- read.csv(file = "test.csv")
train <- read.csv("train.csv")

vars <- names(train)
desc <- c("Unique Identifier", "Outcome - 0 or 1", "Ticket Class",
          "Name", "Gender", "Age", "Num Siblings/Spouses Aboard",
          "Num Parents/Children Aboard", "Ticket Number", "Fare",
          "Cabin Number", "Point Embarked From")

knitr::kable(data.frame(Variable = vars, Description = desc))
```

| Variable | Description |
|---|---|
| PassengerId | Unique Identifier |
| Survived | Outcome - 0 or 1 |
| Pclass | Ticket Class |
| Name | Name |
| Sex | Gender |
| Age | Age |
| SibSp | Num Siblings/Spouses Aboard |
| Parch | Num Parents/Children Aboard |
| Ticket | Ticket Number |
| Fare | Fare |
| Cabin | Cabin Number |
| Embarked | Point Embarked From |

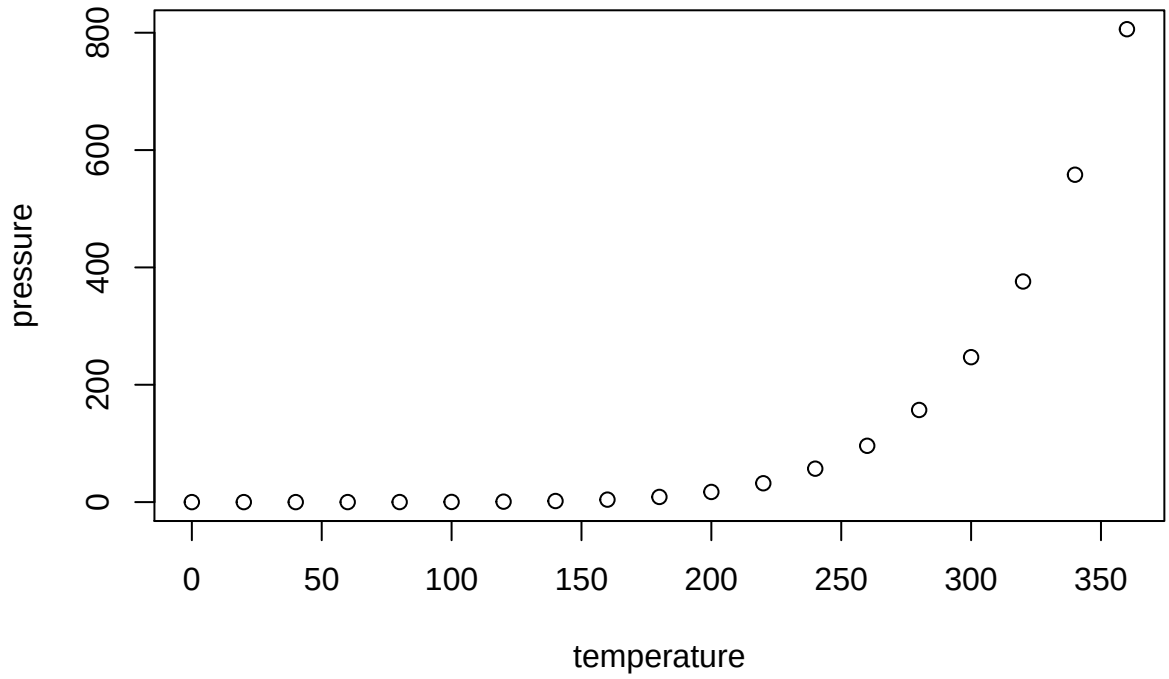
We should also note that we have 891 observations in our training dataset, and we need to use that to predict the survival of the 418 cases in our test dataset.

```
dim(test)
```

```
## [1] 418  11
```

## Including Plots

You can also embed plots, for example:




Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.