Titanic: Machine Learning from disaster

Texas Tech University |

Midterm Project Report

Vivekanand Praturi

[Year]

Contents

1.Introduction

1.1 Problem statement

2.Literature

3.Feature Engineering and Imputing missing Values

4.Data Visualization

5.Application of multivariate tools and techniques

6.Results

Titanic: Machine Learning from Disaster

**1.Introduction**

* 1. **Problem Statement:**

On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. The objective of this task is to complete the analysis of what sorts of people were likely to survive in the sinking of the RMS Titanic is one of the most infamous shipwrecks in history. The need to address this problem is important to avoid such massive tragedies and ensure safety regulations for ships.

**2. Literature**

1. **Exploring the Titanic Dataset**

[**https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic/notebook**](https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic/notebook)

1. **Titanic: Getting Started With R**

[**http://trevorstephens.com/post/72916401642/titanic-getting-started-with-r**](http://trevorstephens.com/post/72916401642/titanic-getting-started-with-r)

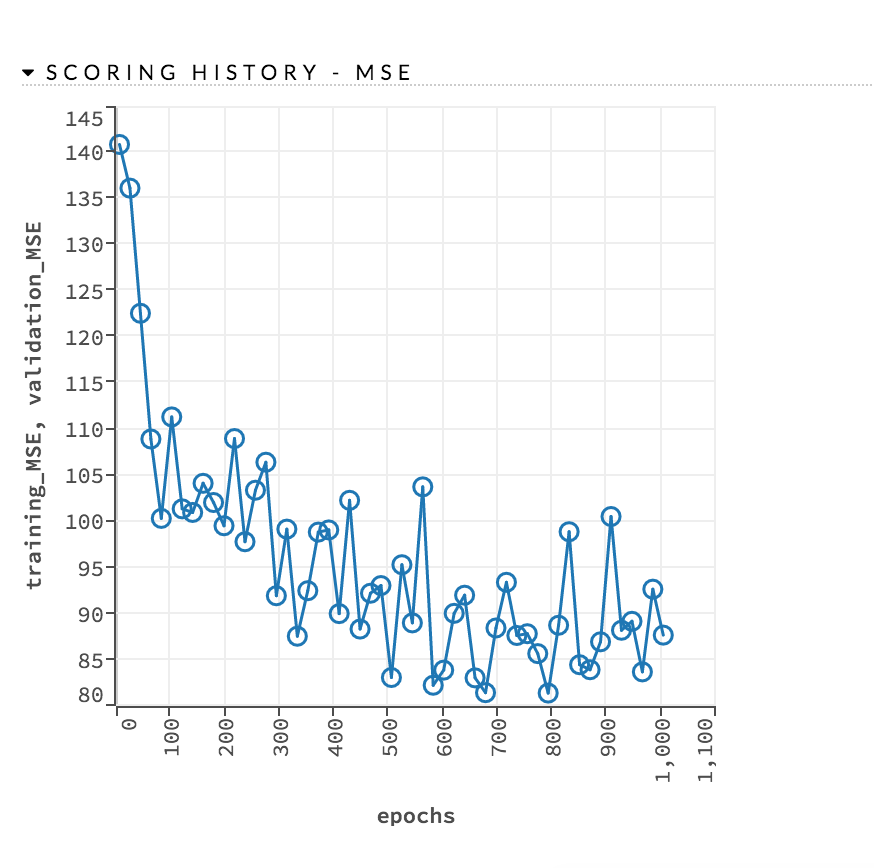
1. **R Libraries**

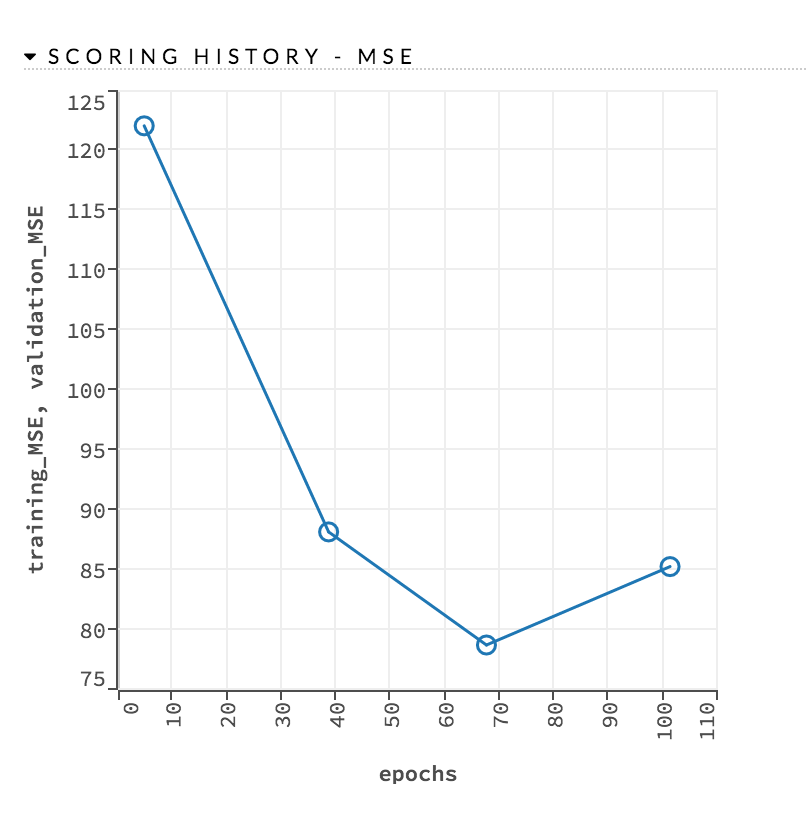
[**library(h2o)**](https://cran.r-project.org/web/packages/h2o/index.html)**,** [**library(randomForest)**](https://cran.r-project.org/web/packages/randomForest/index.html)**,** [**library(plyr)**](https://cran.r-project.org/web/packages/plyr/index.html)**,** [**library(kernlab)**](https://cran.r-project.org/web/packages/kernlab/index.html)**,** [**library(gbm)**](https://cran.r-project.org/web/packages/gbm/index.html)**,**

[**library(nnet)**](https://cran.r-project.org/web/packages/nnet/index.html)**,** [**library(dplyr)**](https://cran.r-project.org/web/packages/plyr/index.html)

**4. Learning from data - Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin.** [**http://amlbook.com/**](http://amlbook.com/)

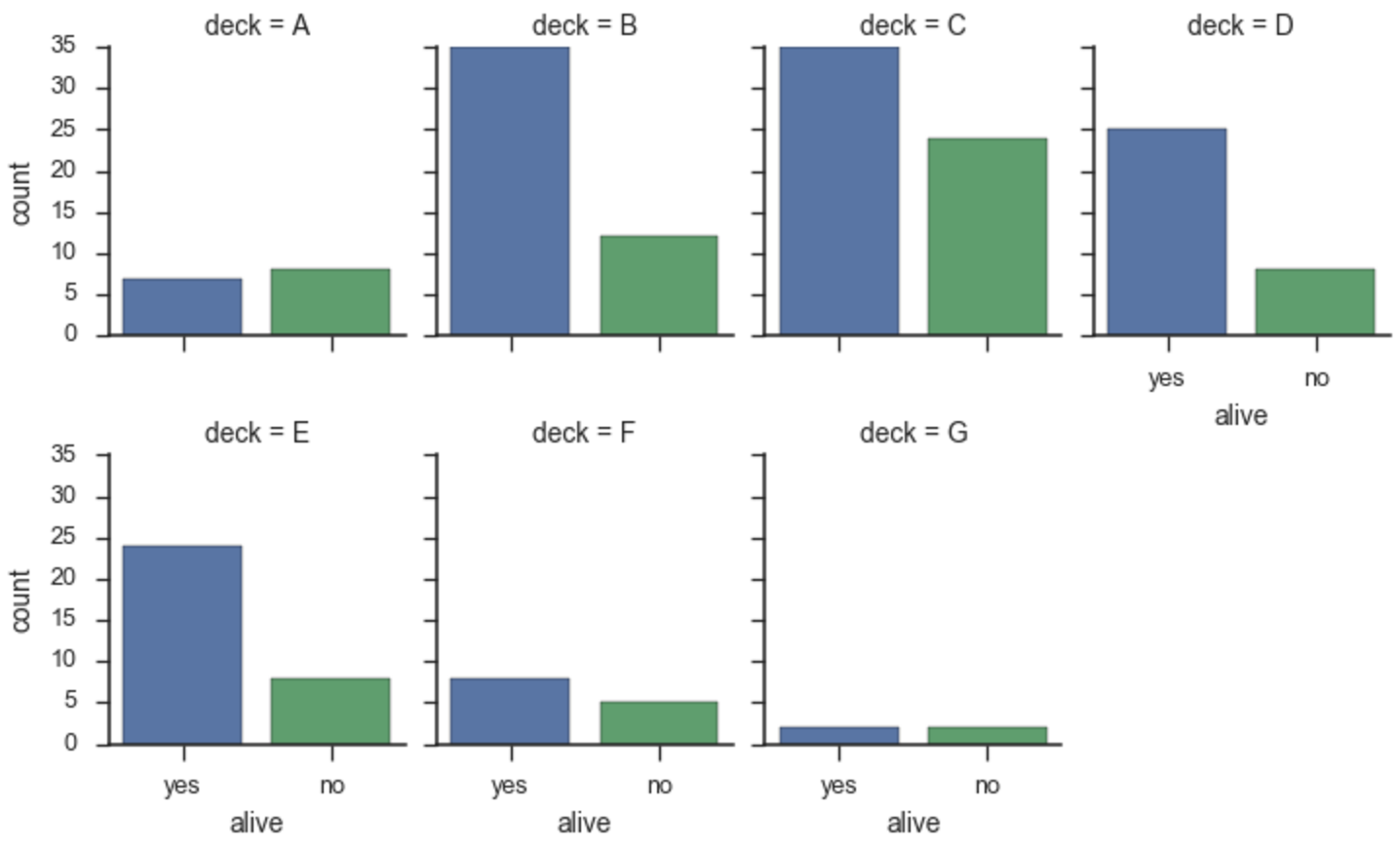
**3. Feature Engineering and imputing missing values**

* Title – From Name
* Family Size – SibSp+Parch+1
* Surname – From Name
* Family Type – Small/Big
* Missing Values
* Age
  + Regression with Rpart, Deep Learning, GLM, Multivariate Imputation by Chained Equations
  + Deep Learning fares badly at regressing Age from available predictor variables
* Fare
  + Replace Na with Median value as the distribution is skewed

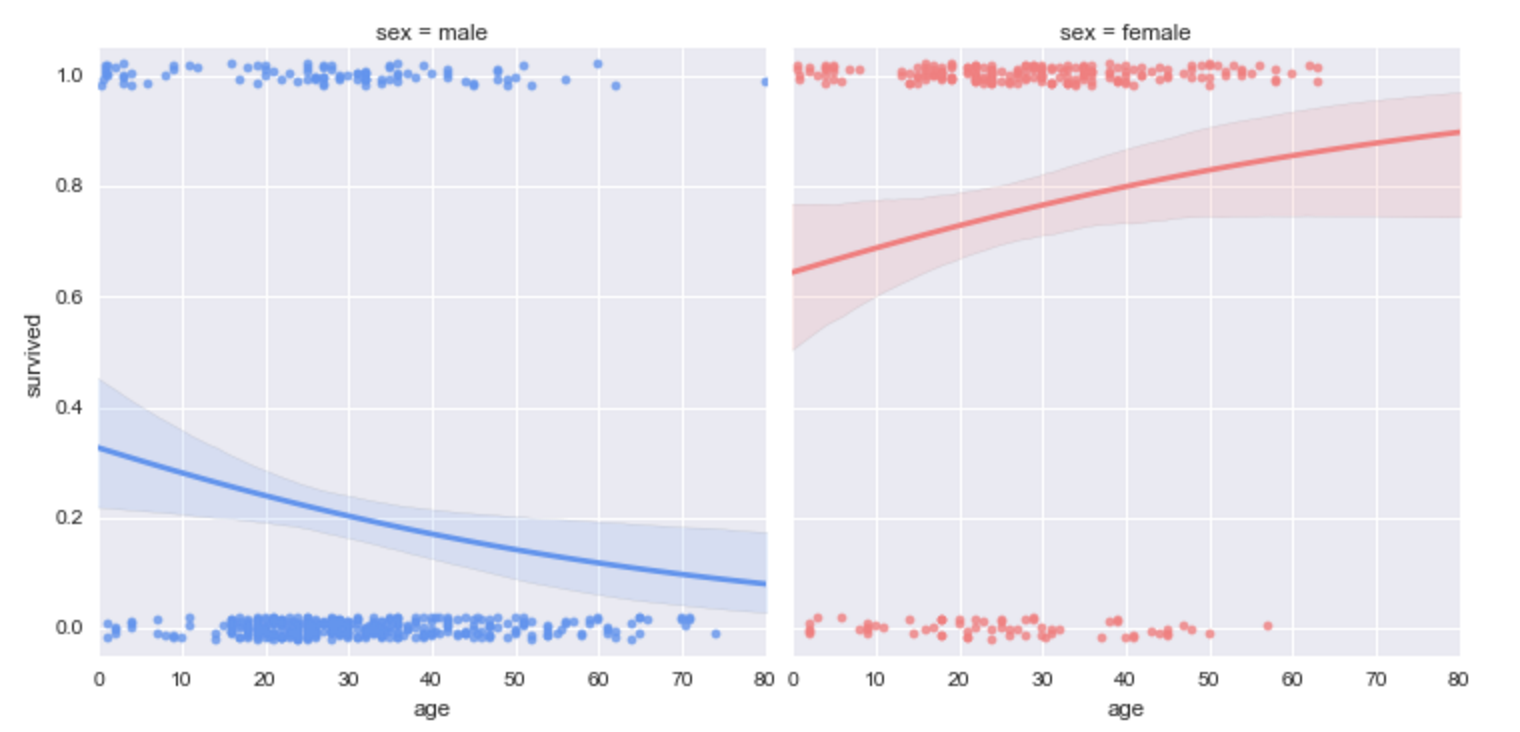


**4.Data Visualization:**

Following are various visualizations of relations between survival and Sex, Survival and Cabin etc.

****

**4.1 Kaggle Submission Result:**

****

**5.Application of Multivariate tools and Techniques:**

* H2O : Deep Learning
* H2O : GBM
* H2O : GLM
* H2O : Random Forest
* Conditional Inference Forest
* Ensemble Methods

Used GBM, GLM, Random Forests as base learners with GBM meta learner.

The models are given random parameters based but there is a room for improvement by choosing the right parameters. So choosing the right parameters for the models is key to get better results.

* Hyper parameter tuning with Grid Search

The model takes the range of parameters and generates all possible models and returns the result for all the models. The best model can be selected by sorting the models based on the Mean square error or Cross Validation error.

And the best models are used in ensembles learning as base learners with GBM as meta learner to get the Kaggle Score of 0.842.

Range of Parameters :

learn\_rate\_opt <- c(0.01, 0.03)

max\_depth\_opt <- c(3, 4, 5, 6, 9)

sample\_rate\_opt <- c(0.7, 0.8, 0.9, 1.0)

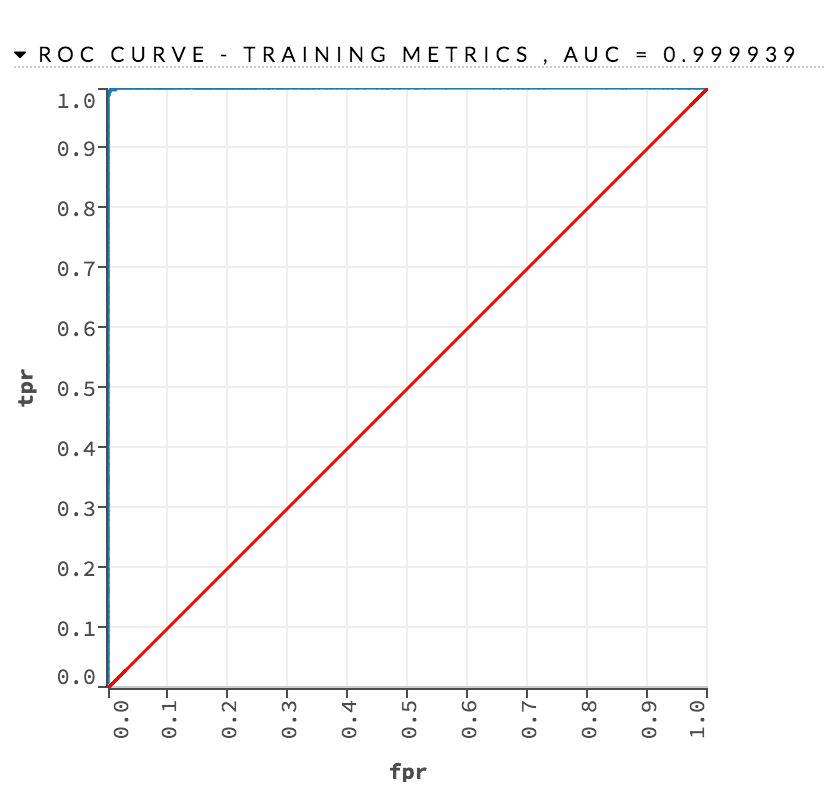
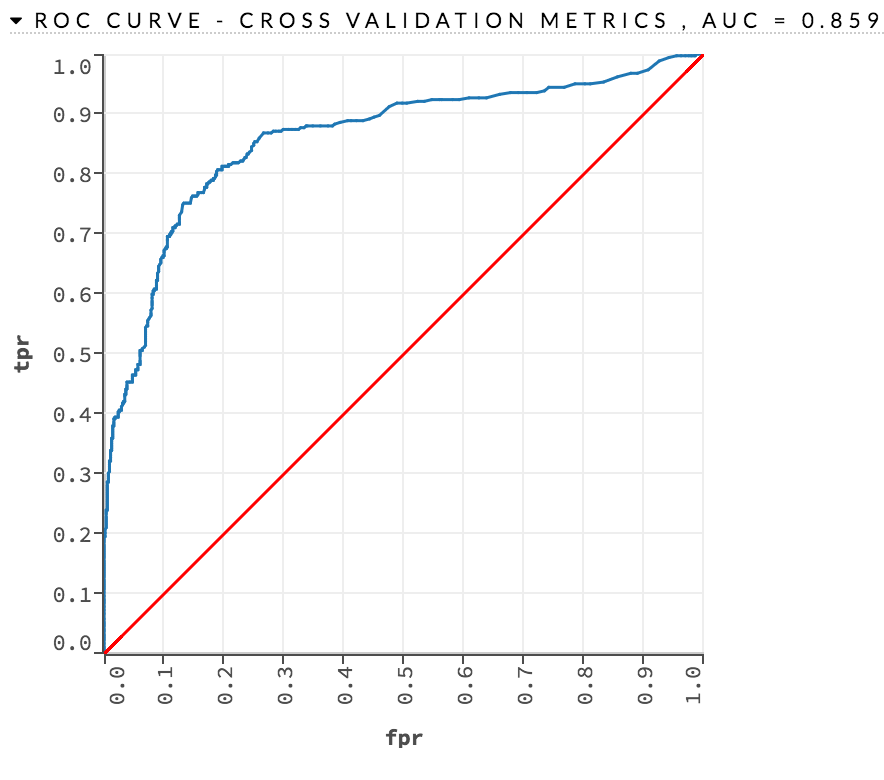
col\_sample\_rate\_opt <- c(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)

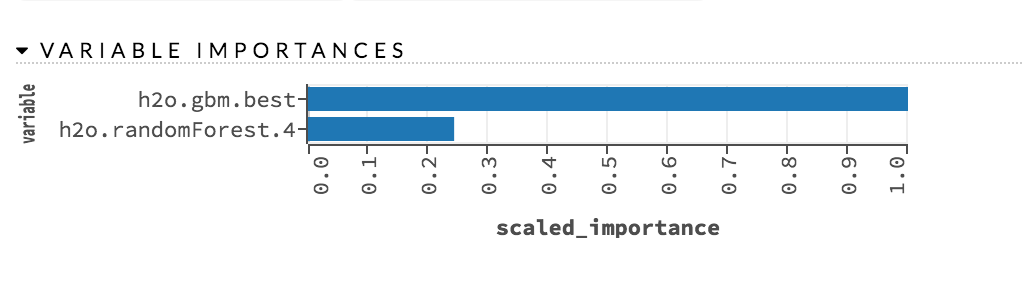
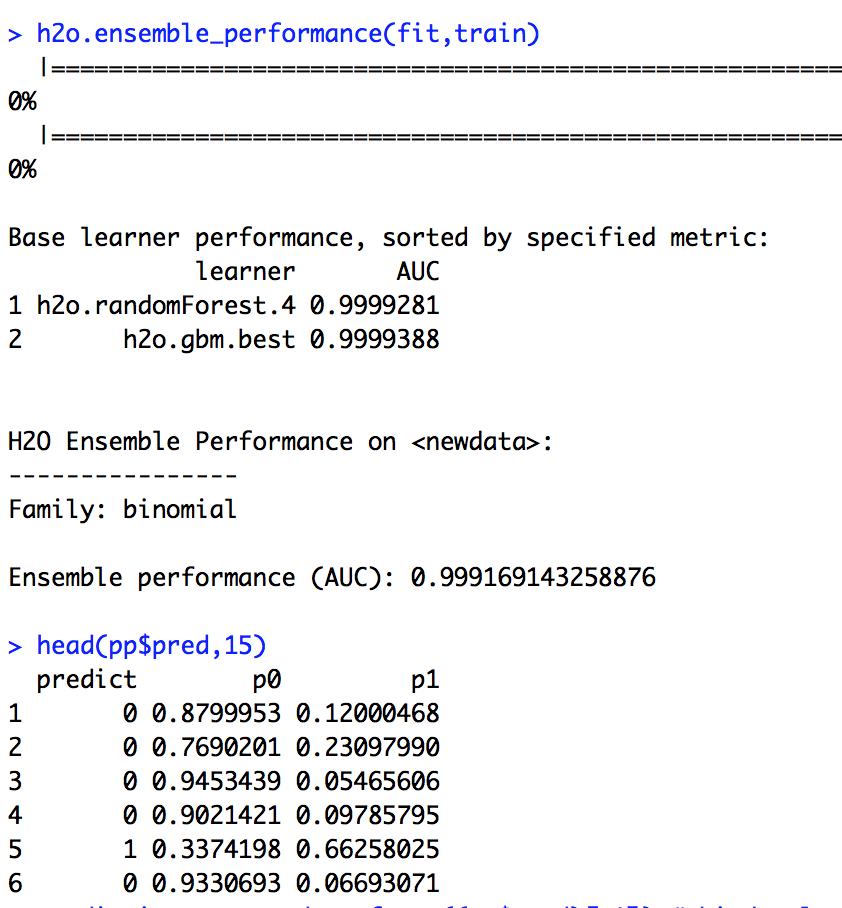
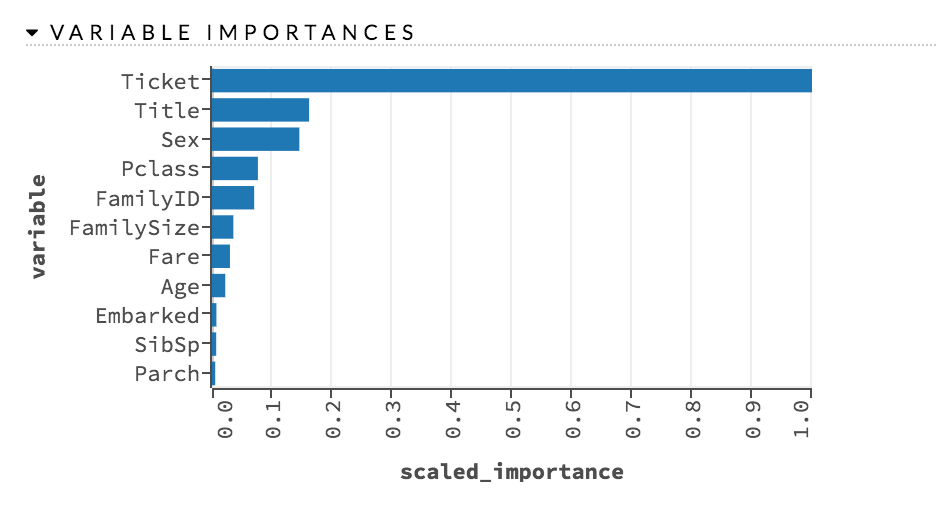
ntrees<-c(100,90,150)

nfolds <- 4

**Results:**

Final Model Performance metrics:





**Rcode:**

