A randomForest Approach to the Kaggle Competition: *Titanic: Machine Learning from Disaster*

Megan Chiovaro[*]

07 May 2019

Abstract

This is a Machine Learning model for the Kaggle Competition: "Titanic: Machine Learning from Disaster". The model was trained to predict survival ("Survived") on the training dataset consisting of 891 observations of 11 features. Missing value imputation was performed using a combination of logical reasoning and Multivariate Imputation by Chained Equations (MICE). Several engineered features were created to improve accuracy. The model was trained using the Machine Learning package randomForest. Predictions were made on the testing dataset consisting of 418 observations and was submitted to Kaggle for evaluation of accuracy. The model achieved 80.04% accuracy in predicting survival of Titanic passengers.

Keywords: Kaggle; Machine Learning; **MICE**; **randomForest**

1 Introduction

This is a proposal for a Machine Learning model for Kaggle's "Titanic: Machine Learning from Disaster" Competition. Training data contained 891 observations of 11 variables:

Variable Name	Description	Levels
Survived	Survived (1) or died (0)	0 = No, 1 = Yes
Pclass	Passenger's class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Passenger's name	
Sex	Passenger's sex	
Age	Passenger's age	
SibSp	Number of siblings/spouses aboard	
Parch	Number of parents/children aboard	
Ticket	Ticket number	
Fare	Fare	
Cabin	Cabin	
Embarked	Port of embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Test data contains 418 observations of all variables with the exception of 'Survived'. The challenge questions the likelihood of survival of passengers across these features. Guidelines request the use of Machine Learning techniques to analyze and create a model predicting 'Survived' on the test set.

This model employs randomForest for its adversion to overfitting, as well as it's easy to use algorithms (Breiman, (2001, Liaw and Wiener (2002)). It is fast to train and quick to make predictions on

[*] megan.chiovaro@uconn.edu; Ph.D. student at Department of Psychological Sciences, University of Connecticut.

small data files such as those used in this competition.

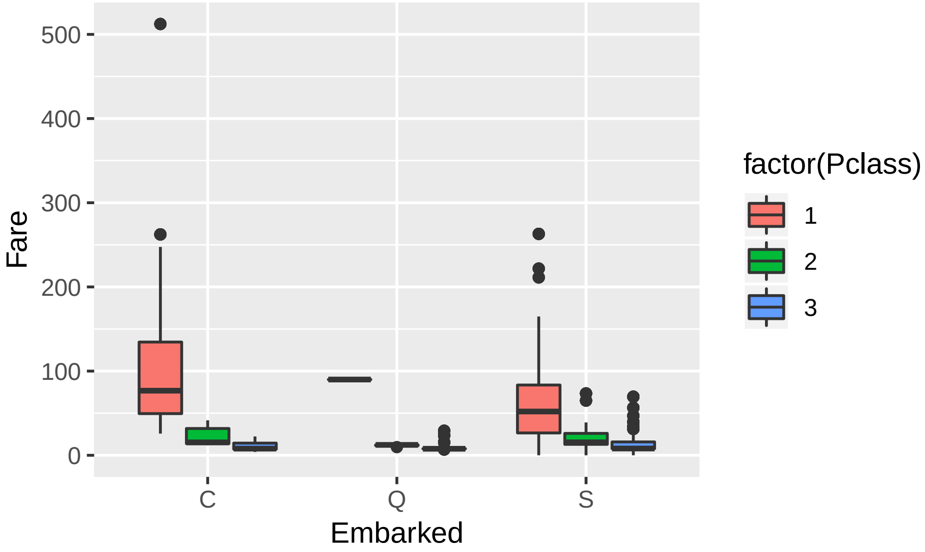
2 Missing Value Imputation

2.1 Embarked

The feature Embarked was missing for two observations.

```
    PassengerId Pclass Fare Embarked
62           62      1   80     <NA>
830         830      1   80     <NA>
```

Features related to Embarked are Fare and Pclass, likely because they both have to do with living location and thus SES. Below is a box plot of these variables.

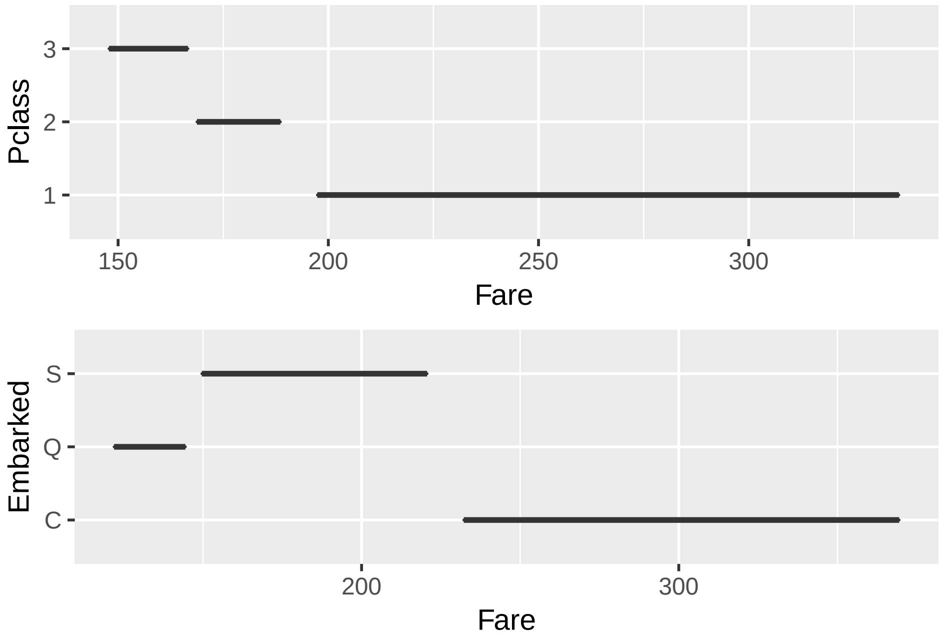


Both observations missing Embarked had a Pclass of 1 and a Fare of 80. Thus for passengers in Pclass 1 with Fare 80, it is clear that location C is the most likely point of Embarkment. Location C was imputed for these missing values.

2.2 Fare

There was one observation (#1044) that was missing a value for Fare. They had Pclass 3 and Embarked S. As demonstrated previously, Fare is highly related to Pclass.

In the figures below, it is obvious that there is clear, non-overlapping price range for individuals in each Pclass and for each Embarked location.



Fare was imputed for observation 1044 using the median of Fares for passengers in Pclass 3 having Embarked from location S.

2.3 Cabin

The feature Cabin was largely unrecorded and was thus unable to be imputed on. For this reason, it was left out of the model.

2.4 Age

The value for Age was missing across 263 observations. Although this is a sizable portion of the data, the variable is logically essential in determining likelihood of survival.

Multivariate Imputation by Chained Equations (*MICE*) imputation was done using the features Pclass, Sex, Age, Fare, Title, SibSp, Parch, Embarked, and Mother (variables Title and Mother were created during feature engineering, see sections 3.3-3.4). All variables used were first transformed into factors to denote the imputation as a classification problem and not a regression problem. Random forest (rf) method was done 30 times to attain good imputations.

3 Feature Engineering

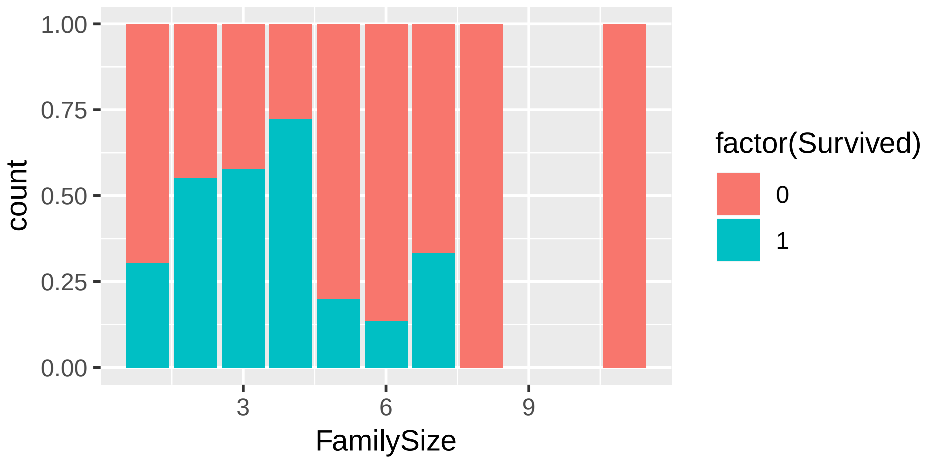
3.1 FamilySize

Using SibSp and Parch, a feature for total family size, FamilySize, was created. This was done by summing the values for SibSp and Parch and then adding one to include that given passenger.

$$FamilySize = SibSp + Parch + 1$$

3.2 FamilyType

Families were further grouped into “Large” (FamilySize > 4), “Medium” (1 < FamilySize <= 4), or “Single” (FamilySize = 1), given the drop in survival for families larger than four and those traveling alone.



This varible is denoted ‘FamilyType’.

3.3 Title

Titles were extracted from the Name feature to get a measure of status for the passengers. Original data contained 18 unique Titles.

```
       Capt Col Don Dona  Dr Jonkheer Lady Major Master Miss Mlle Mme
female    0   0   0    1   1        0    1     0      0  260    2   1
male      1   4   1    0   7        1    0     2     61    0    0   0

        Mr Mrs  Ms Rev Sir the Countess
female   0 197   2   0   0            1
male   757   0   0   8   1            0
```

In the early 1900’s, there were multiple Titles to denote a married or unmarried woman. They were: Mlle, Miss, and Ms for an unmarried woman, and Mme and Mrs for a married woman. These Titles were replaced with Miss and Mrs respectively for simplicity.

Some Titles appeared for a minute number of passengers. Titles used for 8 or fewer passengers were replaced with “Rare Title”, as they could not be used for prediction. Having a “Rare Title” is also an indicator of specialized jobs or work positions that may affect survival probability.

3.4 Mother

Another new feature, Mother, was created and evaluated as ‘yes’ or ‘no’ across observations. A passenger was given the label Mother if they had an adult, feminine Title (“Mrs”, “the Countess”, “Dona”, “Lady”) and Parch was greater than zero.

3.5 Surname

Another feature extracted from the Name feature was Surname. Often used to evaluate ethnicity, last name plays a roll in survival, likely due to ethnic biases amoung the lifeboat workers of the ship.

3.6 Shared_ticket

Some passengers shared their tickets. This was evaluated based on those whom had the same Ticket number. This feature helps identify groups of people who were together on the Titanic, but did not necessarily have the same Surname, including groups of friends and unmarried couples or fiances.

This feature may appear redundant to family membership, but there were many groups of people who shared tickets but did not share family membership. Adding this feature did increase accuracy of the model.

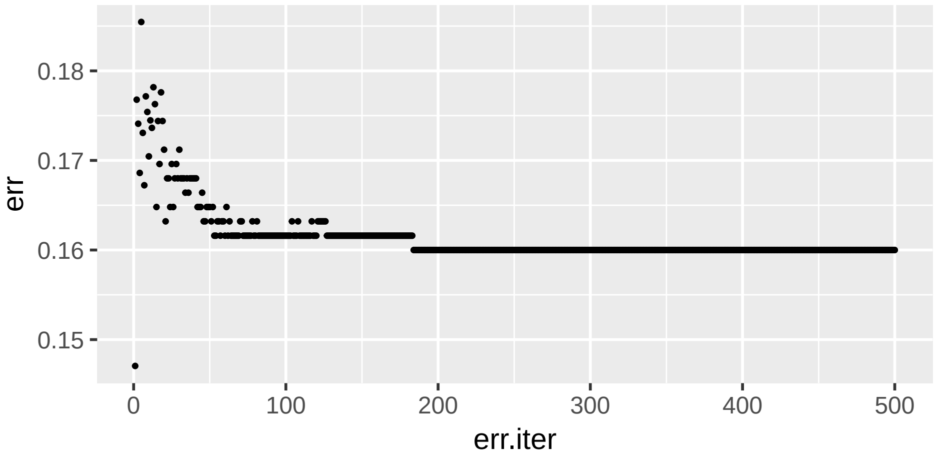
4 The Model

This machine learning model employs *randomForest* to predict Survived based on the features: Pclass, Sex, Age, SibSp, Parch, Fare, Embarked, Title, FamilyType, Shared_ticket, Mother, and MissingAge. The logic for the inclusion of these features is as follows:

- Pclass: Passenger class may have an impact on percieved importance by boat workers, thus giving them higher priority for getting onto a lifeboat.
- Sex: Women were prioritized in getting on lifeboats
- Age: Young children and the elderly would need assistance in getting off the ship, thus potentially decreasing their chances of survival.
- SibSp and Parch: Family relationships lead to heroic acts and sacrifice.
- Fare: A passengers Fare is also an indicator of their status and may also have an impact on their percieved importance by boat workers.
- Title: A passengers title is also an indicator of their status.
- FamilyType: There is a clear survival benefit to being in a medium sized family.
- Shared_ticket: This helps identify groups and thus acts similarly to FamilyType for those traveling with groups of friends.
- Mother: Mothers were prioritized in getting on the life boats.
- MissingAge: Those whose age was missing were commonly ship workers and lower class individuals, who had low priority for getting a spot on a life boat.

5 Results

The model was run on the training set, predicting 418 witheld patient survival results. Running 500 trees, the model converged nicely to a 16% error rate.



The .csv file containing only PassengerId and Survived was submitted to Kaggle for evaluation of Accuracy. Results showed 80.04% accuracy in predicting survival. This was improved from the 78.5% accuracy of the author's previous version of the model.

6 Summary and Discussion

This machine learning model aimed to predict survival of passengers on the Titanic. As per Kaggle's evaluation, the model achieved over 80% accuracy. This placed in the top 10% of competing models on the Kaggle Leaderboard page. Future improvements to the model should include an analysis of ethnicity to account for biases at the time. Identification of young children traveling with their Mothers could also likely improve accuracy, as they were also prioritized in getting onto a lifeboat with Mothers.

Acknowledgment

The author would like to thank Jun Yan and all students in the Data Science in Action course for their feedback and support in this learning process.

References

Breiman, L. ((2001)), "Random Forests," *Machine Learning*, 45, 5–32.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by RandomForest," Tech. rep.