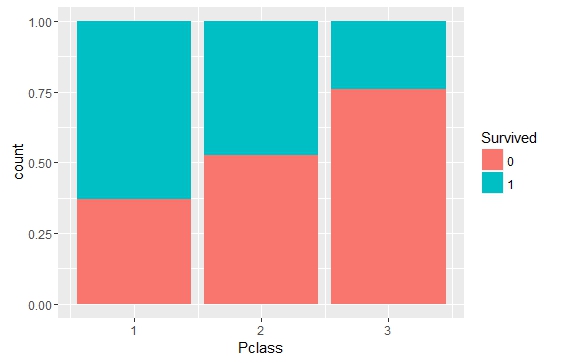
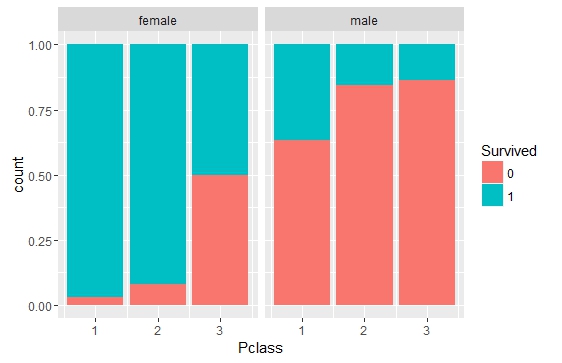


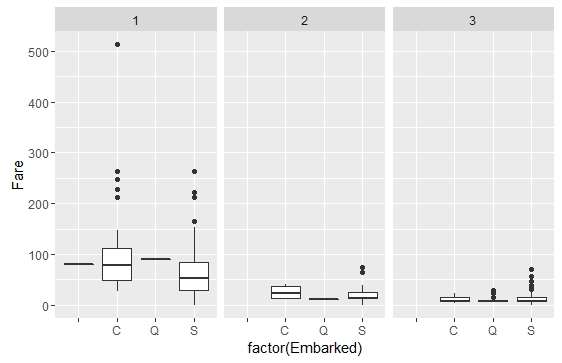
Females in the C category have the best survival rate. Males of this category have the worst rate.



Clearly Pclass affects survival rate



Pclass has the same impact amongst both males and females.



The missing Embarked data is in the first column. We can see that the fare paid by them i.e 80$ is the median fare paid by 1st class passenger embarking from C. So we will update the missing Embark values with C

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 | 2 | 3 | 4 | 5 |
| Error on train | 13.73 | 11.6 | 11.73 | 10.67 |  |
| Error on test | 14.18 | 13.475 | 13.475 | 14.18 |  |
| Kaggle Score | 0.76555 | 0.75598 | 0.76555 |  |  |

Model 1 : Train data divided into train and test. C5.0 Model with Zero Boost

predict\_model <- C5.0(trainingdata,trainingdata\_survived,trials=1) #Not using boosting

(a) (b) <-classified as

---- ----

438 21 (a): class 0

82 209 (b): class 1

13.73% error on self testing on training data set.

Total Observations in Table: 141

| test\_predicted\_survived

testdata$Survived | 0 | 1 | Row Total |

------------------|-----------|-----------|-----------|

0 | 84 | 6 | 90 |

| 7.353 | 16.758 | |

| 0.933 | 0.067 | 0.638 |

| 0.857 | 0.140 | |

| 0.596 | 0.043 | |

------------------|-----------|-----------|-----------|

1 | 14 | 37 | 51 |

| 12.976 | 29.574 | |

| 0.275 | 0.725 | 0.362 |

| 0.143 | 0.860 | |

| 0.099 | 0.262 | |

------------------|-----------|-----------|-----------|

Column Total | 98 | 43 | 141 |

| 0.695 | 0.305 | |

------------------|-----------|-----------|-----------|

20/141 Error rate = 14%

Model 2: Train data divided into train and test. C5.0 Model with Zero Boost

(a) (b) <-classified as

---- ----

431 28 (a): class 0

59 232 (b): class 1

11.6% Error rate on self test on the 750 rows of training data set

Total Observations in Table: 141

| test\_predicted\_survived

testdata$Survived | 0 | 1 | Row Total |

------------------|-----------|-----------|-----------|

0 | 80 | 10 | 90 |

| 9.468 | 16.204 | |

| 0.889 | 0.111 | 0.638 |

| 0.899 | 0.192 | |

| 0.567 | 0.071 | |

------------------|-----------|-----------|-----------|

1 | 9 | 42 | 51 |

| 16.708 | 28.596 | |

| 0.176 | 0.824 | 0.362 |

| 0.101 | 0.808 | |

| 0.064 | 0.298 | |

------------------|-----------|-----------|-----------|

Column Total | 89 | 52 | 141 |

| 0.631 | 0.369 | |

------------------|-----------|-----------|-----------|

13.475%error rate on the test data

Model 3: Train data divided into train and test. C5.0 Model with Zero Boost. Using AgeCategory instead of Age

(a) (b) <-classified as

---- ----

436 23 (a): class 0

65 226 (b): class 1

11.73% error rate on self testing on traing data

| test\_predicted\_survived

testdata$Survived | 0 | 1 | Row Total |

------------------|-----------|-----------|-----------|

0 | 82 | 8 | 90 |

| 8.633 | 16.727 | |

| 0.911 | 0.089 | 0.638 |

| 0.882 | 0.167 | |

| 0.582 | 0.057 | |

------------------|-----------|-----------|-----------|

1 | 11 | 40 | 51 |

| 15.235 | 29.519 | |

| 0.216 | 0.784 | 0.362 |

| 0.118 | 0.833 | |

| 0.078 | 0.284 | |

------------------|-----------|-----------|-----------|

Column Total | 93 | 48 | 141 |

| 0.660 | 0.340 | |

------------------|-----------|-----------|-----------|

13.475% error rate on test data

Model 4: Train data divided into train and test. C5.0 Model with 10 Boost. Using AgeCategory instead of Age

(a) (b) <-classified as

---- ----

428 31 (a): class 0

49 242 (b): class 1

10.67% error rate on self testing

| test\_predicted\_survived

testdata$Survived | 0 | 1 | Row Total |

------------------|-----------|-----------|-----------|

0 | 78 | 12 | 90 |

| 9.726 | 15.208 | |

| 0.867 | 0.133 | 0.638 |

| 0.907 | 0.218 | |

| 0.553 | 0.085 | |

------------------|-----------|-----------|-----------|

1 | 8 | 43 | 51 |

| 17.164 | 26.838 | |

| 0.157 | 0.843 | 0.362 |

| 0.093 | 0.782 | |

| 0.057 | 0.305 | |

------------------|-----------|-----------|-----------|

Column Total | 86 | 55 | 141 |

| 0.610 | 0.390 | |

------------------|-----------|-----------|-----------|

14.184% error rate