

Space Titanic: Kaggle Prediction Model

```
library(tidyverse)
library(knitr)
library(lubridate)
library(rpart)
library(pROC)
library(rattle)
library(randomForest)
library(ggplot2)
library(simputation)
library(naniar)
```

Pre-Processing

```
train = read.csv("train.csv")
```

```
test = read.csv("test.csv")
```

```
miss_var_summary(train)
```

```
## # A tibble: 14 x 3
##   variable      n_miss pct_miss
##   <chr>         <int>   <num>
## 1 ShoppingMall    208     2.39
## 2 VRDeck          188     2.16
## 3 FoodCourt       183     2.11
## 4 Spa             183     2.11
## 5 RoomService     181     2.08
## 6 Age             179     2.06
## 7 PassengerId      0      0
## 8 HomePlanet       0      0
## 9 CryoSleep        0      0
## 10 Cabin           0      0
## 11 Destination     0      0
## 12 VIP             0      0
## 13 Name            0      0
## 14 Transported     0      0
```

Imputing all NA Values in Missing Variables

```
imp_age = lm(Age~ShoppingMall+VRDeck+FoodCourt+Spa+RoomService+HomePlanet+
             CryoSleep+Destination+VIP, data=train)
summary(imp_age)
```

```
##
## Call:
## lm(formula = Age ~ ShoppingMall + VRDeck + FoodCourt + Spa +
##     RoomService + HomePlanet + CryoSleep + Destination + VIP,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.527  -8.737  -2.495   8.505  55.140
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.8169087   2.0822319   13.839 < 2e-16 ***
## ShoppingMall     0.0002736   0.0002701    1.013 0.311221
## VRDeck           0.0002456   0.0001555    1.580 0.114175
## FoodCourt        0.0002309   0.0001098    2.103 0.035534 *
## Spa              0.0005768   0.0001484    3.886 0.000103 ***
## RoomService      0.0009596   0.0002538    3.782 0.000157 ***
## HomePlanetEarth  -2.8511257   1.0590858   -2.692 0.007117 **
## HomePlanetEuropa  5.3846105   1.1045482    4.875 1.11e-06 ***
## HomePlanetMars    0.2977731   1.0993682    0.271 0.786507
## CryoSleepFalse    1.0911023   1.0302207    1.059 0.289590
## CryoSleepTrue     -0.9412191   1.0494040   -0.897 0.369796
## Destination55 Cancr e -2.8014595   1.1607616   -2.413 0.015825 *
## DestinationPS0 J318.5-22 0.1583797   1.2283327    0.129 0.897410
## DestinationTRAPPIST-1e -1.4021401   1.1161152   -1.256 0.209057
## VIPFalse          0.6368065   1.0467319    0.608 0.542956
## VIPTrue           4.4543859   1.4911027    2.987 0.002823 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.94 on 7604 degrees of freedom
## (1073 observations deleted due to missingness)
## Multiple R-squared:  0.07847,    Adjusted R-squared:  0.07665
## F-statistic: 43.17 on 15 and 7604 DF,  p-value: < 2.2e-16
```

```
train2 = train %>% mutate(Age_pred = predict(imp_age,newdata=train))
```

```
train2=train2 %>%mutate(NewAge = case_when(is.na(Age) ~ round(Age_pred, digits = 0),
                                           TRUE ~ Age))
```

```
train2 = train2 %>% mutate(NewAge_Final = case_when(is.na(NewAge) ~
                                                    round(mean(NewAge, na.rm=T), digits=0)
                                                    , TRUE ~ NewAge))
```

```
imp_spa = lm(Spa~ShoppingMall+VRDeck+FoodCourt+Age+RoomService+HomePlanet+
             CryoSleep+Destination+VIP, data=train)
summary(imp_spa)
```

```
##
## Call:
## lm(formula = Spa ~ ShoppingMall + VRDeck + FoodCourt + Age +
##     RoomService + HomePlanet + CryoSleep + Destination + VIP,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2486.0  -360.8  -229.5   200.5 20538.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.711e+02  1.627e+02   1.052 0.292889
## ShoppingMall   -4.654e-02  2.084e-02  -2.233 0.025564 *
## VRDeck         1.031e-02  1.200e-02   0.859 0.390335
## FoodCourt      5.475e-02  8.455e-03  6.476 1.00e-10 ***
## Age           3.436e+00  8.842e-01   3.886 0.000103 ***
## RoomService   -4.517e-02  1.960e-02  -2.305 0.021180 *
## HomePlanetEarth -1.096e+02  8.177e+01  -1.340 0.180290
## HomePlanetEuropa 5.585e+02  8.514e+01  6.560 5.75e-11 ***
## HomePlanetMars  -6.986e+01  8.484e+01  -0.823 0.410292
## CryoSleepFalse  2.122e+02  7.948e+01  2.670 0.007597 **
## CryoSleepTrue   -3.159e+02  8.091e+01  -3.905 9.52e-05 ***
## Destination55 Cancr i e -4.123e+01  8.962e+01  -0.460 0.645457
## DestinationPS0 J318.5-22 -1.401e+01  9.480e+01  -0.148 0.882509
## DestinationTRAPPIST-1e -7.317e+01  8.614e+01  -0.849 0.395679
## VIPFalse       4.868e+00  8.079e+01   0.060 0.951948
## VIPTrue        -4.867e+01  1.151e+02  -0.423 0.672563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1076 on 7604 degrees of freedom
## (1073 observations deleted due to missingness)
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1304
## F-statistic: 77.17 on 15 and 7604 DF, p-value: < 2.2e-16
```

```
train2 = train2 %>% mutate(Spa_pred = predict(imp_spa,newdata=train2))
```

```
train2=train2 %>%mutate(NewSpa = case_when(is.na(Spa) ~ round(Spa_pred, digits = 0),
                                           TRUE ~ Spa))
```

```
train2 = train2 %>% mutate(NewSpa_Final = case_when(is.na(NewSpa) ~
                                                    round(mean(NewSpa, na.rm=T), digits=0)
                                                    , TRUE ~ NewSpa))
```

```
imp_foodcourt = lm(FoodCourt~ShoppingMall+VRDeck+Age+Spa+RoomService+HomePlanet+
                   CryoSleep+Destination+VIP, data=train)
summary(imp_foodcourt)
```

```
##
## Call:
## lm(formula = FoodCourt ~ ShoppingMall + VRDeck + Age + Spa +
```

```
## RoomService + HomePlanet + CryoSleep + Destination + VIP,
## data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3405.1  -475.2  -235.7   377.4 27241.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    144.99316   220.09389   0.659  0.51006
## ShoppingMall    -0.15087    0.02815  -5.360 8.57e-08 ***
## VRDeck          0.08653    0.01620   5.341 9.54e-08 ***
## Age             2.51651    1.19686   2.103  0.03553 *
## Spa             0.10018    0.01547   6.476 1.00e-10 ***
## RoomService    -0.13211    0.02647  -4.990 6.16e-07 ***
## HomePlanetEarth -144.78427  110.60598  -1.309  0.19057
## HomePlanetEuropa 1163.61866  114.71849  10.143 < 2e-16 ***
## HomePlanetMars   -76.91869  114.76830  -0.670  0.50275
## CryoSleepFalse   288.77551  107.50926   2.686  0.00725 **
## CryoSleepTrue   -547.06014  109.38103  -5.001 5.82e-07 ***
## Destination55 Cancr i e 247.06941  121.19374   2.039  0.04152 *
## DestinationPS0 J318.5-22 209.40525  128.21234   1.633  0.10245
## DestinationTRAPPIST-1e  85.18204  116.52748   0.731  0.46480
## VIPFalse        -79.93184  109.27487  -0.731  0.46451
## VIPTrue         497.54903  155.65396   3.197  0.00140 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1455 on 7604 degrees of freedom
## (1073 observations deleted due to missingness)
## Multiple R-squared:  0.221, Adjusted R-squared:  0.2194
## F-statistic: 143.8 on 15 and 7604 DF, p-value: < 2.2e-16
```

```
train2 = train2 %>% mutate(foodcourt_pred = predict(imp_foodcourt,newdata=train2))
```

```
train2=train2 %>%mutate(NewFoodCourt = case_when(is.na(FoodCourt) ~
                                                round(foodcourt_pred, digits = 0),
                                                TRUE ~ FoodCourt))
```

```
train2 = train2 %>% mutate(NewFoodCourt_final = case_when(is.na(NewFoodCourt) ~
                                                            round(mean(NewFoodCourt, na.rm=T)
                                                                , digits=0)
                                                            , TRUE ~ NewFoodCourt))
```

```
imp_VR = lm(VRDeck~ShoppingMall+Spa+FoodCourt+Age+RoomService+HomePlanet+
            CryoSleep+Destination+VIP, data=train)
summary(imp_VR)
```

```
##
## Call:
## lm(formula = VRDeck ~ ShoppingMall + Spa + FoodCourt + Age +
##      RoomService + HomePlanet + CryoSleep + Destination + VIP,
```

```
##      data = train)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -2751.8   -339.7   -196.7    212.8   19071.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.389e+02  1.555e+02   0.893  0.371817
## ShoppingMall   -7.713e-02  1.990e-02  -3.875  0.000107 ***
## Spa            9.413e-03  1.096e-02   0.859  0.390335
## FoodCourt      4.318e-02  8.086e-03   5.341  9.54e-08 ***
## Age            1.336e+00  8.456e-01   1.580  0.114175
## RoomService   -9.314e-02  1.870e-02  -4.981  6.48e-07 ***
## HomePlanetEarth -1.821e+02  7.811e+01  -2.331  0.019778 *
## HomePlanetEuropa 4.980e+02  8.138e+01   6.120  9.85e-10 ***
## HomePlanetMars  -1.796e+02  8.105e+01  -2.216  0.026690 *
## CryoSleepFalse  2.941e+02  7.591e+01   3.875  0.000108 ***
## CryoSleepTrue   -2.406e+02  7.735e+01  -3.110  0.001876 **
## Destination55 Cancr i  1.063e+02  8.563e+01   1.241  0.214617
## DestinationPS0 J318.5-22 1.433e+02  9.057e+01   1.582  0.113594
## DestinationTRAPPIST-1e 6.771e+01  8.232e+01   0.823  0.410808
## VIPFalse       -5.256e+01  7.719e+01  -0.681  0.495988
## VIPTrue        3.665e+02  1.099e+02   3.333  0.000862 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1028 on 7604 degrees of freedom
## (1073 observations deleted due to missingness)
## Multiple R-squared:  0.1466, Adjusted R-squared:  0.1449
## F-statistic: 87.08 on 15 and 7604 DF, p-value: < 2.2e-16
```

```
train2 = train2 %>% mutate(vr_pred = predict(imp_VR,newdata=train2))
```

```
train2=train2 %>%mutate(NewVR = case_when(is.na(VRDeck) ~
                                         round(vr_pred, digits = 0),
                                         TRUE ~ VRDeck))
```

```
train2 = train2 %>% mutate(VRDeck_final = case_when(is.na(NewVR) ~
                                                     round(mean(NewVR, na.rm=T)
                                                         , digits=0)
                                                     , TRUE ~ NewVR))
```

```
imp_RS = lm(RoomService~ShoppingMall+VRDeck+FoodCourt+Age+Spa+HomePlanet+
             CryoSleep+Destination+VIP, data=train)
summary(imp_RS)
```

```
##
## Call:
## lm(formula = RoomService ~ ShoppingMall + VRDeck + FoodCourt +
##     Age + Spa + HomePlanet + CryoSleep + Destination + VIP, data = train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -836.7  -274.1  -154.3   126.9 13984.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.596e+02  9.517e+01   1.677 0.093634 .
## ShoppingMall   -4.038e-02  1.219e-02  -3.313 0.000926 ***
## VRDeck         -3.491e-02  7.009e-03  -4.981 6.48e-07 ***
## FoodCourt      -2.471e-02  4.951e-03  -4.990 6.16e-07 ***
## Age            1.956e+00  5.173e-01   3.782 0.000157 ***
## Spa           -1.546e-02  6.706e-03  -2.305 0.021180 *
## HomePlanetEarth -9.903e+01  4.783e+01  -2.071 0.038423 *
## HomePlanetEuropa  9.454e-01  4.995e+01   0.019 0.984898
## HomePlanetMars   3.435e+02  4.948e+01   6.943 4.15e-12 ***
## CryoSleepFalse  1.793e+02  4.647e+01   3.857 0.000116 ***
## CryoSleepTrue   -2.411e+02  4.730e+01  -5.097 3.54e-07 ***
## Destination55 Cancr i e  9.141e+01  5.242e+01   1.744 0.081231 .
## DestinationPS0 J318.5-22  5.461e+01  5.545e+01   0.985 0.324729
## DestinationTRAPPIST-1e  5.489e+01  5.039e+01   1.089 0.276047
## VIPFalse        -6.832e+01  4.725e+01  -1.446 0.148288
## VIPTrue         3.679e+01  6.736e+01   0.546 0.584955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 629.3 on 7604 degrees of freedom
## (1073 observations deleted due to missingness)
## Multiple R-squared:  0.1353, Adjusted R-squared:  0.1336
## F-statistic: 79.31 on 15 and 7604 DF, p-value: < 2.2e-16
```

```
train2 = train2 %>% mutate(rs_pred = predict(imp_RS,newdata=train2))
```

```
train2=train2 %>%mutate(Newrs = case_when(is.na(RoomService) ~
                                         round(rs_pred, digits = 0),
                                         TRUE ~ RoomService))
```

```
train2 = train2 %>% mutate(RoomService_final = case_when(is.na(Newrs) ~
                                                         round(mean(Newrs, na.rm=T)
                                                         , digits=0)
                                                         , TRUE ~ Newrs))
```

```
imp_sm = lm(ShoppingMall~RoomService+VRDeck+FoodCourt+Age+Spa+HomePlanet+
            CryoSleep+Destination+VIP, data=train)
summary(imp_sm)
```

```
##
## Call:
## lm(formula = ShoppingMall ~ RoomService + VRDeck + FoodCourt +
##     Age + Spa + HomePlanet + CryoSleep + Destination + VIP, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -486.3 -231.9 -126.9 75.8 23142.8
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.504e+01 8.950e+01 -0.503 0.614785
## RoomService -3.570e-02 1.077e-02 -3.313 0.000926 ***
## VRDeck -2.556e-02 6.595e-03 -3.875 0.000107 ***
## FoodCourt -2.495e-02 4.654e-03 -5.360 8.57e-08 ***
## Age 4.930e-01 4.868e-01 1.013 0.311221
## Spa -1.408e-02 6.306e-03 -2.233 0.025564 *
## HomePlanetEarth -4.751e+00 4.498e+01 -0.106 0.915885
## HomePlanetEuropa 1.152e+02 4.695e+01 2.454 0.014169 *
## HomePlanetMars 2.165e+02 4.660e+01 4.646 3.45e-06 ***
## CryoSleepFalse 1.503e+02 4.370e+01 3.439 0.000587 ***
## CryoSleepTrue -1.886e+02 4.450e+01 -4.239 2.27e-05 ***
## Destination55 Cancr e 7.380e+01 4.929e+01 1.497 0.134331
## DestinationPS0 J318.5-22 8.159e+01 5.214e+01 1.565 0.117661
## DestinationTRAPPIST-1e 5.357e+01 4.738e+01 1.131 0.258243
## VIPFalse 8.103e+01 4.443e+01 1.824 0.068208 .
## VIPTrue 4.415e+01 6.334e+01 0.697 0.485728
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 591.7 on 7604 degrees of freedom
## (1073 observations deleted due to missingness)
## Multiple R-squared: 0.06807, Adjusted R-squared: 0.06623
## F-statistic: 37.03 on 15 and 7604 DF, p-value: < 2.2e-16
```

```
train2 = train2 %>% mutate(sm_pred = predict(imp_sm,newdata=train2))
```

```
train2=train2 %>%mutate(Newsm = case_when(is.na(ShoppingMall) ~
                                         round(sm_pred, digits = 0),
                                         TRUE ~ ShoppingMall))
```

```
train2 = train2 %>% mutate(ShoppingMall_final = case_when(is.na(Newsm) ~
                                                         round(mean(Newsm, na.rm=T)
                                                         , digits=0)
                                                         , TRUE ~ Newsm))
```

```
train2 = train2 %>% mutate(NewTransported = ifelse(Transported == "True", 1, 0))
```

```
train_f = subset(train2, select = -c(Age,RoomService,FoodCourt,
                                     ShoppingMall,Spa,VRDeck,
                                     Age_pred,NewAge,Spa_pred,NewSpa,
                                     foodcourt_pred,NewFoodCourt,
                                     vr_pred,NewVR,rs_pred,Newrs,
                                     sm_pred,Newsm, Transported))
```

Creating Logistic Regression Model using Imputed Data

```
logit = glm(NewTransported ~ HomePlanet+CryoSleep+Destination+VIP+
            NewAge_Final+NewSpa_Final+NewFoodCourt_final+VRDeck_final+
            RoomService_final+ShoppingMall_final, family = binomial,
            data = train_f)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logit)
```

```
##
## Call:
## glm(formula = NewTransported ~ HomePlanet + CryoSleep + Destination +
##     VIP + NewAge_Final + NewSpa_Final + NewFoodCourt_final +
##     VRDeck_final + RoomService_final + ShoppingMall_final, family = binomial,
##     data = train_f)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.136e+00  3.593e-01   3.161 0.001573 **
## HomePlanetEarth  -5.912e-01  1.792e-01  -3.300 0.000968 ***
## HomePlanetEuropa  1.629e+00  2.067e-01   7.884 3.16e-15 ***
## HomePlanetMars    1.179e-01  1.871e-01   0.630 0.528663
## CryoSleepFalse   -2.549e-01  1.697e-01  -1.502 0.133193
## CryoSleepTrue     9.797e-01  1.735e-01   5.645 1.65e-08 ***
## Destination55 Cancr i e  7.407e-02  2.035e-01   0.364 0.715925
## DestinationPS0 J318.5-22 -3.543e-01  2.089e-01  -1.696 0.089870 .
## DestinationTRAPPIST-1e -4.107e-01  1.936e-01  -2.121 0.033904 *
## VIPFalse         -8.097e-02  1.862e-01  -0.435 0.663706
## VIPTrue          -5.384e-01  3.004e-01  -1.792 0.073101 .
## NewAge_Final     -7.914e-03  1.999e-03  -3.959 7.52e-05 ***
## NewSpa_Final     -1.921e-03  1.022e-04 -18.798 < 2e-16 ***
## NewFoodCourt_final  5.048e-04  3.865e-05  13.058 < 2e-16 ***
## VRDeck_final     -1.782e-03  9.683e-05 -18.400 < 2e-16 ***
## RoomService_final -1.474e-03  8.999e-05 -16.378 < 2e-16 ***
## ShoppingMall_final  5.310e-04  6.615e-05   8.028 9.92e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12050.6  on 8692  degrees of freedom
## Residual deviance:  7749.2  on 8676  degrees of freedom
## AIC: 7783.2
##
## Number of Fisher Scoring iterations: 7
```

```
p = predict(logit, type = "response")
roc_logit = roc(train_f$NewTransported ~ p)
```

```
## Setting levels: control = 0, case = 1
```



```
## Setting direction: controls < cases
```

```
auc(roc_logit)
```

```
## Area under the curve: 0.8717
```

Processing Test Data

```
imp_age = lm(Age~ShoppingMall+VRDeck+FoodCourt+Spa+RoomService+HomePlanet+
             CryoSleep+Destination+VIP, data=test)
summary(imp_age)
```

```
##
## Call:
## lm(formula = Age ~ ShoppingMall + VRDeck + FoodCourt + Spa +
##     RoomService + HomePlanet + CryoSleep + Destination + VIP,
##     data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.149  -8.485  -2.706   8.140  52.561
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.077e+01  3.067e+00  10.034 < 2e-16 ***
## ShoppingMall    5.619e-04  4.192e-04   1.340  0.18017
## VRDeck          2.968e-04  2.064e-04   1.438  0.15056
## FoodCourt       9.706e-05  1.687e-04   0.575  0.56506
## Spa            5.159e-04  2.169e-04   2.379  0.01741 *
## RoomService     1.114e-03  4.107e-04   2.714  0.00669 **
## HomePlanetEarth -2.772e+00  1.653e+00  -1.677  0.09359 .
## HomePlanetEuropa 5.410e+00  1.703e+00   3.177  0.00150 **
## HomePlanetMars   8.199e-01  1.700e+00   0.482  0.62958
## CryoSleepFalse  -1.677e+00  1.519e+00  -1.104  0.26953
## CryoSleepTrue   -3.120e+00  1.549e+00  -2.014  0.04408 *
## Destination55 Cancr i e  1.330e-01  1.674e+00   0.079  0.93668
## DestinationPS0 J318.5-22 2.594e+00  1.764e+00   1.470  0.14163
## DestinationTRAPPIST-1e  1.632e+00  1.606e+00   1.016  0.30949
## VIPFalse        -1.988e+00  1.534e+00  -1.296  0.19507
## VIPTrue         -1.983e+00  2.290e+00  -0.866  0.38646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.71 on 3724 degrees of freedom
## (537 observations deleted due to missingness)
## Multiple R-squared:  0.06931, Adjusted R-squared:  0.06556
## F-statistic: 18.49 on 15 and 3724 DF, p-value: < 2.2e-16
```

```
test2 = test %>% mutate(Age_pred = predict(imp_age,newdata=test))
```

```
test2=test2 %>%mutate(NewAge = case_when(is.na(Age) ~ round(Age_pred, digits = 0),
                                         TRUE ~ Age))
```

```
test2 = test2 %>% mutate(NewAge_Final = case_when(is.na(NewAge) ~
                                                  round(mean(NewAge, na.rm=T), digits=0)
                                                  , TRUE ~ NewAge))
```

```
imp_spa = lm(Spa~ShoppingMall+VRDeck+FoodCourt+Age+RoomService+HomePlanet+
             CryoSleep+Destination+VIP, data=test)
summary(imp_spa)
```

```
##
## Call:
## lm(formula = Spa ~ ShoppingMall + VRDeck + FoodCourt + Age +
##     RoomService + HomePlanet + CryoSleep + Destination + VIP,
##     data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3324.3  -354.4  -193.1   198.8 18777.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.367e+02  2.346e+02   1.862  0.062740 .
## ShoppingMall   -8.434e-02  3.163e-02  -2.666  0.007701 **
## VRDeck         4.297e-03  1.559e-02   0.276  0.782891
## FoodCourt      8.302e-02  1.266e-02   6.555  6.32e-11 ***
## Age            2.941e+00  1.236e+00   2.379  0.017412 *
## RoomService    -9.795e-02  3.100e-02  -3.160  0.001591 **
## HomePlanetEarth -1.429e+02  1.248e+02  -1.144  0.252492
## HomePlanetEuropa 4.701e+02  1.285e+02   3.658  0.000258 ***
## HomePlanetMars  -8.275e+01  1.283e+02  -0.645  0.519116
## CryoSleepFalse  6.729e+01  1.147e+02   0.587  0.557445
## CryoSleepTrue   -4.708e+02  1.168e+02  -4.032  5.64e-05 ***
## Destination55 Cancr i  1.430e+02  1.264e+02   1.132  0.257762
## DestinationPS0 J318.5-22 9.964e+01  1.333e+02   0.748  0.454669
## DestinationTRAPPIST-1e 1.327e+01  1.213e+02   0.109  0.912884
## VIPFalse        -1.655e+02  1.158e+02  -1.429  0.153105
## VIPTrue         4.691e+02  1.727e+02   2.716  0.006644 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1035 on 3724 degrees of freedom
## (537 observations deleted due to missingness)
## Multiple R-squared:  0.1573, Adjusted R-squared:  0.1539
## F-statistic: 46.33 on 15 and 3724 DF, p-value: < 2.2e-16
```

```
test2 = test2 %>% mutate(Spa_pred = predict(imp_spa,newdata=test2))
```

```
test2=test2 %>%mutate(NewSpa = case_when(is.na(Spa) ~ round(Spa_pred, digits = 0),
                                         TRUE ~ Spa))
```

```
test2 = test2 %>% mutate(NewSpa_Final = case_when(is.na(NewSpa) ~
                                                    round(mean(NewSpa, na.rm=T), digits=0)
                                                    , TRUE ~ NewSpa))
```

```
imp_foodcourt = lm(FoodCourt~ShoppingMall+VRDeck+Age+Spa+RoomService+HomePlanet+
                   CryoSleep+Destination+VIP, data=test)
summary(imp_foodcourt)
```

```
##
## Call:
## lm(formula = FoodCourt ~ ShoppingMall + VRDeck + Age + Spa +
##     RoomService + HomePlanet + CryoSleep + Destination + VIP,
##     data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4114.4  -478.7  -214.9   342.2 23454.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    48.26029   301.92365    0.160  0.87301
## ShoppingMall   -0.03255    0.04073   -0.799  0.42428
## VRDeck         0.15536    0.01990    7.808 7.49e-15 ***
## Age            0.91594    1.59183    0.575  0.56506
## Spa            0.13741    0.02096    6.555 6.32e-11 ***
## RoomService    -0.17386    0.03983   -4.365 1.31e-05 ***
## HomePlanetEarth -225.11735  160.60373   -1.402  0.16109
## HomePlanetEuropa  914.51576  164.98951    5.543 3.18e-08 ***
## HomePlanetMars   -176.59456  165.09911   -1.070  0.28486
## CryoSleepFalse  441.62517  147.38827    2.996  0.00275 **
## CryoSleepTrue   -297.24193  150.47580   -1.975  0.04830 *
## Destination55 Cancr i e  109.88104  162.60075    0.676  0.49923
## DestinationPS0 J318.5-22  215.93313  171.40673    1.260  0.20783
## DestinationTRAPPIST-1e   77.65553  156.03898    0.498  0.61875
## VIPFalse       -13.46613  149.06071   -0.090  0.92802
## VIPTrue        525.76236  222.27676    2.365  0.01806 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1332 on 3724 degrees of freedom
## (537 observations deleted due to missingness)
## Multiple R-squared:  0.233, Adjusted R-squared:  0.2299
## F-statistic: 75.43 on 15 and 3724 DF, p-value: < 2.2e-16
```

```
test2 = test2 %>% mutate(foodcourt_pred = predict(imp_foodcourt,newdata=test2))
```

```
test2=test2 %>%mutate(NewFoodCourt = case_when(is.na(FoodCourt) ~
                                             round(foodcourt_pred, digits = 0),
                                             TRUE ~ FoodCourt))
```

```
test2 = test2 %>% mutate(NewFoodCourt_final = case_when(is.na(NewFoodCourt) ~
                                                         round(mean(NewFoodCourt, na.rm=T)
                                                         , digits=0)
                                                         , TRUE ~ NewFoodCourt))
```

```
imp_VR = lm(VRDeck+ShoppingMall+Spa+FoodCourt+Age+RoomService+HomePlanet+
            CryoSleep+Destination+VIP, data=test)
summary(imp_VR)
```

```
##
```

```
## Call:
## lm(formula = VRDeck ~ ShoppingMall + Spa + FoodCourt + Age +
##     RoomService + HomePlanet + CryoSleep + Destination + VIP,
##     data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2328.2  -369.5  -185.4   204.8 17023.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.189e+02  2.464e+02   2.512  0.01206 *
## ShoppingMall     -1.255e-02  3.327e-02  -0.377  0.70601
## Spa              4.746e-03  1.722e-02   0.276  0.78289
## FoodCourt        1.037e-01  1.328e-02   7.808 7.49e-15 ***
## Age              1.869e+00  1.300e+00   1.438  0.15056
## RoomService      -1.318e-01  3.255e-02  -4.050 5.22e-05 ***
## HomePlanetEarth  -3.844e+02  1.311e+02  -2.933  0.00338 **
## HomePlanetEuropa  2.763e+02  1.353e+02   2.043  0.04112 *
## HomePlanetMars    -3.707e+02  1.348e+02  -2.751  0.00597 **
## CryoSleepFalse    2.151e+02  1.205e+02   1.785  0.07429 .
## CryoSleepTrue     -2.960e+02  1.229e+02  -2.408  0.01608 *
## Destination55 Cancri e -6.741e+01  1.328e+02  -0.508  0.61182
## DestinationPS0 J318.5-22 5.416e+01  1.400e+02   0.387  0.69900
## DestinationTRAPPIST-1e -4.808e+01  1.275e+02  -0.377  0.70605
## VIPFalse          -1.644e+02  1.217e+02  -1.351  0.17690
## VIPTrue           2.426e+02  1.817e+02   1.335  0.18180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1088 on 3724 degrees of freedom
## (537 observations deleted due to missingness)
## Multiple R-squared:  0.1513, Adjusted R-squared:  0.1478
## F-statistic: 44.25 on 15 and 3724 DF, p-value: < 2.2e-16
```

```
test2 = test2 %>% mutate(vr_pred = predict(imp_VR,newdata=test2))
```

```
test2=test2 %>%mutate(NewVR = case_when(is.na(VRDeck) ~
                                         round(vr_pred, digits = 0),
                                         TRUE ~ VRDeck))
```

```
test2 = test2 %>% mutate(VRDeck_final = case_when(is.na(NewVR) ~
                                                    round(mean(NewVR, na.rm=T)
                                                          , digits=0)
                                                    , TRUE ~ NewVR))
```

```
imp_RS = lm(RoomService~ShoppingMall+VRDeck+FoodCourt+Age+Spa+HomePlanet+
             CryoSleep+Destination+VIP, data=test)
summary(imp_RS)
```

```
##
## Call:
```

```
## lm(formula = RoomService ~ ShoppingMall + VRDeck + FoodCourt +
##     Age + Spa + HomePlanet + CryoSleep + Destination + VIP, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -808.4  -272.3  -123.2   124.2 11339.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.999e+01  1.239e+02   0.807 0.419636
## ShoppingMall   -6.374e-02  1.668e-02  -3.821 0.000135 ***
## VRDeck         -3.327e-02  8.213e-03  -4.050 5.22e-05 ***
## FoodCourt      -2.927e-02  6.707e-03  -4.365 1.31e-05 ***
## Age            1.771e+00  6.526e-01   2.714 0.006685 **
## Spa            -2.730e-02  8.639e-03  -3.160 0.001591 **
## HomePlanetEarth -1.159e+02  6.589e+01  -1.759 0.078656 .
## HomePlanetEuropa -2.932e+01  6.798e+01  -0.431 0.666268
## HomePlanetMars   3.198e+02  6.755e+01   4.734 2.28e-06 ***
## CryoSleepFalse   1.055e+02  6.053e+01   1.743 0.081433 .
## CryoSleepTrue   -3.227e+02  6.155e+01  -5.243 1.66e-07 ***
## Destination55 Cancr i e  9.354e+01  6.671e+01   1.402 0.160913
## DestinationPS0 J318.5-22  9.156e+01  7.033e+01   1.302 0.193088
## DestinationTRAPPIST-1e  5.029e+01  6.403e+01   0.785 0.432222
## VIPFalse        9.601e+01  6.114e+01   1.570 0.116466
## VIPTrue         2.351e+02  9.120e+01   2.578 0.009971 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 546.4 on 3724 degrees of freedom
## (537 observations deleted due to missingness)
## Multiple R-squared:  0.1749, Adjusted R-squared:  0.1716
## F-statistic: 52.63 on 15 and 3724 DF, p-value: < 2.2e-16
```

```
test2 = test2 %>% mutate(rs_pred = predict(imp_RS,newdata=test2))
```

```
test2=test2 %>%mutate(Newrs = case_when(is.na(RoomService) ~
                                     round(rs_pred, digits = 0),
                                     TRUE ~ RoomService))
```

```
test2 = test2 %>% mutate(RoomService_final = case_when(is.na(Newrs) ~
                                                         round(mean(Newrs, na.rm=T)
                                                         , digits=0)
                                                         , TRUE ~ Newrs))
```

```
imp_sm = lm(ShoppingMall~RoomService+VRDeck+FoodCourt+Age+Spa+HomePlanet+
            CryoSleep+Destination+VIP, data=test)
summary(imp_sm)
```

```
##
## Call:
## lm(formula = ShoppingMall ~ RoomService + VRDeck + FoodCourt +
##     Age + Spa + HomePlanet + CryoSleep + Destination + VIP, data = test)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -516.9 -226.2 -149.2   70.1 7997.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.935e+02  1.214e+02   2.418 0.015647 *
## RoomService      -6.127e-02  1.603e-02  -3.821 0.000135 ***
## VRDeck            -3.044e-03  8.070e-03  -0.377 0.706007
## FoodCourt         -5.267e-03  6.592e-03  -0.799 0.424277
## Age               8.582e-01  6.402e-01   1.340 0.180172
## Spa              -2.259e-02  8.473e-03  -2.666 0.007701 **
## HomePlanetEarth  -1.893e+02  6.455e+01  -2.932 0.003388 **
## HomePlanetEuropa -1.281e+02  6.661e+01  -1.923 0.054528 .
## HomePlanetMars    6.902e+01  6.642e+01   1.039 0.298814
## CryoSleepFalse    1.245e+02  5.933e+01   2.099 0.035928 *
## CryoSleepTrue     -2.078e+02  6.047e+01  -3.437 0.000594 ***
## Destination55 Cancr i e  7.535e+00  6.542e+01   0.115 0.908309
## DestinationPS0 J318.5-22 1.541e+01  6.897e+01   0.223 0.823242
## DestinationTRAPPIST-1e -2.354e+01  6.277e+01  -0.375 0.707648
## VIPFalse          8.689e+00  5.996e+01   0.145 0.884793
## VIPTrue           9.259e+01  8.947e+01   1.035 0.300808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 535.7 on 3724 degrees of freedom
## (537 observations deleted due to missingness)
## Multiple R-squared:  0.09125,    Adjusted R-squared:  0.08759
## F-statistic: 24.93 on 15 and 3724 DF,  p-value: < 2.2e-16
```

```
test2 = test2 %>% mutate(sm_pred = predict(imp_sm,newdata=test2))
```

```
test2=test2 %>%mutate(Newsm = case_when(is.na(ShoppingMall) ~
                                     round(sm_pred, digits = 0),
                                     TRUE ~ ShoppingMall))
```

```
test2 = test2 %>% mutate(ShoppingMall_final = case_when(is.na(Newsm) ~
                                                         round(mean(Newsm, na.rm=T)
                                                         , digits=0)
                                                         , TRUE ~ Newsm))
```

```
test_f = subset(test2, select = -c(Age,RoomService,FoodCourt,
                                   ShoppingMall,Spa,VRDeck,
                                   Age_pred,NewAge,Spa_pred,NewSpa,
                                   foodcourt_pred,NewFoodCourt,
                                   vr_pred,NewVR,rs_pred,Newsrs,
                                   sm_pred,Newsm))
```

Predicting Test Data

```
test_f = test_f %>% mutate(prediction =  
  predict(logit, type = "response", newdata = test_f)) %>%  
  mutate(Transported = case_when(prediction>0.5~"True",  
    TRUE ~"False") )  
  
submission = test_f %>% select(PassengerId, Transported)  
  
write.csv(submission, file="submission.csv", row.names=FALSE)
```