Exploration of Missing Values

```
set.seed(123)
data=read.csv("hotel_bookings.csv")
originalData=data
#Checking for missing values (NA). Observed 4 missing values in the children column.
data[rowSums(is.na(data))>0,]
```

```
##            hotel is_canceled lead_time arrival_date_year arrival_date_month
## 40601 City Hotel           1         2              2015             August
## 40668 City Hotel           1         1              2015             August
## 40680 City Hotel           1         1              2015             August
## 41161 City Hotel           1         8              2015             August
##       arrival_date_week_number arrival_date_day_of_month
## 40601                       32                         3
## 40668                       32                         5
## 40680                       32                         5
## 41161                       33                        13
##       stays_in_weekend_nights stays_in_week_nights adults children babies meal
## 40601                       1                    0      2       NA      0   BB
## 40668                       0                    2      2       NA      0   BB
## 40680                       0                    2      3       NA      0   BB
## 41161                       2                    5      2       NA      0   BB
##       country market_segment distribution_channel is_repeated_guest
## 40601     PRT      Undefined            Undefined                 0
## 40668     PRT         Direct            Undefined                 0
## 40680     PRT      Undefined            Undefined                 0
## 41161     PRT      Online TA            Undefined                 0
##       previous_cancellations previous_bookings_not_canceled reserved_room_type
## 40601                      0                              0                  B
## 40668                      0                              0                  B
## 40680                      0                              0                  B
## 41161                      0                              0                  B
##       assigned_room_type booking_changes deposit_type agent company
## 40601                  B               0   No Deposit  NULL    NULL
## 40668                  B               0   No Deposit    14    NULL
## 40680                  B               0   No Deposit  NULL    NULL
## 41161                  B               0   No Deposit     9    NULL
##       days_in_waiting_list   customer_type  adr required_car_parking_spaces
## 40601                    0 Transient-Party 12.0                           0
## 40668                    0 Transient-Party 12.0                           0
## 40680                    0 Transient-Party 18.0                           0
## 41161                    0 Transient-Party 76.5                           0
##       total_of_special_requests reservation_status reservation_status_date
## 40601                         1           Canceled              2015-08-01
## 40668                         1           Canceled              2015-08-04
## 40680                         2           Canceled              2015-08-04
## 41161                         1           Canceled              2015-08-09
```

```
#Removing these 4 instances as there is a lot of observations
data=na.omit(data)
```

Contingency table of all the columns

```
#lapply(data,table) Commented out as it's too big of a print.
```

It’s observed that there are NULL values in the data. The columns with NULL values are company, agent, and country.

```
colSums(data=="NULL")
```

```
##                          hotel                    is_canceled 
##                              0                              0 
##                      lead_time              arrival_date_year 
##                              0                              0 
##             arrival_date_month       arrival_date_week_number 
##                              0                              0 
##      arrival_date_day_of_month        stays_in_weekend_nights 
##                              0                              0 
##           stays_in_week_nights                         adults 
##                              0                              0 
##                       children                         babies 
##                              0                              0 
##                           meal                        country 
##                              0                            488 
##                 market_segment           distribution_channel 
##                              0                              0 
##              is_repeated_guest         previous_cancellations 
##                              0                              0 
## previous_bookings_not_canceled             reserved_room_type 
##                              0                              0 
##             assigned_room_type                booking_changes 
##                              0                              0 
##                   deposit_type                          agent 
##                              0                          16338 
##                        company           days_in_waiting_list 
##                         112589                              0 
##                  customer_type                            adr 
##                              0                              0 
##    required_car_parking_spaces      total_of_special_requests 
##                              0                              0 
##             reservation_status        reservation_status_date 
##                              0                              0
```

The contigency table for the company feature.

```
#table(data$company) Commented out as it's too big of a print.
```

It is observed that the most common element is the NULL value with 112589 observations which is much more than 50% of the data. This is most likely due to a majority of the hotel bookings not be associated with a company booking. As a result, this implys that the NULL values are important so they will be renamed to “No Company”

```
data=data%>%mutate(company=ifelse(company=="NULL","No Company",company))
```

The agent feature has 16338 NULL values. As the agent number is related to the distribution channel of the booking, we will investigate the distribution channel.

```
#table(data$agent) Commented out as it's too big of a print.
```

```
agentNullData=data%>% filter(agent=="NULL")
#table(agentNullData$agent,agentNullData$distribution_channel) Commented out as it's too big of a print.
```

Of the 16338 NULL values in the agent field, 13168 (5543+7625) of them belong to the corporate and direct distribution channels which have no agents as they directly contact the hotel for the booking. We will fill these with “No Travel Agency” as they don’t use any travel agency. There is 3167 NULL values with TA/TO distribution channels. We will fill in these with “TA/TO No Agent Number” as they have travel agents but have no agent id. The remaining 3 NULL values will be removed as they are only 3 of them.

```
data=data%>%mutate(agent=ifelse(distribution_channel %in% c("Corporate","Direct") & agent=='NULL','No Travel Agency',agent))
data=data%>%mutate(agent=ifelse(distribution_channel=="TA/TO" & agent=="NULL","TA/TO No Agent Number",agent))
data=data%>%filter(agent!="NULL")
```

Looking at the Contingency table of the country column we see that there is 488 NULL values.

```
#table(data$country) Commented out as it's too big of a print.
```

```
countryNulldata=data%>% filter(country=="NULL")
x=table(countryNulldata$country,countryNulldata$agent)
#x["NULL",] Commented out as it's too big of a print.
```

It is observed that majority of the observations with NULL for countries also had no agents which are now “No Travel Agency” and “TA/TO No Agent Number”. We will fill these with countries with “Unknown”. For all the other NULL countries, we will remove them as there is a small amount of them.

```
data=data%>%mutate(country=ifelse(agent %in% c("No Travel Agency","TA/TO No Agent Number") & country=='NULL','Unknown',country))
data=data%>%filter(data$country!="NULL")
```

```
#lapply(data,table)
```

It is observed that there is 1168 undefined columns in the meal feature. As the other options are BB (Bed and Breakfast), FB(Full Board), HB(Half Board), and SC (Self Catering) it is observed that there is no option for no meal services. As a result, we will fill these undefined values with “Other”

```
data=data%>%mutate(meal=ifelse(meal=='Undefined','Other',meal))
table(data$meal)
```

```
## 
##    BB    FB    HB Other    SC 
## 92164   798 14450  1168 10649
```

```
head(data)
```

```
##          hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 Resort Hotel           0       342              2015               July
## 2 Resort Hotel           0       737              2015               July
## 3 Resort Hotel           0         7              2015               July
## 4 Resort Hotel           0        13              2015               July
## 5 Resort Hotel           0        14              2015               July
## 6 Resort Hotel           0        14              2015               July
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       27                         1                       0
## 2                       27                         1                       0
## 3                       27                         1                       0
## 4                       27                         1                       0
## 5                       27                         1                       0
## 6                       27                         1                       0
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    0      2        0      0   BB     PRT         Direct
## 2                    0      2        0      0   BB     PRT         Direct
## 3                    1      1        0      0   BB     GBR         Direct
## 4                    1      1        0      0   BB     GBR      Corporate
## 5                    2      2        0      0   BB     GBR      Online TA
## 6                    2      2        0      0   BB     GBR      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                      0
## 2               Direct                 0                      0
## 3               Direct                 0                      0
## 4            Corporate                 0                      0
## 5                TA/TO                 0                      0
## 6                TA/TO                 0                      0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                  C                  C
## 2                              0                  C                  C
## 3                              0                  A                  C
## 4                              0                  A                  A
## 5                              0                  A                  A
## 6                              0                  A                  A
##   booking_changes deposit_type            agent    company days_in_waiting_list
## 1               3   No Deposit No Travel Agency No Company                    0
## 2               4   No Deposit No Travel Agency No Company                    0
## 3               0   No Deposit No Travel Agency No Company                    0
## 4               0   No Deposit              304 No Company                    0
## 5               0   No Deposit              240 No Company                    0
## 6               0   No Deposit              240 No Company                    0
##   customer_type adr required_car_parking_spaces total_of_special_requests
## 1     Transient   0                           0                         0
## 2     Transient   0                           0                         0
## 3     Transient  75                           0                         0
## 4     Transient  75                           0                         0
## 5     Transient  98                           0                         1
## 6     Transient  98                           0                         1
##   reservation_status reservation_status_date
## 1          Check-Out              2015-07-01
## 2          Check-Out              2015-07-01
## 3          Check-Out              2015-07-02
## 4          Check-Out              2015-07-02
## 5          Check-Out              2015-07-03
## 6          Check-Out              2015-07-03
```

```
write.csv(data,"data.csv",row.names = FALSE) # Writing out for easier factor conversion
```

```
data=read.csv("data.csv",stringsAsFactors = TRUE)
data$is_canceled=as.factor(data$is_canceled)
file.remove("data.csv")
```

```
## [1] TRUE
```

```
data=subset(data,select=-reservation_status) #Dropping variables that are observed after a hotel booking is finalized (Canceled, No Show, etc)
data=subset(data,select=-reservation_status_date)
```

```
originalData=data
remove_rare_levels <- function(factor_var, threshold = 0.001) {
  freq_table <- table(factor_var)
  total_count <- sum(freq_table)
  proportions <- freq_table / total_count
  levels_to_keep <- names(proportions[proportions >= threshold])

  return(factor(factor_var, levels = levels_to_keep))
}

# Apply to all factor columns in the data frame
clean_factors <- function(data, threshold = 0.001) {
  data[sapply(data, is.factor)] <- lapply(
    data[sapply(data, is.factor)],
    remove_rare_levels,
    threshold = threshold
  )
  return(data)
}
data_cleaned <- clean_factors(data, threshold = 0.001)


data=na.omit(data_cleaned)
```

```
partition=createDataPartition(data$is_canceled,p=0.75,list=FALSE)
data_train=data[partition,]
data_test=data[-partition,]
n=length(data_test$is_canceled)
z=1.96
```

Train test split

Random Forest

```
gridRF=expand.grid(mtry=5,splitrule="gini",min.node.size=1)
rg=train(is_canceled~.,data=data_train,method="ranger",importance = "impurity",num.trees=1000,trControl = trainControl(method = "none"),tuneGrid=gridRF)
#gridRF=expand.grid(mtry=5),splitrule="gini",min.node.size=1)
#control=trainControl(method="cv",number=5,verboseIter=TRUE)
#rg=train(is_canceled~.,data=data_train,method="ranger",tuneGrid=gridRF,trControl=control,importance = "impurity",num.trees=1000)

rg
```

```
## Random Forest
##
## 80511 samples
##    29 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: None
```

```
rgPreds=predict(rg,newdata=data_test)
```

Variable Importance

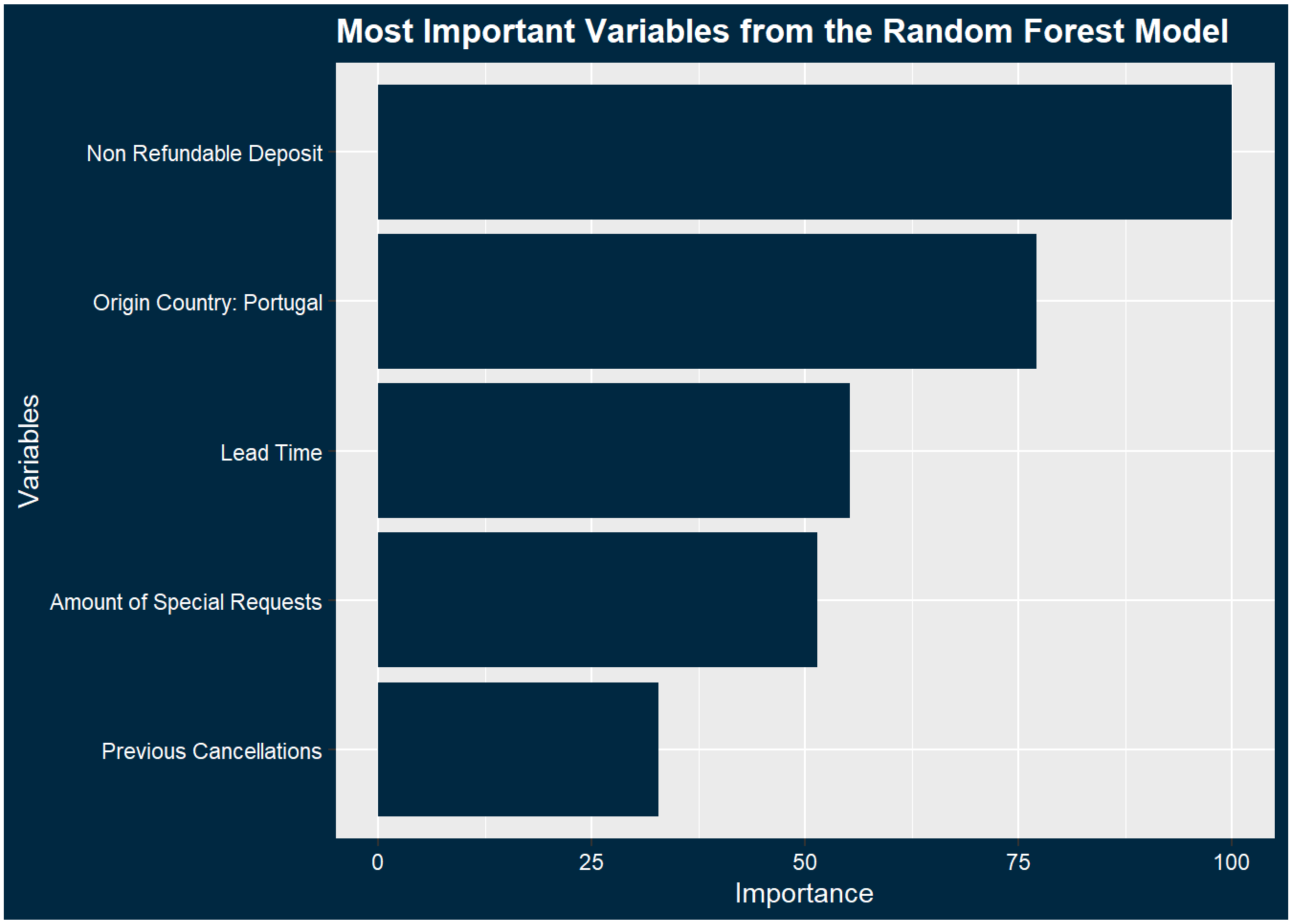
```
rfImportance=varImp(rg)
Top5RfImportance=rfImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% arrange(desc(Overall))%>%head(5)
Top5RfImportance
```

```
##                       Feature   Overall
## 1    deposit_typeNon Refund 100.00000
## 2                 countryPRT  77.14135
## 3                  lead_time  55.29687
## 4 total_of_special_requests  51.47270
## 5    previous_cancellations  32.80224
```

```
#Found Deposit type:Non refundable, country:Portugal, lead_time, total of special requests, and previous_cancellations important
```

Variable Importance Plot

```
ggplot(data=Top5RfImportance,mapping=aes(x=Overall,y= reorder(Feature, Overall)))+geom_bar(stat
="identity",fill="#002845")+scale_y_discrete(labels=c("Previous Cancellations","Amount of Specia
l Requests","Lead Time","Origin Country: Portugal","Non Refundable Deposit"))+xlab("Importance")
+ylab("Variables")+ggtitle("Most Important Variables from the Random Forest Model")+ theme(plot.
background = element_rect(fill = "#002845"),axis.text=element_text(color = "white"),axis.title =
element_text(color = "white"),plot.title=element_text(face = "bold",color = "white"))
```



Confusion Matrixs

```
table(rgPreds,data_test$is_canceled)
```

```
##
## rgPreds     0     1
##       0 16128  3934
##       1   493  6281
```

Accuracy

```
(rfAccuracy=mean(rgPreds==data_test$is_canceled))
```

```
## [1] 0.835035
```

```
rfse=sqrt(rfAccuracy*(1 - rfAccuracy)/n)
rfLowerbound=rfAccuracy-z*rfse
rfUpperbound=rfAccuracy+z*rfse
```

Decision Tree

```
tree=rpart(is_canceled~.,data=data_train, method = "class")
```

Accuracy

```
treePreds=predict(tree,newdata=data_test,type="class")
(treeAccuracy=mean(treePreds==data_test$is_canceled))
```

```
## [1] 0.801647
```

```
treese=sqrt(treeAccuracy*(1 - treeAccuracy)/n)
treeLowerbound=treeAccuracy-z*treese
treeUpperbound=treeAccuracy+z*treese
```

Confusion Matrix

```
table(treePreds,data_test$is_canceled)
```

```
##          
## treePreds     0     1
##         0 15231  3933
##         1  1390  6282
```

Variable Importance

```
tree_Imp=as.data.frame(tree$variable.importance)

tree_Imp=tree_Imp%>%rownames_to_column()
names(tree_Imp)=c("Variable","Importance")

tree_Imp=tree_Imp%>%arrange(desc(Importance))%>%head(5)
tree_Imp#Important vars are deposit type, agent, market segment, total of special requests and c
ountry.
```

```
##                    Variable Importance
## 1              deposit_type   8615.008
## 2                     agent   4847.048
## 3            market_segment   2068.328
## 4 total_of_special_requests   1540.051
## 5                 lead_time   1161.706
```

Logistic Regression

```
lg=train(is_canceled~.,data=data_train,trControl=trainControl(method="none"),method="glm",trace=
FALSE)
```

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

Accuracy

```
lgPreds=predict(lg,newdata=data_test)
(lgAccuracy=mean(lgPreds==data_test$is_canceled))
```

```
## [1] 0.8267998
```

```
lgse=sqrt(lgAccuracy*(1 - lgAccuracy)/n)
lgLowerbound=lgAccuracy-z*lgse
lgUpperbound=lgAccuracy+z*lgse
```

Confusion Matrix

```
table(lgPreds,data_test$is_canceled)
```

```
## 
## lgPreds     0     1
##       0 14893  2920
##       1  1728  7295
```

Variable Importance

```
lgImportance=varImp(lg)
Top5lgImportance=lgImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% a
rrange(desc(Overall))%>%head(5)
Top5lgImportance
```

```
##                      Feature   Overall
## 1 total_of_special_requests 100.00000
## 2                 lead_time  58.43158
## 3  `deposit_typeNon Refund`  49.97554
## 4    previous_cancellations  46.35755
## 5       assigned_room_typeD  43.86662
```

```
#Found Deposit type:Agent 252, total_of_special_requests, lead_time, deposit_type non refund, pr
evious cancellations
```

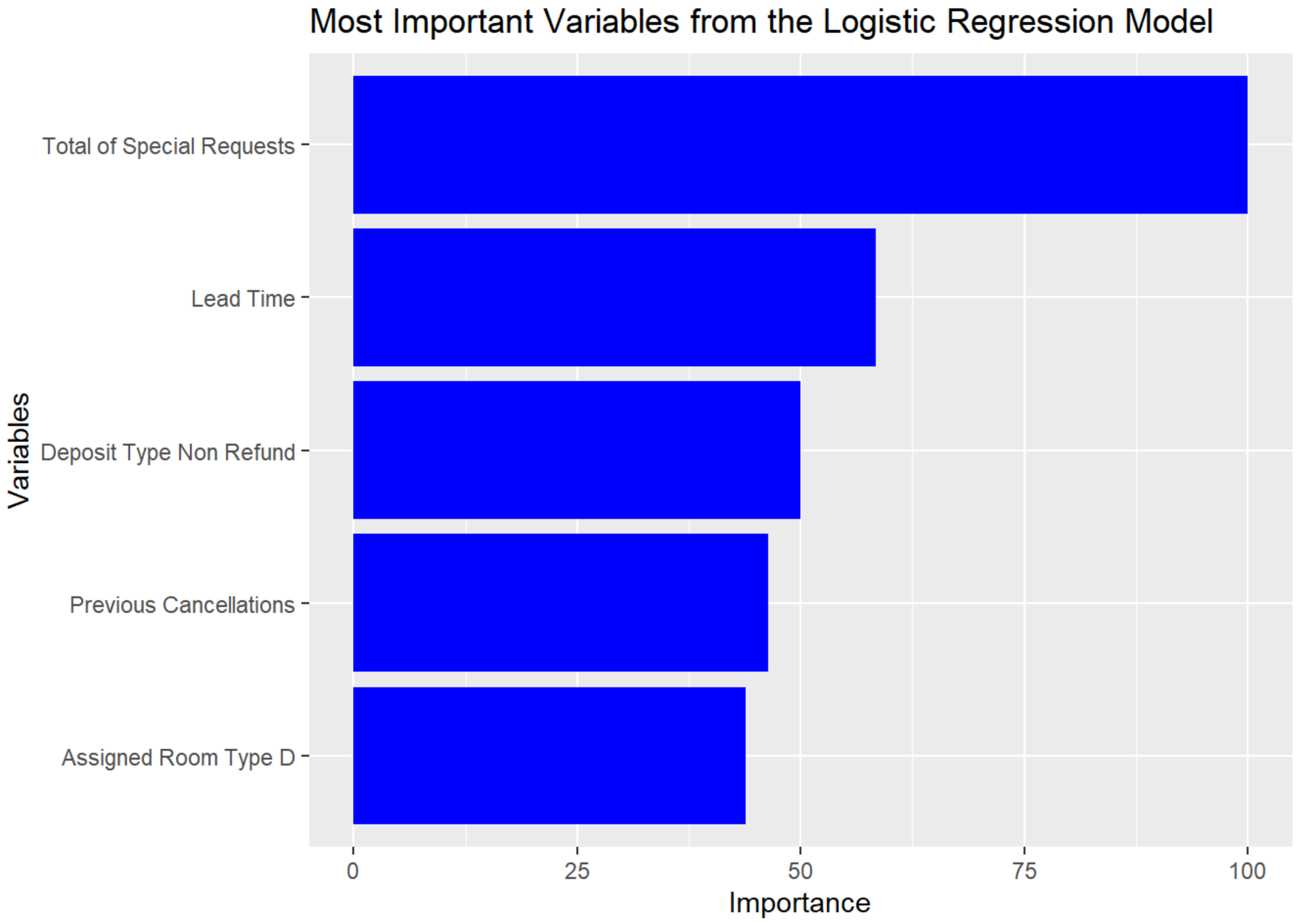
Important variables by pvalue of logistic regression

```r
lg_summary=summary(lg)
coefs=lg_summary$coefficients
pvalues=coefs[,"Pr(>|z|)"]
rownames(coefs)[pvalues<0.05]
```

```
##   [1] "(Intercept)"                    "`hotelResort Hotel`"
##   [3] "lead_time"                      "arrival_date_year"
##   [5] "arrival_date_monthMarch"        "stays_in_weekend_nights"
##   [7] "stays_in_week_nights"           "adults"
##   [9] "children"                       "mealFB"
##  [11] "mealHB"                         "mealOther"
##  [13] "mealSC"                         "countryARG"
##  [15] "countryAUS"                     "countryAUT"
##  [17] "countryBEL"                     "countryBRA"
##  [19] "countryCHE"                     "countryCHN"
##  [21] "countryCN"                      "countryCZE"
##  [23] "countryDEU"                     "countryDNK"
##  [25] "countryESP"                     "countryFIN"
##  [27] "countryFRA"                     "countryGBR"
##  [29] "countryGRC"                     "countryHUN"
##  [31] "countryIND"                     "countryIRL"
##  [33] "countryISR"                     "countryITA"
##  [35] "countryJPN"                     "countryKOR"
##  [37] "countryLUX"                     "countryMAR"
##  [39] "countryNLD"                     "countryNOR"
##  [41] "countryPOL"                     "countryROU"
##  [43] "countryRUS"                     "countrySWE"
##  [45] "countryTUR"                     "countryUSA"
##  [47] "distribution_channelDirect"     "is_repeated_guest"
##  [49] "previous_cancellations"         "previous_bookings_not_canceled"
##  [51] "reserved_room_typeB"            "reserved_room_typeC"
##  [53] "reserved_room_typeD"            "reserved_room_typeE"
##  [55] "reserved_room_typeF"            "reserved_room_typeG"
##  [57] "reserved_room_typeH"            "assigned_room_typeB"
##  [59] "assigned_room_typeC"            "assigned_room_typeD"
##  [61] "assigned_room_typeE"            "assigned_room_typeF"
##  [63] "assigned_room_typeG"            "assigned_room_typeH"
##  [65] "assigned_room_typeI"            "assigned_room_typeK"
##  [67] "booking_changes"                "`deposit_typeNon Refund`"
##  [69] "deposit_typeRefundable"         "agent11"
##  [71] "agent119"                       "agent132"
##  [73] "agent134"                       "agent138"
##  [75] "agent14"                        "agent142"
##  [77] "agent143"                       "agent147"
##  [79] "agent152"                       "agent16"
##  [81] "agent168"                       "agent17"
##  [83] "agent171"                       "agent177"
##  [85] "agent19"                        "agent191"
##  [87] "agent208"                       "agent22"
##  [89] "agent229"                       "agent234"
##  [91] "agent240"                       "agent241"
##  [93] "agent242"                       "agent243"
##  [95] "agent248"                       "agent250"
##  [97] "agent26"                        "agent27"
##  [99] "agent28"                        "agent29"
## [101] "agent30"                        "agent315"
## [103] "agent37"                        "agent38"
## [105] "agent40"                        "agent42"
## [107] "agent56"                        "agent58"
## [109] "agent67"                        "agent68"
## [111] "agent7"                         "agent8"
## [113] "agent85"                        "agent9"
## [115] "`agentNo Travel Agency`"        "`agentTA/TO No Agent Number`"
## [117] "customer_typeTransient"         "`customer_typeTransient-Party`"
## [119] "adr"                            "total_of_special_requests"
```

Variable Importance Plot

```
ggplot(data=Top5lgImportance,mapping=aes(x=Overall,y= reorder(Feature, Overall)))+geom_bar(stat="identity",fill="blue")+scale_y_discrete(labels=c("Assigned Room Type D","Previous Cancellations","Deposit Type Non Refund","Lead Time","Total of Special Requests"))+xlab("Importance")+ylab("Variables")+ggtitle("Most Important Variables from the Logistic Regression Model")
```



#

Neural Net

```
nn_trainControl=trainControl(method="none")
nn_tuneGrid=expand.grid(size=5, decay = 0.01)
nnModel=train(is_canceled~.,data=data_train,method="nnet",trControl=nn_trainControl,tuneGrid=nn_tuneGrid, trace = FALSE)
```

```
nnPreds=predict(nnModel,newdata=data_test)
```

Confusion Matrix

```
table(nnPreds,data_test$is_canceled)
```

```
## 
## nnPreds     0     1
##       0 14508  3081
##       1  2113  7134
```

Accuracy

```
(nnaccuracy=mean(nnPreds==data_test$is_canceled))
```

```
## [1] 0.806454
```

```
nnse=sqrt(nnaccuracy*(1 - nnaccuracy)/n)
nnLowerbound=nnaccuracy-z*nnse
nnUpperbound=nnaccuracy+z*nnse
```

Important variables

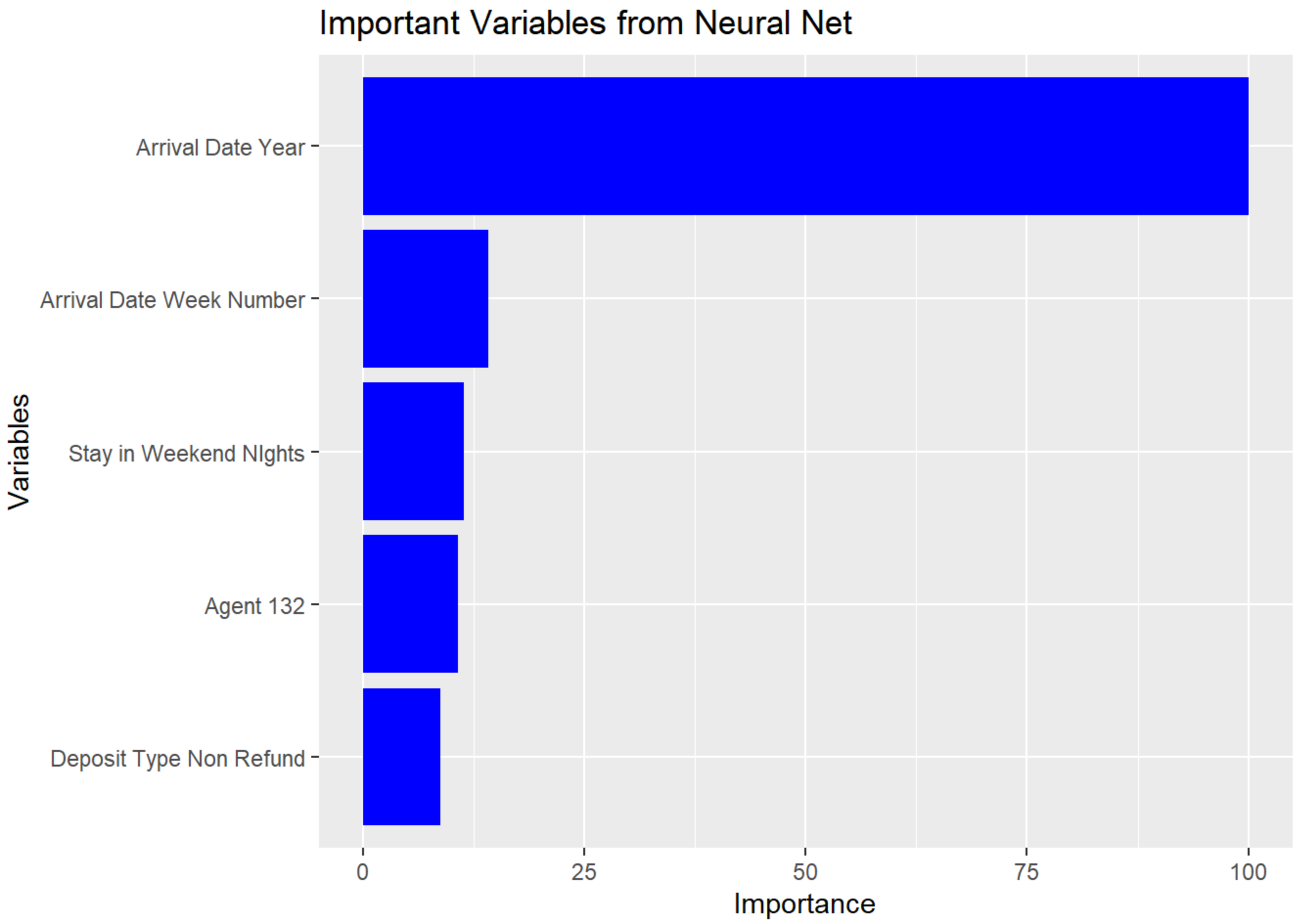
```
nnImportance=varImp(nnModel)
Top5nnImportance=nnImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% a
rrange(desc(Overall))%>%head(5)
Top5nnImportance
```

```
##                       Feature    Overall
## 1           arrival_date_year 100.000000
## 2    arrival_date_week_number  14.136013
## 3     stays_in_weekend_nights  11.360916
## 4                    agent132  10.737572
## 5 arrival_date_monthSeptember   8.720424
```

```
#Found deposit_typeNon Refund, market segment Complementary, agent 253, agent 94, and agent 281
important
```

Important Variables Plot

```
ggplot(data=Top5nnImportance,mapping=aes(x=Overall,y= reorder(Feature, Overall)))+geom_bar(stat
="identity",fill="blue")+scale_y_discrete(labels=c("Deposit Type Non Refund","Agent 132","Stay i
n Weekend NIghts","Arrival Date Week Number","Arrival Date Year"))+xlab("Importance")+ylab("Vari
ables")+ggtitle("Important Variables from Neural Net")
```

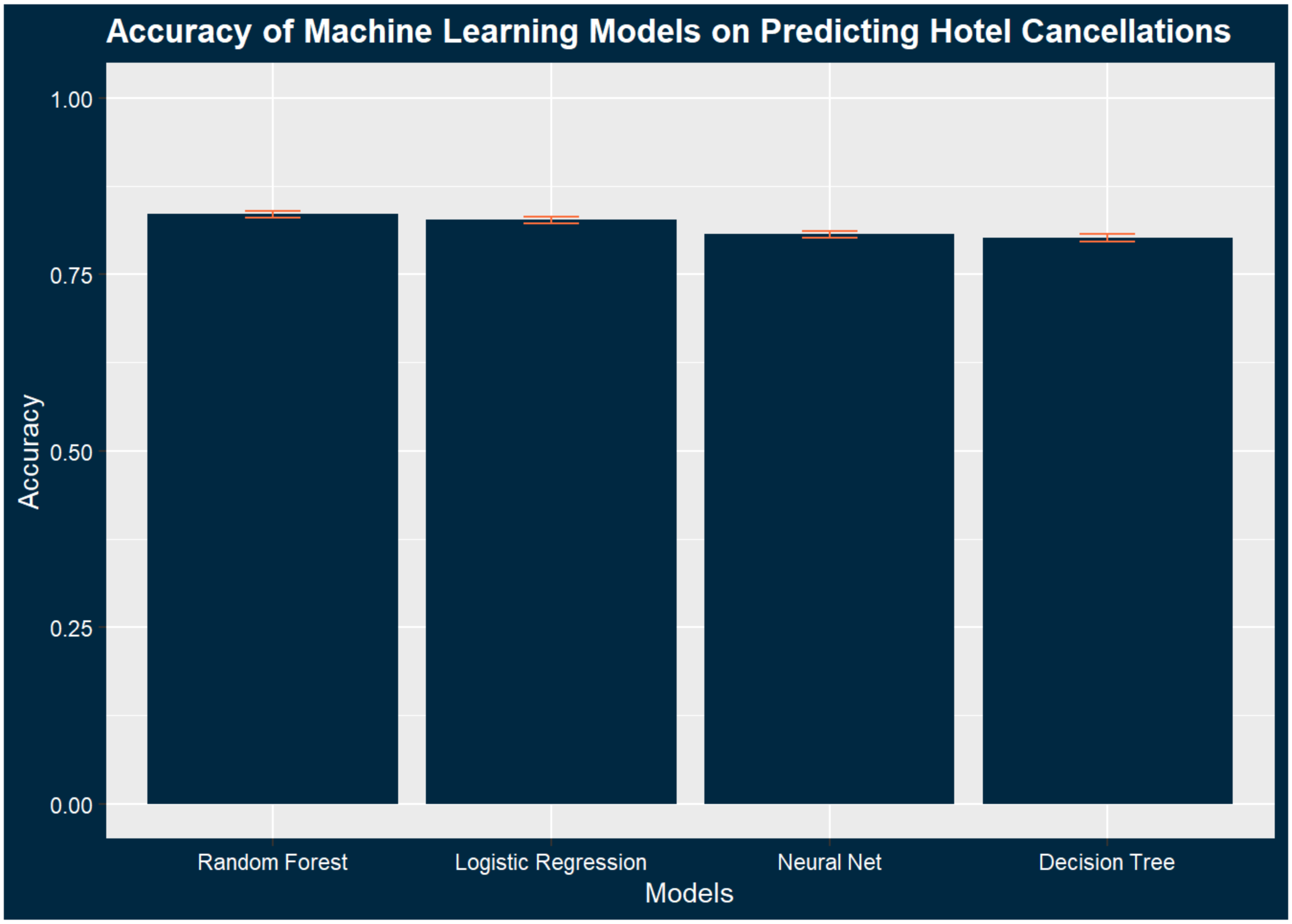


```
accuracy=as.data.frame(rbind(rfAccuracy,treeAccuracy,lgAccuracy,nnaccuracy))
upperbound=as.data.frame(rbind(rfUpperbound,treeUpperbound,lgUpperbound,nnUpperbound))
lowerbound=as.data.frame(rbind(rfLowerbound,treeLowerbound,lgLowerbound,nnLowerbound))
accuracy=cbind(accuracy,upperbound,lowerbound)
#accuracy
names(accuracy) = c("Accuracy","Upper Bound","Lower Bound")
```

```
accuracy=accuracy%>%rownames_to_column()
```

```
names(accuracy) = c("Model", "Accuracy","Upper Bound","Lower Bound")
accuracy=accuracy[order(-accuracy$Accuracy),]
```

```
ggplot(data=accuracy,mapping=aes(x=reorder(Model,-Accuracy),y=Accuracy))+geom_bar(stat="identity",fill="#002845")+
  geom_errorbar(aes(ymin = `Lower Bound`, ymax = `Upper Bound`), width = 0.2,col="#fc723f")+scale_x_discrete(labels=c("rfAccuracy"="Random Forest","lgAccuracy"="Logistic Regression","treeAccuracy"="Decision Tree","nnaccuracy"="Neural Net"))+ggtitle("Accuracy of Machine Learning Models on Predicting Hotel Cancellations")+xlab("Models")+ylab("Accuracy")+ theme(plot.background = element_rect(fill = "#002845"),axis.text=element_text(color = "white"),axis.title = element_text(color = "white"),plot.title=element_text(face = "bold",color = "white"))+coord_cartesian(ylim=c(0,1))
```


Accuracy of Machine Learning Models on Predicting Hotel Cancellations
Accuracy
1.00
0.75
0.50
0.25
0.00
Random Forest
Logistic Regression
Neural Net
Decision Tree
Models