Did a simple train test split of 75%/25% for faster computation. Abandoned the CV file from before. Still takes a while to run (5-10 min).

In terms of accuracy Random forest > Logistic Regression > Decision Tree > Neural Network on Variables from Trees > Neural Network on Variables from Logisitic Regression.

Important variables are Deposit Type, Country, Lead Time, Total of Special Requests, Previous Cancellations in the random forest model.

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=='NUL
L','No Travel Agency',agent))
data=data%>%mutate(agent=ifelse(distribution_channel=="TA/TO" & agent=="NULL","TA/TO No Agent Nu
mber",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") & cou
ntry=='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
```

Correlation Exploration

```
library(corrplot)
numericData=data[sapply(data,is.numeric)]
corr=cor(numericData)
corrplot(corr,method="color")
```

As their isn't any highly correlated columns, no columns will be removed.

Creating New Features

Binning the lead time into quartiles

```
q1LeadTime=quantile(data$lead_time,0.25)
q2LeadTime=quantile(data$lead_time,0.50)
q3LeadTime=quantile(data$lead_time,0.75)
data$lead_timeCategories=cut(data$lead_time,breaks=c(-Inf,q1LeadTime,q2LeadTime,q3LeadTime,Inf),
labels=c("Very Low Lead Time", "Below Average Lead Time", "Above Average Lead Time", "High Lead
Time"))
```

Making a continent Column

```
data$Continent=countrycode(data$country, origin = "iso3c", destination = "continent")
```

```
## Warning: Some values were not matched unambiguously: ATF, CN, TMP, UMI, Unknown
```

```
southAmerica=c("ARG", "BRA", "CHL", "PER", "COL", "VEN", "SUR", "ECU", "GUY", "PRY", "BOL", "GU
Y")

#Manually fixing continent values that the country code couldn't define

#South America is linked together as Americas with North America
data$Continent=ifelse(data$country %in% southAmerica & data$Continent == "Americas","South Ameri
ca",data$Continent)
data$Continent=ifelse(data$country == "ATF", "None",data$Continent)  #French South Territories i
sn't associated with a continent
data$Continent=ifelse(data$country == "CN","Asia",data$Continent) #China
data$Continent=ifelse(data$country == "TMP","Asia",data$Continent) #East Timor, part of ASIA
data$Continent=ifelse(data$country == "UMI","None",data$Continent) #United States Minor Outlying
Islands isn't associated with a continent
data$Continent=ifelse(data$country == "Unknown","Unknown",data$Continent)
```

Making a holiday seasons column (Summer, Chirstmas, New years)

```
data$ArrivalHolidaySeason=cut(data$arrival_date_week_number,breaks=c(-Inf,1,20,26,47,51,Inf),lab
els=c("New Year","Regular","Summer","Regular","Chirstmas","New Year"))
```

Making a seasonal column

```
data=data%>%mutate(ArrivalSeason=case_when(
    arrival_date_month %in% c("December", "January", "February") ~ "Winter",
    arrival_date_month %in% c("March", "April", "May") ~ "Spring",
    arrival_date_month %in% c("June", "July", "August") ~ "Summer",
    arrival_date_month %in% c("September", "October", "November") ~ "Fall")
  )
data$ArrivalSeason=as.factor(data$ArrivalSeason)
```

```
originalData=data#Before removing columns stored original with features engineered for later us
e.

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)

data=subset(data,select=-arrival_date_week_number)#Dropping arrival week number as I used it to
create the Seasonal columns
```

Splitting the data for ML

```
data$is_canceled=as.factor(data$is_canceled)
data$Continent=as.factor(data$Continent)
partition=createDataPartition(data$is_canceled,p=0.75,list=FALSE)
data_train=data[partition,]
data_test=data[-partition,]
```

Train test split

Random Forest

```
rg=train(is_canceled~.,data=data_train,method="ranger",importance = "impurity",num.trees=1000,tr
Control = trainControl(method = "none"))
```

```
## Growing trees.. Progress: 62%. Estimated remaining time: 19 seconds.
## Growing trees.. Progress: 100%. Estimated remaining time: 0 seconds.
```

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

Variable Importance

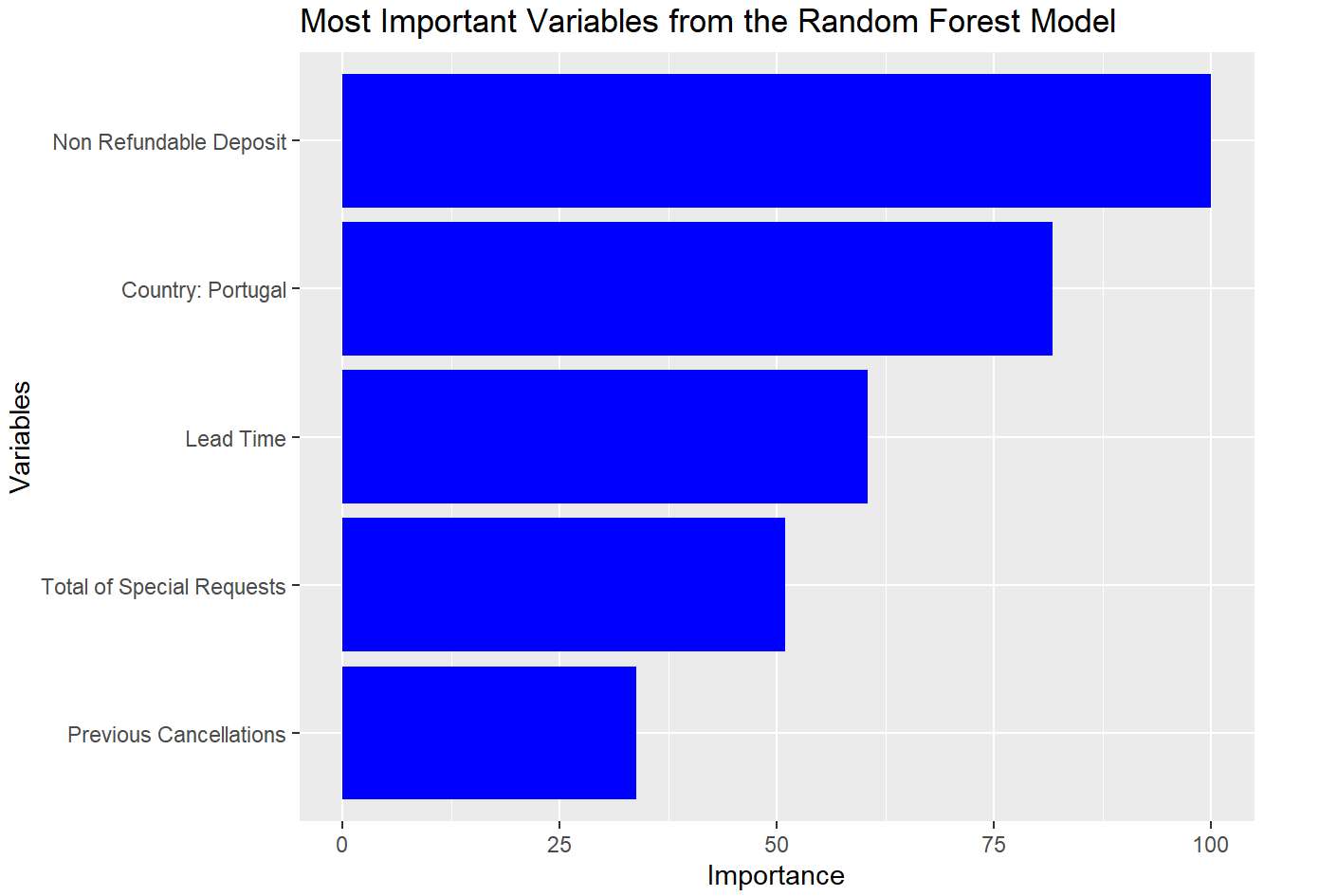
```
rfImportance=varImp(rg)
Top5RfImportance=rfImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% a
rrange(desc(Overall))%>%head(5)
Top5RfImportance
```

```
##                     Feature   Overall
## 1    deposit_typeNon Refund 100.00000
## 2                countryPRT  81.77147
## 3                 lead_time  60.48215
## 4 total_of_special_requests  51.01191
## 5    previous_cancellations  33.82327
```

```
#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="blue")+scale_y_discrete(labels=c("Previous Cancellations","Total of Special Requests","Lead Time","Country: Portugal","Non Refundable Deposit"))+xlab("Importance")+ylab("Variables")+ggtitle("Most Important Variables from the Random Forest Model")
```



Confusion Matrixs

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

```
##
## rgPreds     0     1
##       0 17924  3160
##       1   836  7887
```

Accuracy

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

```
## [1] 0.8659375
```

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] 0.8121582
```

Confusion Matrix

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

```
##
## treePreds     0     1
##         0 17364  4203
##         1  1396  6844
```

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   9606.412
## 2                     agent   5754.454
## 3            market_segment   2584.130
## 4 total_of_special_requests   2063.877
## 5                   country   1236.755
```

Logistic Regression

```
lg=train(is_canceled~.,data=data_train,trControl=trainControl(method="none"),method="multinom",t
race=FALSE)
```

Accuracy

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

```
## [1] 0.8408092
```

Confusion Matrix

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

```
##
## lgPreds     0     1
##       0 16891  2876
##       1  1869  8171
```

Variable Importance

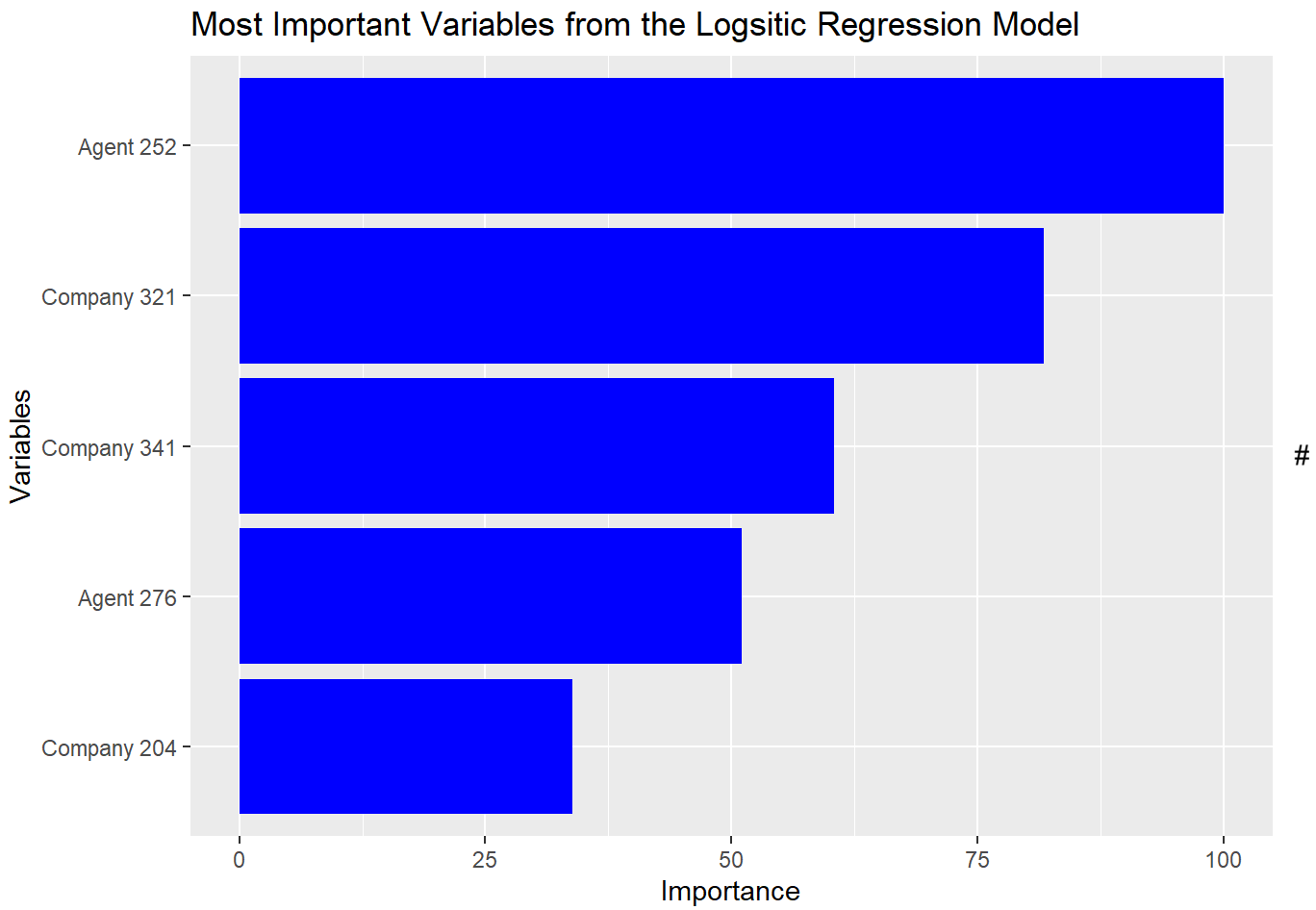
```
lgImportance=varImp(lg)
Top5lgImportance=lgImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% a
rrange(desc(Overall))%>%head(5)
Top5lgImportance
```

```
##      Feature   Overall
## 1   agent252 100.00000
## 2 company321  84.82284
## 3   agent341  79.43282
## 4   agent276  78.54034
## 5 company204  78.12741
```

```
#Found Deposit type:Agent 252, Company 321, Agent 341, Agent 276, Company 204 important
```

Variable Importance Plot

```
ggplot(data=Top5RfImportance,mapping=aes(x=Overall,y= reorder(Feature, Overall)))+geom_bar(stat
="identity",fill="blue")+scale_y_discrete(labels=c("Company 204","Agent 276","Company 341","Comp
any 321","Agent 252"))+xlab("Importance")+ylab("Variables")+ggtitle("Most Important Variables fr
om the Logsitic Regression Model")
```



Neural Net using variables from tree methods

```
nn_trainControl=trainControl(method="none")
nn_tuneGrid=expand.grid(size=1, decay = 0.01)
nnModel=train(is_canceled~lead_time+deposit_type+country+total_of_special_requests+previous_canc
ellations+agent+market_segment,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 16935  3869
##       1  1825  7178
```

Accuracy

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

```
## [1] 0.808971
```

Important variables

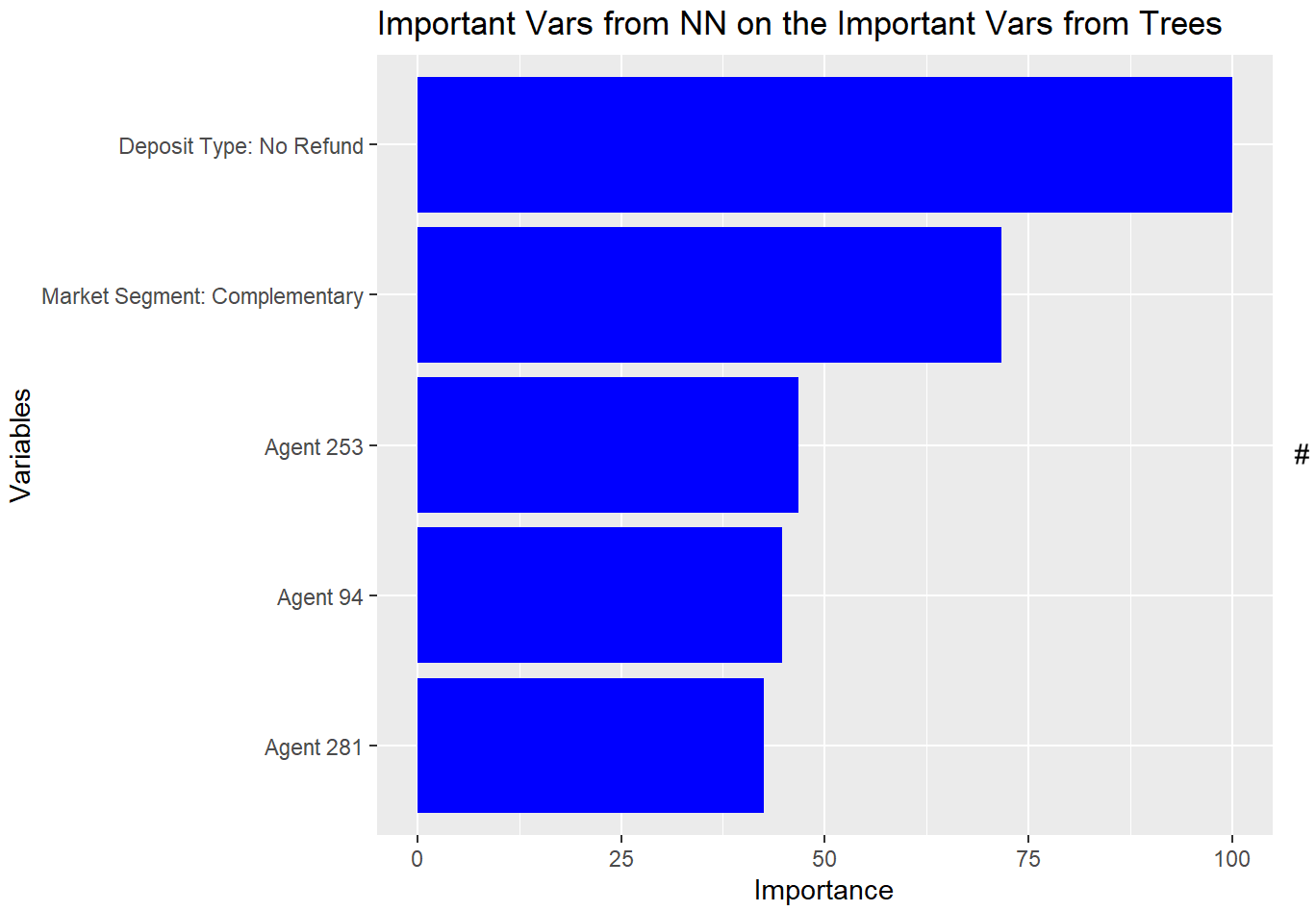
```
nnImportance=varImp(nnModel)
Top5nnImportance=nnImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% a
rrange(desc(Overall))%>%head(5)
Top5nnImportance
```

```
##                  Feature   Overall
## 1   market_segmentDirect 100.00000
## 2 deposit_typeNon Refund  71.75672
## 3 previous_cancellations  46.84072
## 4                agent12  44.78708
## 5                agent40  42.51549
```

```
#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("Agent 281","Agent 94","Agent 253","Market Se
gment: Complementary","Deposit Type: No Refund"))+xlab("Importance")+ylab("Variables")+ggtitle
("Important Vars from NN on the Important Vars from Trees")
```



Neural Net using variables from Logistic Regression

```
NN_trainControl=trainControl(method="none")
NN_tuneGrid=expand.grid(size=1, decay = 0.1)
NNModel=train(is_canceled~company+agent,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 16761  7021
##       1  1999  4026
```

Accuracy

```
(NNAccuracy=mean(NNPreds==data_test$is_canceled))
```

```
## [1] 0.6973865
```

Important variables

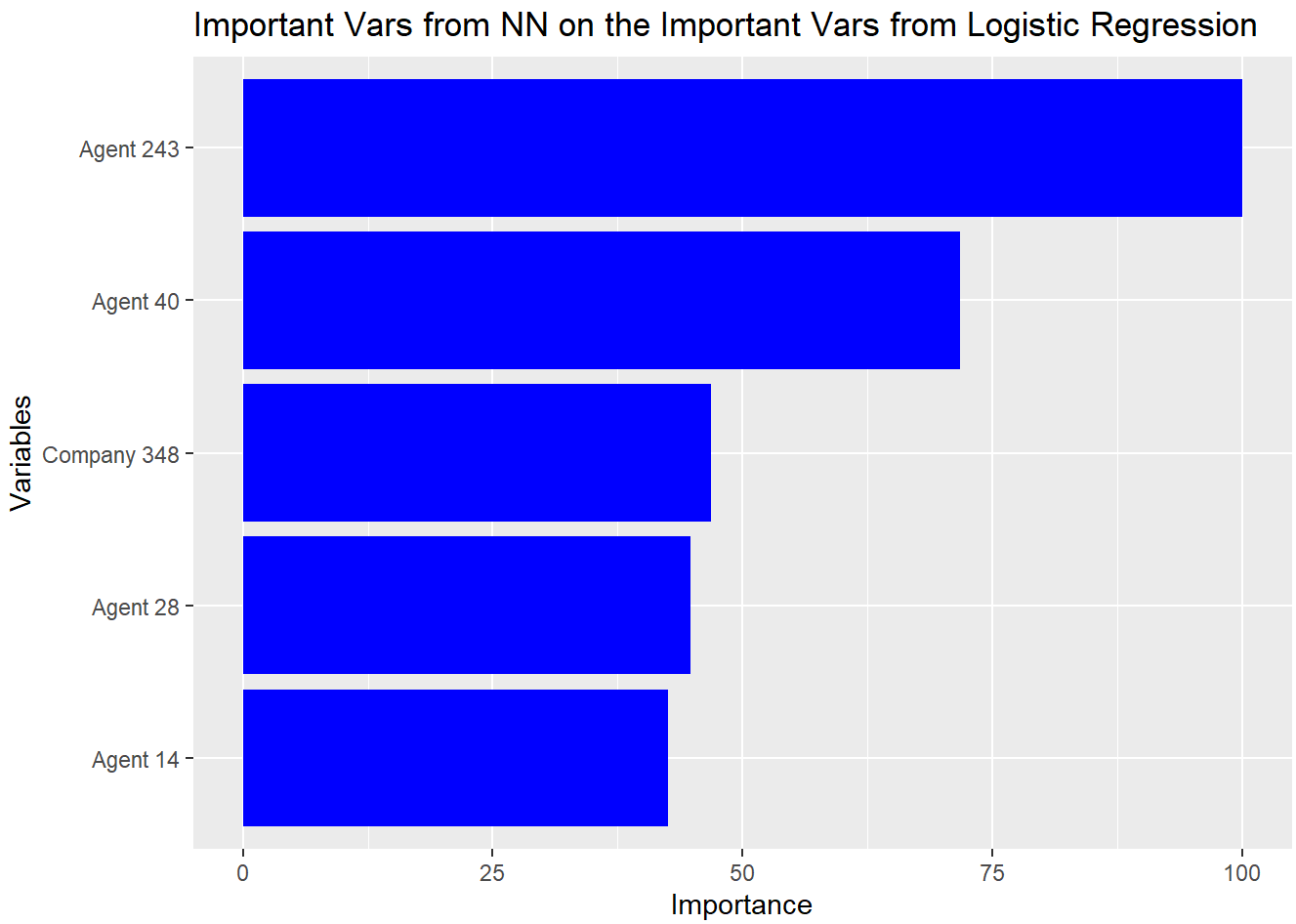
```
NNImportance=varImp(NNModel)
Top5NNImportance=NNImportance$importance%>%as.data.frame()%>%rownames_to_column("Feature") %>% a
rrange(desc(Overall))%>%head(5)
Top5NNImportance
```

```
##                 Feature   Overall
## 1 agentNo Travel Agency 100.00000
## 2            company281  75.49280
## 3     companyNo Company  68.94139
## 4                agent7  55.87487
## 5               agent28  50.40730
```

```
#Found agent 243, agent 40, company 348, agent 28 and agent 14 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("Agent 14","Agent 28","Company 348","Agent 4
0","Agent 243"))+xlab("Importance")+ylab("Variables")+ggtitle("Important Vars from NN on the Imp
ortant Vars from Logistic Regression")
```

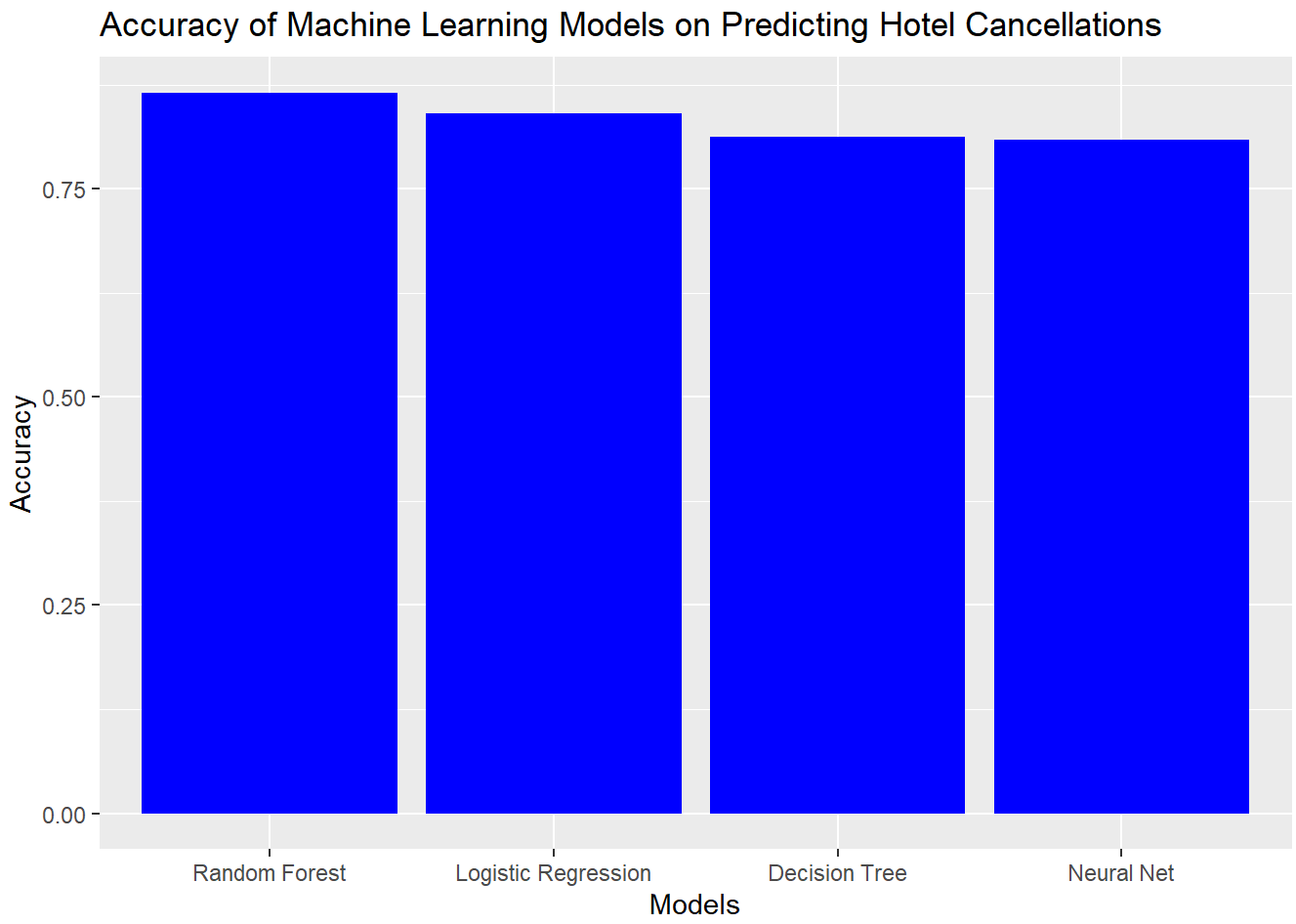



```
accuracy=as.data.frame(rbind(rfAccuracy,treeAccuracy,lgAccuracy,nnaccuracy))

accuracy=accuracy%>%rownames_to_column()
names(accuracy) = c("Model", "Accuracy")
accuracy=accuracy[order(-accuracy$Accuracy),]
accuracy
```

```
##          Model  Accuracy
## 1   rfAccuracy 0.8659375
## 3   lgAccuracy 0.8408092
## 2 treeAccuracy 0.8121582
## 4   nnaccuracy 0.8089710
```

```
ggplot(data=accuracy,mapping=aes(x=reorder(Model,-Accuracy),y=Accuracy))+geom_bar(stat="identit
y",fill="blue")+scale_x_discrete(labels=c("rfAccuracy"="Random Forest","lgAccuracy"="Logistic Re
gression","treeAccuracy"="Decision Tree","nnaccuracy"="Neural Net"))+ggtitle("Accuracy of Machin
e Learning Models on Predicting Hotel Cancellations")+xlab("Models")+ylab("Accuracy")
```


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