Part 2

Kevin Gilmore

2/9/2020

# 

## Create Binary Responses & Identify Important Factors

### Step 1

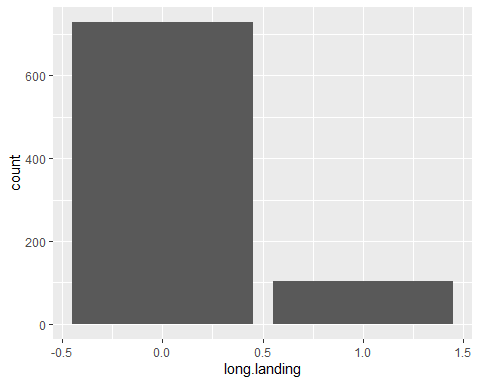
long.landing <- ifelse(filtered$distance > 2500, 1, 0)  
risky.landing <- ifelse(filtered$distance > 3000, 1, 0)  
filtered\_2 <- cbind(filtered[-8], long.landing, risky.landing)

Two new binary variables are created and the data for “distance” is eliminated.

### Step 2

Below is a histogram for distribution of long.landing.

ggplot(filtered\_2,aes(x=long.landing)) + geom\_bar()



### Step 3

variables <- c("aircraft", "duration","no\_pasg" ,"speed\_ground", "speed\_air" ,"height" ,"pitch", "risky.landing")   
coefficient <- rep(NA,length(variables))  
p\_values <- rep(NA,length(variables))  
  
fitted <- data.frame(variables,coefficient,p\_values)  
  
for (i in seq\_along(variables))  
{  
 fitted$coefficient[i] <- summary(glm(filtered\_2$long.landing ~ filtered\_2[[variables[i]]],family=binomial))$coefficients[,1][2]  
 fitted$p\_values[i] <- summary(glm(filtered\_2$long.landing ~ filtered\_2[[variables[i]]],family=binomial))$coefficients[,4][2]  
 }  
  
single\_factor\_table <- fitted %>%   
 mutate(Odds\_ratio= exp(coefficient)) %>%  
 mutate(Direction = ifelse(coefficient >0,'Positive','Negative'))%>%   
 arrange(p\_values)  
single\_factor\_table

## variables coefficient p\_values Odds\_ratio Direction  
## 1 speed\_ground 0.472345755 3.935329e-14 1.603752e+00 Positive  
## 2 speed\_air 0.512321765 4.334124e-11 1.669162e+00 Positive  
## 3 aircraft 0.860999732 8.894411e-05 2.365524e+00 Positive  
## 4 pitch 0.401253005 4.635611e-02 1.493695e+00 Positive  
## 5 height 0.008661495 4.200538e-01 1.008699e+00 Positive  
## 6 no\_pasg -0.007334497 6.021288e-01 9.926923e-01 Negative  
## 7 duration -0.001070492 6.305122e-01 9.989301e-01 Negative  
## 8 risky.landing 21.420072621 9.795377e-01 2.007333e+09 Positive

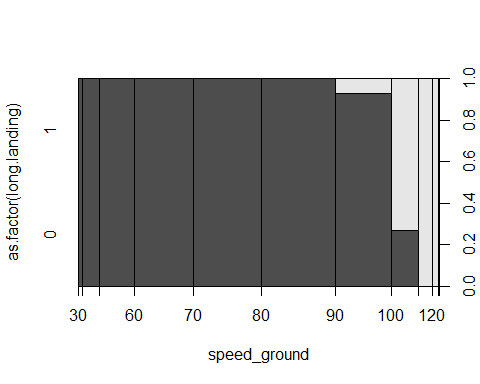
Based on the results; ground speed, air speed, aircraft, and pitch are seen as significant factors.

### Step 4

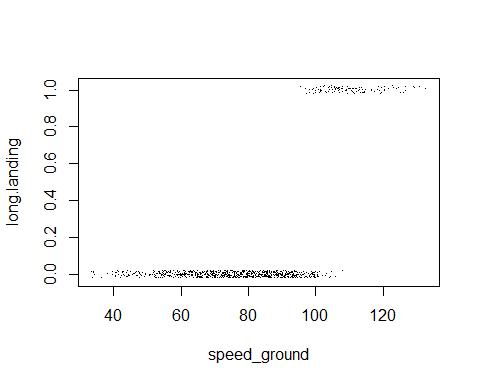
Below are the assocations of significant factors with “long.landing.”

Ground speed

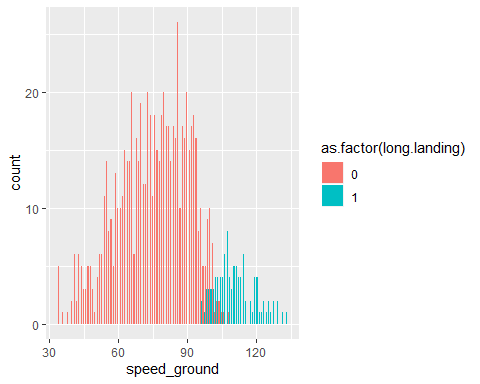
plot(as.factor(long.landing)~ speed\_ground,data = filtered\_2)



plot(jitter(long.landing,0.1)~jitter(speed\_ground),filtered\_2,xlab="speed\_ground",  
 ylab="long.landing",pch=".")

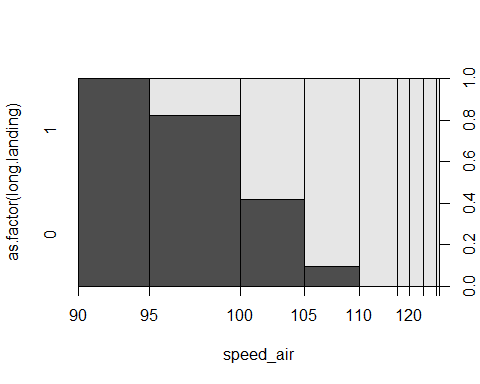


ggplot(filtered\_2,aes(x=speed\_ground,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

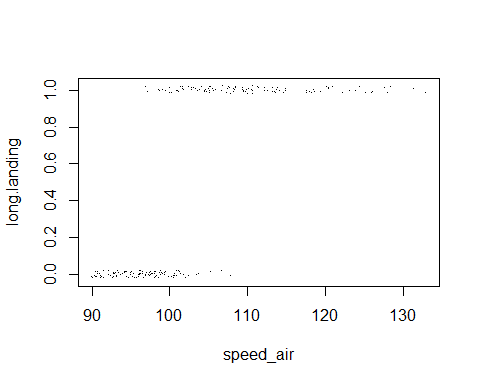


Air Speed

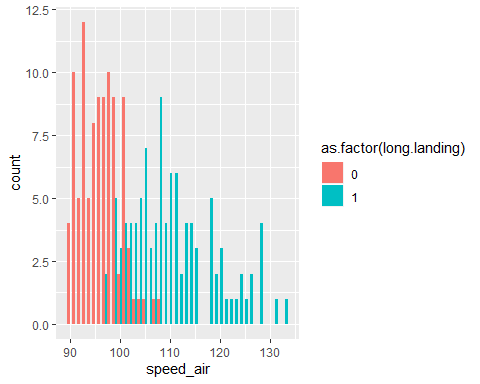
plot(as.factor(long.landing)~ speed\_air,data = filtered\_2)



plot(jitter(long.landing,0.1)~jitter(speed\_air),filtered\_2,xlab="speed\_air",  
 ylab="long.landing",pch=".")

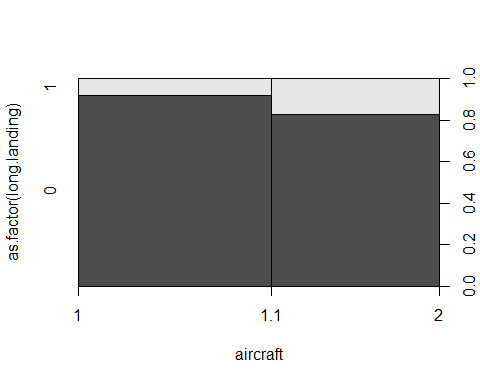


ggplot(filtered\_2,aes(x=speed\_air,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

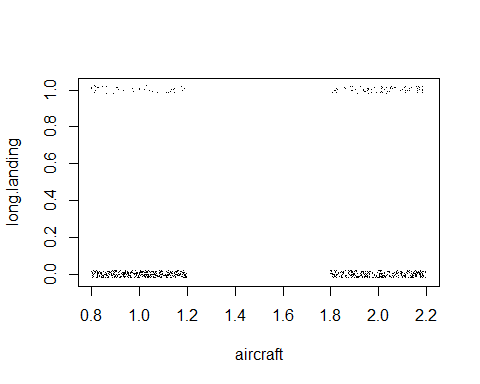


Aircraft

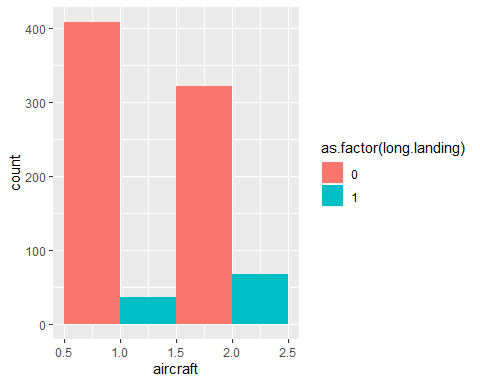
plot(as.factor(long.landing)~ aircraft,data = filtered\_2)



plot(jitter(long.landing,0.1)~jitter(aircraft),filtered\_2,xlab="aircraft",  
 ylab="long.landing",pch=".")

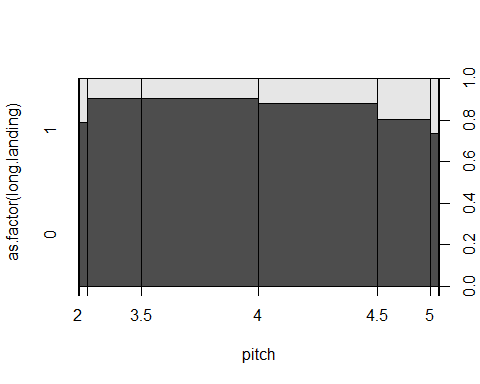


ggplot(filtered\_2,aes(x=aircraft,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

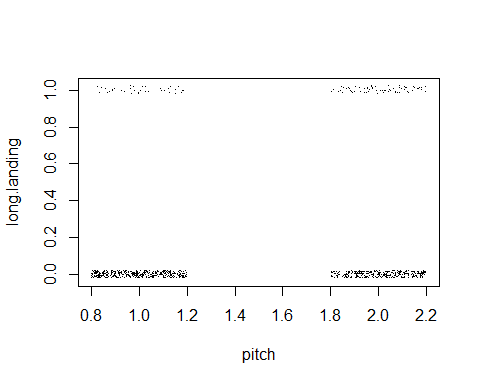


Pitch

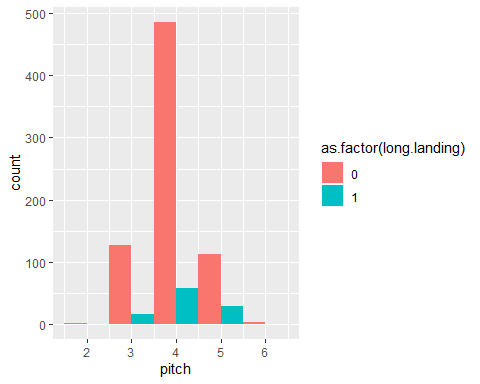
plot(as.factor(long.landing)~ pitch,data = filtered\_2)



plot(jitter(long.landing,0.1)~jitter(aircraft),filtered\_2,xlab="pitch",  
 ylab="long.landing",pch=".")



ggplot(filtered\_2,aes(x=pitch,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)



### Step 5

Air speed is removed because there are over 600 missing values. The full model contains the significant factors of ground speed, aircraft, and pitch in relation to long.landing.

full\_model <- glm(long.landing ~ speed\_ground + aircraft + pitch,  
 family = binomial,data = filtered\_2)  
summary(full\_model)

##   
## Call:  
## glm(formula = long.landing ~ speed\_ground + aircraft + pitch,   
## family = binomial, data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.11589 -0.01114 -0.00026 0.00000 2.40741   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -70.97203 10.88816 -6.518 7.11e-11 \*\*\*  
## speed\_ground 0.61471 0.09184 6.694 2.18e-11 \*\*\*  
## aircraft 3.04348 0.73345 4.150 3.33e-05 \*\*\*  
## pitch 1.06599 0.60389 1.765 0.0775 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 623.043 on 831 degrees of freedom  
## Residual deviance: 81.309 on 828 degrees of freedom  
## AIC: 89.309  
##   
## Number of Fisher Scoring iterations: 10

### Step 6

step\_f <- glm(long.landing~ aircraft+ duration + no\_pasg + speed\_ground + speed\_air + height + pitch, data = filtered\_2, family= "binomial")  
model\_aic <- step(step\_f,trace = 0,direction = "forward")  
summary(model\_aic)

##   
## Call:  
## glm(formula = long.landing ~ aircraft + duration + no\_pasg +   
## speed\_ground + speed\_air + height + pitch, family = "binomial",   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.48853 -0.01367 0.00000 0.00047 1.56917   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.051e+02 5.852e+01 -3.506 0.000455 \*\*\*  
## aircraft 8.784e+00 2.623e+00 3.349 0.000811 \*\*\*  
## duration 3.031e-04 1.048e-02 0.029 0.976919   
## no\_pasg -7.359e-02 7.009e-02 -1.050 0.293744   
## speed\_ground -2.255e-01 3.845e-01 -0.587 0.557471   
## speed\_air 1.985e+00 7.080e-01 2.804 0.005051 \*\*   
## height 4.226e-01 1.429e-01 2.956 0.003116 \*\*   
## pitch 1.469e+00 1.055e+00 1.392 0.163818   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 270.199 on 194 degrees of freedom  
## Residual deviance: 32.909 on 187 degrees of freedom  
## (637 observations deleted due to missingness)  
## AIC: 48.909  
##   
## Number of Fisher Scoring iterations: 10

This model is different from the table obtained in step 3.In this model, aircraft, speed\_air, and height are seen as significant.

### Step 7

model\_bic <- step(step\_f, trace = 0,direction = "forward",criterion = "BIC")  
summary(model\_bic)

##   
## Call:  
## glm(formula = long.landing ~ aircraft + duration + no\_pasg +   
## speed\_ground + speed\_air + height + pitch, family = "binomial",   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.48853 -0.01367 0.00000 0.00047 1.56917   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.051e+02 5.852e+01 -3.506 0.000455 \*\*\*  
## aircraft 8.784e+00 2.623e+00 3.349 0.000811 \*\*\*  
## duration 3.031e-04 1.048e-02 0.029 0.976919   
## no\_pasg -7.359e-02 7.009e-02 -1.050 0.293744   
## speed\_ground -2.255e-01 3.845e-01 -0.587 0.557471   
## speed\_air 1.985e+00 7.080e-01 2.804 0.005051 \*\*   
## height 4.226e-01 1.429e-01 2.956 0.003116 \*\*   
## pitch 1.469e+00 1.055e+00 1.392 0.163818   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 270.199 on 194 degrees of freedom  
## Residual deviance: 32.909 on 187 degrees of freedom  
## (637 observations deleted due to missingness)  
## AIC: 48.909  
##   
## Number of Fisher Scoring iterations: 10

The results are the same as that in the previous step, as height is also seen as significant, altough it is not in step 3.

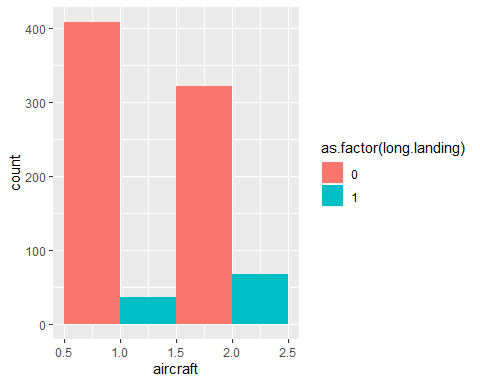
### Step 8

The model I would select will include ground speed, aircraft, and height. Below is the full model and histograms with this specific variables.

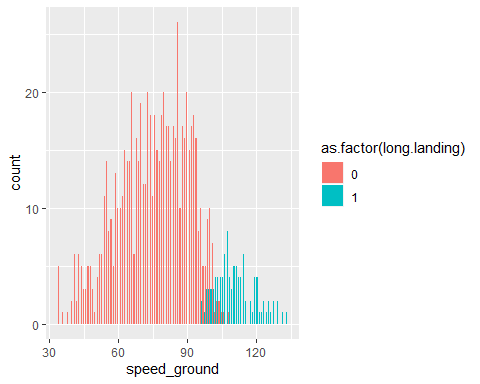
selected\_model <- glm(long.landing ~ speed\_ground + aircraft + height, family = binomial,  
 data = filtered\_2)  
summary(selected\_model)

##   
## Call:  
## glm(formula = long.landing ~ speed\_ground + aircraft + height,   
## family = binomial, data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.43442 -0.00116 0.00000 0.00000 2.57435   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -108.00250 20.13151 -5.365 8.10e-08 \*\*\*  
## speed\_ground 0.92657 0.17242 5.374 7.70e-08 \*\*\*  
## aircraft 5.04813 1.11520 4.527 5.99e-06 \*\*\*  
## height 0.23106 0.05959 3.877 0.000106 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 623.043 on 831 degrees of freedom  
## Residual deviance: 57.047 on 828 degrees of freedom  
## AIC: 65.047  
##   
## Number of Fisher Scoring iterations: 11

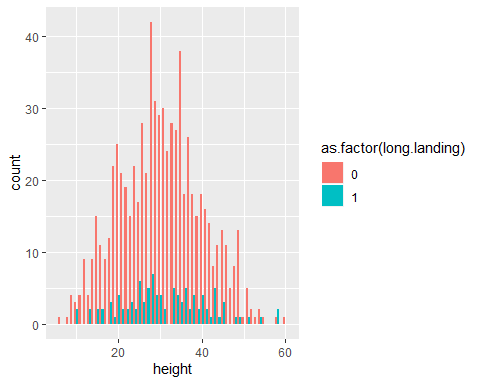
ggplot(filtered\_2,aes(x=aircraft,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)



ggplot(filtered\_2,aes(x=speed\_ground,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)



ggplot(filtered\_2,aes(x=height,fill=as.factor(long.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

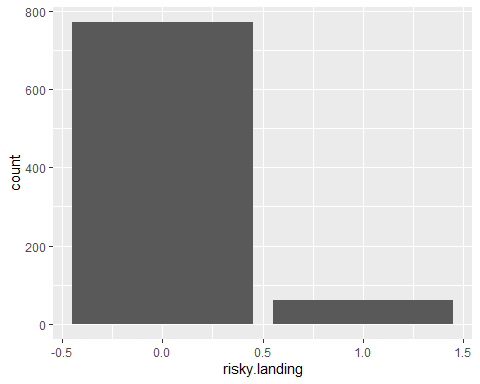


* If I met with a FAA agent I would try to stay away from raw numbers and show figures of the risk factors.
* The chances of long landing are higher in a Boeing
* There are only 103 instances of long landing, and 729 without a long landing
* Once AIC and BIC were examined, it became clear that height was a significant factor
* High ground speed results in long landing distance, which could be a huge risk

## Risky Landing

### Step 9

ggplot(filtered\_2,aes(x=risky.landing)) + geom\_bar()

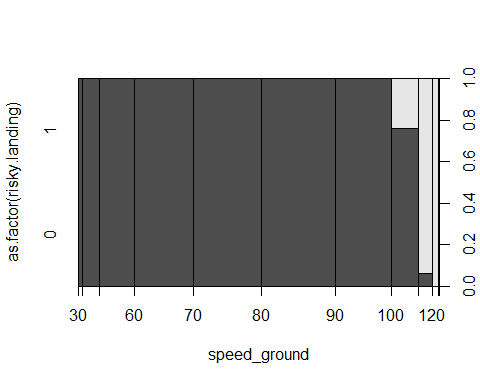


fitted\_risky <- data.frame(variables,coefficient,p\_values)  
  
for (i in seq\_along(variables))  
{  
 fitted\_risky$coefficient[i] <- summary(glm(filtered\_2$risky.landing ~ filtered\_2[[variables[i]]],family=binomial))$coefficients[,1][2]  
 fitted\_risky$p\_values[i] <- summary(glm(filtered\_2$risky.landing ~ filtered\_2[[variables[i]]],family=binomial))$coefficients[,4][2]  
}  
  
single\_factor\_table\_risky <- fitted\_risky %>%   
 mutate(Odds\_ratio= exp(coefficient)) %>%  
 mutate(Direction = ifelse(coefficient >0,'Positive','Negative'))%>%   
 arrange(p\_values)  
single\_factor\_table\_risky

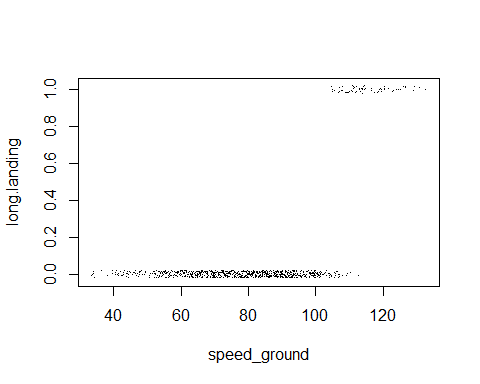
## variables coefficient p\_values Odds\_ratio Direction  
## 1 speed\_ground 0.614218747 6.897994e-08 1.848212e+00 Positive  
## 2 speed\_air 0.870401900 3.728032e-06 2.387870e+00 Positive  
## 3 aircraft 0.998880972 4.733996e-04 2.715242e+00 Positive  
## 4 pitch 0.371773979 1.427375e-01 1.450305e+00 Positive  
## 5 no\_pasg -0.025468021 1.523329e-01 9.748536e-01 Negative  
## 6 duration -0.001151836 6.801987e-01 9.988488e-01 Negative  
## 7 height -0.002193042 8.721356e-01 9.978094e-01 Negative  
## 8 risky.landing 53.132132924 9.991050e-01 1.188481e+23 Positive

Ground speed

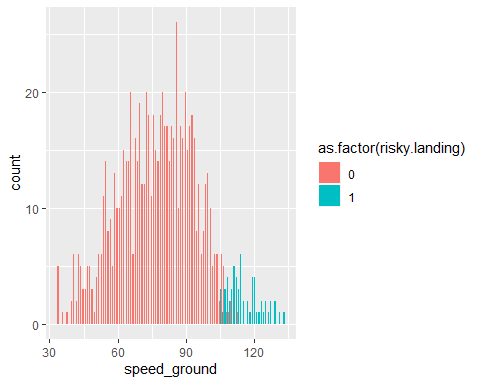
plot(as.factor(risky.landing)~ speed\_ground,data = filtered\_2)



plot(jitter(risky.landing,0.1)~jitter(speed\_ground),filtered\_2,xlab="speed\_ground",  
 ylab="long.landing",pch=".")

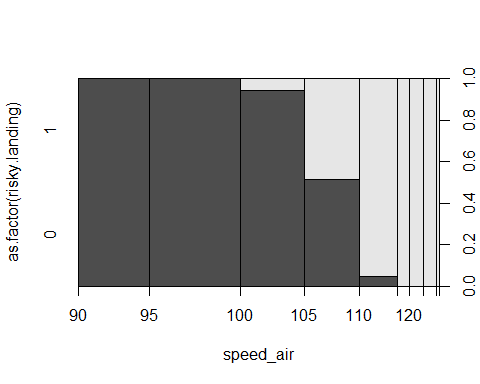


ggplot(filtered\_2,aes(x=speed\_ground,fill=as.factor(risky.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

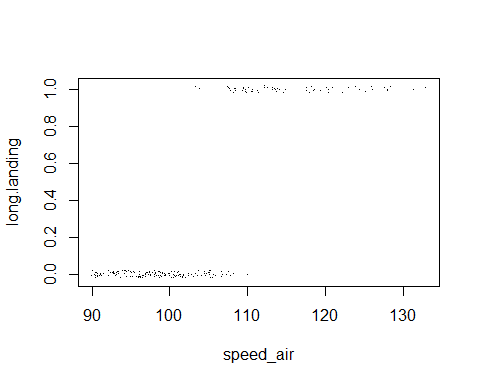


Air Speed

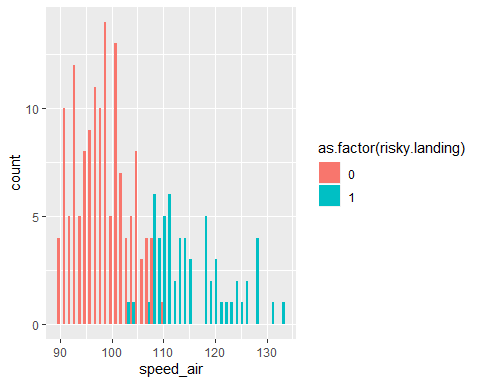
plot(as.factor(risky.landing)~ speed\_air,data = filtered\_2)



plot(jitter(risky.landing,0.1)~jitter(speed\_air),filtered\_2,xlab="speed\_air",  
 ylab="long.landing",pch=".")

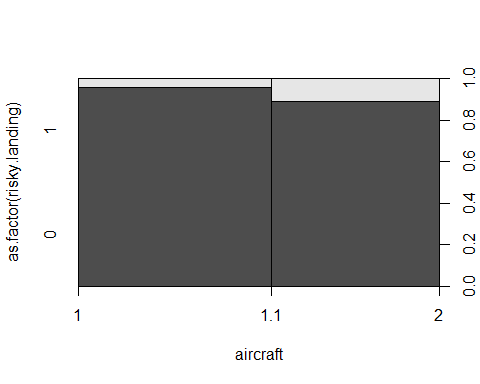


ggplot(filtered\_2,aes(x=speed\_air,fill=as.factor(risky.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

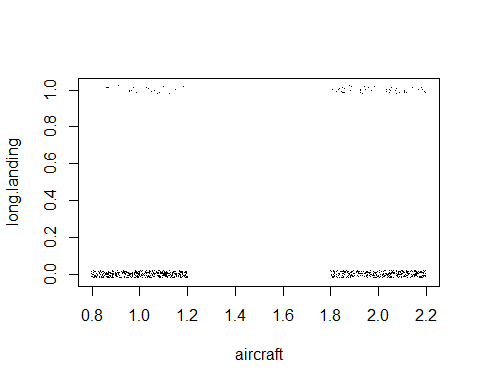


Aircraft

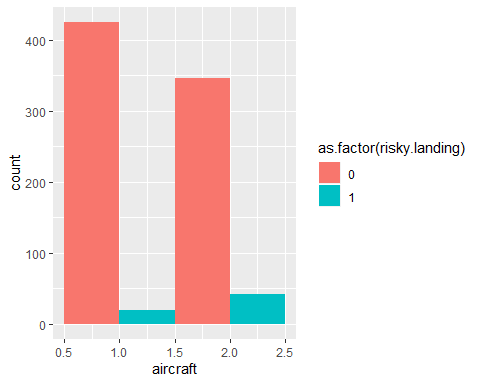
plot(as.factor(risky.landing)~ aircraft,data = filtered\_2)



plot(jitter(risky.landing,0.1)~jitter(aircraft),filtered\_2,xlab="aircraft",  
 ylab="long.landing",pch=".")

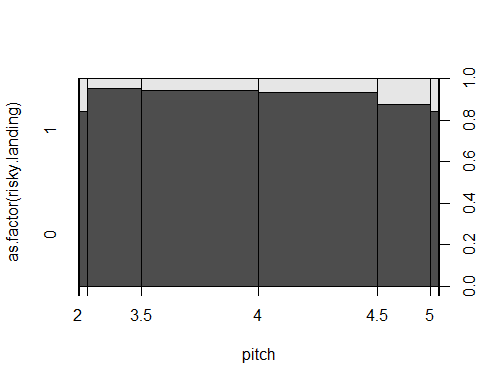


ggplot(filtered\_2,aes(x=aircraft,fill=as.factor(risky.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)

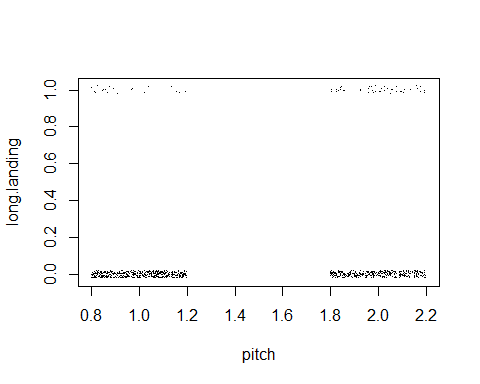


Pitch

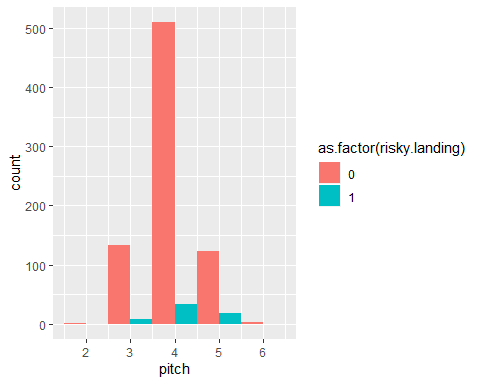
plot(as.factor(risky.landing)~ pitch,data = filtered\_2)



plot(jitter(risky.landing,0.1)~jitter(aircraft),filtered\_2,xlab="pitch",  
 ylab="long.landing",pch=".")



ggplot(filtered\_2,aes(x=pitch,fill=as.factor(risky.landing)))+  
 geom\_histogram(position="dodge",binwidth = 1)



Risky Full Model

full\_model\_2 <- glm(risky.landing ~ speed\_ground + aircraft,  
 family = binomial,data = filtered\_2)  
summary(full\_model\_2)

##   
## Call:  
## glm(formula = risky.landing ~ speed\_ground + aircraft, family = binomial,   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.24398 -0.00011 0.00000 0.00000 1.61021   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -106.0963 25.7459 -4.121 3.77e-05 \*\*\*  
## speed\_ground 0.9263 0.2248 4.121 3.78e-05 \*\*\*  
## aircraft 4.0190 1.2494 3.217 0.0013 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 436.195 on 831 degrees of freedom  
## Residual deviance: 40.097 on 829 degrees of freedom  
## AIC: 46.097  
##   
## Number of Fisher Scoring iterations: 12

Forward

step\_f\_risky <- glm(risky.landing~ aircraft+ duration + no\_pasg + speed\_ground + speed\_air + height + pitch,  
 data = filtered\_2, family= "binomial")  
model\_aic\_risky <- step(step\_f\_risky,trace = 0,direction = "forward")  
summary(model\_aic\_risky)

##   
## Call:  
## glm(formula = risky.landing ~ aircraft + duration + no\_pasg +   
## speed\_ground + speed\_air + height + pitch, family = "binomial",   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.95653 -0.00291 -0.00017 0.00001 2.23576   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -156.74968 50.67835 -3.093 0.00198 \*\*  
## aircraft 7.33037 3.00197 2.442 0.01461 \*   
## duration 0.00198 0.01587 0.125 0.90070   
## no\_pasg -0.12011 0.09589 -1.253 0.21034   
## speed\_ground -0.16366 0.49825 -0.328 0.74256   
## speed\_air 1.61745 0.65439 2.472 0.01345 \*   
## height 0.04535 0.05768 0.786 0.43179   
## pitch -1.31605 1.42985 -0.920 0.35736   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 240.724 on 194 degrees of freedom  
## Residual deviance: 22.144 on 187 degrees of freedom  
## (637 observations deleted due to missingness)  
## AIC: 38.144  
##   
## Number of Fisher Scoring iterations: 10

The full model and the forward AIC show that aircraft and speed\_air are the significant factors in both instances.

BIC

model\_bic\_risky <- step(step\_f\_risky, trace = 0,direction = "forward",criterion = "BIC")  
summary(model\_bic\_risky)

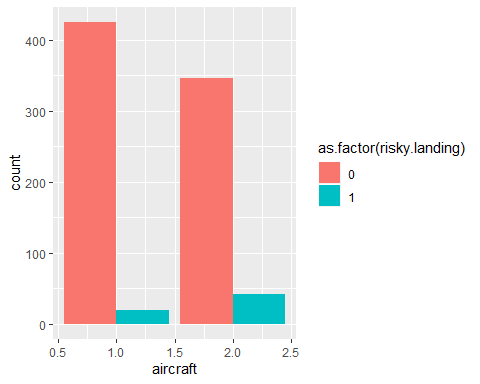
##   
## Call:  
## glm(formula = risky.landing ~ aircraft + duration + no\_pasg +   
## speed\_ground + speed\_air + height + pitch, family = "binomial",   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.95653 -0.00291 -0.00017 0.00001 2.23576   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -156.74968 50.67835 -3.093 0.00198 \*\*  
## aircraft 7.33037 3.00197 2.442 0.01461 \*   
## duration 0.00198 0.01587 0.125 0.90070   
## no\_pasg -0.12011 0.09589 -1.253 0.21034   
## speed\_ground -0.16366 0.49825 -0.328 0.74256   
## speed\_air 1.61745 0.65439 2.472 0.01345 \*   
## height 0.04535 0.05768 0.786 0.43179   
## pitch -1.31605 1.42985 -0.920 0.35736   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 240.724 on 194 degrees of freedom  
## Residual deviance: 22.144 on 187 degrees of freedom  
## (637 observations deleted due to missingness)  
## AIC: 38.144  
##   
## Number of Fisher Scoring iterations: 10

The results in the AIC and BIC are the same for risky.landing.

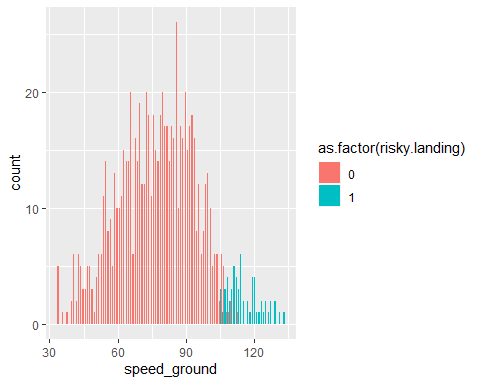
### Step 10

The best model for risky.landing contains ground speed and aircraft.

full\_model\_2 <- glm(risky.landing ~ speed\_ground + aircraft,  
 family = binomial,data = filtered\_2)  
  
ggplot(filtered\_2,aes(x=aircraft,fill=as.factor(risky.landing)))+  
 geom\_bar(position="dodge")



ggplot(filtered\_2,aes(x=speed\_ground,fill=as.factor(risky.landing)))+  
 geom\_histogram(position="dodge",binwidth=1)



count(risky.landing)

## x freq  
## 1 0 771  
## 2 1 61

* We recieve the same model for both AIC and BIC in risky.landing.
* There were 61 entries of long landings and 771 with normal landings
* Boeing shows more chance of long landing

## Model Comparison

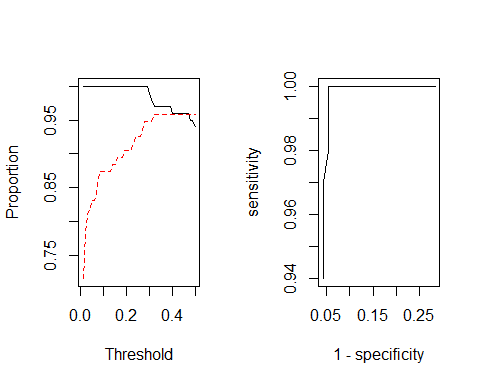
### Step 11

* Model 1 shows three signficant factors of risk: aircraft, ground speed, and height.
* Model two shows only two significant factors of risk: ground\_speed and aircaft.
* The first model decision was changed based on the AIC and BIC criterion, however model 2 stayed consistent.
* AIC for model 2 is 38.144, while it is 65.047 for model 1.

### Step 12

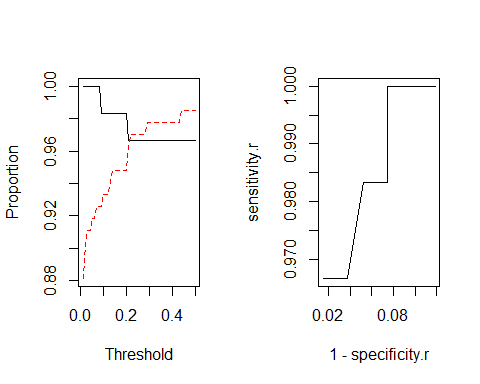
long.landing

thresh <- seq(0.01,0.5,0.01)  
predprob <- predict(model\_aic, type = "response")  
predprob\_risky <- predict(model\_aic\_risky, type = "response")  
long.landing\_na <- long.landing[!is.na(filtered\_2$speed\_air) & !is.na(filtered\_2$duration)]  
  
sensitivity <- specificity<-rep(NA,length(thresh))  
  
for(j in seq(along=thresh)) {  
 pp<-ifelse(predprob < thresh[j], "no", "yes")  
 xx<-xtabs(~long.landing\_na + pp, filtered\_2)  
 specificity[j] <- xx[1,1]/(xx[1,1] + xx[1,2])  
 sensitivity[j] <- xx[2,2]/(xx[2,1] + xx[2,2])  
}  
  
par(mfrow= c(1,2))  
matplot(thresh,cbind(sensitivity, specificity), type="l", xlab="Threshold", ylab="Proportion", lty=1:2)  
plot(1-specificity, sensitivity, type="l");abline(0, 1, lty=2)



Risky.landing

thresh <- seq(0.01, 0.5, 0.01)  
predprob <- predict(model\_aic, type = "response")  
predprob\_risky <- predict(model\_aic\_risky, type = "response")  
risky.landing\_na <- risky.landing[!is.na(filtered\_2$speed\_air) & !is.na(filtered\_2$duration)]  
  
sensitivity.r <- specificity.r <- rep(NA,length(thresh))  
  
for(j in seq(along=thresh)) {  
 pp <- ifelse(predprob\_risky < thresh[j], "no", "yes")  
 xx <- xtabs(~risky.landing\_na + pp, filtered\_2)  
 specificity.r[j] <- xx[1, 1]/(xx[1, 1] + xx[1, 2])  
 sensitivity.r[j] <- xx[2, 2]/(xx[2, 1] + xx[2, 2])  
}  
par(mfrow=c(1,2))  
matplot(thresh,cbind(sensitivity.r,specificity.r),type="l",xlab="Threshold",ylab="Proportion",lty=1:2)  
plot(1-specificity.r,sensitivity.r,type="l");abline(0,1,lty=2)



There seems to be a relationship between the two plots, as when one increases, the other decreases.

### Step 13

ilogit <- function (x) {  
 exp(x)/(1+exp(x))  
}  
new\_ind\_risky<-data.frame(aircraft=2, duration=200, no\_pasg=80, speed\_ground=115, speed\_air=120, height=40, pitch=4)

Predicted probability for risky landing

predict(full\_model\_2,newdata=new\_ind\_risky,type="link",se=T)

## $fit  
## 1   
## 8.463332   
##   
## $se.fit  
## [1] 2.089366  
##   
## $residual.scale  
## [1] 1

round(ilogit(c(8.463332-1.96\*2.089367,8.463332+1.96\*2.089367)),3)

## [1] 0.987 1.000

Long Landing

new\_ind\_long<-data.frame(aircraft=1, duration=200, no\_pasg=80, speed\_ground=115, speed\_air=120, height=40, pitch=4)

Predicted probability for long landing

predict(full\_model,newdata=new\_ind\_long,type="link",se=T)

## $fit  
## 1   
## 7.026765   
##   
## $se.fit  
## [1] 1.211391  
##   
## $residual.scale  
## [1] 1

round(ilogit(c(7.026765-1.96\*1.211391,7.026765+1.96\*1.211391)),3)

## [1] 0.991 1.000

## Different link functions

### Step 14

risky\_logit <- glm(risky.landing ~ aircraft + speed\_ground, data = filtered\_2, family = binomial(link = logit))  
  
risky\_probit <- glm(risky.landing ~ aircraft + speed\_ground, data = filtered\_2, family = binomial(link = probit))  
  
risky\_cloglog<-glm(risky.landing ~ speed\_ground + aircraft,family=binomial(link=cloglog),filtered\_2)  
  
risky\_cloglog\_summary <- summary(risky\_cloglog)  
risky\_cloglog\_summary

##   
## Call:  
## glm(formula = risky.landing ~ speed\_ground + aircraft, family = binomial(link = cloglog),   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.24103 -0.00182 -0.00004 0.00000 1.67963   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -72.1639 15.3685 -4.696 2.66e-06 \*\*\*  
## speed\_ground 0.6221 0.1326 4.690 2.74e-06 \*\*\*  
## aircraft 2.8984 0.8002 3.622 0.000292 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 436.195 on 831 degrees of freedom  
## Residual deviance: 41.443 on 829 degrees of freedom  
## AIC: 47.443  
##   
## Number of Fisher Scoring iterations: 13

risky\_logit\_summary <- summary(risky\_logit)  
risky\_logit\_summary

##   
## Call:  
## glm(formula = risky.landing ~ aircraft + speed\_ground, family = binomial(link = logit),   
## data = filtered\_2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.24398 -0.00011 0.00000 0.00000 1.61021   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -106.0963 25.7459 -4.121 3.77e-05 \*\*\*  
## aircraft 4.0190 1.2494 3.217 0.0013 \*\*   
## speed\_ground 0.9263 0.2248 4.121 3.78e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 436.195 on 831 degrees of freedom  
## Residual deviance: 40.097 on 829 degrees of freedom  
## AIC: 46.097  
##   
## Number of Fisher Scoring iterations: 12

risky\_prohibit\_summary <- risky\_probit  
risky\_prohibit\_summary

##   
## Call: glm(formula = risky.landing ~ aircraft + speed\_ground, family = binomial(link = probit),   
## data = filtered\_2)  
##   
## Coefficients:  
## (Intercept) aircraft speed\_ground   
## -61.0498 2.3567 0.5322   
##   
## Degrees of Freedom: 831 Total (i.e. Null); 829 Residual  
## Null Deviance: 436.2   
## Residual Deviance: 39.44 AIC: 45.44

round(coefficients(risky\_logit\_summary),4)

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -106.0963 25.7459 -4.1209 0.0000  
## aircraft 4.0190 1.2494 3.2168 0.0013  
## speed\_ground 0.9263 0.2248 4.1206 0.0000

round(coefficients(risky\_prohibit\_summary),4)

## (Intercept) aircraft speed\_ground   
## -61.0498 2.3567 0.5322

round(coefficients(risky\_cloglog\_summary),4)

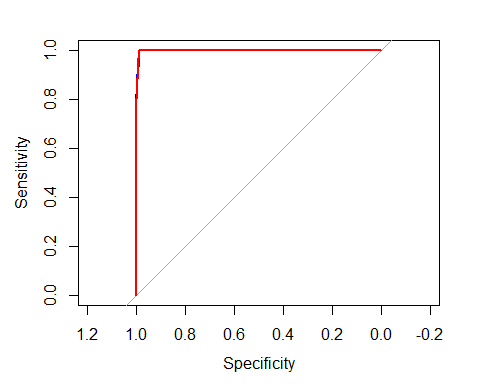
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -72.1639 15.3685 -4.6956 0e+00  
## speed\_ground 0.6221 0.1326 4.6898 0e+00  
## aircraft 2.8984 0.8002 3.6221 3e-04

The logit model coefficients seem to be nearly double of prohibit. The null deviance is nearly identical in all three models.

### Step 15

pred\_logit <- predict(risky\_logit, type = "response")  
pred\_probit <- predict(risky\_probit, type = "response")  
pred\_cloglog <- predict(risky\_cloglog, type = "response")

ROC\_logit <- roc(filtered\_2$risky.landing, pred\_logit)  
ROC\_probit <- roc(filtered\_2$risky.landing, pred\_probit)  
ROC\_cloglog <- roc(filtered\_2$risky.landing, pred\_cloglog)  
  
plot(ROC\_logit)  
lines(ROC\_probit, col= 'blue')  
lines(ROC\_cloglog, col= "red")



All three ROC curves are very closely overlapping, thus making it hard to differentiate.

### Step 16

logit\_index <- sort(as.numeric(names(tail(sort(pred\_logit), 5))))  
logit\_sort <- filtered\_2[logit\_index,1:7]  
logit\_sort

## aircraft duration no\_pasg speed\_ground speed\_air height pitch  
## 64 2 161.89247 72 129.2649 128.4177 33.94900 4.139951  
## 307 2 154.52460 67 129.3072 127.5933 23.97850 5.154699  
## 362 2 63.32952 52 132.7847 132.9115 18.17703 4.110664  
## 387 2 153.83445 61 126.8393 126.1186 20.54783 4.334558  
## 408 1 131.73110 60 131.0352 131.3379 28.27797 3.660194

prohibit\_index <-sort(as.numeric(names(tail(sort(pred\_probit),5))))  
prohibit\_sort <- filtered\_2[prohibit\_index,1:7]  
prohibit\_sort

## aircraft duration no\_pasg speed\_ground speed\_air height pitch  
## 362 2 63.32952 52 132.7847 132.9115 18.17703 4.110664  
## 383 2 99.68150 61 121.8371 120.9534 33.18460 3.867476  
## 387 2 153.83445 61 126.8393 126.1186 20.54783 4.334558  
## 408 1 131.73110 60 131.0352 131.3379 28.27797 3.660194  
## 643 1 137.58573 66 126.2443 127.9371 35.17570 2.701924

cloglog\_index <- sort(as.numeric(names(tail(sort(pred\_cloglog), 5))))  
cloglog\_sort <- filtered\_2[cloglog\_index,1:7]  
cloglog\_sort

## aircraft duration no\_pasg speed\_ground speed\_air height pitch  
## 643 1 137.58573 66 126.2443 127.9371 35.17570 2.701924  
## 669 1 140.45311 75 120.4189 118.4847 31.26345 2.796731  
## 751 1 175.51443 49 125.2123 125.1385 22.52478 4.365772  
## 765 1 220.05713 61 120.5579 118.2882 15.66566 4.111265  
## 769 1 98.50031 66 123.3105 124.3908 22.32718 4.276710

risky.df <- cbind(logit\_index,prohibit\_index,cloglog\_index)  
colnames(risky.df)<- c("Logit","Prohibit","Cloglog")  
risky.df

## Logit Prohibit Cloglog  
## [1,] 64 362 643  
## [2,] 307 383 669  
## [3,] 362 387 751  
## [4,] 387 408 765  
## [5,] 408 643 769

In logit and prohibit, 362, 387, and 408 share commmon values. In prohibit and cloglog, 643 is the only value.

### Step 17

predict(risky\_logit,newdata = new\_ind\_risky,type = "link",se=T)

## $fit  
## 1   
## 8.463332   
##   
## $se.fit  
## [1] 2.089366  
##   
## $residual.scale  
## [1] 1

round(ilogit(c(8.463332-1.96\*2.089366,8.463332+1.96\*2.089366)),3)

## [1] 0.987 1.000

predict(risky\_probit,newdata = new\_ind\_risky,type = "link",se=T)

## $fit  
## 1   
## 4.872041   
##   
## $se.fit  
## [1] 1.127891  
##   
## $residual.scale  
## [1] 1

round(ilogit(c(4.872041-1.96\*1.127891,4.872041+1.96\*1.27891)),3)

## [1] 0.935 0.999

predict(risky\_cloglog,newdata = new\_ind\_risky,type = "link",se=T)

## $fit  
## 1   
## 5.16944   
##   
## $se.fit  
## [1] 1.173423  
##   
## $residual.scale  
## [1] 1

round(ilogit(c(5.16944-1.96\*1.173423,5.16944+1.96\*1.173423)),3)

## [1] 0.946 0.999

Logit has the same confidence interval, while prohibit and cloglog are 0.935 and 0.946 instead of 0.987.