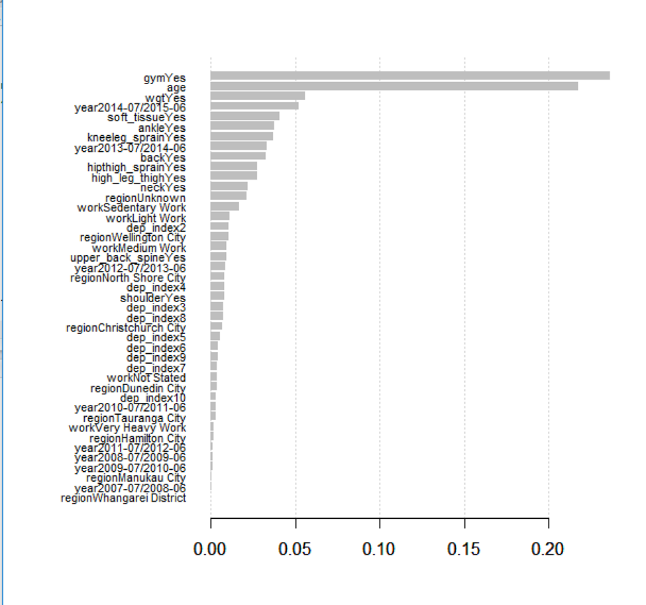
**High level summary:**

There are some insights gained from the data which could be easily used in the app to target clients having a future gym injury. An obvious indicator would be if a person had previous gym or weightlifting injuries (highest indicator in data) .The other top injury factors as per dataset shows are if they had a soft tissue or ankle injury previously.

Another logical find is that people who had made an ACC claim recently in the last 2 financial years would have a high likelihood of having a gym injury in the future. The target age group for this app should be at a younger age as age.

Other less / influencing variables but could also be used are back, neck or hip thigh sprains as a flag.

The graph below shows the ranking of the other variables in terms of importance.

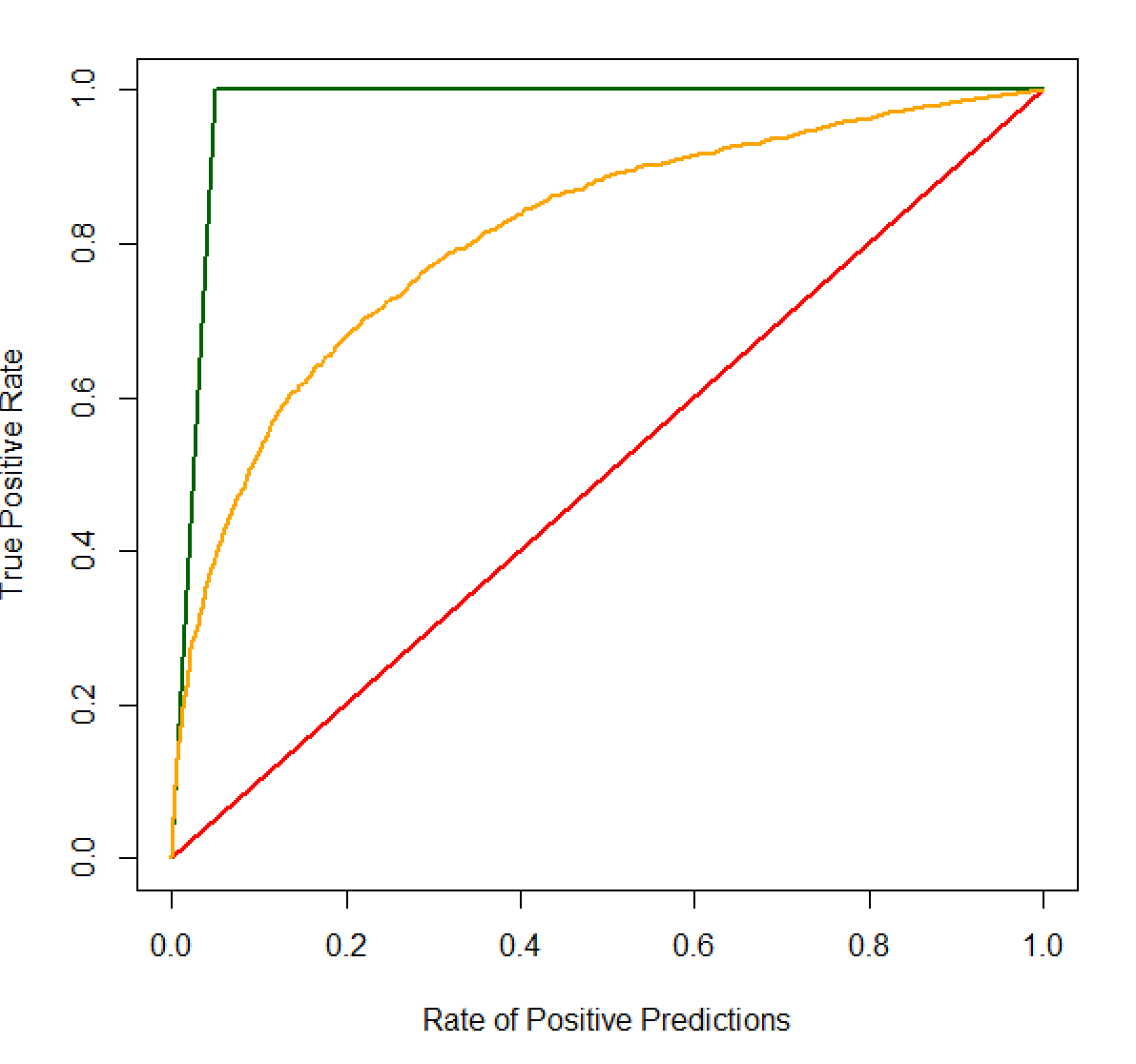


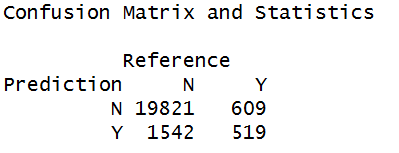
If we randomly chose 4% of a target list of clients without any model –4% of the selection would have an injury in the next year.

Now if we were to deploy the created model into a campaign management system of all ACC clients– what we would find is if we take the top scored 4% ~ 1000 individuals, 30% would have a gym injury in the next year. Therefore this is a lift of 7.5.

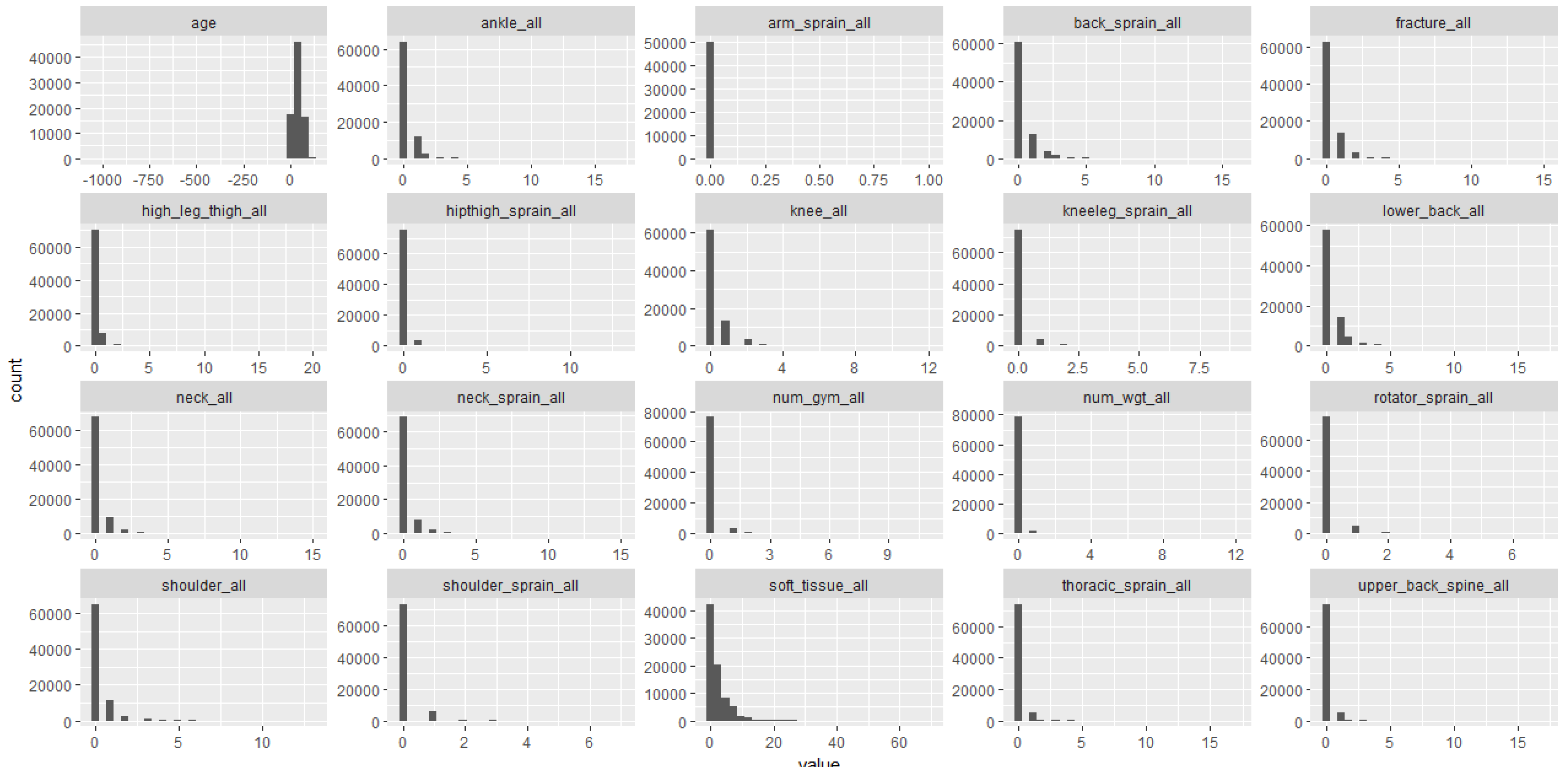
See graph below where the orange is the fitted model, green is the perfect model where roughly 5% of the population have all gym injuries and the red line is if we don’t do any model and leave it to chance.

Based on this, using a model for the campaign management system to target clients would 7.5 times better than randomly selection hence would save cost.





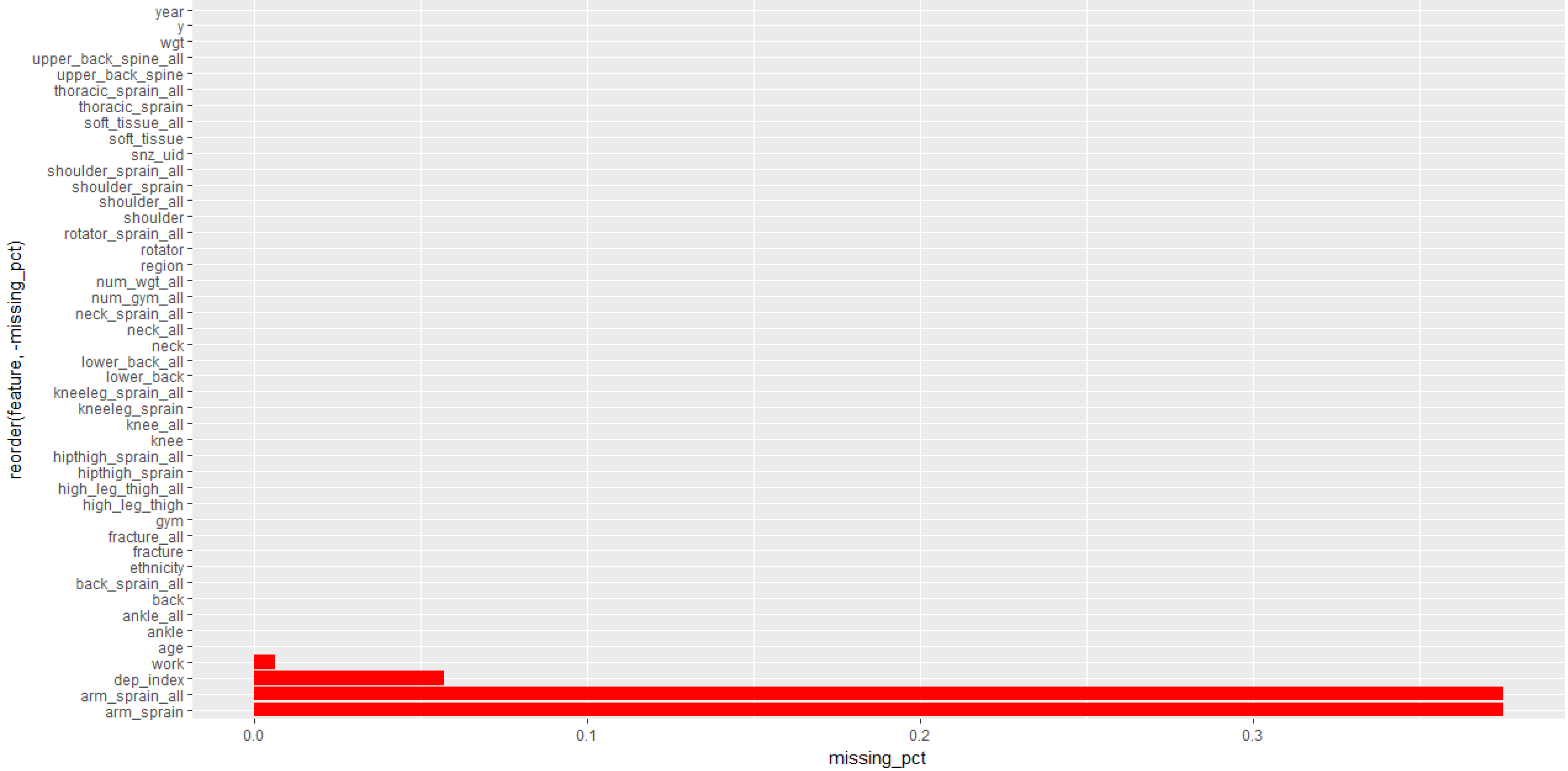
**Appendix/ Methodology**



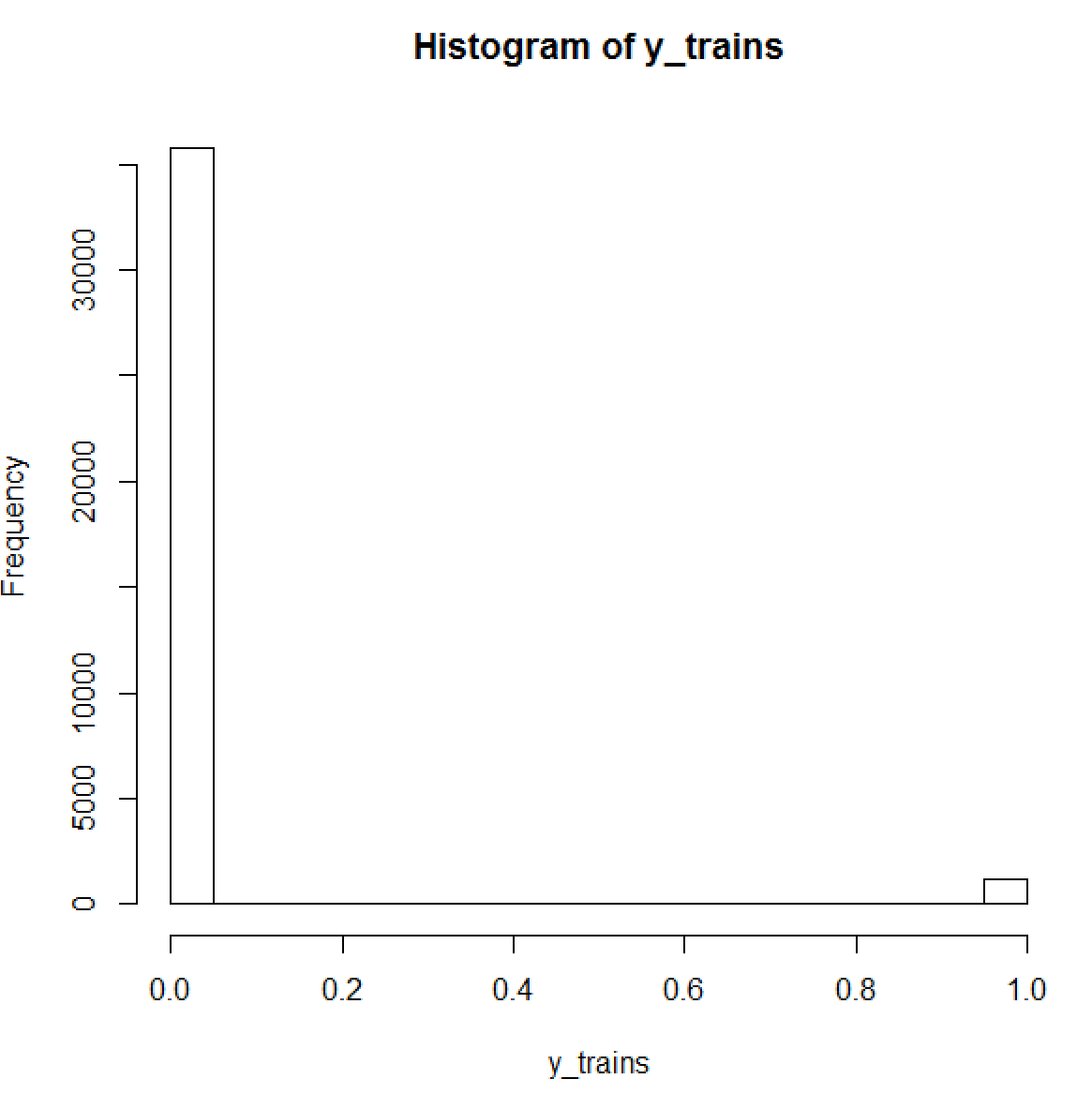
Raw data – of Dataset – Numerical variables to identify outliers

* Noticed lot of weird data especially for age – will apply scale smoothing here i.e trim means

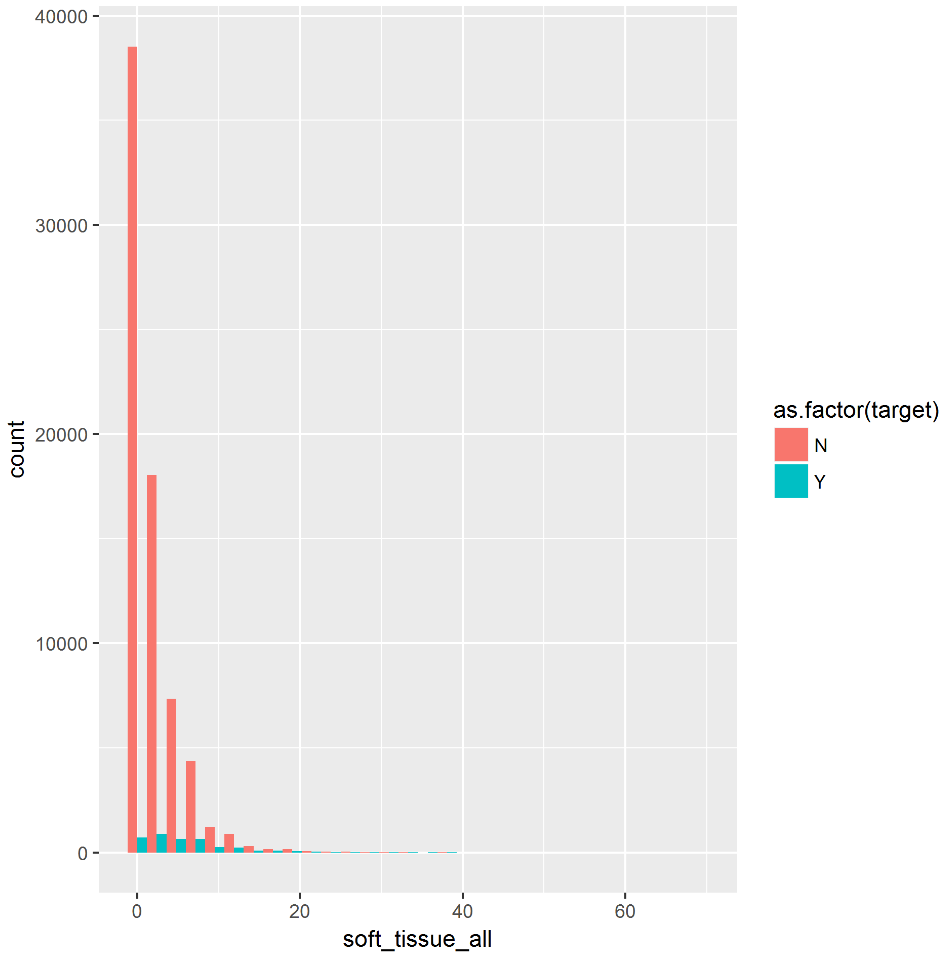
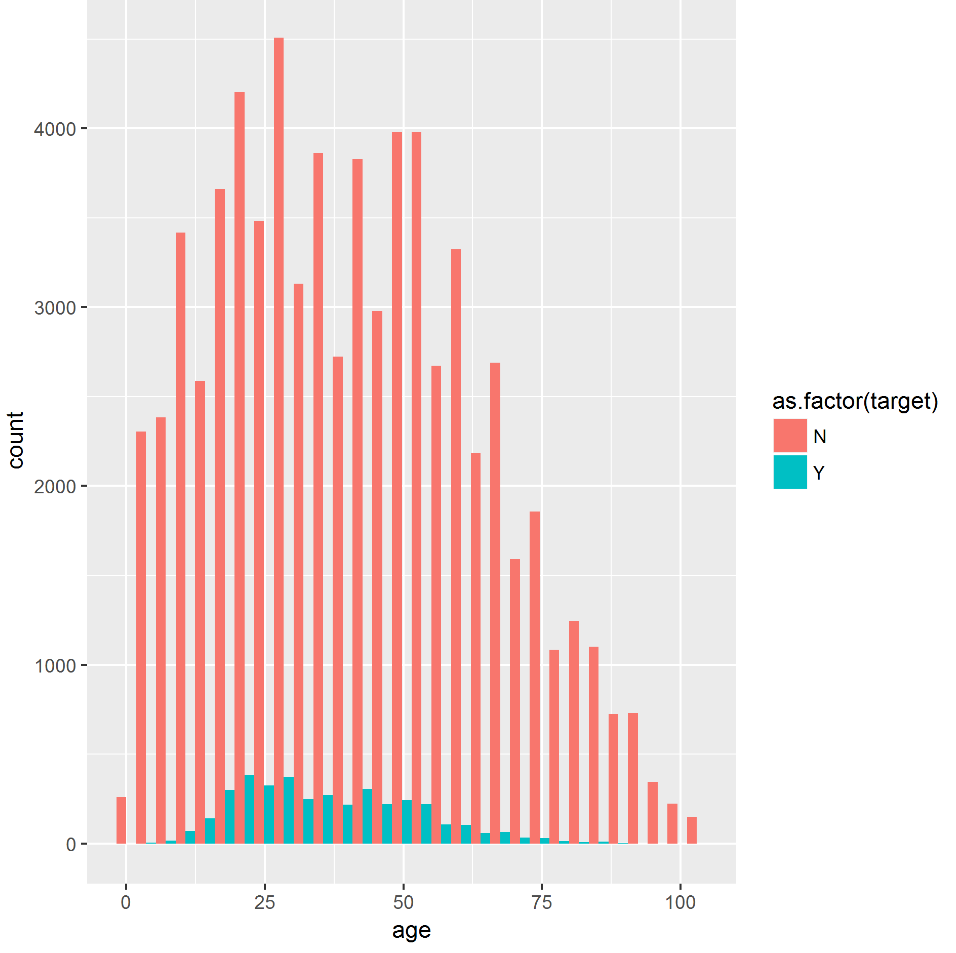
Looking at all data – we notice here as well that Arm sprain all contains a lot of missing values – will remove this variable rather than try and impute it

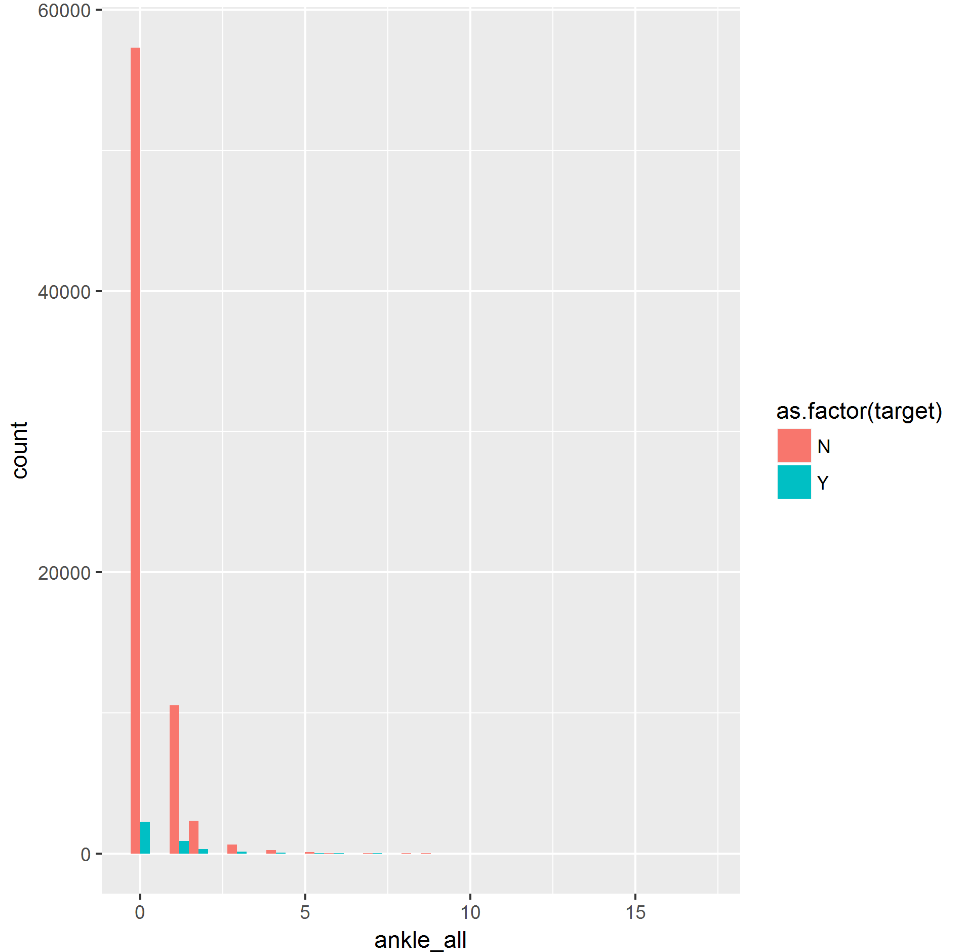


Looking at the rate of injuries – numbers are quite low so dealing with an unbalanced dataset – decided here to do SMOTE which deals with unbalanced datasets.

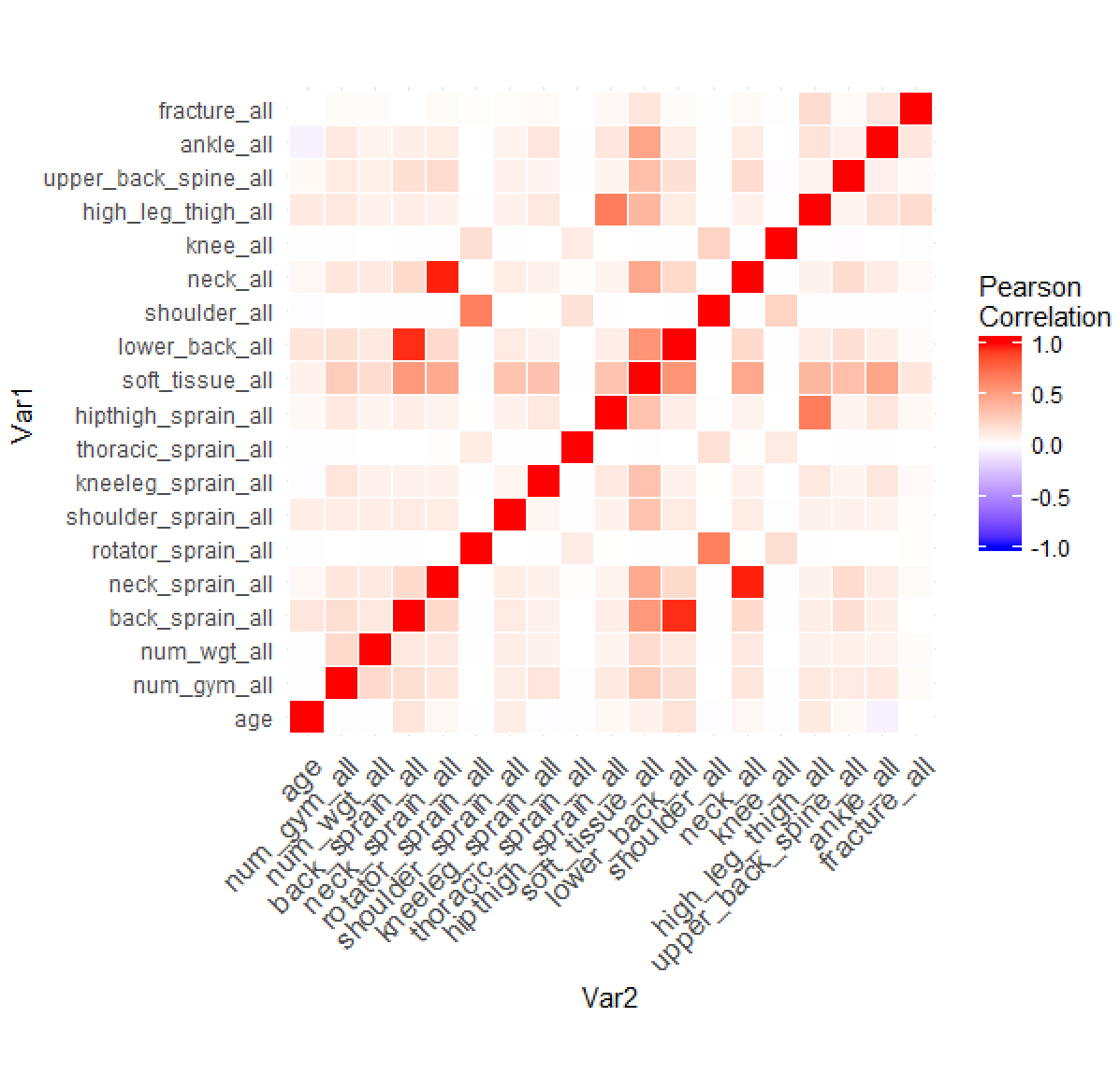


Did a quick comparison of the distributions of those who had Gym injuries vs those who don’t in terms of numerical variables – younger population and certain type of injuries seem to have a different distribution here



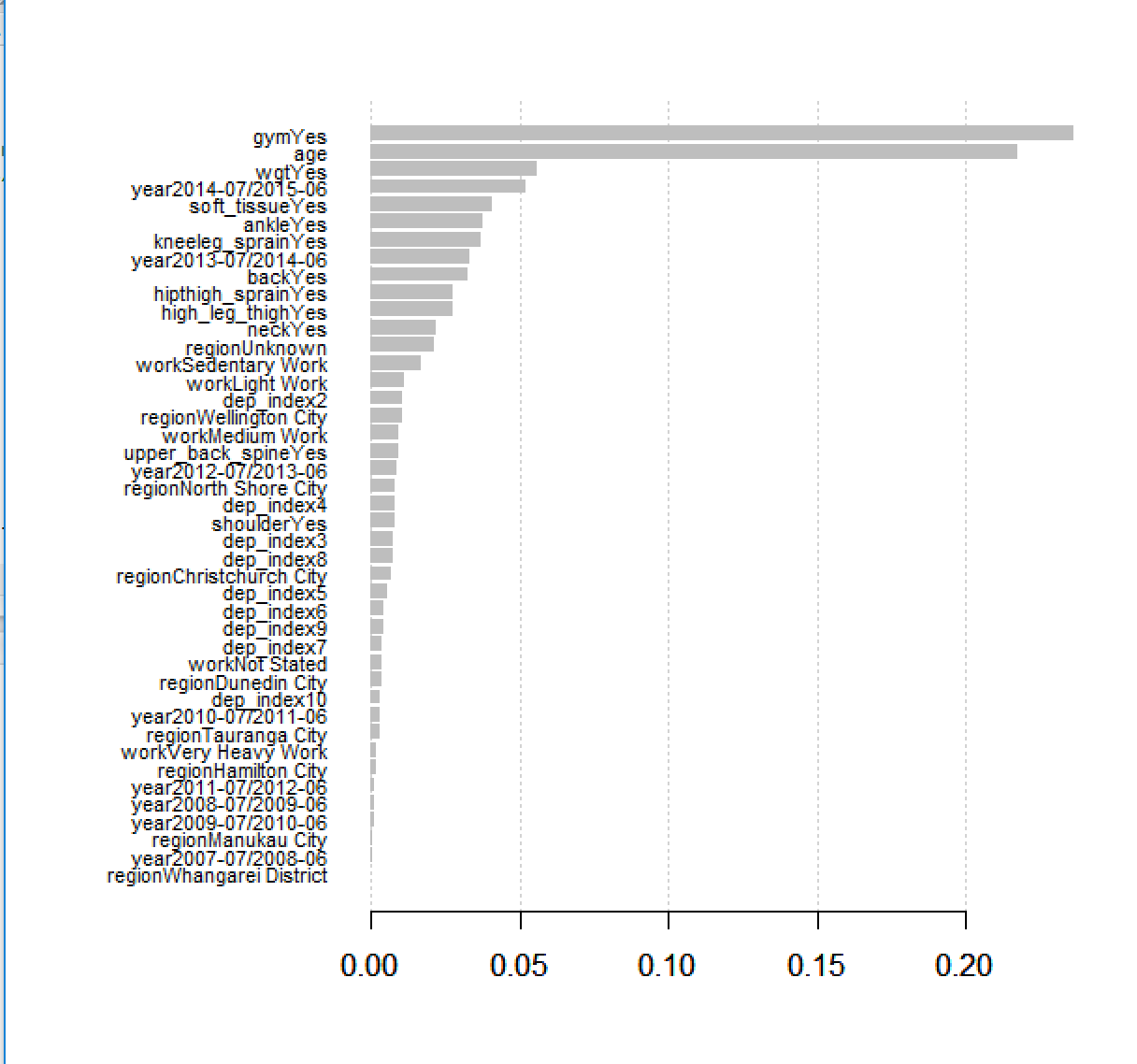


Looking at the numerical data we want to remove variables that are correlated to each other in order to avoid multicollinearity



Next splitting data into training and validation set we then now run a gradient boosting with random forest to determine which variables are of importance if we are to create a predictive targeting tool.

If we look at the variables to see which variables would have the biggest explanatory power of injuries – we could see whether someone has had a Gym and Weights, followed by Age as an important indicator. We note as well that there were a– we could therefore create the app based on those factors –



Identifying which way the variables go based on coeffecients

> yy<-glm(y ~ . , family=binomial,data=train\_data)

> summary(yy)

Call:

glm(formula = y ~ ., family = binomial, data = train\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.9850 -0.3168 -0.2194 -0.1433 3.5543

Coefficients: (1 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.550330 0.309128 -14.720 < 2e-16 \*\*\*

age -0.016455 0.001240 -13.275 < 2e-16 \*\*\*

ethnicityEuropean 0.081014 0.081598 0.993 0.320787

ethnicityMaori 0.003554 0.102434 0.035 0.972324

ethnicityOther Ethnicity 0.054519 0.124280 0.439 0.660894

ethnicityPacific Peoples -0.028365 0.117197 -0.242 0.808757

ethnicityResidual Categories 0.112137 0.156231 0.718 0.472903

regionChristchurch City -0.304286 0.086689 -3.510 0.000448 \*\*\*

regionDunedin City -0.225195 0.134585 -1.673 0.094278 .

regionHamilton City -0.269386 0.120698 -2.232 0.025621 \*

regionManukau City -0.250209 0.160533 -1.559 0.119088

regionNorth Shore City -0.020300 0.127235 -0.160 0.873240

regionTauranga City -0.251700 0.121541 -2.071 0.038368 \*

regionUnknown -0.413354 0.055938 -7.390 1.47e-13 \*\*\*

regionWellington City 0.123001 0.098341 1.251 0.211020

regionWhangarei District -0.268024 0.161858 -1.656 0.097739 .

workLight Work 0.442033 0.088141 5.015 5.30e-07 \*\*\*

workMedium Work 0.177146 0.085550 2.071 0.038389 \*

workNot Stated 0.117787 0.221128 0.533 0.594267

workSedentary Work 0.020894 0.076172 0.274 0.783857

workVery Heavy Work -0.358551 0.157159 -2.281 0.022522 \*

dep\_index2 -0.024053 0.089658 -0.268 0.788485

dep\_index3 -0.082209 0.093303 -0.881 0.378267

dep\_index4 -0.052517 0.093805 -0.560 0.575579

dep\_index5 -0.259570 0.097223 -2.670 0.007589 \*\*

dep\_index6 -0.295907 0.100391 -2.948 0.003203 \*\*

dep\_index7 -0.262039 0.096450 -2.717 0.006591 \*\*

dep\_index8 -0.246234 0.095895 -2.568 0.010237 \*

dep\_index9 -0.264301 0.096743 -2.732 0.006295 \*\*

dep\_index10 -0.516410 0.109396 -4.721 2.35e-06 \*\*\*

num\_gym\_all 0.391667 0.071608 5.470 4.51e-08 \*\*\*

num\_wgt\_all 0.051244 0.117519 0.436 0.662802

back\_sprain\_all 0.028893 0.023437 1.233 0.217645

neck\_sprain\_all 0.014539 0.032916 0.442 0.658716

shoulder\_sprain\_all 0.080461 0.045445 1.771 0.076643 .

kneeleg\_sprain\_all -0.037835 0.066665 -0.568 0.570341

thoracic\_sprain\_all -0.199799 0.152456 -1.311 0.190015

hipthigh\_sprain\_all 0.018068 0.073297 0.247 0.805295

shoulder\_all -0.076668 0.063458 -1.208 0.226980

knee\_all -0.007521 0.053450 -0.141 0.888103

upper\_back\_spine\_all 0.008491 0.069994 0.121 0.903445

ankle\_all 0.016285 0.032951 0.494 0.621140

fracture\_all -0.087757 0.052357 -1.676 0.093717 .

year2006-07/2007-06 0.024588 0.363855 0.068 0.946123

year2007-07/2008-06 0.096557 0.335272 0.288 0.773350

year2008-07/2009-06 0.177940 0.320373 0.555 0.578610

year2009-07/2010-06 -0.066362 0.328528 -0.202 0.839918

year2010-07/2011-06 0.697575 0.295141 2.364 0.018102 \*

year2011-07/2012-06 0.497824 0.294990 1.688 0.091489 .

year2012-07/2013-06 0.809992 0.284041 2.852 0.004349 \*\*

year2013-07/2014-06 1.065756 0.277528 3.840 0.000123 \*\*\*

year2014-07/2015-06 1.130374 0.274963 4.111 3.94e-05 \*\*\*

year2015-07/2016-06 1.303624 0.273105 4.773 1.81e-06 \*\*\*

gymYes 0.917185 0.108625 8.444 < 2e-16 \*\*\*

wgtYes 1.052350 0.170347 6.178 6.50e-10 \*\*\*

backYes 0.518174 0.130987 3.956 7.62e-05 \*\*\*

rotatorYes 0.029202 0.105625 0.276 0.782186

shoulder\_sprainYes NA NA NA NA

kneeleg\_sprainYes 0.625408 0.105707 5.916 3.29e-09 \*\*\*

thoracic\_sprainYes 0.199789 0.199906 0.999 0.317594

hipthigh\_sprainYes 0.100472 0.137040 0.733 0.463459

soft\_tissueYes 0.925393 0.102215 9.053 < 2e-16 \*\*\*

lower\_backYes -0.085890 0.123786 -0.694 0.487769

shoulderYes 0.174806 0.100040 1.747 0.080576 .

neckYes 0.358081 0.068897 5.197 2.02e-07 \*\*\*

kneeYes 0.011180 0.090976 0.123 0.902198

high\_leg\_thighYes 0.261498 0.075236 3.476 0.000509 \*\*\*

upper\_back\_spineYes -0.090280 0.114538 -0.788 0.430571

ankleYes 0.230840 0.068993 3.346 0.000820 \*\*\*

fractureYes 0.123312 0.088528 1.393 0.163646

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1