BAN502CourseProjectM3WEG

Winslow Goins

3/6/2022

library(tidyverse)  
library(tidymodels)  
library(mice)   
library(VIM)   
library(naniar)   
library(skimr)   
library(GGally)  
library(lmtest)  
library(MASS)  
library(car)  
library(ggcorrplot)  
library(glmnet)  
library(lubridate)  
library(splines)  
library(gridExtra)  
library(e1071)  
library(ROCR)  
library(ggcorrplot)  
library(lmtest)  
library(caret)  
library(rpart)  
library(rpart.plot)  
library(rattle)  
library(RColorBrewer)  
library(gridExtra)  
library(vip)  
library(ranger)  
library(skimr)  
library(e1071)  
library(xgboost)  
library(usemodels)  
library(nnet)  
library(stacks)  
library(leaps)   
library(splines)

ames\_student<- read\_csv("ames\_student.csv")

summary(ames\_student)

## MS\_SubClass MS\_Zoning Lot\_Frontage Lot\_Area   
## Length:2053 Length:2053 Min. : 0.00 Min. : 1300   
## Class :character Class :character 1st Qu.: 43.00 1st Qu.: 7500   
## Mode :character Mode :character Median : 62.00 Median : 9548   
## Mean : 57.38 Mean : 10258   
## 3rd Qu.: 78.00 3rd Qu.: 11600   
## Max. :313.00 Max. :215245   
## Street Alley Lot\_Shape Land\_Contour   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Utilities Lot\_Config Land\_Slope Neighborhood   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Condition\_1 Condition\_2 Bldg\_Type House\_Style   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Overall\_Qual Overall\_Cond Year\_Built Year\_Remod\_Add  
## Length:2053 Length:2053 Min. :1875 Min. :1950   
## Class :character Class :character 1st Qu.:1953 1st Qu.:1965   
## Mode :character Mode :character Median :1972 Median :1993   
## Mean :1971 Mean :1984   
## 3rd Qu.:2000 3rd Qu.:2004   
## Max. :2010 Max. :2010   
## Roof\_Style Roof\_Matl Exterior\_1st Exterior\_2nd   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Mas\_Vnr\_Type Mas\_Vnr\_Area Exter\_Qual Exter\_Cond   
## Length:2053 Min. : 0.0 Length:2053 Length:2053   
## Class :character 1st Qu.: 0.0 Class :character Class :character   
## Mode :character Median : 0.0 Mode :character Mode :character   
## Mean : 103.8   
## 3rd Qu.: 164.0   
## Max. :1600.0   
## Foundation Bsmt\_Qual Bsmt\_Cond Bsmt\_Exposure   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## BsmtFin\_Type\_1 BsmtFin\_SF\_1 BsmtFin\_Type\_2 BsmtFin\_SF\_2   
## Length:2053 Min. :1.00 Length:2053 Min. : 0.00   
## Class :character 1st Qu.:3.00 Class :character 1st Qu.: 0.00   
## Mode :character Median :3.00 Mode :character Median : 0.00   
## Mean :4.21 Mean : 52.57   
## 3rd Qu.:7.00 3rd Qu.: 0.00   
## Max. :7.00 Max. :1526.00   
## Bsmt\_Unf\_SF Total\_Bsmt\_SF Heating Heating\_QC   
## Min. : 0.0 Min. : 0 Length:2053 Length:2053   
## 1st Qu.: 226.0 1st Qu.: 793 Class :character Class :character   
## Median : 460.0 Median : 988 Mode :character Mode :character   
## Mean : 561.2 Mean :1055   
## 3rd Qu.: 801.0 3rd Qu.:1304   
## Max. :2336.0 Max. :5095   
## Central\_Air Electrical First\_Flr\_SF Second\_Flr\_SF   
## Length:2053 Length:2053 Min. : 432 Min. : 0.0   
## Class :character Class :character 1st Qu.: 882 1st Qu.: 0.0   
## Mode :character Mode :character Median :1088 Median : 0.0   
## Mean :1168 Mean : 326.1   
## 3rd Qu.:1402 3rd Qu.: 701.0   
## Max. :5095 Max. :1862.0   
## Low\_Qual\_Fin\_SF Gr\_Liv\_Area Bsmt\_Full\_Bath Bsmt\_Half\_Bath   
## Min. : 0.000 Min. : 480 Min. :0.0000 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.:1137 1st Qu.:0.0000 1st Qu.:0.00000   
## Median : 0.000 Median :1447 Median :0.0000 Median :0.00000   
## Mean : 4.973 Mean :1499 Mean :0.4301 Mean :0.05796   
## 3rd Qu.: 0.000 3rd Qu.:1737 3rd Qu.:1.0000 3rd Qu.:0.00000   
## Max. :1064.000 Max. :5095 Max. :3.0000 Max. :2.00000   
## Full\_Bath Half\_Bath Bedroom\_AbvGr Kitchen\_AbvGr   
## Min. :0.000 Min. :0.0000 Min. :0.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000 1st Qu.:1.000   
## Median :2.000 Median :0.0000 Median :3.000 Median :1.000   
## Mean :1.564 Mean :0.3751 Mean :2.855 Mean :1.047   
## 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :4.000 Max. :2.0000 Max. :6.000 Max. :3.000   
## Kitchen\_Qual TotRms\_AbvGrd Functional Fireplaces   
## Length:2053 Min. : 3.000 Length:2053 Min. :0.000   
## Class :character 1st Qu.: 5.000 Class :character 1st Qu.:0.000   
## Mode :character Median : 6.000 Mode :character Median :1.000   
## Mean : 6.442 Mean :0.603   
## 3rd Qu.: 7.000 3rd Qu.:1.000   
## Max. :15.000 Max. :4.000   
## Fireplace\_Qu Garage\_Type Garage\_Finish Garage\_Cars   
## Length:2053 Length:2053 Length:2053 Min. :0.000   
## Class :character Class :character Class :character 1st Qu.:1.000   
## Mode :character Mode :character Mode :character Median :2.000   
## Mean :1.774   
## 3rd Qu.:2.000   
## Max. :5.000   
## Garage\_Area Garage\_Qual Garage\_Cond Paved\_Drive   
## Min. : 0 Length:2053 Length:2053 Length:2053   
## 1st Qu.: 320 Class :character Class :character Class :character   
## Median : 478 Mode :character Mode :character Mode :character   
## Mean : 472   
## 3rd Qu.: 576   
## Max. :1488   
## Wood\_Deck\_SF Open\_Porch\_SF Enclosed\_Porch Three\_season\_porch  
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.00 Median : 27.00 Median : 0.00 Median : 0.000   
## Mean : 93.52 Mean : 48.17 Mean : 23.02 Mean : 2.799   
## 3rd Qu.: 168.00 3rd Qu.: 72.00 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :1424.00 Max. :742.00 Max. :584.00 Max. :407.000   
## Screen\_Porch Pool\_Area Pool\_QC Fence   
## Min. : 0.00 Min. : 0.000 Length:2053 Length:2053   
## 1st Qu.: 0.00 1st Qu.: 0.000 Class :character Class :character   
## Median : 0.00 Median : 0.000 Mode :character Mode :character   
## Mean : 16.68 Mean : 1.339   
## 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :576.00 Max. :800.000   
## Misc\_Feature Misc\_Val Mo\_Sold Year\_Sold   
## Length:2053 Min. : 0.00 Min. : 1.000 Min. :2006   
## Class :character 1st Qu.: 0.00 1st Qu.: 4.000 1st Qu.:2007   
## Mode :character Median : 0.00 Median : 6.000 Median :2008   
## Mean : 60.12 Mean : 6.189 Mean :2008   
## 3rd Qu.: 0.00 3rd Qu.: 8.000 3rd Qu.:2009   
## Max. :17000.00 Max. :12.000 Max. :2010   
## Sale\_Type Sale\_Condition Longitude Latitude   
## Length:2053 Length:2053 Min. :-93.69 Min. :41.99   
## Class :character Class :character 1st Qu.:-93.66 1st Qu.:42.02   
## Mode :character Mode :character Median :-93.64 Median :42.03   
## Mean :-93.64 Mean :42.03   
## 3rd Qu.:-93.62 3rd Qu.:42.05   
## Max. :-93.58 Max. :42.06   
## Above\_Median   
## Length:2053   
## Class :character   
## Mode :character   
##   
##   
##

ames = ames\_student %>% mutate(MS\_SubClass = as\_factor(MS\_SubClass)) %>%  
 mutate(MS\_Zoning = as\_factor(MS\_Zoning)) %>%  
 mutate(Street = as\_factor(Street)) %>%  
 mutate(Alley = as\_factor(Alley)) %>%  
 mutate(Lot\_Shape = as\_factor(Lot\_Shape)) %>%  
 mutate(Land\_Contour = as\_factor(Land\_Contour)) %>%  
 mutate(Utilities = as\_factor(Utilities)) %>%  
 mutate(Lot\_Config = as\_factor(Lot\_Config)) %>%  
 mutate(Land\_Slope = as\_factor(Land\_Slope)) %>%  
 mutate(Neighborhood = as\_factor(Neighborhood)) %>%  
 mutate(Condition\_1 = as\_factor(Condition\_1)) %>%  
 mutate(Condition\_2 = as\_factor(Condition\_2)) %>%  
 mutate(Bldg\_Type = as\_factor(Bldg\_Type)) %>%  
 mutate(House\_Style = as\_factor(House\_Style)) %>%  
 mutate(Overall\_Qual = as\_factor(Overall\_Qual)) %>%  
 mutate(Exter\_Cond = as\_factor(Exter\_Cond)) %>%  
 mutate(Year\_Built = as\_factor(Year\_Built)) %>%  
 mutate(Year\_Remod\_Add = as\_factor(Year\_Remod\_Add)) %>%  
 mutate(Roof\_Style = as\_factor(Roof\_Style)) %>%  
 mutate(Roof\_Matl = as\_factor(Roof\_Matl)) %>%  
 mutate(Exterior\_1st = as\_factor(Exterior\_1st)) %>%  
 mutate(Exterior\_2nd = as\_factor(Exterior\_2nd)) %>%  
 mutate(Mas\_Vnr\_Type = as\_factor(Mas\_Vnr\_Type)) %>%  
 mutate(Exter\_Qual = as\_factor(Exter\_Qual)) %>%  
 mutate(Exter\_Cond = as\_factor(Exter\_Cond)) %>%  
 mutate(Foundation = as\_factor(Foundation)) %>%  
 mutate(Bsmt\_Qual = as\_factor(Bsmt\_Qual)) %>%  
 mutate(Bsmt\_Cond = as\_factor(Bsmt\_Cond)) %>%  
 mutate(Bsmt\_Exposure = as\_factor(Bsmt\_Exposure)) %>%  
 mutate(BsmtFin\_Type\_1 = as\_factor(BsmtFin\_Type\_1)) %>%  
 mutate(MS\_SubClass = as\_factor(MS\_SubClass)) %>%  
 mutate(BsmtFin\_Type\_2 = as\_factor(BsmtFin\_Type\_2)) %>%  
 mutate(Heating = as\_factor(Heating)) %>%  
 mutate(Heating\_QC = as\_factor(Heating\_QC)) %>%  
 mutate(Central\_Air = as\_factor(Central\_Air)) %>%  
 mutate(Electrical = as\_factor(Electrical)) %>%  
 mutate(Bsmt\_Full\_Bath = as\_factor(Bsmt\_Full\_Bath)) %>%  
 mutate(Bsmt\_Half\_Bath = as\_factor(Bsmt\_Half\_Bath)) %>%  
 mutate(Full\_Bath = as\_factor(Full\_Bath)) %>%  
 mutate(Half\_Bath = as\_factor(Half\_Bath)) %>%  
 mutate(Bedroom\_AbvGr = as\_factor(Bedroom\_AbvGr)) %>%  
 mutate(Kitchen\_AbvGr = as\_factor(Kitchen\_AbvGr)) %>%  
 mutate(Kitchen\_Qual = as\_factor(Kitchen\_Qual)) %>%  
 mutate(Functional = as\_factor(Functional)) %>%  
 mutate(Fireplaces = as\_factor(Fireplaces)) %>%  
 mutate(Fireplace\_Qu = as\_factor(Fireplace\_Qu)) %>%  
 mutate(Garage\_Type = as\_factor(Garage\_Type)) %>%  
 mutate(Garage\_Finish = as\_factor(Garage\_Finish)) %>%  
 mutate(Garage\_Cars = as\_factor(Garage\_Cars)) %>%  
 mutate(Garage\_Qual = as\_factor(Garage\_Qual)) %>%  
 mutate(Garage\_Cond = as\_factor(Garage\_Cond)) %>%  
 mutate(Paved\_Drive = as\_factor(Paved\_Drive)) %>%  
 mutate(Pool\_QC = as\_factor(Pool\_QC)) %>%  
 mutate(Fence = as\_factor(Fence)) %>%  
 mutate(Misc\_Feature = as\_factor(Misc\_Feature)) %>%  
 mutate(Mo\_Sold = as\_factor(Mo\_Sold)) %>%  
 mutate(Year\_Sold = as\_factor(Year\_Sold)) %>%  
 mutate(Sale\_Type = as\_factor(Sale\_Type)) %>%  
 mutate(Sale\_Condition = as\_factor(Sale\_Condition)) %>%  
 mutate(Above\_Median = as\_factor(Above\_Median)) %>%  
 mutate(Overall\_Cond = as\_factor(Overall\_Cond))

summary(ames)

## MS\_SubClass MS\_Zoning   
## One\_Story\_1946\_and\_Newer\_All\_Styles :772 Residential\_Low\_Density :1600   
## Two\_Story\_1946\_and\_Newer :383 Residential\_High\_Density : 20   
## One\_and\_Half\_Story\_Finished\_All\_Ages:204 Floating\_Village\_Residential: 87   
## One\_Story\_PUD\_1946\_and\_Newer :129 Residential\_Medium\_Density : 326   
## One\_Story\_1945\_and\_Older : 98 C\_all : 17   
## Two\_Story\_1945\_and\_Older : 95 A\_agr : 2   
## (Other) :372 I\_all : 1   
## Lot\_Frontage Lot\_Area Street Alley   
## Min. : 0.00 Min. : 1300 Pave:2046 No\_Alley\_Access:1914   
## 1st Qu.: 43.00 1st Qu.: 7500 Grvl: 7 Paved : 45   
## Median : 62.00 Median : 9548 Gravel : 94   
## Mean : 57.38 Mean : 10258   
## 3rd Qu.: 78.00 3rd Qu.: 11600   
## Max. :313.00 Max. :215245   
##   
## Lot\_Shape Land\_Contour Utilities Lot\_Config   
## Slightly\_Irregular : 714 Lvl:1833 AllPub:2052 Corner : 359   
## Regular :1275 HLS: 94 NoSewr: 1 Inside :1495   
## Moderately\_Irregular: 53 Bnk: 81 CulDSac: 135   
## Irregular : 11 Low: 45 FR2 : 56   
## FR3 : 8   
##   
##   
## Land\_Slope Neighborhood Condition\_1 Condition\_2 Bldg\_Type   
## Gtl:1951 North\_Ames : 327 Norm :1771 Norm :2027 OneFam :1706   
## Mod: 89 College\_Creek: 183 Feedr : 113 Feedr : 12 TwnhsE : 157   
## Sev: 13 Old\_Town : 181 Artery : 67 PosA : 4 Twnhs : 67   
## Edwards : 129 RRAn : 35 Artery : 4 Duplex : 76   
## Somerset : 119 PosN : 24 PosN : 3 TwoFmCon: 47   
## Gilbert : 109 RRAe : 19 RRNn : 1   
## (Other) :1005 (Other): 24 (Other): 2   
## House\_Style Overall\_Qual Overall\_Cond   
## One\_Story :1052 Average :587 Average :1143   
## Two\_Story : 590 Above\_Average:518 Above\_Average: 376   
## One\_and\_Half\_Fin: 225 Good :411 Good : 286   
## SLvl : 90 Very\_Good :237 Very\_Good : 98   
## SFoyer : 56 Below\_Average:169 Below\_Average: 73   
## Two\_and\_Half\_Unf: 19 Excellent : 70 Fair : 35   
## (Other) : 21 (Other) : 61 (Other) : 42   
## Year\_Built Year\_Remod\_Add Roof\_Style Roof\_Matl Exterior\_1st  
## 2005 : 104 1950 : 256 Hip : 404 CompShg:2023 VinylSd:705   
## 2006 : 93 2006 : 147 Gable :1607 WdShake: 8 MetalSd:319   
## 2007 : 76 2007 : 116 Mansard: 9 Tar&Grv: 17 Wd Sdng:313   
## 2003 : 62 2005 : 94 Gambrel: 14 WdShngl: 3 HdBoard:303   
## 2004 : 60 2004 : 80 Shed : 5 Roll : 1 Plywood:151   
## 1977 : 40 2003 : 70 Flat : 14 Metal : 1 CemntBd: 90   
## (Other):1618 (Other):1290 (Other):172   
## Exterior\_2nd Mas\_Vnr\_Type Mas\_Vnr\_Area Exter\_Qual   
## VinylSd:699 Stone : 166 Min. : 0.0 Typical :1272   
## MetalSd:317 None :1231 1st Qu.: 0.0 Good : 682   
## Wd Sdng:302 BrkFace: 638 Median : 0.0 Excellent: 78   
## HdBoard:277 BrkCmn : 17 Mean : 103.8 Fair : 21   
## Plywood:190 CBlock : 1 3rd Qu.: 164.0   
## CmentBd: 90 Max. :1600.0   
## (Other):178   
## Exter\_Cond Foundation Bsmt\_Qual Bsmt\_Cond   
## Typical :1787 CBlock:880 Typical :911 Good : 80   
## Good : 213 PConc :911 Good :849 Typical :1833   
## Fair : 43 Wood : 4 Excellent :178 Poor : 4   
## Excellent: 9 BrkTil:216 No\_Basement: 57 No\_Basement: 57   
## Poor : 1 Slab : 36 Fair : 57 Fair : 76   
## Stone : 6 Poor : 1 Excellent : 3   
##   
## Bsmt\_Exposure BsmtFin\_Type\_1 BsmtFin\_SF\_1 BsmtFin\_Type\_2  
## Gd : 199 BLQ :196 Min. :1.00 Unf :1740   
## No :1331 Rec :216 1st Qu.:3.00 LwQ : 64   
## Av : 284 ALQ :298 Median :3.00 BLQ : 47   
## Mn : 179 GLQ :578 Mean :4.21 Rec : 79   
## No\_Basement: 60 Unf :602 3rd Qu.:7.00 GLQ : 23   
## LwQ :106 Max. :7.00 No\_Basement: 58   
## No\_Basement: 57 ALQ : 42   
## BsmtFin\_SF\_2 Bsmt\_Unf\_SF Total\_Bsmt\_SF Heating   
## Min. : 0.00 Min. : 0.0 Min. : 0 GasA :2019   
## 1st Qu.: 0.00 1st Qu.: 226.0 1st Qu.: 793 GasW : 21   
## Median : 0.00 Median : 460.0 Median : 988 Grav : 6   
## Mean : 52.57 Mean : 561.2 Mean :1055 Wall : 5   
## 3rd Qu.: 0.00 3rd Qu.: 801.0 3rd Qu.:1304 Floor: 1   
## Max. :1526.00 Max. :2336.0 Max. :5095 OthW : 1   
##   
## Heating\_QC Central\_Air Electrical First\_Flr\_SF Second\_Flr\_SF   
## Fair : 61 Y:1916 SBrkr :1887 Min. : 432 Min. : 0.0   
## Typical : 618 N: 137 FuseA : 126 1st Qu.: 882 1st Qu.: 0.0   
## Excellent:1040 FuseF : 33 Median :1088 Median : 0.0   
## Good : 333 FuseP : 6 Mean :1168 Mean : 326.1   
## Poor : 1 Unknown: 1 3rd Qu.:1402 3rd Qu.: 701.0   
## Max. :5095 Max. :1862.0   
##   
## Low\_Qual\_Fin\_SF Gr\_Liv\_Area Bsmt\_Full\_Bath Bsmt\_Half\_Bath Full\_Bath  
## Min. : 0.000 Min. : 480 0:1201 0:1936 0: 10   
## 1st Qu.: 0.000 1st Qu.:1137 1: 823 1: 115 1: 920   
## Median : 0.000 Median :1447 2: 27 2: 2 2:1080   
## Mean : 4.973 Mean :1499 3: 2 3: 41   
## 3rd Qu.: 0.000 3rd Qu.:1737 4: 2   
## Max. :1064.000 Max. :5095   
##   
## Half\_Bath Bedroom\_AbvGr Kitchen\_AbvGr Kitchen\_Qual TotRms\_AbvGrd   
## 0:1300 0: 7 1:1959 Typical :1070 Min. : 3.000   
## 1: 736 1: 73 2: 92 Good : 790 1st Qu.: 5.000   
## 2: 17 2: 527 3: 2 Excellent: 142 Median : 6.000   
## 3:1105 Fair : 50 Mean : 6.442   
## 4: 297 Poor : 1 3rd Qu.: 7.000   
## 5: 32 Max. :15.000   
## 6: 12   
## Functional Fireplaces Fireplace\_Qu Garage\_Type   
## Typ :1896 0:993 Good :538 Attchd :1204   
## Min2 : 54 1:891 No\_Fireplace:993 BuiltIn : 127   
## Min1 : 51 2:161 Typical :409 Basment : 29   
## Mod : 27 3: 7 Poor : 36 Detchd : 549   
## Maj1 : 15 4: 1 Excellent : 21 No\_Garage : 108   
## Maj2 : 6 Fair : 56 CarPort : 15   
## (Other): 4 More\_Than\_Two\_Types: 21   
## Garage\_Finish Garage\_Cars Garage\_Area Garage\_Qual Garage\_Cond   
## Fin :509 0: 108 Min. : 0 Typical :1839 Typical :1872   
## Unf :872 1: 539 1st Qu.: 320 No\_Garage: 109 No\_Garage: 109   
## RFn :563 2:1131 Median : 478 Fair : 85 Fair : 53   
## No\_Garage:109 3: 261 Mean : 472 Good : 16 Excellent: 1   
## 4: 13 3rd Qu.: 576 Excellent: 2 Poor : 8   
## 5: 1 Max. :1488 Poor : 2 Good : 10   
##   
## Paved\_Drive Wood\_Deck\_SF Open\_Porch\_SF Enclosed\_Porch   
## Partial\_Pavement: 42 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## Paved :1848 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00   
## Dirt\_Gravel : 163 Median : 0.00 Median : 27.00 Median : 0.00   
## Mean : 93.52 Mean : 48.17 Mean : 23.02   
## 3rd Qu.: 168.00 3rd Qu.: 72.00 3rd Qu.: 0.00   
## Max. :1424.00 Max. :742.00 Max. :584.00   
##   
## Three\_season\_porch Screen\_Porch Pool\_Area Pool\_QC   
## Min. : 0.000 Min. : 0.00 Min. : 0.000 No\_Pool :2047   
## 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.000 Excellent: 2   
## Median : 0.000 Median : 0.00 Median : 0.000 Typical : 2   
## Mean : 2.799 Mean : 16.68 Mean : 1.339 Fair : 1   
## 3rd Qu.: 0.000 3rd Qu.: 0.00 3rd Qu.: 0.000 Good : 1   
## Max. :407.000 Max. :576.00 Max. :800.000   
##   
## Fence Misc\_Feature Misc\_Val Mo\_Sold   
## No\_Fence :1661 None:1978 Min. : 0.00 6 :352   
## Minimum\_Privacy : 225 Gar2: 5 1st Qu.: 0.00 7 :320   
## Good\_Privacy : 81 Shed: 66 Median : 0.00 5 :275   
## Good\_Wood : 77 Othr: 3 Mean : 60.12 4 :187   
## Minimum\_Wood\_Wire: 9 Elev: 1 3rd Qu.: 0.00 8 :169   
## Max. :17000.00 3 :164   
## (Other):586   
## Year\_Sold Sale\_Type Sale\_Condition Longitude Latitude   
## 2006:442 WD :1789 Normal :1712 Min. :-93.69 Min. :41.99   
## 2007:499 New : 163 Partial: 169 1st Qu.:-93.66 1st Qu.:42.02   
## 2008:445 COD : 54 Family : 30 Median :-93.64 Median :42.03   
## 2009:456 ConLD : 16 Abnorml: 121 Mean :-93.64 Mean :42.03   
## 2010:211 ConLI : 8 Alloca : 16 3rd Qu.:-93.62 3rd Qu.:42.05   
## CWD : 8 AdjLand: 5 Max. :-93.58 Max. :42.06   
## (Other): 15   
## Above\_Median  
## Yes:1043   
## No :1010   
##   
##   
##   
##   
##

skim(ames)

Data summary

|  |  |
| --- | --- |
| Name | ames |
| Number of rows | 2053 |
| Number of columns | 81 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 59 |
| numeric | 22 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| MS\_SubClass | 0 | 1 | FALSE | 16 | One: 772, Two: 383, One: 204, One: 129 |
| MS\_Zoning | 0 | 1 | FALSE | 7 | Res: 1600, Res: 326, Flo: 87, Res: 20 |
| Street | 0 | 1 | FALSE | 2 | Pav: 2046, Grv: 7 |
| Alley | 0 | 1 | FALSE | 3 | No\_: 1914, Gra: 94, Pav: 45 |
| Lot\_Shape | 0 | 1 | FALSE | 4 | Reg: 1275, Sli: 714, Mod: 53, Irr: 11 |
| Land\_Contour | 0 | 1 | FALSE | 4 | Lvl: 1833, HLS: 94, Bnk: 81, Low: 45 |
| Utilities | 0 | 1 | FALSE | 2 | All: 2052, NoS: 1 |
| Lot\_Config | 0 | 1 | FALSE | 5 | Ins: 1495, Cor: 359, Cul: 135, FR2: 56 |
| Land\_Slope | 0 | 1 | FALSE | 3 | Gtl: 1951, Mod: 89, Sev: 13 |
| Neighborhood | 0 | 1 | FALSE | 28 | Nor: 327, Col: 183, Old: 181, Edw: 129 |
| Condition\_1 | 0 | 1 | FALSE | 9 | Nor: 1771, Fee: 113, Art: 67, RRA: 35 |
| Condition\_2 | 0 | 1 | FALSE | 8 | Nor: 2027, Fee: 12, Pos: 4, Art: 4 |
| Bldg\_Type | 0 | 1 | FALSE | 5 | One: 1706, Twn: 157, Dup: 76, Twn: 67 |
| House\_Style | 0 | 1 | FALSE | 8 | One: 1052, Two: 590, One: 225, SLv: 90 |
| Overall\_Qual | 0 | 1 | FALSE | 10 | Ave: 587, Abo: 518, Goo: 411, Ver: 237 |
| Overall\_Cond | 0 | 1 | FALSE | 9 | Ave: 1143, Abo: 376, Goo: 286, Ver: 98 |
| Year\_Built | 0 | 1 | FALSE | 114 | 200: 104, 200: 93, 200: 76, 200: 62 |
| Year\_Remod\_Add | 0 | 1 | FALSE | 61 | 195: 256, 200: 147, 200: 116, 200: 94 |
| Roof\_Style | 0 | 1 | FALSE | 6 | Gab: 1607, Hip: 404, Gam: 14, Fla: 14 |
| Roof\_Matl | 0 | 1 | FALSE | 6 | Com: 2023, Tar: 17, WdS: 8, WdS: 3 |
| Exterior\_1st | 0 | 1 | FALSE | 16 | Vin: 705, Met: 319, Wd : 313, HdB: 303 |
| Exterior\_2nd | 0 | 1 | FALSE | 17 | Vin: 699, Met: 317, Wd : 302, HdB: 277 |
| Mas\_Vnr\_Type | 0 | 1 | FALSE | 5 | Non: 1231, Brk: 638, Sto: 166, Brk: 17 |
| Exter\_Qual | 0 | 1 | FALSE | 4 | Typ: 1272, Goo: 682, Exc: 78, Fai: 21 |
| Exter\_Cond | 0 | 1 | FALSE | 5 | Typ: 1787, Goo: 213, Fai: 43, Exc: 9 |
| Foundation | 0 | 1 | FALSE | 6 | PCo: 911, CBl: 880, Brk: 216, Sla: 36 |
| Bsmt\_Qual | 0 | 1 | FALSE | 6 | Typ: 911, Goo: 849, Exc: 178, No\_: 57 |
| Bsmt\_Cond | 0 | 1 | FALSE | 6 | Typ: 1833, Goo: 80, Fai: 76, No\_: 57 |
| Bsmt\_Exposure | 0 | 1 | FALSE | 5 | No: 1331, Av: 284, Gd: 199, Mn: 179 |
| BsmtFin\_Type\_1 | 0 | 1 | FALSE | 7 | Unf: 602, GLQ: 578, ALQ: 298, Rec: 216 |
| BsmtFin\_Type\_2 | 0 | 1 | FALSE | 7 | Unf: 1740, Rec: 79, LwQ: 64, No\_: 58 |
| Heating | 0 | 1 | FALSE | 6 | Gas: 2019, Gas: 21, Gra: 6, Wal: 5 |
| Heating\_QC | 0 | 1 | FALSE | 5 | Exc: 1040, Typ: 618, Goo: 333, Fai: 61 |
| Central\_Air | 0 | 1 | FALSE | 2 | Y: 1916, N: 137 |
| Electrical | 0 | 1 | FALSE | 5 | SBr: 1887, Fus: 126, Fus: 33, Fus: 6 |
| Bsmt\_Full\_Bath | 0 | 1 | FALSE | 4 | 0: 1201, 1: 823, 2: 27, 3: 2 |
| Bsmt\_Half\_Bath | 0 | 1 | FALSE | 3 | 0: 1936, 1: 115, 2: 2 |
| Full\_Bath | 0 | 1 | FALSE | 5 | 2: 1080, 1: 920, 3: 41, 0: 10 |
| Half\_Bath | 0 | 1 | FALSE | 3 | 0: 1300, 1: 736, 2: 17 |
| Bedroom\_AbvGr | 0 | 1 | FALSE | 7 | 3: 1105, 2: 527, 4: 297, 1: 73 |
| Kitchen\_AbvGr | 0 | 1 | FALSE | 3 | 1: 1959, 2: 92, 3: 2 |
| Kitchen\_Qual | 0 | 1 | FALSE | 5 | Typ: 1070, Goo: 790, Exc: 142, Fai: 50 |
| Functional | 0 | 1 | FALSE | 8 | Typ: 1896, Min: 54, Min: 51, Mod: 27 |
| Fireplaces | 0 | 1 | FALSE | 5 | 0: 993, 1: 891, 2: 161, 3: 7 |
| Fireplace\_Qu | 0 | 1 | FALSE | 6 | No\_: 993, Goo: 538, Typ: 409, Fai: 56 |
| Garage\_Type | 0 | 1 | FALSE | 7 | Att: 1204, Det: 549, Bui: 127, No\_: 108 |
| Garage\_Finish | 0 | 1 | FALSE | 4 | Unf: 872, RFn: 563, Fin: 509, No\_: 109 |
| Garage\_Cars | 0 | 1 | FALSE | 6 | 2: 1131, 1: 539, 3: 261, 0: 108 |
| Garage\_Qual | 0 | 1 | FALSE | 6 | Typ: 1839, No\_: 109, Fai: 85, Goo: 16 |
| Garage\_Cond | 0 | 1 | FALSE | 6 | Typ: 1872, No\_: 109, Fai: 53, Goo: 10 |
| Paved\_Drive | 0 | 1 | FALSE | 3 | Pav: 1848, Dir: 163, Par: 42 |
| Pool\_QC | 0 | 1 | FALSE | 5 | No\_: 2047, Exc: 2, Typ: 2, Fai: 1 |
| Fence | 0 | 1 | FALSE | 5 | No\_: 1661, Min: 225, Goo: 81, Goo: 77 |
| Misc\_Feature | 0 | 1 | FALSE | 5 | Non: 1978, She: 66, Gar: 5, Oth: 3 |
| Mo\_Sold | 0 | 1 | FALSE | 12 | 6: 352, 7: 320, 5: 275, 4: 187 |
| Year\_Sold | 0 | 1 | FALSE | 5 | 200: 499, 200: 456, 200: 445, 200: 442 |
| Sale\_Type | 0 | 1 | FALSE | 10 | WD: 1789, New: 163, COD: 54, Con: 16 |
| Sale\_Condition | 0 | 1 | FALSE | 6 | Nor: 1712, Par: 169, Abn: 121, Fam: 30 |
| Above\_Median | 0 | 1 | FALSE | 2 | Yes: 1043, No: 1010 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Lot\_Frontage | 0 | 1 | 57.38 | 33.20 | 0.00 | 43.00 | 62.00 | 78.00 | 313.00 | ▇▇▁▁▁ |
| Lot\_Area | 0 | 1 | 10258.40 | 8427.38 | 1300.00 | 7500.00 | 9548.00 | 11600.00 | 215245.00 | ▇▁▁▁▁ |
| Mas\_Vnr\_Area | 0 | 1 | 103.75 | 183.59 | 0.00 | 0.00 | 0.00 | 164.00 | 1600.00 | ▇▁▁▁▁ |
| BsmtFin\_SF\_1 | 0 | 1 | 4.21 | 2.24 | 1.00 | 3.00 | 3.00 | 7.00 | 7.00 | ▅▆▁▁▇ |
| BsmtFin\_SF\_2 | 0 | 1 | 52.57 | 175.99 | 0.00 | 0.00 | 0.00 | 0.00 | 1526.00 | ▇▁▁▁▁ |
| Bsmt\_Unf\_SF | 0 | 1 | 561.19 | 441.72 | 0.00 | 226.00 | 460.00 | 801.00 | 2336.00 | ▇▅▂▁▁ |
| Total\_Bsmt\_SF | 0 | 1 | 1054.57 | 435.33 | 0.00 | 793.00 | 988.00 | 1304.00 | 5095.00 | ▇▇▁▁▁ |
| First\_Flr\_SF | 0 | 1 | 1167.52 | 391.79 | 432.00 | 882.00 | 1088.00 | 1402.00 | 5095.00 | ▇▃▁▁▁ |
| Second\_Flr\_SF | 0 | 1 | 326.07 | 422.44 | 0.00 | 0.00 | 0.00 | 701.00 | 1862.00 | ▇▂▂▁▁ |
| Low\_Qual\_Fin\_SF | 0 | 1 | 4.97 | 49.09 | 0.00 | 0.00 | 0.00 | 0.00 | 1064.00 | ▇▁▁▁▁ |
| Gr\_Liv\_Area | 0 | 1 | 1498.56 | 487.84 | 480.00 | 1137.00 | 1447.00 | 1737.00 | 5095.00 | ▇▇▁▁▁ |
| TotRms\_AbvGrd | 0 | 1 | 6.44 | 1.54 | 3.00 | 5.00 | 6.00 | 7.00 | 15.00 | ▅▇▃▁▁ |
| Garage\_Area | 0 | 1 | 471.96 | 213.43 | 0.00 | 320.00 | 478.00 | 576.00 | 1488.00 | ▃▇▂▁▁ |
| Wood\_Deck\_SF | 0 | 1 | 93.52 | 127.71 | 0.00 | 0.00 | 0.00 | 168.00 | 1424.00 | ▇▁▁▁▁ |
| Open\_Porch\_SF | 0 | 1 | 48.17 | 69.51 | 0.00 | 0.00 | 27.00 | 72.00 | 742.00 | ▇▁▁▁▁ |
| Enclosed\_Porch | 0 | 1 | 23.02 | 60.59 | 0.00 | 0.00 | 0.00 | 0.00 | 584.00 | ▇▁▁▁▁ |
| Three\_season\_porch | 0 | 1 | 2.80 | 25.65 | 0.00 | 0.00 | 0.00 | 0.00 | 407.00 | ▇▁▁▁▁ |
| Screen\_Porch | 0 | 1 | 16.68 | 57.94 | 0.00 | 0.00 | 0.00 | 0.00 | 576.00 | ▇▁▁▁▁ |
| Pool\_Area | 0 | 1 | 1.34 | 27.74 | 0.00 | 0.00 | 0.00 | 0.00 | 800.00 | ▇▁▁▁▁ |
| Misc\_Val | 0 | 1 | 60.12 | 662.76 | 0.00 | 0.00 | 0.00 | 0.00 | 17000.00 | ▇▁▁▁▁ |
| Longitude | 0 | 1 | -93.64 | 0.03 | -93.69 | -93.66 | -93.64 | -93.62 | -93.58 | ▅▅▇▇▁ |
| Latitude | 0 | 1 | 42.03 | 0.02 | 41.99 | 42.02 | 42.03 | 42.05 | 42.06 | ▂▂▇▇▇ |

## Data Clean, Data Split, Recip

amesclean = ames %>% dplyr::select(-Alley, -Street, -Utilities, -Heating, -Bsmt\_Half\_Bath, -Pool\_QC, -Mo\_Sold, -Year\_Sold, -Pool\_Area, -Longitude, -Latitude)

set.seed(12345)  
ames\_split = initial\_split(amesclean, prop = 0.65, strata = Above\_Median)  
train = training(ames\_split)  
test = testing(ames\_split)

amesABMV\_recipe = recipe(Above\_Median ~., train) #recipe 1  
  
amesABMV\_recipe2 = recipe(Above\_Median ~ MS\_SubClass + Lot\_Shape + Neighborhood + Bldg\_Type + House\_Style + Overall\_Qual + Overall\_Cond + Year\_Built + Year\_Remod\_Add + Foundation + Bsmt\_Qual + Bsmt\_Cond + Bsmt\_Full\_Bath + Full\_Bath + Half\_Bath + Kitchen\_Qual + Fireplaces + Garage\_Cars + Garage\_Type + Sale\_Type + Sale\_Condition + Lot\_Area + Total\_Bsmt\_SF + First\_Flr\_SF + Second\_Flr\_SF + Gr\_Liv\_Area+ TotRms\_AbvGrd + Garage\_Area, train) #recipe2  
  
amesABMV\_recipe3 = recipe(Above\_Median ~ Neighborhood + House\_Style + Overall\_Qual + Overall\_Cond + Full\_Bath + Half\_Bath + Kitchen\_Qual + Fireplaces + Garage\_Cars + Sale\_Type + Lot\_Area + Total\_Bsmt\_SF + Gr\_Liv\_Area+ Garage\_Area, train) #recipe3  
  
ctrl\_grid = control\_stack\_grid() #necessary for working with the stacks package  
ctrl\_res = control\_stack\_resamples() #necessary for working with the stacks package

## Model 3 Random Forest

ames\_recipe4 = amesABMV\_recipe %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_model = rand\_forest() %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_wflow4 =   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(ames\_recipe4)

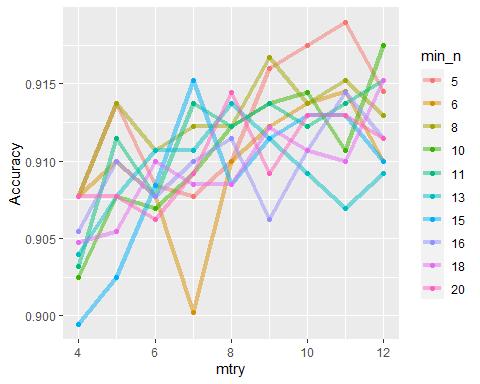
set.seed(123)  
rf\_folds = vfold\_cv(train, v = 5)

ames\_recipe5 = amesABMV\_recipe %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_model = rand\_forest(mtry = tune(), min\_n = tune(), trees = 100) %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_wflow5 =   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(ames\_recipe5)  
  
rf\_grid = grid\_regular(  
 mtry(range = c(4, 12)),   
 min\_n(range = c(5, 20)),   
 levels = 10  
)  
  
set.seed(123)  
rf\_res\_tuned = tune\_grid(  
 ames\_wflow5,  
 resamples = rf\_folds,  
 grid = rf\_grid   
)

saveRDS(rf\_res\_tuned,"RandomF.rds")

rf\_res\_tuned = readRDS("RandomF.rds")

rf\_res\_tuned %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 mutate(min\_n = factor(min\_n)) %>%  
 ggplot(aes(mtry, mean, color = min\_n)) +  
 geom\_line(alpha = 0.5, size = 1.5) +  
 geom\_point() +  
 labs(y = "Accuracy")



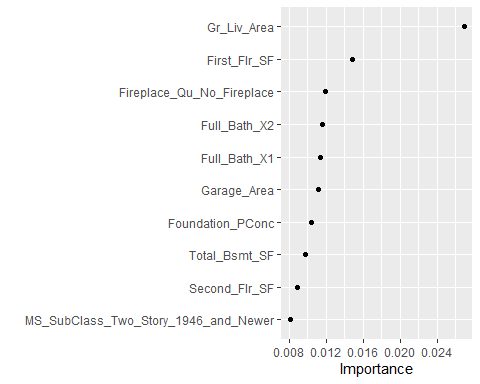
best\_rf = select\_best(rf\_res\_tuned, "accuracy")  
  
final\_rf = finalize\_workflow(  
 ames\_wflow5,  
 best\_rf  
)  
  
final\_rf

## == Workflow ====================================================================  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## -- Preprocessor ----------------------------------------------------------------  
## 1 Recipe Step  
##   
## \* step\_dummy()  
##   
## -- Model -----------------------------------------------------------------------  
## Random Forest Model Specification (classification)  
##   
## Main Arguments:  
## mtry = 11  
## trees = 100  
## min\_n = 5  
##   
## Engine-Specific Arguments:  
## importance = permutation  
##   
## Computational engine: ranger

final\_rf\_fit = fit(final\_rf, train)

saveRDS(final\_rf\_fit,"ABV\_raforest.rds")

final\_rf\_fit %>% extract\_fit\_parsnip() %>% vip(geom = "point")



trainpredrf = predict(final\_rf\_fit, train)  
head(trainpredrf)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 No   
## 3 No   
## 4 No   
## 5 No   
## 6 No

Confusion matrix

confusionMatrix(trainpredrf$.pred\_class, train$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 673 8  
## No 4 648  
##   
## Accuracy : 0.991   
## 95% CI : (0.9843, 0.9953)  
## No Information Rate : 0.5079   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.982   
##   
## Mcnemar's Test P-Value : 0.3865   
##   
## Sensitivity : 0.9941   
## Specificity : 0.9878   
## Pos Pred Value : 0.9883   
## Neg Pred Value : 0.9939   
## Prevalence : 0.5079   
## Detection Rate : 0.5049   
## Detection Prevalence : 0.5109   
## Balanced Accuracy : 0.9909   
##   
## 'Positive' Class : Yes   
##

Predictions on test

testpredrf = predict(final\_rf\_fit, test)  
head(testpredrf)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 Yes   
## 2 Yes   
## 3 Yes   
## 4 No   
## 5 Yes   
## 6 No

confusionMatrix(testpredrf$.pred\_class, test$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 332 28  
## No 34 326  
##   
## Accuracy : 0.9139   
## 95% CI : (0.891, 0.9333)  
## No Information Rate : 0.5083   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8278   
##   
## Mcnemar's Test P-Value : 0.5254   
##   
## Sensitivity : 0.9071   
## Specificity : 0.9209   
## Pos Pred Value : 0.9222   
## Neg Pred Value : 0.9056   
## Prevalence : 0.5083   
## Detection Rate : 0.4611   
## Detection Prevalence : 0.5000   
## Balanced Accuracy : 0.9140   
##   
## 'Positive' Class : Yes   
##

The first random forest model is very accurate for the training data, but it has a drop in accuracy for the test data. Its overfitted to the training data, but still not a bad model to use.

## RF2

ames\_recipe4A = amesABMV\_recipe2 %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_modelA = rand\_forest() %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_wflow4A =   
 workflow() %>%   
 add\_model(rf\_modelA) %>%   
 add\_recipe(ames\_recipe4A)

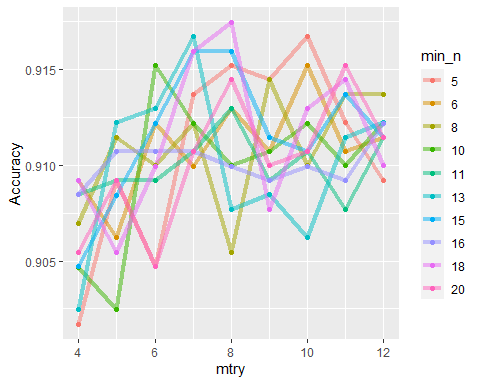
set.seed(123)  
rf\_foldsA = vfold\_cv(train, v = 5)

ames\_recipe5A = amesABMV\_recipe2 %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_modelA = rand\_forest(mtry = tune(), min\_n = tune(), trees = 100) %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_wflow5A =   
 workflow() %>%   
 add\_model(rf\_modelA) %>%   
 add\_recipe(ames\_recipe5A)  
  
rf\_gridA = grid\_regular(  
 mtry(range = c(4, 12)),   
 min\_n(range = c(5, 20)),   
 levels = 10  
)  
  
set.seed(123)  
rf\_res\_tunedA = tune\_grid(  
 ames\_wflow5A,  
 resamples = rf\_foldsA,  
 grid = rf\_gridA  
)

saveRDS(rf\_res\_tunedA,"RandomF2.rds")

rf\_res\_tunedA = readRDS("RandomF2.rds")

rf\_res\_tunedA %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 mutate(min\_n = factor(min\_n)) %>%  
 ggplot(aes(mtry, mean, color = min\_n)) +  
 geom\_line(alpha = 0.5, size = 1.5) +  
 geom\_point() +  
 labs(y = "Accuracy")

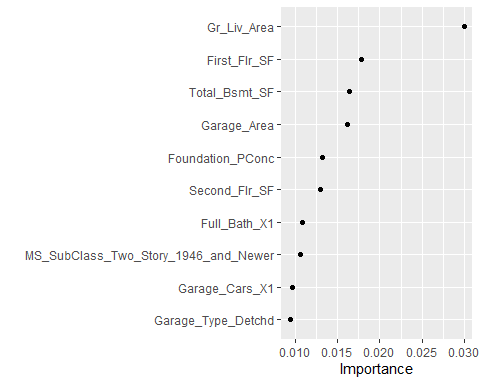


best\_rfA = select\_best(rf\_res\_tunedA, "accuracy")  
  
final\_rfA = finalize\_workflow(  
 ames\_wflow5A,  
 best\_rfA  
)  
  
final\_rfA

## == Workflow ====================================================================  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## -- Preprocessor ----------------------------------------------------------------  
## 1 Recipe Step  
##   
## \* step\_dummy()  
##   
## -- Model -----------------------------------------------------------------------  
## Random Forest Model Specification (classification)  
##   
## Main Arguments:  
## mtry = 8  
## trees = 100  
## min\_n = 18  
##   
## Engine-Specific Arguments:  
## importance = permutation  
##   
## Computational engine: ranger

final\_rf\_fitA = fit(final\_rfA, train)

final\_rf\_fitA %>% extract\_fit\_parsnip() %>% vip(geom = "point")



trainpredrfA = predict(final\_rf\_fitA, train)  
head(trainpredrfA)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 No   
## 3 No   
## 4 No   
## 5 No   
## 6 No

Confusion matrix

confusionMatrix(trainpredrfA$.pred\_class, train$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 648 22  
## No 29 634  
##   
## Accuracy : 0.9617   
## 95% CI : (0.95, 0.9714)  
## No Information Rate : 0.5079   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9235   
##   
## Mcnemar's Test P-Value : 0.4008   
##   
## Sensitivity : 0.9572   
## Specificity : 0.9665   
## Pos Pred Value : 0.9672   
## Neg Pred Value : 0.9563   
## Prevalence : 0.5079   
## Detection Rate : 0.4861   
## Detection Prevalence : 0.5026   
## Balanced Accuracy : 0.9618   
##   
## 'Positive' Class : Yes   
##

Predictions on test

testpredrfA = predict(final\_rf\_fitA, test)  
head(testpredrfA)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 Yes   
## 2 Yes   
## 3 Yes   
## 4 No   
## 5 Yes   
## 6 No

confusionMatrix(testpredrfA$.pred\_class, test$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 331 25  
## No 35 329  
##   
## Accuracy : 0.9167   
## 95% CI : (0.894, 0.9358)  
## No Information Rate : 0.5083   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8334   
##   
## Mcnemar's Test P-Value : 0.2453   
##   
## Sensitivity : 0.9044   
## Specificity : 0.9294   
## Pos Pred Value : 0.9298   
## Neg Pred Value : 0.9038   
## Prevalence : 0.5083   
## Detection Rate : 0.4597   
## Detection Prevalence : 0.4944   
## Balanced Accuracy : 0.9169   
##   
## 'Positive' Class : Yes   
##

The second random forest model is a little less accurate on the training data and has similiar results on the testing data. So its slightly less overfitted.

## RF3

ames\_recipe4B = amesABMV\_recipe3 %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_modelB = rand\_forest() %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_wflow4B =   
 workflow() %>%   
 add\_model(rf\_modelB) %>%   
 add\_recipe(ames\_recipe4B)

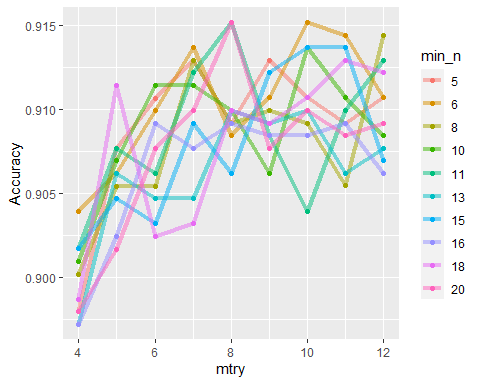
set.seed(123)  
rf\_foldsB = vfold\_cv(train, v = 5)

ames\_recipe5B = amesABMV\_recipe3 %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_modelB = rand\_forest(mtry = tune(), min\_n = tune(), trees = 100) %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_wflow5B =   
 workflow() %>%   
 add\_model(rf\_modelB) %>%   
 add\_recipe(ames\_recipe5B)  
  
rf\_gridB = grid\_regular(  
 mtry(range = c(4, 12)),   
 min\_n(range = c(5, 20)),   
 levels = 10  
)  
  
set.seed(123)  
rf\_res\_tunedB = tune\_grid(  
 ames\_wflow5B,  
 resamples = rf\_foldsB,  
 grid = rf\_gridB   
)

saveRDS(rf\_res\_tunedB,"RandomF3.rds")

rf\_res\_tunedB = readRDS("RandomF3.rds")

rf\_res\_tunedB %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 mutate(min\_n = factor(min\_n)) %>%  
 ggplot(aes(mtry, mean, color = min\_n)) +  
 geom\_line(alpha = 0.5, size = 1.5) +  
 geom\_point() +  
 labs(y = "Accuracy")

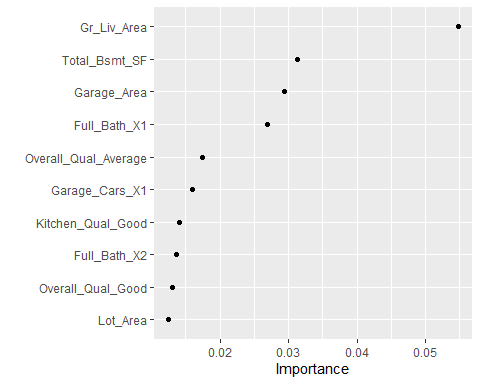


best\_rfB = select\_best(rf\_res\_tunedB, "accuracy")  
  
final\_rfB = finalize\_workflow(  
 ames\_wflow5B,  
 best\_rfB  
)  
  
final\_rfB

## == Workflow ====================================================================  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## -- Preprocessor ----------------------------------------------------------------  
## 1 Recipe Step  
##   
## \* step\_dummy()  
##   
## -- Model -----------------------------------------------------------------------  
## Random Forest Model Specification (classification)  
##   
## Main Arguments:  
## mtry = 8  
## trees = 100  
## min\_n = 11  
##   
## Engine-Specific Arguments:  
## importance = permutation  
##   
## Computational engine: ranger

final\_rf\_fitB = fit(final\_rfB, train)

final\_rf\_fitB %>% extract\_fit\_parsnip() %>% vip(geom = "point")



trainpredrfB = predict(final\_rf\_fitB, train)  
head(trainpredrfB)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 No   
## 3 No   
## 4 No   
## 5 No   
## 6 No

Confusion matrix

confusionMatrix(trainpredrfB$.pred\_class, train$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 657 23  
## No 20 633  
##   
## Accuracy : 0.9677   
## 95% CI : (0.9568, 0.9766)  
## No Information Rate : 0.5079   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9355   
##   
## Mcnemar's Test P-Value : 0.7604   
##   
## Sensitivity : 0.9705   
## Specificity : 0.9649   
## Pos Pred Value : 0.9662   
## Neg Pred Value : 0.9694   
## Prevalence : 0.5079   
## Detection Rate : 0.4929   
## Detection Prevalence : 0.5101   
## Balanced Accuracy : 0.9677   
##   
## 'Positive' Class : Yes   
##

Predictions on test

testpredrfB = predict(final\_rf\_fitB, test)  
head(testpredrfB)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 Yes   
## 2 Yes   
## 3 Yes   
## 4 No   
## 5 Yes   
## 6 No

confusionMatrix(testpredrfB$.pred\_class, test$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 343 28  
## No 23 326  
##   
## Accuracy : 0.9292   
## 95% CI : (0.9079, 0.9468)  
## No Information Rate : 0.5083   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8583   
##   
## Mcnemar's Test P-Value : 0.5754   
##   
## Sensitivity : 0.9372   
## Specificity : 0.9209   
## Pos Pred Value : 0.9245   
## Neg Pred Value : 0.9341   
## Prevalence : 0.5083   
## Detection Rate : 0.4764   
## Detection Prevalence : 0.5153   
## Balanced Accuracy : 0.9290   
##   
## 'Positive' Class : Yes   
##

The third random forest model is better than the second and has a better result for the test data the either of the other two.