final\_code Writeup

## Data Read

## Data Checking

## Using unique fixes the issue  
raw\_policy\_data %>%  
 unique() %>%  
 count(policy\_id) %>%  
 filter(n > 1) %>%  
 arrange(desc(n))

## # A tibble: 0 x 2  
## # i 2 variables: policy\_id <chr>, n <int>

## Going to see how variables are correlated with fraud  
check\_fraud\_pct <- function(x) {  
 combined\_data %>%   
 group\_by(!!sym(x)) %>%   
 summarise(n = n()  
 ,fraud\_ind = sum(fraud\_ind)  
 ,fraud\_pct = fraud\_ind / n) %>%   
 arrange(desc(n))  
}  
  
check\_fraud\_pct("triple\_exclamation") ## Extremely predictive ... although few data points

## # A tibble: 2 x 4  
## triple\_exclamation n fraud\_ind fraud\_pct  
## <dbl> <int> <dbl> <dbl>  
## 1 0 103319 16326 0.158  
## 2 1 782 0 0

check\_fraud\_pct("fraud\_note\_ind") ## Extremely predictive ... although few data points

## # A tibble: 5 x 4  
## fraud\_note\_ind n fraud\_ind fraud\_pct  
## <fct> <int> <dbl> <dbl>  
## 1 no mention of fraud 96335 14482 0.150   
## 2 no signs of fraud 4178 564 0.135   
## 3 suspected fraud 2027 916 0.452   
## 4 probably not fraud 1292 95 0.0735  
## 5 fraud 269 269 1

## Checking the numeric cols  
check\_fraud\_pct\_num2("num\_drivers") ## less drivers seems more fraudulent ... fix those with 17

## # A tibble: 11 x 4  
## num\_drivers n fraud\_ind fraud\_pct  
## <dbl> <int> <dbl> <dbl>  
## 1 1 36936 6485 0.176  
## 2 2 26695 4101 0.154  
## 3 3 19502 2820 0.145  
## 4 4 10753 1487 0.138  
## 5 5 4735 630 0.133  
## 6 6 1736 226 0.130  
## 7 7 480 60 0.125  
## 8 8 121 16 0.132  
## 9 9 15 2 0.133  
## 10 10 3 1 0.333  
## 11 17 3098 495 0.160

check\_fraud\_pct\_num2("clms\_flt1") ## Seems very predictive ... 1+ is very fraudulent ... fix those with 17

## # A tibble: 9 x 4  
## clms\_flt1 n fraud\_ind fraud\_pct  
## <dbl> <int> <dbl> <dbl>  
## 1 0 86967 11918 0.137  
## 2 1 8510 2036 0.239  
## 3 2 3701 1161 0.314  
## 4 3 1303 507 0.389  
## 5 4 390 156 0.4   
## 6 5 86 42 0.488  
## 7 6 18 8 0.444  
## 8 7 1 0 0   
## 9 17 3098 495 0.160

check\_fraud\_pct\_num2("viol\_mjr2") ## Seems very predictive ... 1+ is very fraudulent ... fix those with 17

## # A tibble: 9 x 4  
## viol\_mjr2 n fraud\_ind fraud\_pct  
## <dbl> <int> <dbl> <dbl>  
## 1 0 87079 11757 0.135  
## 2 1 8400 2118 0.252  
## 3 2 3736 1185 0.317  
## 4 3 1272 518 0.407  
## 5 4 391 199 0.509  
## 6 5 86 45 0.523  
## 7 6 15 8 0.533  
## 8 7 1 1 1   
## 9 17 3098 495 0.160

## Fixing the columns that were identified as having issues above  
combined\_data <- combined\_data %>%   
 mutate(num\_drivers\_new = na\_if(num\_drivers ,17)  
 ,clms\_flt1\_new = na\_if(clms\_flt1 ,17)  
 ,late\_90d\_new = na\_if(late\_90d ,500)  
 ,outs\_bal\_new = na\_if(outs\_bal ,99)  
 ,viol\_mjr2\_new = na\_if(viol\_mjr2 ,17)  
 ,time\_bet10pm2am\_new = na\_if(time\_bet10pm2am ,0.75)  
 ,credit\_score\_new = na\_if(credit\_score ,1.5)  
 ,report\_lag\_new = na\_if(report\_lag ,99)  
 ,pop\_density = na\_if(pop\_density ,5) %>% as.factor()  
 )

#########################################################  
## Exploratory Data Analysis - Character Columns  
#########################################################  
## Checking their relationships to fraud  
check\_fraud\_pct\_num2("gender") ## Need to fix these fields

## # A tibble: 6 x 4  
## gender n fraud\_ind fraud\_pct  
## <chr> <int> <dbl> <dbl>  
## 1 Boy 8220 1298 0.158  
## 2 F 7626 1201 0.157  
## 3 Female 34727 5375 0.155  
## 4 Girl 7609 1189 0.156  
## 5 M 8348 1357 0.163  
## 6 Male 37547 5903 0.157

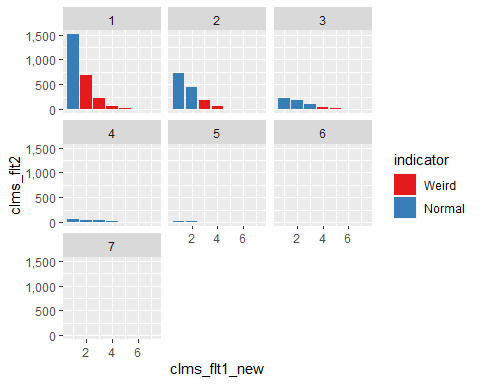
## Fixing the inconsistent character fields  
combined\_data <- combined\_data %>%   
 mutate(gender\_new = case\_when(gender %in% c('Boy' ,'M' ,'Male') ~ 'Male'  
 ,gender %in% c('F' ,'Female' ,'Girl') ~ 'Female') %>%   
 factor(levels = c('Male' ,'Female'))  
 ,education\_new = factor(education ,levels = c('Some High School' ,'High School or GED'  
 ,'Bachelors' ,'Masters' ,'Doctorate'))  
 ,limits\_new = str\_replace\_all(limits ,pattern = "k" ,replacement = "000") %>%   
 str\_replace\_all(pattern = "000$" ,replacement = "k") %>%   
 factor(levels = c("15k" ,"20k" ,"25k" ,"50k" ,"100k" ,"200k" ,"250k" ,"300k" ,"500k"))  
 ,limits\_numeric = limits\_new %>%   
 str\_replace(pattern = "k" ,replacement = "000") %>%   
 as.numeric()  
 ,marital\_status\_new = case\_when(marital\_status %in% c('M' ,'Marr' ,'Married') ~ 'Married'  
 ,marital\_status %in% c('S' ,'Single') ~ 'Single') %>%   
 as.factor()  
 ,num\_cars\_new = case\_when(num\_cars == "Four" ~ "4"  
 ,num\_cars == "Three" ~ "3"  
 ,num\_cars == "Two" ~ "2"  
 ,TRUE ~ num\_cars  
 ) %>%   
 as.factor()  
 ,num\_cars\_band\_new = case\_when(num\_cars == "1" ~ "1"  
 ,num\_cars == "2" ~ "2"  
 ,num\_cars == "3" ~ "3"  
 ,num\_cars == "4" ~ "4"  
 ,TRUE ~ "5+") %>%   
 as.factor()   
 ,seat\_belt\_new = seat\_belt %>%  
 factor(levels = c('Never' ,'Rarely' ,'Occasionally' ,'Usually' ,'Always'   
 ,'Unknown'))  
 ,income\_new = case\_when(income %in% c('Mid' ,'Middle') ~ 'Middle'  
 ,income %in% c('Working' ,'Wrk') ~ 'Working'  
 ,TRUE ~ income) %>%   
 factor(levels = c('Poverty' ,'Working' ,'Middle' ,'Upper'))  
 ,note\_type\_new = case\_when(!(note\_type %in% c(":" ,"Fraud Suspected:" ,"Hit-and-run incident:")) ~ "All Other"  
 ,TRUE ~ note\_type) %>%   
 factor(levels = c(":" ,"All Other" ,"Hit-and-run incident:" ,"Fraud Suspected:"))  
 ,pedestrian\_type\_new = pedestrian\_type %>% tolower() %>% factor(levels = c("other" ,"cyclist" ,"pedestrian"))  
 ,injury\_type\_new = injury\_type %>% as.factor()  
 ,police\_type\_new = police\_type %>%   
 coalesce("no mention of police") %>%   
 factor(levels = c("officer on site" ,"police report"   
 ,"police and medical assistance"  
 ,"police notified" ,"no mention of police"))  
 ,alcohol\_or\_drugs\_new = alcohol\_or\_drugs %>% coalesce("All Other") %>% as.factor()  
 ,claim\_greater\_than\_limit = as.numeric(claimamount > limits\_numeric)  
 )   
  
## Confirming the new fields are looking better  
# check\_fraud\_pct\_num2("gender\_new")  
# check\_fraud\_pct\_num2("education\_new")  
# check\_fraud\_pct\_num2("limits\_new")  
# check\_fraud\_pct\_num2("marital\_status\_new")  
# check\_fraud\_pct\_num2("num\_cars\_new")  
# check\_fraud\_pct\_num2("num\_cars\_band\_new")  
# check\_fraud\_pct\_num2("seat\_belt\_new")  
# check\_fraud\_pct\_num2("income\_new")  
# check\_fraud\_pct\_num2("pedestrian\_type\_new")  
# check\_fraud\_pct\_num2("injury\_type\_new")  
# check\_fraud\_pct\_num2("police\_type\_new")  
# check\_fraud\_pct\_num2("alcohol\_or\_drugs\_new")  
# check\_fraud\_pct\_num2("claim\_greater\_than\_limit")  
  
## Fixing the columns that were identified as having issues above  
combined\_data <- combined\_data %>%   
 mutate(num\_cars\_new = na\_if(num\_cars\_new ,"17"))

#########################################################  
## Exploratory Data Analysis - Date Columns  
#########################################################  
## Converting to dates  
combined\_data <- combined\_data %>%   
 mutate(policy\_orig\_eff\_date\_new = ymd(policy\_orig\_eff\_date)  
 ,accident\_date\_new = ymd(accident\_date)  
 ,policy\_month = month(policy\_orig\_eff\_date\_new)  
 ,policy\_day = day(policy\_orig\_eff\_date\_new)  
 ,policy\_day\_of\_week = wday(policy\_orig\_eff\_date\_new ,label = TRUE)  
 ,acc\_month = month(accident\_date\_new)  
 ,acc\_day = day(accident\_date\_new)  
 ,acc\_day\_of\_week = wday(accident\_date\_new ,label = TRUE)   
 ,report\_date = accident\_date\_new + days(report\_lag)  
 )

## Warning: There were 2 warnings in `mutate()`.  
## The first warning was:  
## i In argument: `policy\_orig\_eff\_date\_new = ymd(policy\_orig\_eff\_date)`.  
## Caused by warning:  
## ! 574 failed to parse.  
## i Run `dplyr::last\_dplyr\_warnings()` to see the 1 remaining warning.

## Date variables that are missing:  
## 1) Report\_date - re-created it with the report lag

#########################################################  
## Variable Relationships/Bucketing  
#########################################################  
## Creating a function to see the relationship b/w different variables  
plot\_x\_y\_vars <- function(x , y) {  
 combined\_data %>%  
 filter(!!sym(x) != 0 & !!sym(y) != 0) %>%   
 mutate(indicator = if\_else(!!sym(x) > !!sym(y)  
 ,"Weird"  
 ,"Normal") %>%   
 factor(levels = c("Weird" ,"Normal"))  
 ) %>%   
 ggplot(aes(x = !!sym(x) ,fill = indicator)) +  
 geom\_bar() +  
 scale\_fill\_brewer(palette = "Set1") +  
 facet\_wrap(str\_c("~" ,y ,sep = " ")) +  
 scale\_y\_continuous(label = comma) +  
 ylab(y)   
}  
  
  
## Looking at one of the variables that has x years  
plot\_x\_y\_vars("clms\_flt1\_new" ,"clms\_flt2")



# plot\_x\_y\_vars("clms\_flt1\_new" ,"clms\_flt3")  
# plot\_x\_y\_vars("clms\_flt1\_new" ,"clms\_flt4")  
# plot\_x\_y\_vars("clms\_flt1\_new" ,"clms\_flt5")  
# plot\_x\_y\_vars("clms\_flt2" ,"clms\_flt3")  
# plot\_x\_y\_vars("clms\_flt2" ,"clms\_flt4")  
# plot\_x\_y\_vars("clms\_flt2" ,"clms\_flt5")  
# plot\_x\_y\_vars("clms\_flt3" ,"clms\_flt4")  
# plot\_x\_y\_vars("clms\_flt3" ,"clms\_flt5")  
# plot\_x\_y\_vars("clms\_flt4" ,"clms\_flt5")