Final Project

Mitali Yadav (3034158469)

```
library("spls")
```

```
## Warning: package 'spls' was built under R version 4.0.4

## Sparse Partial Least Squares (SPLS) Regression and
## Classification (version 2.2-3)
```

```
library("plsr")
```

```
## Warning: package 'plsr' was built under R version 4.0.4

## Be aware that plsr 0.0.1 contains experimental and partly untested code.
## Use cautiously.

##
## Attaching package: 'plsr'

## The following object is masked from 'package:stats':
##
##     loadings
```

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## Warning: package 'dplyr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library("caret")
```

```
## Warning: package 'caret' was built under R version 4.0.4

## Loading required package: lattice

## Registered S3 methods overwritten by 'caret':
##   method         from
##   predict.splsda spls
##   print.splsda   spls

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## The following object is masked from 'package:spls':
##
##     splsda
```

```r
library("glmnet")
```

```
## Warning: package 'glmnet' was built under R version 4.0.4

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-1
```

```r
library("rpart")
```

```
## Warning: package 'rpart' was built under R version 4.0.4
```

```r
library("rpart.plot")
```

```
## Warning: package 'rpart.plot' was built under R version 4.0.4
```

```r
library("ipred")
```

```
## Warning: package 'ipred' was built under R version 4.0.4
```

```r
library("randomForest")
```

```
## Warning: package 'randomForest' was built under R version 4.0.4
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library("stats")
library("stargazer")
```

```
## Warning: package 'stargazer' was built under R version 4.0.3
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library("moderndive")
```

```
## Warning: package 'moderndive' was built under R version 4.0.4
```

```r
library("readxl")
```

```r
#importing the train and test dataset
train_set = read_csv('train.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   TMC = col_double(),
##   Severity = col_double(),
##   Start_Time = col_datetime(format = ""),
##   End_Time = col_datetime(format = ""),
##   Start_Lat = col_double(),
##   Start_Lng = col_double(),
##   End_Lat = col_double(),
##   End_Lng = col_double(),
##   Distance.mi. = col_double(),
##   Number = col_double(),
##   Weather_Timestamp = col_datetime(format = ""),
##   Temperature.F. = col_double(),
##   Wind_Chill.F. = col_double(),
##   Humidity... = col_double(),
##   Pressure.in. = col_double(),
##   Visibility.mi. = col_double(),
##   Wind_Speed.mph. = col_double(),
##   Precipitation.in. = col_double(),
##   Amenity = col_logical(),
##   Bump = col_logical()
##   # ... with 11 more columns
## )

## See spec(...) for full column specifications.
```

```r
test_set = read.csv('test.csv')

head(train_set)
```

```
## # A tibble: 6 x 49
##   ID    Source   TMC Severity Start_Time          End_Time            Start_Lat
##   <chr> <chr>  <dbl>    <dbl> <dttm>              <dttm>                  <dbl>
## 1 A-20~ MapQu~   201        2 2018-07-19 20:30:23 2018-07-19 21:14:11      34.2
## 2 A-33~ Bing      NA        2 2020-12-27 13:22:48 2020-12-27 15:02:42      40.3
## 3 A-32~ Bing      NA        2 2020-12-19 20:27:52 2020-12-19 22:23:39      30.0
## 4 A-27~ Bing      NA        3 2016-09-27 17:29:27 2016-09-27 23:29:27      39.0
## 5 A-37~ Bing      NA        2 2020-02-11 19:22:00 2020-02-11 23:22:00      45.7
## 6 A-40~ MapQu~   201        2 2017-04-08 07:42:02 2017-04-08 08:10:29      34.0
## # ... with 42 more variables: Start_Lng <dbl>, End_Lat <dbl>, End_Lng <dbl>,
## #   Distance.mi. <dbl>, Description <chr>, Number <dbl>, Street <chr>,
## #   Side <chr>, City <chr>, County <chr>, State <chr>, Zipcode <chr>,
## #   Country <chr>, Timezone <chr>, Airport_Code <chr>,
## #   Weather_Timestamp <dttm>, Temperature.F. <dbl>, Wind_Chill.F. <dbl>,
## #   Humidity... <dbl>, Pressure.in. <dbl>, Visibility.mi. <dbl>,
## #   Wind_Direction <chr>, Wind_Speed.mph. <dbl>, Precipitation.in. <dbl>,
## #   Weather_Condition <chr>, Amenity <lgl>, Bump <lgl>, Crossing <lgl>,
## #   Give_Way <lgl>, Junction <lgl>, No_Exit <lgl>, Railway <lgl>,
## #   Roundabout <lgl>, Station <lgl>, Stop <lgl>, Traffic_Calming <lgl>,
## #   Traffic_Signal <lgl>, Turning_Loop <lgl>, Sunrise_Sunset <chr>,
## #   Civil_Twilight <chr>, Nautical_Twilight <chr>, Astronomical_Twilight <chr>
```

```r
head(test_set)
```

```
##     ID   Source TMC          Start_Time            End_Time Start_Lat Start_Lng
## 1  A-1 MapQuest 201 2016-02-08 05:46:00 2016-02-08 11:00:00  39.86515 -84.05872
## 2  A-5 MapQuest 201 2016-02-08 07:39:07 2016-02-08 08:09:07  39.62778 -84.18835
## 3  A-7 MapQuest 201 2016-02-08 07:59:35 2016-02-08 08:29:35  39.75827 -84.23051
## 4 A-14 MapQuest 201 2016-02-08 08:37:07 2016-02-08 09:07:07  39.79076 -84.24155
## 5 A-22 MapQuest 201 2016-02-08 10:24:27 2016-02-08 10:54:27  39.77335 -84.22469
## 6 A-39 MapQuest 201 2016-02-09 05:17:08 2016-02-09 05:47:08  39.78258 -84.17869
##   End_Lat End_Lng Distance.mi.
## 1      NA      NA         0.01
## 2      NA      NA         0.01
## 3      NA      NA         0.00
## 4      NA      NA         0.01
## 5      NA      NA         0.00
## 6      NA      NA         0.01
##                                                                              Description
## 1 Right lane blocked due to accident on I-70 Eastbound at Exit 41 OH-235 State Route 4.
## 2            Accident on McEwen Rd at OH-725 Miamisburg Centerville Rd. Expect delays.
## 3                              Accident on Oakridge Dr at Woodward Ave. Expect delays.
## 4               Accident on Salem Ave at Hillcrest Ave / Kensington Dr. Expect delays.
## 5                               Accident on Princeton Dr at Catalpa Dr. Expect delays.
## 6                                       Accident on Leo St at Kiser St. Expect delays.
##   Number                    Street Side   City     County State    Zipcode
## 1     NA                    I-70 E    R Dayton Montgomery    OH      45424
## 2     NA Miamisburg Centerville Rd    R Dayton Montgomery    OH      45459
## 3    376            N Woodward Ave    R Dayton Montgomery    OH 45417-2476
## 4   3198                 Salem Ave    L Dayton Montgomery    OH 45406-2708
## 5   1391              Princeton Dr    R Dayton Montgomery    OH 45406-4736
## 6    898                  Kiser St    R Dayton Montgomery    OH 45404-1672
##   Country   Timezone Airport_Code   Weather_Timestamp Temperature.F.
## 1      US US/Eastern         KFFO 2016-02-08 05:58:00           36.9
## 2      US US/Eastern         KMGY 2016-02-08 07:53:00           36.0
## 3      US US/Eastern         KDAY 2016-02-08 07:56:00           34.0
## 4      US US/Eastern         KDAY 2016-02-08 08:56:00           36.0
## 5      US US/Eastern         KDAY 2016-02-08 09:56:00           36.0
## 6      US US/Eastern         KFFO 2016-02-09 04:58:00           22.8
##   Wind_Chill.F. Humidity... Pressure.in. Visibility.mi. Wind_Direction
## 1            NA          91        29.68             10           Calm
## 2          33.3          89        29.65              6             SW
## 3          31.0         100        29.66              7            WSW
## 4          31.1          89        29.65             10             NW
## 5          30.3          89        29.65             10           West
## 6          11.5          89        29.69              4             SW
##   Wind_Speed.mph. Precipitation.in. Weather_Condition Amenity  Bump Crossing
## 1              NA              0.02        Light Rain   False False    False
## 2             3.5                NA     Mostly Cloudy   False False    False
## 3             3.5                NA          Overcast   False False    False
## 4             5.8                NA     Mostly Cloudy   False False    False
## 5             6.9                NA     Mostly Cloudy   False False    False
## 6            11.5              0.00        Light Snow   False False    False
##   Give_Way Junction No_Exit Railway Roundabout Station  Stop Traffic_Calming
## 1    False    False   False   False      False   False False           False
## 2    False    False   False   False      False   False False           False
## 3    False    False   False   False      False   False False           False
## 4    False    False   False   False      False   False False           False
## 5    False    False   False   False      False   False False           False
## 6    False    False   False   False      False   False False           False
##   Traffic_Signal Turning_Loop Sunrise_Sunset Civil_Twilight Nautical_Twilight
## 1          False        False          Night          Night             Night
## 2           True         False             Day            Day              Day
## 3          False         False             Day            Day              Day
## 4           True         False             Day            Day              Day
## 5          False         False             Day            Day              Day
## 6          False         False           Night          Night            Night
##   Astronomical_Twilight
## 1                 Night
## 2                   Day
## 3                   Day
## 4                   Day
## 5                   Day
## 6                 Night
```

```
names(train_set)
```

```
##  [1] "ID"                    "Source"                "TMC"
##  [4] "Severity"              "Start_Time"            "End_Time"
##  [7] "Start_Lat"             "Start_Lng"             "End_Lat"
## [10] "End_Lng"               "Distance.mi."          "Description"
## [13] "Number"                "Street"                "Side"
## [16] "City"                  "County"                "State"
## [19] "Zipcode"               "Country"               "Timezone"
## [22] "Airport_Code"          "Weather_Timestamp"     "Temperature.F."
## [25] "Wind_Chill.F."         "Humidity..."           "Pressure.in."
## [28] "Visibility.mi."        "Wind_Direction"        "Wind_Speed.mph."
## [31] "Precipitation.in."     "Weather_Condition"     "Amenity"
## [34] "Bump"                  "Crossing"              "Give_Way"
## [37] "Junction"              "No_Exit"               "Railway"
## [40] "Roundabout"            "Station"               "Stop"
## [43] "Traffic_Calming"       "Traffic_Signal"        "Turning_Loop"
## [46] "Sunrise_Sunset"        "Civil_Twilight"        "Nautical_Twilight"
## [49] "Astronomical_Twilight"
```

EDA

1. Starting with changing the target variable'
   If Severity <2 => 1
   Else Severity => 0

```
train_set['Y'] = as.integer(train_set$Severity > 2)
names(train_set)
```

```
##  [1] "ID"                    "Source"                "TMC"
##  [4] "Severity"              "Start_Time"            "End_Time"
##  [7] "Start_Lat"             "Start_Lng"             "End_Lat"
## [10] "End_Lng"               "Distance.mi."          "Description"
## [13] "Number"                "Street"                "Side"
## [16] "City"                  "County"                "State"
## [19] "Zipcode"               "Country"               "Timezone"
## [22] "Airport_Code"          "Weather_Timestamp"     "Temperature.F."
## [25] "Wind_Chill.F."         "Humidity..."           "Pressure.in."
## [28] "Visibility.mi."        "Wind_Direction"        "Wind_Speed.mph."
## [31] "Precipitation.in."     "Weather_Condition"     "Amenity"
## [34] "Bump"                  "Crossing"              "Give_Way"
## [37] "Junction"              "No_Exit"               "Railway"
## [40] "Roundabout"            "Station"               "Stop"
## [43] "Traffic_Calming"       "Traffic_Signal"        "Turning_Loop"
## [46] "Sunrise_Sunset"        "Civil_Twilight"        "Nautical_Twilight"
## [49] "Astronomical_Twilight" "Y"
```

```
dim(train_set)
```

```
## [1] 2962779      50
```

```
#getting number of nulls in each column
na_count = sapply(train_set, function(x) { round(length(which(is.na(x)))/nrow(train_set),3)})
na_count_df = data.frame(na_count)
na_count_df
```

```
##                       na_count
## ID                       0.000
## Source                   0.000
## TMC                      0.358
## Severity                 0.000
## Start_Time               0.000
## End_Time                 0.000
## Start_Lat                0.000
## Start_Lng                0.000
## End_Lat                  0.642
## End_Lng                  0.642
## Distance.mi.             0.000
## Description              0.000
## Number                   0.635
## Street                   0.000
## Side                     0.000
## City                     0.000
## County                   0.000
## State                    0.000
## Zipcode                  0.000
## Country                  0.000
## Timezone                 0.001
## Airport_Code             0.002
## Weather_Timestamp        0.015
## Temperature.F.           0.021
## Wind_Chill.F.            0.448
## Humidity...              0.023
## Pressure.in.             0.018
## Visibility.mi.           0.023
## Wind_Direction           0.020
## Wind_Speed.mph.          0.113
## Precipitation.in.        0.488
## Weather_Condition        0.023
## Amenity                  0.000
## Bump                     0.000
## Crossing                  0.000
## Give_Way                  0.000
## Junction                  0.000
## No_Exit                   0.000
## Railway                   0.000
## Roundabout                0.000
## Station                   0.000
## Stop                      0.000
## Traffic_Calming           0.000
## Traffic_Signal            0.000
## Turning_Loop              0.000
## Sunrise_Sunset            0.000
## Civil_Twilight            0.000
## Nautical_Twilight         0.000
## Astronomical_Twilight     0.000
## Y                         0.000
```

```
#column- diff in lat and long
#columns with <60% data missing should be eliminated?
```