Advaned Analytics and Machine Learning Assignment 1

15th-March-2024

Introduction

The essence of this assignment was to deal with an exceptionally large dataset. As for the workflow, we were required to do the following things. 1. Follow the instructions and generate a large dataset which should not be lower than 30 MBs in size. 2. Perform an exploratory analysis of the dataset accompanied by a zero analysis. 3. Make relevant visualization to observe any patterns. 4. Divide the dataset into four on the basis of the variable called product.field_description and perform the following tasks for each of the subset: Lasso Regression, Logistic Regression, Linear Discriminant Analysis (LDA). 5. Compare the results of each Logistic Regression and LDA.

Dataset Generation

As per the instructions, the generated dataset should have been around 300 – 400 MBs. However, if a lower size was generated, that was accepted if the size is not lower than 30 MB. My dataset generated was 154 MBs with 552,422 observations. I did not need to change any other line of code except to add my student number.

Codes

Following are all the pre processing up to this point.

```
#Load libraries
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.3
```

```
## Loaded glmnet 4.1-8
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.2.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.3
```

```
##
## Attaching package: 'MASS'
```

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

```
#set seed
set.seed(1)

#Load dataset
df <- read.csv("C:/Users/User/Desktop/QUB Classes/Semester 2/Advanced Analytics/assignment 1/datasets/output_404266
85.csv")

#summarize dataset
dim(df)
```

```
## [1] 552422     77
```

```
head(df)
```

```
##   ID_non_uniq date_event last_year_all_product_codes_num_uniq
## 1     p860004   20-11-12                                   1
## 2     p860004   22-11-12                                   1
## 3     p860004   27-11-12                                   1
## 4     p860004   27-11-12                                   1
## 5     p860004   13-12-12                                   1
## 6     p860004   01-01-13                                   1
##   last_year_all_product_codes_most_freq last_year_brand_name_num_uniq
## 1                                  1458                             1
## 2                                  1458                             1
## 3                                  1458                             1
## 4                                  1458                             1
## 5                                  1458                             1
## 6                                  1458                             1
##   last_year_brand_name_most_freq last_year_classification0_num_uniq
## 1                           4789                                  2
## 2                           4789                                  2
## 3                           4789                                  4
## 4                           4789                                  4
## 5                           4789                                  2
## 6                           4789                                 12
##   last_year_classification1_num_uniq last_year_classification2_num_uniq
## 1                                  0                                  0
## 2                                  0                                  0
## 3                                  0                                  0
## 4                                  0                                  0
## 5                                  0                                  0
## 6                                  0                                  0
##   last_year_company_name_num_uniq last_year_company_name_most_freq
## 1                               1                              349
## 2                               1                              349
## 3                               1                              349
## 4                               1                              349
## 5                               1                              349
## 6                               1                              349
##   last_year_reason_for_legal_announcement_num_uniq
## 1                                                0
## 2                                                0
## 3                                                0
## 4                                                0
## 5                                                0
## 6                                                0
##   last_year_reason_for_legal_announcement_most_freq
## 1                                                 0
## 2                                                 0
## 3                                                 0
## 4                                                 0
## 5                                                 0
## 6                                                 0
##   last_year_legal_announcementing_firm_num_uniq
## 1                                             1
## 2                                             1
## 3                                             1
## 4                                             1
## 5                                             1
## 6                                             1
##   last_year_legal_announcementing_firm_most_freq
## 1                                            292
## 2                                            292
## 3                                            292
## 4                                            292
## 5                                            292
```

```
## 6                                          292
##   last_year_root_cause_description_num_uniq
## 1                                         1
## 2                                         1
## 3                                         1
## 4                                         1
## 5                                         1
## 6                                         1
##   last_year_root_cause_description_most_freq
## 1                                         19
## 2                                         19
## 3                                         19
## 4                                         19
## 5                                         19
## 6                                         19
##   last_year_product_quantity_average_num_uniq
## 1                                           1
## 2                                           1
## 3                                           1
## 4                                           1
## 5                                           1
## 6                                           1
##   last_year_product_quantity_average_max
## 1                                174700
## 2                                174700
## 3                                174700
## 4                                174700
## 5                                174700
## 6                                174700
##   last_year_product_quantity_average_average
## 1                                     174700
## 2                                     174700
## 3                                     174700
## 4                                     174700
## 5                                     174700
## 6                                     174700
##   last_year_decision_date_max_changes_in_product
## 1                                             17
## 2                                             17
## 3                                             17
## 4                                             17
## 5                                             15
## 6                                             15
##   last_year_decision_date_average_changes_in_product
## 1                                                 17
## 2                                                 17
## 3                                                 17
## 4                                                 17
## 5                                                 15
## 6                                                 15
##   last_two_years_all_product_codes_num_uniq
## 1                                         1
## 2                                         1
## 3                                         1
## 4                                         1
## 5                                         1
## 6                                         1
##   last_two_years_all_product_codes_most_freq last_two_years_brand_name_num_uniq
## 1                                       1458                                  1
## 2                                       1458                                  1
## 3                                       1458                                  1
## 4                                       1458                                  1
```

```
## 5                                    1458                                    1
## 6                                    1458                                    1
##   last_two_years_brand_name_most_freq last_two_years_classification0_num_uniq
## 1                                4789                                       2
## 2                                4789                                       2
## 3                                4789                                       4
## 4                                4789                                       4
## 5                                4789                                       2
## 6                                4789                                      12
##   last_two_years_classification1_num_uniq
## 1                                       0
## 2                                       0
## 3                                       0
## 4                                       0
## 5                                       0
## 6                                       0
##   last_two_years_classification2_num_uniq last_two_years_company_name_num_uniq
## 1                                       0                                    1
## 2                                       0                                    1
## 3                                       0                                    1
## 4                                       0                                    1
## 5                                       0                                    1
## 6                                       0                                    1
##   last_two_years_company_name_most_freq
## 1                                   349
## 2                                   349
## 3                                   349
## 4                                   349
## 5                                   349
## 6                                   349
##   last_two_years_reason_for_legal_announcement_num_uniq
## 1                                                     0
## 2                                                     0
## 3                                                     0
## 4                                                     0
## 5                                                     0
## 6                                                     0
##   last_two_years_reason_for_legal_announcement_most_freq
## 1                                                      0
## 2                                                      0
## 3                                                      0
## 4                                                      0
## 5                                                      0
## 6                                                      0
##   last_two_years_legal_announcementing_firm_num_uniq
## 1                                                  1
## 2                                                  1
## 3                                                  1
## 4                                                  1
## 5                                                  1
## 6                                                  1
##   last_two_years_legal_announcementing_firm_most_freq
## 1                                                 292
## 2                                                 292
## 3                                                 292
## 4                                                 292
## 5                                                 292
## 6                                                 292
##   last_two_years_root_cause_description_num_uniq
## 1                                              1
## 2                                              1
## 3                                              1
```

```
## 4                                          1
## 5                                          1
## 6                                          1
##   last_two_years_root_cause_description_most_freq
## 1                                          19
## 2                                          19
## 3                                          19
## 4                                          19
## 5                                          19
## 6                                          19
##   last_two_years_product_quantity_average_num_uniq
## 1                                           1
## 2                                           1
## 3                                           1
## 4                                           1
## 5                                           1
## 6                                           1
##   last_two_years_product_quantity_average_max
## 1                                      174700
## 2                                      174700
## 3                                      174700
## 4                                      174700
## 5                                      174700
## 6                                      174700
##   last_two_years_product_quantity_average_average
## 1                                        174700
## 2                                        174700
## 3                                        174700
## 4                                        174700
## 5                                        174700
## 6                                        174700
##   last_two_years_decision_date_max_changes_in_product
## 1                                                34
## 2                                                34
## 3                                                33
## 4                                                33
## 5                                                33
## 6                                                35
##   last_two_years_decision_date_average_changes_in_product
## 1                                                      34
## 2                                                      34
## 3                                                      33
## 4                                                      33
## 5                                                      33
## 6                                                      35
##   last_four_years_all_product_codes_num_uniq
## 1                                        1
## 2                                        1
## 3                                        1
## 4                                        1
## 5                                        1
## 6                                        1
##   last_four_years_all_product_codes_most_freq
## 1                                      1458
## 2                                      1458
## 3                                      1458
## 4                                      1458
## 5                                      1458
## 6                                      1458
##   last_four_years_brand_name_num_uniq last_four_years_brand_name_most_freq
## 1                                   1                                4789
## 2                                   1                                4789
```

```
## 3                                  1                                4789
## 4                                  1                                4789
## 5                                  1                                4789
## 6                                  1                                4789
##   last_four_years_classification0_num_uniq
## 1                                        2
## 2                                        2
## 3                                        4
## 4                                        4
## 5                                        2
## 6                                       12
##   last_four_years_classification1_num_uniq
## 1                                        0
## 2                                        0
## 3                                        0
## 4                                        0
## 5                                        0
## 6                                        0
##   last_four_years_classification2_num_uniq
## 1                                        0
## 2                                        0
## 3                                        0
## 4                                        0
## 5                                        0
## 6                                        0
##   last_four_years_company_name_num_uniq last_four_years_company_name_most_freq
## 1                                     1                                     349
## 2                                     1                                     349
## 3                                     1                                     349
## 4                                     1                                     349
## 5                                     1                                     349
## 6                                     1                                     349
##   last_four_years_reason_for_legal_announcement_num_uniq
## 1                                                      0
## 2                                                      0
## 3                                                      0
## 4                                                      0
## 5                                                      0
## 6                                                      0
##   last_four_years_reason_for_legal_announcement_most_freq
## 1                                                       0
## 2                                                       0
## 3                                                       0
## 4                                                       0
## 5                                                       0
## 6                                                       0
##   last_four_years_legal_announcementing_firm_num_uniq
## 1                                                   1
## 2                                                   1
## 3                                                   1
## 4                                                   1
## 5                                                   1
## 6                                                   1
##   last_four_years_legal_announcementing_firm_most_freq
## 1                                                  292
## 2                                                  292
## 3                                                  292
## 4                                                  292
## 5                                                  292
## 6                                                  292
##   last_four_years_root_cause_description_num_uniq
## 1                                               1
```

```
## 2                                          1
## 3                                          1
## 4                                          1
## 5                                          1
## 6                                          1
##   last_four_years_root_cause_description_most_freq
## 1                                               19
## 2                                               19
## 3                                               19
## 4                                               19
## 5                                               19
## 6                                               19
##   last_four_years_product_quantity_average_num_uniq
## 1                                                 1
## 2                                                 1
## 3                                                 1
## 4                                                 1
## 5                                                 1
## 6                                                 1
##   last_four_years_product_quantity_average_max
## 1                                       174700
## 2                                       174700
## 3                                       174700
## 4                                       174700
## 5                                       174700
## 6                                       174700
##   last_four_years_product_quantity_average_average
## 1                                           174700
## 2                                           174700
## 3                                           174700
## 4                                           174700
## 5                                           174700
## 6                                           174700
##   last_four_years_decision_date_max_changes_in_product
## 1                                                   65
## 2                                                   65
## 3                                                   65
## 4                                                   65
## 5                                                   66
## 6                                                   68
##   last_four_years_decision_date_average_changes_in_product
## 1                                                       65
## 2                                                       65
## 3                                                       65
## 4                                                       65
## 5                                                       66
## 6                                                       68
##   Product.issue.consequence manufacturer_contact_address_1 product.brand_name
## 1                    Injury                            9476             281286
## 2                    Injury                            9476             281286
## 3                    Injury                            9476             281286
## 4                Malfunction                           9476             281286
## 5                Malfunction                           9476             281286
## 6                    Injury                            9476             281286
##   product.generic_name product.issue.type type_of_report.1 reporter_job_code
## 1                73852               438                0                32
## 2                73852               629                1                42
## 3                73852               456                1                32
## 4                73852               599                1                32
## 5                73852               599                1                32
## 6                73852               906                0                32
##   source_type product.manufacturer_name product.product_operator
```

```
## 1           6                    18383                       15
## 2          10                    19327                       15
## 3           6                    18383                       15
## 4           5                    18383                       15
## 5           5                    19327                       15
## 6          12                    18383                       15
##   product.manufacturer_city product.manufacturer_state
## 1                      4513                        48
## 2                      5990                        32
## 3                      4513                        48
## 4                      4513                        48
## 5                      5990                        32
## 6                      4513                        48
##   product.manufacturer_country product.field_description
## 1                          126                   Unknown
## 2                          126                   Unknown
## 3                          126                   Unknown
## 4                          126                   Unknown
## 5                          126                   Unknown
## 6                          126                   Unknown
##   product.product_report_product_code
## 1                                 LKK
## 2                                 LKK
## 3                                 LKK
## 4                                 LKK
## 5                                 LKK
## 6                                 LKK
```

```
summary(df)
```

```
##  ID_non_uniq         date_event        last_year_all_product_codes_num_uniq
##  Length:552422      Length:552422      Min.   :0.000
##  Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :1.000
##                                        Mean   :1.438
##                                        3rd Qu.:2.000
##                                        Max.   :8.000
##  last_year_all_product_codes_most_freq last_year_brand_name_num_uniq
##  Min.   :   0                          Min.   :0.00
##  1st Qu.:1465                          1st Qu.:1.00
##  Median :1472                          Median :2.00
##  Mean   :1458                          Mean   :1.85
##  3rd Qu.:1480                          3rd Qu.:2.00
##  Max.   :3566                          Max.   :8.00
##  last_year_brand_name_most_freq last_year_classification0_num_uniq
##  Min.   :   0                   Min.   :  0.000
##  1st Qu.:2216                   1st Qu.:  0.000
##  Median :4056                   Median :  0.000
##  Mean   :3342                   Mean   :  8.766
##  3rd Qu.:4789                   3rd Qu.:  6.000
##  Max.   :4789                   Max.   :300.000
##  last_year_classification1_num_uniq last_year_classification2_num_uniq
##  Min.   :  0.00                     Min.   : 0.000
##  1st Qu.:  4.00                     1st Qu.: 0.000
##  Median : 12.00                     Median : 0.000
##  Mean   : 19.97                     Mean   : 1.889
##  3rd Qu.: 25.00                     3rd Qu.: 0.000
##  Max.   :300.00                     Max.   :85.000
##  last_year_company_name_num_uniq last_year_company_name_most_freq
##  Min.   :0.0000                  Min.   :  0.0
##  1st Qu.:1.0000                  1st Qu.:349.0
##  Median :1.0000                  Median :349.0
##  Mean   :0.9887                  Mean   :344.3
##  3rd Qu.:1.0000                  3rd Qu.:349.0
##  Max.   :2.0000                  Max.   :489.0
##  last_year_reason_for_legal_announcement_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :2.996
##  3rd Qu.:4.000
##  Max.   :8.000
##  last_year_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
##  1st Qu.:  52.0
##  Median : 698.0
##  Mean   : 708.1
##  3rd Qu.:1143.0
##  Max.   :1401.0
##  last_year_legal_announcementing_firm_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :1.689
##  3rd Qu.:2.000
##  Max.   :3.000
##  last_year_legal_announcementing_firm_most_freq
##  Min.   :  0.0
##  1st Qu.:143.0
##  Median :292.0
##  Mean   :253.6
##  3rd Qu.:292.0
```

```
##  Max.   :440.0
##  last_year_root_cause_description_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.216
##  3rd Qu.:3.000
##  Max.   :5.000
##  last_year_root_cause_description_most_freq
##  Min.   : 0.0
##  1st Qu.: 4.0
##  Median :28.0
##  Mean   :22.7
##  3rd Qu.:28.0
##  Max.   :40.0
##  last_year_product_quantity_average_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.497
##  3rd Qu.:3.000
##  Max.   :6.000
##  last_year_product_quantity_average_max
##  Min.   :     0
##  1st Qu.:  5463
##  Median : 20286
##  Mean   :103998
##  3rd Qu.: 22298
##  Max.   :636572
##  last_year_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  4791
##  Median :  6892
##  Mean   : 43638
##  3rd Qu.: 11154
##  Max.   :333373
##  last_year_decision_date_max_changes_in_product
##  Min.   : 0.00
##  1st Qu.:16.00
##  Median :20.00
##  Mean   :20.28
##  3rd Qu.:24.00
##  Max.   :33.00
##  last_year_decision_date_average_changes_in_product
##  Min.   : 0.00
##  1st Qu.:16.00
##  Median :20.00
##  Mean   :20.28
##  3rd Qu.:24.00
##  Max.   :33.00
##  last_two_years_all_product_codes_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.454
##  3rd Qu.:2.000
##  Max.   :8.000
##  last_two_years_all_product_codes_most_freq last_two_years_brand_name_num_uniq
##  Min.   :   0                               Min.   :0.000
##  1st Qu.:1466                               1st Qu.:1.000
##  Median :1473                               Median :2.000
##  Mean   :1475                               Mean   :1.871
```

```
##  3rd Qu.:1480                              3rd Qu.:2.000
##  Max.   :6593                              Max.   :8.000
##  last_two_years_brand_name_most_freq last_two_years_classification0_num_uniq
##  Min.   :   0                        Min.   :   0.00
##  1st Qu.:2817                        1st Qu.:   0.00
##  Median :4056                        Median :   6.00
##  Mean   :3378                        Mean   :  17.29
##  3rd Qu.:4789                        3rd Qu.:  20.00
##  Max.   :4789                        Max.   :1449.00
##  last_two_years_classification1_num_uniq
##  Min.   :  0.00
##  1st Qu.:  8.00
##  Median : 24.00
##  Mean   : 40.49
##  3rd Qu.: 54.00
##  Max.   :665.00
##  last_two_years_classification2_num_uniq last_two_years_company_name_num_uniq
##  Min.   : 0.000                          Min.   :0.0000
##  1st Qu.: 0.000                          1st Qu.:1.0000
##  Median : 0.000                          Median :1.0000
##  Mean   : 2.339                          Mean   :0.9999
##  3rd Qu.: 0.000                          3rd Qu.:1.0000
##  Max.   :85.000                          Max.   :2.0000
##  last_two_years_company_name_most_freq
##  Min.   :  0.0
##  1st Qu.:349.0
##  Median :349.0
##  Mean   :348.2
##  3rd Qu.:349.0
##  Max.   :489.0
##  last_two_years_reason_for_legal_announcement_num_uniq
##  Min.   : 0.000
##  1st Qu.: 5.000
##  Median : 5.000
##  Mean   : 5.803
##  3rd Qu.: 7.000
##  Max.   :11.000
##  last_two_years_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
##  1st Qu.:  52.0
##  Median :1028.0
##  Mean   : 827.7
##  3rd Qu.:1143.0
##  Max.   :1401.0
##  last_two_years_legal_announcementing_firm_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.081
##  3rd Qu.:2.000
##  Max.   :4.000
##  last_two_years_legal_announcementing_firm_most_freq
##  Min.   :  0.0
##  1st Qu.:292.0
##  Median :292.0
##  Mean   :283.1
##  3rd Qu.:292.0
##  Max.   :440.0
##  last_two_years_root_cause_description_num_uniq
##  Min.   :0.000
##  1st Qu.:3.000
##  Median :3.000
```

```
##  Mean   :3.712
##  3rd Qu.:5.000
##  Max.   :6.000
##  last_two_years_root_cause_description_most_freq
##  Min.   : 0.00
##  1st Qu.: 4.00
##  Median :28.00
##  Mean   :24.53
##  3rd Qu.:40.00
##  Max.   :40.00
##  last_two_years_product_quantity_average_num_uniq
##  Min.   :0.000
##  1st Qu.:4.000
##  Median :5.000
##  Mean   :4.686
##  3rd Qu.:6.000
##  Max.   :8.000
##  last_two_years_product_quantity_average_max
##  Min.   :     0
##  1st Qu.: 13784
##  Median : 22298
##  Mean   :196851
##  3rd Qu.:636572
##  Max.   :636572
##  last_two_years_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  4959
##  Median :  7499
##  Mean   : 65481
##  3rd Qu.:167634
##  Max.   :333373
##  last_two_years_decision_date_max_changes_in_product
##  Min.   : 0.00
##  1st Qu.:34.00
##  Median :40.00
##  Mean   :39.28
##  3rd Qu.:44.00
##  Max.   :55.00
##  last_two_years_decision_date_average_changes_in_product
##  Min.   : 0.00
##  1st Qu.:34.00
##  Median :40.00
##  Mean   :39.28
##  3rd Qu.:44.00
##  Max.   :55.00
##  last_four_years_all_product_codes_num_uniq
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.455
##  3rd Qu.:2.000
##  Max.   :8.000
##  last_four_years_all_product_codes_most_freq
##  Min.   : 279
##  1st Qu.:1466
##  Median :1473
##  Mean   :1476
##  3rd Qu.:1480
##  Max.   :6593
##  last_four_years_brand_name_num_uniq last_four_years_brand_name_most_freq
##  Min.   :1.000                       Min.   :  97
##  1st Qu.:1.000                       1st Qu.:2817
```

```
##  Median :2.000                       Median :4056
##  Mean   :1.871                       Mean   :3379
##  3rd Qu.:2.000                       3rd Qu.:4789
##  Max.   :8.000                       Max.   :4789
##  last_four_years_classification0_num_uniq
##  Min.   :   0.00
##  1st Qu.:  12.00
##  Median :  24.00
##  Mean   :  43.64
##  3rd Qu.:  60.00
##  Max.   :1674.00
##  last_four_years_classification1_num_uniq
##  Min.   :   0.00
##  1st Qu.:  14.00
##  Median :  40.00
##  Mean   :  67.52
##  3rd Qu.:  96.00
##  Max.   :3348.00
##  last_four_years_classification2_num_uniq last_four_years_company_name_num_uniq
##  Min.   : 0.000                           Min.   :1
##  1st Qu.: 0.000                           1st Qu.:1
##  Median : 0.000                           Median :1
##  Mean   : 2.498                           Mean   :1
##  3rd Qu.: 0.000                           3rd Qu.:1
##  Max.   :85.000                           Max.   :2
##  last_four_years_company_name_most_freq
##  Min.   :111.0
##  1st Qu.:349.0
##  Median :349.0
##  Mean   :348.2
##  3rd Qu.:349.0
##  Max.   :489.0
##  last_four_years_reason_for_legal_announcement_num_uniq
##  Min.   : 0.00
##  1st Qu.: 9.00
##  Median :11.00
##  Mean   :10.25
##  3rd Qu.:13.00
##  Max.   :16.00
##  last_four_years_reason_for_legal_announcement_most_freq
##  Min.   :   0
##  1st Qu.:1028
##  Median :1028
##  Mean   :1038
##  3rd Qu.:1028
##  Max.   :1361
##  last_four_years_legal_announcementing_firm_num_uniq
##  Min.   :1.0
##  1st Qu.:2.0
##  Median :2.0
##  Mean   :2.3
##  3rd Qu.:3.0
##  Max.   :5.0
##  last_four_years_legal_announcementing_firm_most_freq
##  Min.   : 40
##  1st Qu.:292
##  Median :292
##  Mean   :292
##  3rd Qu.:292
##  Max.   :371
##  last_four_years_root_cause_description_num_uniq
##  Min.   :1.0
```

```
##  1st Qu.:5.0
##  Median :6.0
##  Mean   :5.2
##  3rd Qu.:6.0
##  Max.   :6.0
##  last_four_years_root_cause_description_most_freq
##  Min.   : 4.00
##  1st Qu.: 4.00
##  Median : 4.00
##  Mean   :15.08
##  3rd Qu.:40.00
##  Max.   :40.00
##  last_four_years_product_quantity_average_num_uniq
##  Min.   : 0.000
##  1st Qu.: 7.000
##  Median : 8.000
##  Mean   : 7.917
##  3rd Qu.:10.000
##  Max.   :11.000
##  last_four_years_product_quantity_average_max
##  Min.   :     0
##  1st Qu.: 22298
##  Median :636572
##  Mean   :437967
##  3rd Qu.:636572
##  Max.   :636572
##  last_four_years_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  7499
##  Median :123061
##  Mean   :112255
##  3rd Qu.:168517
##  Max.   :333373
##  last_four_years_decision_date_max_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 68.00
##  Median : 71.00
##  Mean   : 74.29
##  3rd Qu.: 79.00
##  Max.   :107.00
##  last_four_years_decision_date_average_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 68.00
##  Median : 71.00
##  Mean   : 74.29
##  3rd Qu.: 79.00
##  Max.   :107.00
##  Product.issue.consequence manufacturer_contact_address_1 product.brand_name
##  Length:552422             Min.   : 2179                  Min.   :   130
##  Class :character          1st Qu.: 9476                  1st Qu.:281286
##  Mode  :character          Median : 9476                  Median :281286
##                            Mean   : 9556                  Mean   :276447
##                            3rd Qu.: 9476                  3rd Qu.:281286
##                            Max.   :13145                  Max.   :344588
##  product.generic_name product.issue.type type_of_report.1 reporter_job_code
##  Min.   :    77       Min.   :  2.0      Min.   :0.0000   Min.   : 1.00
##  1st Qu.: 73852       1st Qu.:566.0      1st Qu.:0.0000   1st Qu.:32.00
##  Median : 73852       Median :599.0      Median :0.0000   Median :32.00
##  Mean   : 73600       Mean   :577.4      Mean   :0.2831   Mean   :34.48
##  3rd Qu.: 73852       3rd Qu.:629.0      3rd Qu.:1.0000   3rd Qu.:42.00
##  Max.   :101028       Max.   :964.0      Max.   :1.0000   Max.   :52.00
##   source_type     product.manufacturer_name product.product_operator
```

```
##  Min.   : 3.000   Min.   : 7480            Min.   : 0.00
##  1st Qu.: 3.000   1st Qu.:18383            1st Qu.:15.00
##  Median : 4.000   Median :18383            Median :15.00
##  Mean   : 6.406   Mean   :18818            Mean   :15.12
##  3rd Qu.:11.000   3rd Qu.:19408            3rd Qu.:15.00
##  Max.   :23.000   Max.   :31471            Max.   :41.00
##  product.manufacturer_city product.manufacturer_state
##  Min.   : 1375             Min.   : 8.00
##  1st Qu.: 4513             1st Qu.:48.00
##  Median : 4513             Median :48.00
##  Mean   : 4658             Mean   :46.22
##  3rd Qu.: 4513             3rd Qu.:48.00
##  Max.   :10778             Max.   :63.00
##  product.manufacturer_country product.field_description
##  Min.   :109                  Length:552422
##  1st Qu.:126                  Class :character
##  Median :126                  Mode  :character
##  Mean   :126
##  3rd Qu.:126
##  Max.   :135
##  product.product_report_product_code
##  Length:552422
##  Class :character
##  Mode  :character
##
##
##
```

```
print("ID_non_uniq" %in% names(df))
```

```
## [1] TRUE
```

```
#converting to factor
df1 <- df %>%
  mutate(
    ID_non_uniq = factor(ID_non_uniq),
    Product.issue.consequence = factor(Product.issue.consequence),
    manufacturer_contact_address_1 = factor(manufacturer_contact_address_1),
    product.brand_name = factor(product.brand_name),
    product.generic_name = factor(product.generic_name),
    product.issue.type = factor(product.issue.type),
    type_of_report.1 = factor(type_of_report.1),
    reporter_job_code = factor(reporter_job_code),
    source_type = factor(source_type),
    product.manufacturer_name = factor(product.manufacturer_name),
    product.product_operator = factor(product.product_operator),
    product.manufacturer_city = factor(product.manufacturer_city),
    product.manufacturer_state = factor(product.manufacturer_state),
    product.manufacturer_country = factor(product.manufacturer_country),
    product.field_description = factor(product.field_description),
    product.product_report_product_code = factor(product.product_report_product_code)
  )

#check null values in data
sum(!complete.cases(df1))
```

```
## [1] 0
```

```
#Checking for Unknown values
sum(df1 == "Unknown", na.rm = TRUE)
```

```
## [1] 552251
```

```
sum(is.na(df))
```

```
## [1] 0
```

The data did not have any missing values but almost all the observations had 0s in them. Variables were converted to categorical.

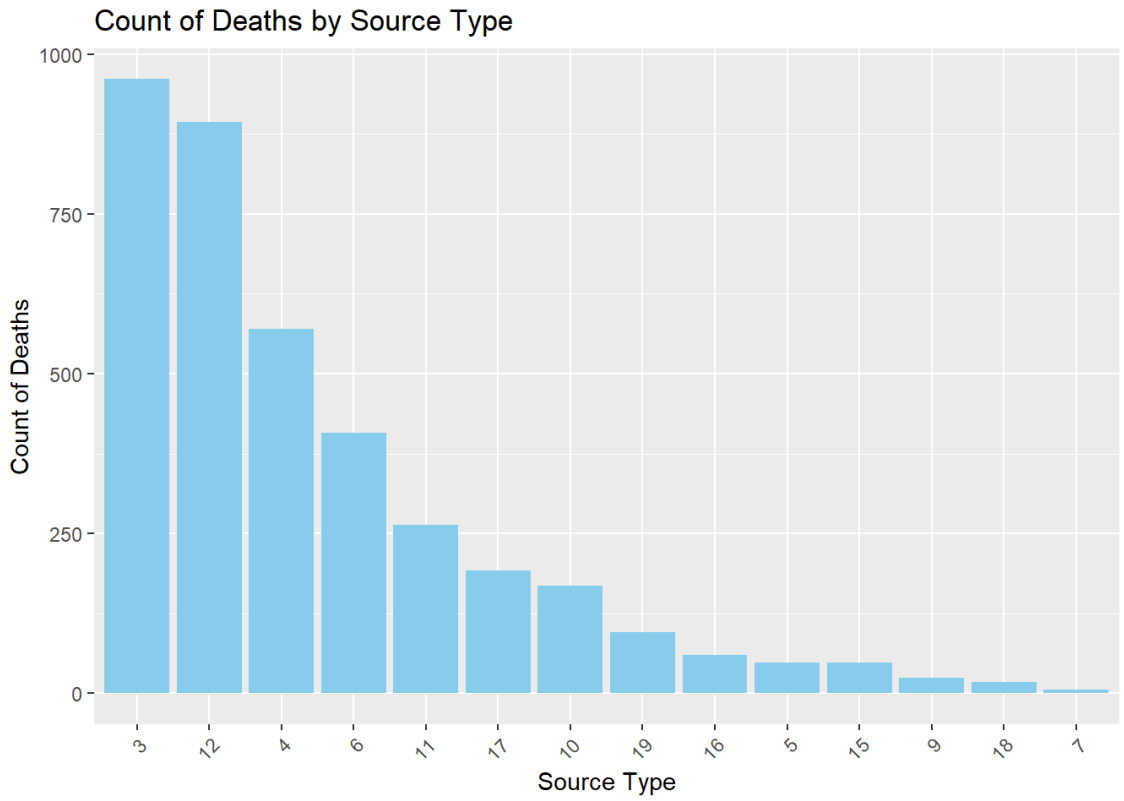
##Exploring each product field

The results here suggest that ID_non_uniq for highest deaths are p080012, p860004, p890055.

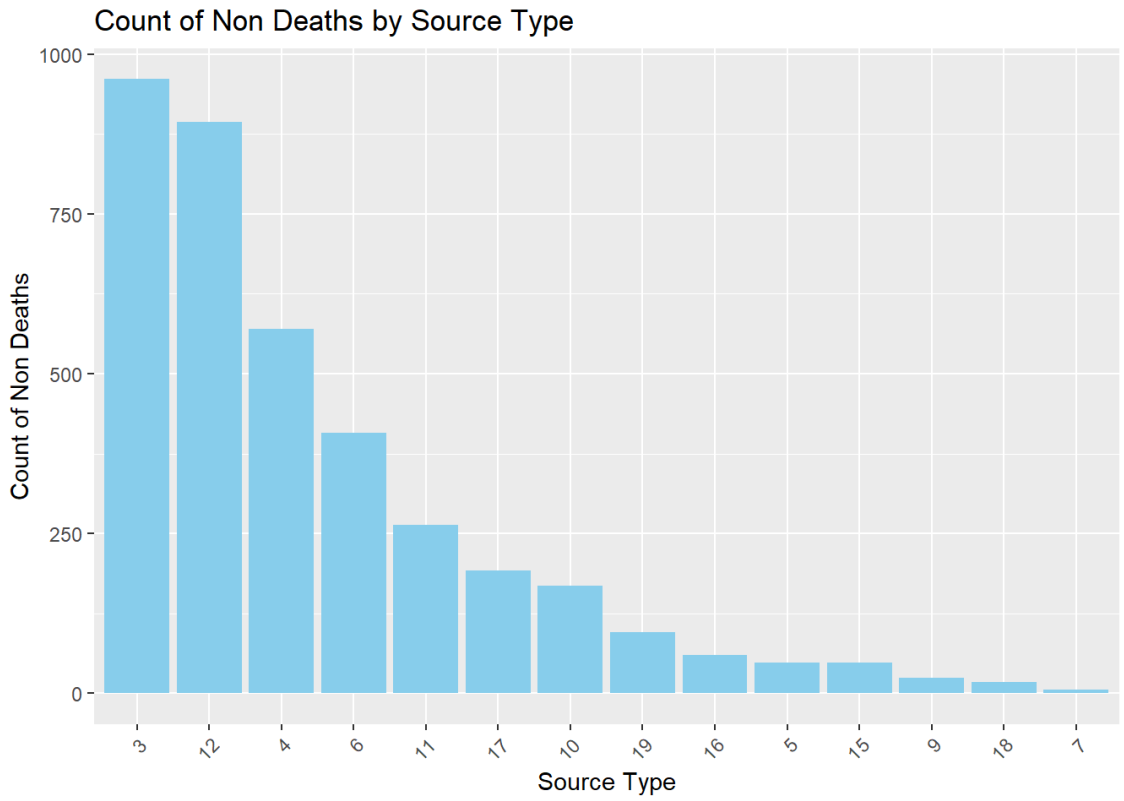
``` { r Product Fields echo=TRUE} df1 %>% filter(product.field_description == “Unknown”) %>% nrow() df1 %>% filter(product.field_description == “Unknown”) %>% group_by(ID_non_uniq) %>% summarise( count = n(), count_death = sum(Product.issue.consequence == “Death”, na.rm = TRUE), count_injury = sum(Product.issue.consequence == “Injury”, na.rm = TRUE), count_malfunction = sum(Product.issue.consequence == “Malfunction”, na.rm = TRUE) ) df1 %>% filter(product.field_description == “General, Plastic Surgery”) %>% group_by(ID_non_uniq) %>% summarise( count= n(), count_death = sum(Product.issue.consequence == “Death”, na.rm = TRUE), count_injury = sum(Product.issue.consequence == “Injury”, na.rm = TRUE), count_malfunction = sum(Product.issue.consequence == “Malfunction”, na.rm = TRUE) )

df1 %>% filter(product.field_description == “Immunology”) %>% group_by(ID_non_uniq) %>% summarise( count = n(), count_death = sum(Product.issue.consequence == “Death”, na.rm = TRUE), count_injury = sum(Product.issue.consequence == “Injury”, na.rm = TRUE), count_malfunction = sum(Product.issue.consequence == “Malfunction”, na.rm = TRUE) )

#exploring product.field_description summary(df1$product.field_description) #! All the Unknowns are in product.field_description

```
Counts of Death and Non deaths by source type

Source_type 3, 12, 4, 6, 11 should be investigated as highest deaths. Majority consequences for Unknown are deaths.

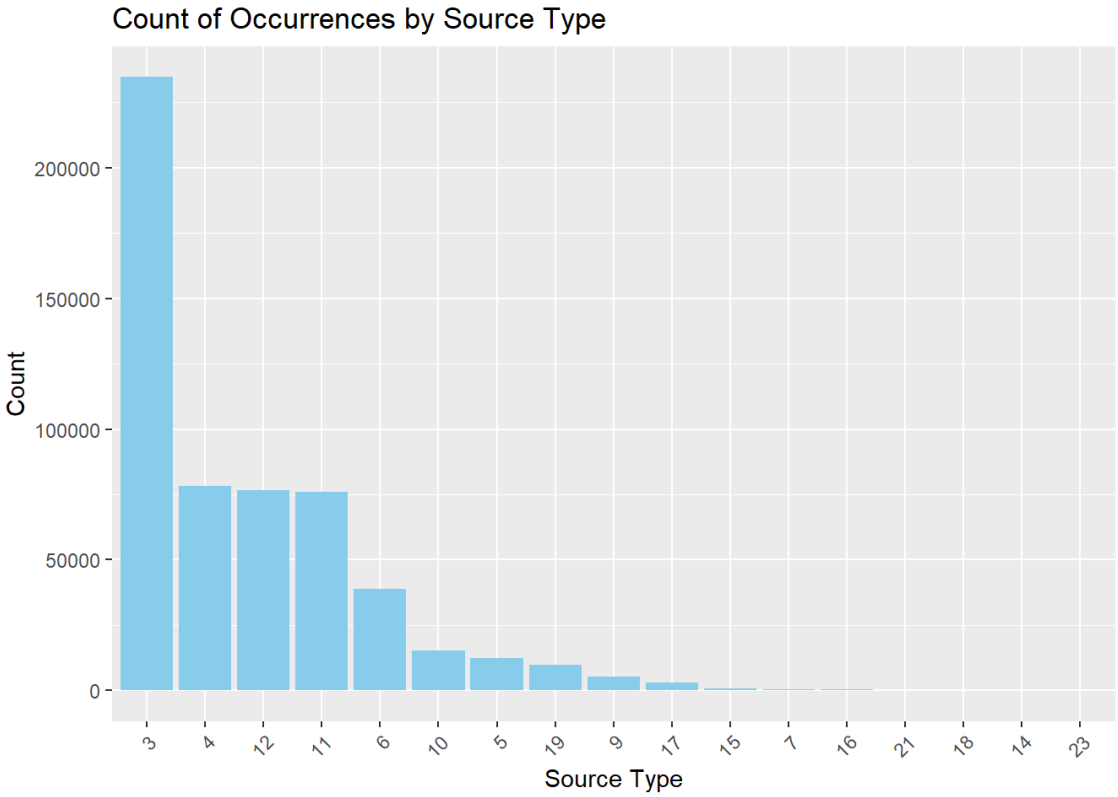
```r
#! We can assume this graph to be for product.field_description == Unknown as only it has deaths
death_data <- df1 %>%
  filter(Product.issue.consequence == "Death") %>%
  group_by(source_type) %>%
  summarise(count = n())
ggplot(death_data, aes(x = reorder(source_type, -count), y = count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(x = "Source Type", y = "Count of Deaths", title = "Count of Deaths by Source Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
non_death_data <- df1 %>%
  filter(Product.issue.consequence == "Death") %>%
  group_by(source_type) %>%
  summarise(count = n())
ggplot(non_death_data, aes(x = reorder(source_type, -count), y = count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(x = "Source Type", y = "Count of Non Deaths", title = "Count of Non Deaths by Source Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
count_by_source <- df1 %>%
  count(source_type) %>%
  arrange(desc(n))
ggplot(count_by_source, aes(x = reorder(source_type, -n), y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(x = "Source Type", y = "Count", title = "Count of Occurrences by Source Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Observing changes of variables over four years

Summaries for columns related to years when grouped by ID can tell how the pattern of change for each specific ID. This can be very useful to predict the behaviors by IDs for each next subsequent year. We see a sharp drop in legal announcements, product quantities and change in products.

```
#defining columns for years

last_year_cols <- c(
  "last_year_all_product_codes_num_uniq", "last_year_all_product_codes_most_freq",
  "last_year_brand_name_num_uniq", "last_year_brand_name_most_freq",
  "last_year_classification0_num_uniq", "last_year_classification1_num_uniq",
  "last_year_classification2_num_uniq", "last_year_company_name_num_uniq",
  "last_year_company_name_most_freq", "last_year_reason_for_legal_announcement_num_uniq",
  "last_year_reason_for_legal_announcement_most_freq", "last_year_legal_announcementing_firm_num_uniq",
  "last_year_legal_announcementing_firm_most_freq", "last_year_root_cause_description_num_uniq",
  "last_year_root_cause_description_most_freq", "last_year_product_quantity_average_num_uniq",
  "last_year_product_quantity_average_max", "last_year_product_quantity_average_average",
  "last_year_decision_date_max_changes_in_product", "last_year_decision_date_average_changes_in_product"
)
last_two_years_cols <- c(
  "last_two_years_all_product_codes_num_uniq", "last_two_years_all_product_codes_most_freq",
  "last_two_years_brand_name_num_uniq", "last_two_years_brand_name_most_freq",
  "last_two_years_classification0_num_uniq", "last_two_years_classification1_num_uniq",
  "last_two_years_classification2_num_uniq", "last_two_years_company_name_num_uniq",
  "last_two_years_company_name_most_freq", "last_two_years_reason_for_legal_announcement_num_uniq",
  "last_two_years_reason_for_legal_announcement_most_freq", "last_two_years_legal_announcementing_firm_num_uniq",
  "last_two_years_legal_announcementing_firm_most_freq", "last_two_years_root_cause_description_num_uniq",
  "last_two_years_root_cause_description_most_freq", "last_two_years_product_quantity_average_num_uniq",
  "last_two_years_product_quantity_average_max", "last_two_years_product_quantity_average_average",
  "last_two_years_decision_date_max_changes_in_product", "last_two_years_decision_date_average_changes_in_product"
)
last_four_years_cols <- c(
  "last_four_years_all_product_codes_num_uniq", "last_four_years_all_product_codes_most_freq",
  "last_four_years_brand_name_num_uniq", "last_four_years_brand_name_most_freq",
  "last_four_years_classification0_num_uniq", "last_four_years_classification1_num_uniq",
  "last_four_years_classification2_num_uniq", "last_four_years_company_name_num_uniq",
  "last_four_years_company_name_most_freq", "last_four_years_reason_for_legal_announcement_num_uniq",
  "last_four_years_reason_for_legal_announcement_most_freq", "last_four_years_legal_announcementing_firm_num_uniq",
  "last_four_years_legal_announcementing_firm_most_freq", "last_four_years_root_cause_description_num_uniq",
  "last_four_years_root_cause_description_most_freq", "last_four_years_product_quantity_average_num_uniq",
  "last_four_years_product_quantity_average_max", "last_four_years_product_quantity_average_average",
  "last_four_years_decision_date_max_changes_in_product", "last_four_years_decision_date_average_changes_in_produc
t"
)
other_cols <- c(
  "ID_non_uniq", "date_event", "manufacturer_contact_address_1",
  "product.brand_name", "product.generic_name", "product.issue.type",
  "type_of_report.1", "reporter_job_code", "source_type",
  "product.manufacturer_name", "product.product_operator",
  "product.manufacturer_city", "product.manufacturer_state",
  "product.manufacturer_country", "product.field_description",
  "product.product_report_product_code"
)
# groups of columns for last year, last two years, and last four years
last_year_vars <- c(
  "last_year_decision_date_average_changes_in_product",
  "last_year_product_quantity_average_average",
  "last_year_root_cause_description_most_freq",
  "last_year_legal_announcementing_firm_most_freq",
  "last_year_all_product_codes_most_freq",
  "last_year_brand_name_most_freq"
)

last_two_years_vars <- c(
  "last_two_years_decision_date_average_changes_in_product",
  "last_two_years_product_quantity_average_average",
  "last_two_years_root_cause_description_most_freq",
```

```
  "last_two_years_legal_announcementing_firm_most_freq",
  "last_two_years_all_product_codes_most_freq",
  "last_two_years_brand_name_most_freq"
)

last_four_years_vars <- c(
  "last_four_years_decision_date_average_changes_in_product",
  "last_four_years_product_quantity_average_average",
  "last_four_years_root_cause_description_most_freq",
  "last_four_years_legal_announcementing_firm_most_freq",
  "last_four_years_all_product_codes_most_freq",
  "last_four_years_brand_name_most_freq"
)

calculate_averages <- function(df, vars) {
  sapply(vars, function(var) mean(df[[var]], na.rm = TRUE))
}

df_specific_id <- df %>%
  filter(ID_non_uniq == "p860004")

average_last_year <- calculate_averages(df_specific_id, last_year_vars)
average_last_two_years <- calculate_averages(df_specific_id, last_two_years_vars)
average_last_four_years <- calculate_averages(df_specific_id, last_four_years_vars)

combined_averages <- rbind(average_last_year, average_last_two_years, average_last_four_years)

row_labels <- c("Decision Date Average Change in Product",
                "Product Quantity Average",
                "Root Cause Description",
                "Legal Announcement",
                "All Product Codes",
                "Brand Name")

colnames(combined_averages) <- last_year_vars

results_df <- as.data.frame(t(combined_averages))
names(results_df) <- c("Last Year", "Last Two Years", "Last Four Years")
results_df$Row <- row_labels

results_df <- results_df[, c("Row", "Last Year", "Last Two Years", "Last Four Years")]

print(results_df)
```

```
##                                                                                 Row
## last_year_decision_date_average_changes_in_product Decision Date Average Change in Product
## last_year_product_quantity_average_average                       Product Quantity Average
## last_year_root_cause_description_most_freq                         Root Cause Description
## last_year_legal_announcementing_firm_most_freq                         Legal Announcement
## last_year_all_product_codes_most_freq                                   All Product Codes
## last_year_brand_name_most_freq                                                 Brand Name
##                                                    Last Year Last Two Years
## last_year_decision_date_average_changes_in_product    20.44364       39.58426
## last_year_product_quantity_average_average         43944.61509    65893.24419
## last_year_root_cause_description_most_freq            22.68648       24.48282
## last_year_legal_announcementing_firm_most_freq       253.75540      283.18433
## last_year_all_product_codes_most_freq               1457.32306     1472.91667
## last_year_brand_name_most_freq                      3345.61798     3381.41667
##                                                   Last Four Years
## last_year_decision_date_average_changes_in_product        74.85855
## last_year_product_quantity_average_average            112690.41002
## last_year_root_cause_description_most_freq                15.03071
## last_year_legal_announcementing_firm_most_freq           292.00000
## last_year_all_product_codes_most_freq                   1472.91667
## last_year_brand_name_most_freq                          3381.41667
```

```
df1 %>%
  group_by(ID_non_uniq) %>%
  summarize(across(all_of(last_year_cols), mean, na.rm = TRUE))
```

```
## Warning: There was 1 warning in `summarize()`.
## ℹ In argument: `across(all_of(last_year_cols), mean, na.rm = TRUE)`.
## ℹ In group 1: `ID_non_uniq = p000021`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
## # A tibble: 12 × 21
##    ID_non_uniq last_year_all_product_codes_num_uniq last_year_all_product_code…¹
##    <fct>                                     <dbl>                        <dbl>
##  1 p000021                                   1                             3549
##  2 p000027                                   0.312                         1112.
##  3 p080012                                   0.984                         1472.
##  4 p120005                                   0                                0
##  5 p840001                                   4.08                          1356.
##  6 p860004                                   1.44                          1457.
##  7 p890055                                   1.05                          1444.
##  8 p930027                                   1                             3550.
##  9 p950021                                   0.64                          1493.
## 10 p960004                                   1.05                           292.
## 11 p990034                                   1                             1467
## 12 p990056                                   0.705                         2493.
## # i abbreviated name: ¹last_year_all_product_codes_most_freq
## # i 18 more variables: last_year_brand_name_num_uniq <dbl>,
## #   last_year_brand_name_most_freq <dbl>,
## #   last_year_classification0_num_uniq <dbl>,
## #   last_year_classification1_num_uniq <dbl>,
## #   last_year_classification2_num_uniq <dbl>,
## #   last_year_company_name_num_uniq <dbl>, …
```

```
# Summaries for last_two_years_cols
df1 %>%
  group_by(ID_non_uniq) %>%
  summarize(across(all_of(last_two_years_cols), mean, na.rm = TRUE))
```

```
## # A tibble: 12 × 21
##    ID_non_uniq last_two_years_all_product_codes_num_uniq last_two_years_all_pr…¹
##    <fct>                                           <dbl>                   <dbl>
##  1 p000021                                         1                        3549
##  2 p000027                                         0.844                    3004.
##  3 p080012                                         1                        1496
##  4 p120005                                         0.8                      2183.
##  5 p840001                                         4.08                     1356.
##  6 p860004                                         1.46                     1473.
##  7 p890055                                         1.08                     1479.
##  8 p930027                                         1                        3550.
##  9 p950021                                         1.10                     2559.
## 10 p960004                                         1.05                      292.
## 11 p990034                                         1                        1467
## 12 p990056                                         0.977                    3459.
## # i abbreviated name: ¹last_two_years_all_product_codes_most_freq
## # i 18 more variables: last_two_years_brand_name_num_uniq <dbl>,
## #   last_two_years_brand_name_most_freq <dbl>,
## #   last_two_years_classification0_num_uniq <dbl>,
## #   last_two_years_classification1_num_uniq <dbl>,
## #   last_two_years_classification2_num_uniq <dbl>,
## #   last_two_years_company_name_num_uniq <dbl>, …
```

```
# Summaries for last_four_years_cols
df1 %>%
  group_by(ID_non_uniq) %>%
  summarize(across(all_of(last_four_years_cols), mean, na.rm = TRUE))
```

```
## # A tibble: 12 × 21
##    ID_non_uniq last_four_years_all_product_codes_num_uniq last_four_years_all_…¹
##    <fct>                                           <dbl>                  <dbl>
##  1 p000021                                          1                     3549
##  2 p000027                                          1                     3560.
##  3 p080012                                          1                     1496
##  4 p120005                                          1.2                   3275.
##  5 p840001                                          4.08                  1356.
##  6 p860004                                          1.46                  1473.
##  7 p890055                                          1.08                  1479.
##  8 p930027                                          1                     3550.
##  9 p950021                                          1.14                  2666.
## 10 p960004                                          1.05                   292.
## 11 p990034                                          1                     1467
## 12 p990056                                          1                     3539
## # ℹ abbreviated name: ¹last_four_years_all_product_codes_most_freq
## # ℹ 18 more variables: last_four_years_brand_name_num_uniq <dbl>,
## #   last_four_years_brand_name_most_freq <dbl>,
## #   last_four_years_classification0_num_uniq <dbl>,
## #   last_four_years_classification1_num_uniq <dbl>,
## #   last_four_years_classification2_num_uniq <dbl>,
## #   last_four_years_company_name_num_uniq <dbl>, …
```

```
# ! gives us an understanding of how variables have altered over the years
```

Creating Subsets for each product field

We see that we don't have enough observations for each product field except for "Unknown". Since Product.issue.consequence is the target variable with special focus on product.field_description. We can see that for every field but "Unknown" there is only malfunction. Focusing on the "Unknown" field, we can see that the ID p860004 is the only account dominantly responsible for all outputs like death, injury and malfunction primarily because of having the highest count as well.

```
#creating subsets for each product field
df1 %>%
  filter(product.field_description == "Unknown") %>%
  group_by(ID_non_uniq) %>%
  summarise(
    count = n(),
    count_death = sum(Product.issue.consequence == "Death", na.rm = TRUE),
    count_injury = sum(Product.issue.consequence == "Injury", na.rm = TRUE),
    count_malfunction = sum(Product.issue.consequence == "Malfunction", na.rm = TRUE)
  )
```

```
## # A tibble: 11 × 5
##    ID_non_uniq  count count_death count_injury count_malfunction
##    <fct>        <int>       <int>        <int>              <int>
##  1 p000021          1           0            0                  1
##  2 p000027        160           0            0                160
##  3 p080012       4470         240         2700               1530
##  4 p840001        208           0          130                 78
##  5 p860004     546336        3504       364128             178704
##  6 p890055        559          13          429                117
##  7 p930027        130           0            0                130
##  8 p950021        154           0            0                154
##  9 p960004         19           0           19                  0
## 10 p990034         38           0           28                 10
## 11 p990056        176           0            0                176
```

```
df1 %>%
  filter(product.field_description == "General, Plastic Surgery") %>%
  group_by(ID_non_uniq) %>%
  summarise(
    count= n(),
    count_death = sum(Product.issue.consequence == "Death", na.rm = TRUE),
    count_injury = sum(Product.issue.consequence == "Injury", na.rm = TRUE),
    count_malfunction = sum(Product.issue.consequence == "Malfunction", na.rm = TRUE)
  )
```

```
## # A tibble: 1 × 5
##   ID_non_uniq count count_death count_injury count_malfunction
##   <fct>       <int>       <int>        <int>             <int>
## 1 p120005       100           0            0               100
```

```
df1 %>%
  filter(product.field_description == "Immunology") %>%
  group_by(ID_non_uniq) %>%
  summarise(
    count = n(),
    count_death = sum(Product.issue.consequence == "Death", na.rm = TRUE),
    count_injury = sum(Product.issue.consequence == "Injury", na.rm = TRUE),
    count_malfunction = sum(Product.issue.consequence == "Malfunction", na.rm = TRUE)
  )
```

```
## # A tibble: 2 × 5
##   ID_non_uniq count count_death count_injury count_malfunction
##   <fct>       <int>       <int>        <int>             <int>
## 1 p120005        50           0            0                50
## 2 p950021        21           0            0                21
```

Dummy Coding and Normalization

For all the categorical variables except for date and product issue consequences, dummy coding was done so that they could be input in forward selection, regression and lda. Product issue consequence of death was encoded to a new variable called death_or_not as the final output variable.

```r
#normalization
normalize <- function(x) {
  return((x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)))
}

columns_to_normalize <- c("last_year_decision_date_average_changes_in_product",
                                            "last_year_product_quantity_average_average",
                                            "last_year_root_cause_description_most_freq",
                                            "last_year_legal_announcementing_firm_most_freq",
                                            "last_year_all_product_codes_most_freq",
                                            "last_year_brand_name_most_freq",
                                            "last_year_classification0_num_uniq",
                                            "last_year_company_name_most_freq",
                                            "last_year_product_quantity_average_max",
                                            "last_year_decision_date_max_changes_in_product",
                                            "last_two_years_decision_date_average_changes_in_product",
                                            "last_two_years_product_quantity_average_average",
                                            "last_two_years_root_cause_description_most_freq",
                                            "last_two_years_legal_announcementing_firm_most_freq",
                                            "last_two_years_all_product_codes_most_freq",
                                            "last_two_years_brand_name_most_freq",
                                            "last_two_years_classification0_num_uniq",
                                            "last_two_years_company_name_most_freq",
                                            "last_two_years_product_quantity_average_max",
                                            "last_two_years_decision_date_max_changes_in_product",
                                            "last_four_years_decision_date_average_changes_in_product",
                                            "last_four_years_product_quantity_average_average",
                                            "last_four_years_root_cause_description_most_freq",
                                            "last_four_years_legal_announcementing_firm_most_freq",
                                            "last_four_years_all_product_codes_most_freq",
                                            "last_four_years_brand_name_most_freq",
                                            "last_four_years_classification0_num_uniq",
                                            "last_four_years_company_name_most_freq")

df1_normalized <- df1
df1_normalized[columns_to_normalize] <- lapply(df1[columns_to_normalize], normalize)

#dummy / one hot coding
df2 <- fastDummies::dummy_cols(df1_normalized, select_columns = c("ID_non_uniq",
                                                    "manufacturer_contact_address_1",
                                                    "product.brand_name",
                                                    "product.generic_name",
                                                    "product.issue.type",
                                                    "type_of_report.1",
                                                    "reporter_job_code",
                                                    "source_type",
                                                    "product.manufacturer_name",
                                                    "product.product_operator",
                                                    "product.manufacturer_city",
                                                    "product.manufacturer_state",
                                                    "product.manufacturer_country",
                                                    "product.field_description",
                                                    "product.product_report_product_code"),remove_selected_colum
ns = TRUE)


df2 <- df2 %>%
  mutate(death_or_not = ifelse(Product.issue.consequence == "Death", 1, 0))

df2 <- subset(df2, select = -Product.issue.consequence)

df2 <- subset(df2, select = -date_event)
```

```
#subseting prepared dataset for further use
df2_general <- df2 %>%
  filter(`product.field_description_General, Plastic Surgery` == "1")

df2_immunology <- df2 %>%
  filter(product.field_description_Immunology == "1")

df2_unknown <- df2 %>%
  filter(product.field_description_Unknown == "1")
df2_unknown_subset <- sample_n(df2_unknown, size = 5000)
```

Alternative subsetting for further Analysis

We will select product.product_report_product_code_LKK, product.issue.type_599, manufacturer_contact_address_1_9476 as they have a strong majority in the variables that they belong in. Analysis in the case of this dataset cannot be done on the different product fields because their negligable counts are not enough for any sort of analysis.

```
summary(df1)
```

```
##   ID_non_uniq      date_event        last_year_all_product_codes_num_uniq
##  p860004:546336   Length:552422      Min.   :0.000
##  p080012:  4470   Class :character   1st Qu.:1.000
##  p890055:   559   Mode  :character   Median :1.000
##  p840001:   208                      Mean   :1.438
##  p990056:   176                      3rd Qu.:2.000
##  p950021:   175                      Max.   :8.000
##  (Other):   498
##  last_year_all_product_codes_most_freq last_year_brand_name_num_uniq
##  Min.   :   0                          Min.   :0.00
##  1st Qu.:1465                          1st Qu.:1.00
##  Median :1472                          Median :2.00
##  Mean   :1458                          Mean   :1.85
##  3rd Qu.:1480                          3rd Qu.:2.00
##  Max.   :3566                          Max.   :8.00
##
##  last_year_brand_name_most_freq last_year_classification0_num_uniq
##  Min.   :   0                   Min.   :  0.000
##  1st Qu.:2216                   1st Qu.:  0.000
##  Median :4056                   Median :  0.000
##  Mean   :3342                   Mean   :  8.766
##  3rd Qu.:4789                   3rd Qu.:  6.000
##  Max.   :4789                   Max.   :300.000
##
##  last_year_classification1_num_uniq last_year_classification2_num_uniq
##  Min.   :  0.00                     Min.   : 0.000
##  1st Qu.:  4.00                     1st Qu.: 0.000
##  Median : 12.00                     Median : 0.000
##  Mean   : 19.97                     Mean   : 1.889
##  3rd Qu.: 25.00                     3rd Qu.: 0.000
##  Max.   :300.00                     Max.   :85.000
##
##  last_year_company_name_num_uniq last_year_company_name_most_freq
##  Min.   :0.0000                  Min.   :  0.0
##  1st Qu.:1.0000                  1st Qu.:349.0
##  Median :1.0000                  Median :349.0
##  Mean   :0.9887                  Mean   :344.3
##  3rd Qu.:1.0000                  3rd Qu.:349.0
##  Max.   :2.0000                  Max.   :489.0
##
##  last_year_reason_for_legal_announcement_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :2.996
##  3rd Qu.:4.000
##  Max.   :8.000
##
##  last_year_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
##  1st Qu.:  52.0
##  Median : 698.0
##  Mean   : 708.1
##  3rd Qu.:1143.0
##  Max.   :1401.0
##
##  last_year_legal_announcementing_firm_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :1.689
##  3rd Qu.:2.000
```

```
##  Max.   :3.000
##
##  last_year_legal_announcementing_firm_most_freq
##  Min.   :  0.0
##  1st Qu.:143.0
##  Median :292.0
##  Mean   :253.6
##  3rd Qu.:292.0
##  Max.   :440.0
##
##  last_year_root_cause_description_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.216
##  3rd Qu.:3.000
##  Max.   :5.000
##
##  last_year_root_cause_description_most_freq
##  Min.   : 0.0
##  1st Qu.: 4.0
##  Median :28.0
##  Mean   :22.7
##  3rd Qu.:28.0
##  Max.   :40.0
##
##  last_year_product_quantity_average_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.497
##  3rd Qu.:3.000
##  Max.   :6.000
##
##  last_year_product_quantity_average_max
##  Min.   :     0
##  1st Qu.:  5463
##  Median : 20286
##  Mean   :103998
##  3rd Qu.: 22298
##  Max.   :636572
##
##  last_year_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  4791
##  Median :  6892
##  Mean   : 43638
##  3rd Qu.: 11154
##  Max.   :333373
##
##  last_year_decision_date_max_changes_in_product
##  Min.   : 0.00
##  1st Qu.:16.00
##  Median :20.00
##  Mean   :20.28
##  3rd Qu.:24.00
##  Max.   :33.00
##
##  last_year_decision_date_average_changes_in_product
##  Min.   : 0.00
##  1st Qu.:16.00
##  Median :20.00
```

```
##  Mean   :20.28
##  3rd Qu.:24.00
##  Max.   :33.00
##
##  last_two_years_all_product_codes_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.454
##  3rd Qu.:2.000
##  Max.   :8.000
##
##  last_two_years_all_product_codes_most_freq last_two_years_brand_name_num_uniq
##  Min.   :   0                               Min.   :0.000
##  1st Qu.:1466                               1st Qu.:1.000
##  Median :1473                               Median :2.000
##  Mean   :1475                               Mean   :1.871
##  3rd Qu.:1480                               3rd Qu.:2.000
##  Max.   :6593                               Max.   :8.000
##
##  last_two_years_brand_name_most_freq last_two_years_classification0_num_uniq
##  Min.   :   0                        Min.   :   0.00
##  1st Qu.:2817                        1st Qu.:   0.00
##  Median :4056                        Median :   6.00
##  Mean   :3378                        Mean   :  17.29
##  3rd Qu.:4789                        3rd Qu.:  20.00
##  Max.   :4789                        Max.   :1449.00
##
##  last_two_years_classification1_num_uniq
##  Min.   :  0.00
##  1st Qu.:  8.00
##  Median : 24.00
##  Mean   : 40.49
##  3rd Qu.: 54.00
##  Max.   :665.00
##
##  last_two_years_classification2_num_uniq last_two_years_company_name_num_uniq
##  Min.   : 0.000                          Min.   :0.0000
##  1st Qu.: 0.000                          1st Qu.:1.0000
##  Median : 0.000                          Median :1.0000
##  Mean   : 2.339                          Mean   :0.9999
##  3rd Qu.: 0.000                          3rd Qu.:1.0000
##  Max.   :85.000                          Max.   :2.0000
##
##  last_two_years_company_name_most_freq
##  Min.   :  0.0
##  1st Qu.:349.0
##  Median :349.0
##  Mean   :348.2
##  3rd Qu.:349.0
##  Max.   :489.0
##
##  last_two_years_reason_for_legal_announcement_num_uniq
##  Min.   : 0.000
##  1st Qu.: 5.000
##  Median : 5.000
##  Mean   : 5.803
##  3rd Qu.: 7.000
##  Max.   :11.000
##
##  last_two_years_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
```

```
##  1st Qu.:  52.0
##  Median :1028.0
##  Mean   : 827.7
##  3rd Qu.:1143.0
##  Max.   :1401.0
##
##  last_two_years_legal_announcementing_firm_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.081
##  3rd Qu.:2.000
##  Max.   :4.000
##
##  last_two_years_legal_announcementing_firm_most_freq
##  Min.   :  0.0
##  1st Qu.:292.0
##  Median :292.0
##  Mean   :283.1
##  3rd Qu.:292.0
##  Max.   :440.0
##
##  last_two_years_root_cause_description_num_uniq
##  Min.   :0.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :3.712
##  3rd Qu.:5.000
##  Max.   :6.000
##
##  last_two_years_root_cause_description_most_freq
##  Min.   : 0.00
##  1st Qu.: 4.00
##  Median :28.00
##  Mean   :24.53
##  3rd Qu.:40.00
##  Max.   :40.00
##
##  last_two_years_product_quantity_average_num_uniq
##  Min.   :0.000
##  1st Qu.:4.000
##  Median :5.000
##  Mean   :4.686
##  3rd Qu.:6.000
##  Max.   :8.000
##
##  last_two_years_product_quantity_average_max
##  Min.   :     0
##  1st Qu.: 13784
##  Median : 22298
##  Mean   :196851
##  3rd Qu.:636572
##  Max.   :636572
##
##  last_two_years_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  4959
##  Median :  7499
##  Mean   : 65481
##  3rd Qu.:167634
##  Max.   :333373
##
```

```
##  last_two_years_decision_date_max_changes_in_product
##  Min.   : 0.00
##  1st Qu.:34.00
##  Median :40.00
##  Mean   :39.28
##  3rd Qu.:44.00
##  Max.   :55.00
##
##  last_two_years_decision_date_average_changes_in_product
##  Min.   : 0.00
##  1st Qu.:34.00
##  Median :40.00
##  Mean   :39.28
##  3rd Qu.:44.00
##  Max.   :55.00
##
##  last_four_years_all_product_codes_num_uniq
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.455
##  3rd Qu.:2.000
##  Max.   :8.000
##
##  last_four_years_all_product_codes_most_freq
##  Min.   : 279
##  1st Qu.:1466
##  Median :1473
##  Mean   :1476
##  3rd Qu.:1480
##  Max.   :6593
##
##  last_four_years_brand_name_num_uniq last_four_years_brand_name_most_freq
##  Min.   :1.000                       Min.   :  97
##  1st Qu.:1.000                       1st Qu.:2817
##  Median :2.000                       Median :4056
##  Mean   :1.871                       Mean   :3379
##  3rd Qu.:2.000                       3rd Qu.:4789
##  Max.   :8.000                       Max.   :4789
##
##  last_four_years_classification0_num_uniq
##  Min.   :   0.00
##  1st Qu.:  12.00
##  Median :  24.00
##  Mean   :  43.64
##  3rd Qu.:  60.00
##  Max.   :1674.00
##
##  last_four_years_classification1_num_uniq
##  Min.   :   0.00
##  1st Qu.:  14.00
##  Median :  40.00
##  Mean   :  67.52
##  3rd Qu.:  96.00
##  Max.   :3348.00
##
##  last_four_years_classification2_num_uniq last_four_years_company_name_num_uniq
##  Min.   : 0.000                           Min.   :1
##  1st Qu.: 0.000                           1st Qu.:1
##  Median : 0.000                           Median :1
##  Mean   : 2.498                           Mean   :1
##  3rd Qu.: 0.000                           3rd Qu.:1
```

```
##  Max.   :85.000                          Max.   :2
##
##  last_four_years_company_name_most_freq
##  Min.   :111.0
##  1st Qu.:349.0
##  Median :349.0
##  Mean   :348.2
##  3rd Qu.:349.0
##  Max.   :489.0
##
##  last_four_years_reason_for_legal_announcement_num_uniq
##  Min.   : 0.00
##  1st Qu.: 9.00
##  Median :11.00
##  Mean   :10.25
##  3rd Qu.:13.00
##  Max.   :16.00
##
##  last_four_years_reason_for_legal_announcement_most_freq
##  Min.   :   0
##  1st Qu.:1028
##  Median :1028
##  Mean   :1038
##  3rd Qu.:1028
##  Max.   :1361
##
##  last_four_years_legal_announcementing_firm_num_uniq
##  Min.   :1.0
##  1st Qu.:2.0
##  Median :2.0
##  Mean   :2.3
##  3rd Qu.:3.0
##  Max.   :5.0
##
##  last_four_years_legal_announcementing_firm_most_freq
##  Min.   : 40
##  1st Qu.:292
##  Median :292
##  Mean   :292
##  3rd Qu.:292
##  Max.   :371
##
##  last_four_years_root_cause_description_num_uniq
##  Min.   :1.0
##  1st Qu.:5.0
##  Median :6.0
##  Mean   :5.2
##  3rd Qu.:6.0
##  Max.   :6.0
##
##  last_four_years_root_cause_description_most_freq
##  Min.   : 4.00
##  1st Qu.: 4.00
##  Median : 4.00
##  Mean   :15.08
##  3rd Qu.:40.00
##  Max.   :40.00
##
##  last_four_years_product_quantity_average_num_uniq
##  Min.   : 0.000
##  1st Qu.: 7.000
##  Median : 8.000
```

```
##  Mean   : 7.917
##  3rd Qu.:10.000
##  Max.   :11.000
##
##  last_four_years_product_quantity_average_max
##  Min.   :     0
##  1st Qu.: 22298
##  Median :636572
##  Mean   :437967
##  3rd Qu.:636572
##  Max.   :636572
##
##  last_four_years_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  7499
##  Median :123061
##  Mean   :112255
##  3rd Qu.:168517
##  Max.   :333373
##
##  last_four_years_decision_date_max_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 68.00
##  Median : 71.00
##  Mean   : 74.29
##  3rd Qu.: 79.00
##  Max.   :107.00
##
##  last_four_years_decision_date_average_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 68.00
##  Median : 71.00
##  Mean   : 74.29
##  3rd Qu.: 79.00
##  Max.   :107.00
##
##  Product.issue.consequence manufacturer_contact_address_1 product.brand_name
##  Death      :  3757        9476   :532655                 281286 :516825
##  Injury     :367434        13145  : 14668                 344588 :  6650
##  Malfunction:181231        7455   :  3834                 43203  :  5616
##                            10387  :   258                 154949 :  3600
##                            5835   :   182                 153901 :  2424
##                            9040   :   150                 238102 :  1806
##                            (Other):   675                 (Other): 15501
##  product.generic_name product.issue.type type_of_report.1 reporter_job_code
##  73852  :546966       599    :177799     0:396032         42     :241122
##  44428  :  3300       629    : 67704     1:156390         32     :193443
##  45330  :   469       849    : 38932                      17     : 61111
##  44436  :   390       566    : 27314                      28     : 18618
##  44437  :   198       624    : 22415                      52     : 16176
##  89386  :   155       906    : 16442                      1      : 12276
##  (Other):   944       (Other):201816                      (Other):  9676
##   source_type     product.manufacturer_name product.product_operator
##  3      :234691   18383  :292546            15     :548701
##  4      : 78159   19408  :202917            41     :  2604
##  12     : 76584   19327  : 48854            0      :   516
##  11     : 75853   12347  :  2166            21     :   432
##  6      : 38814   12348  :  1686            18     :   116
##  10     : 15352   24392  :  1610            32     :    18
##  (Other): 32969   (Other):  2643            (Other):    35
##  product.manufacturer_city product.manufacturer_state
##  4513   :487373            48     :487591
```

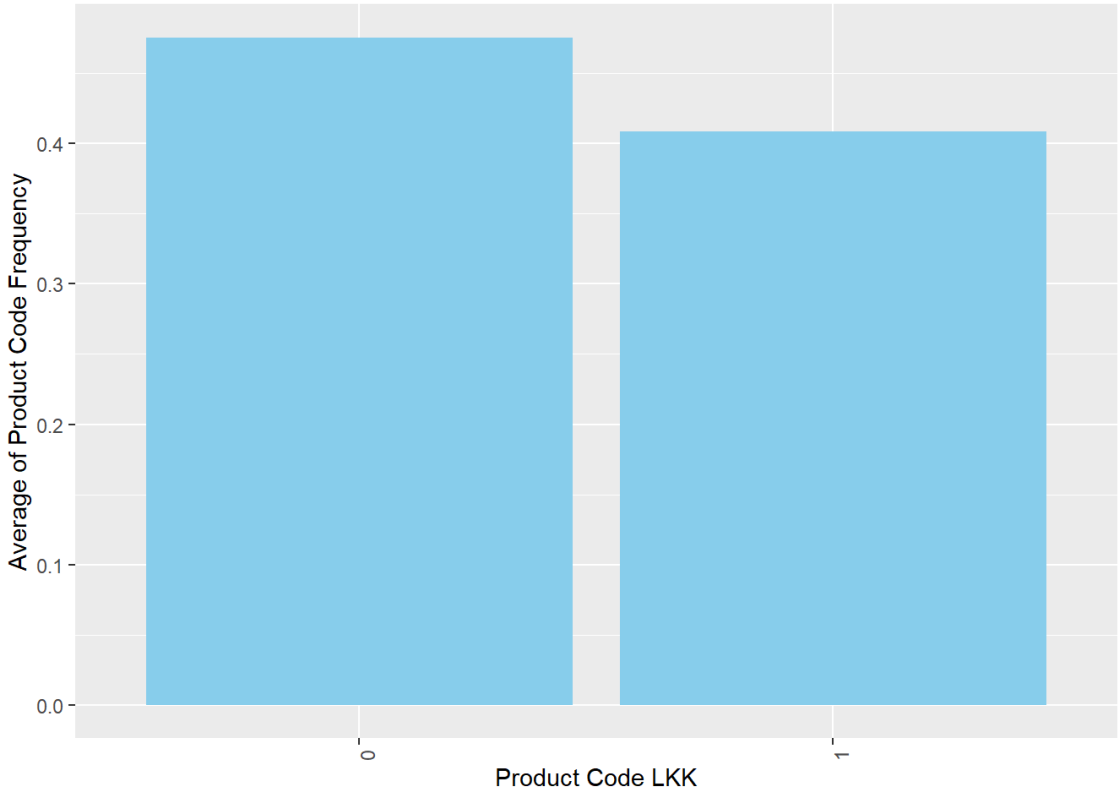
```
##  5990   : 48684            32     : 58698
##  4517   :  8210            40     :  4338
##  6271   :  4260            63     :   921
##  3153   :  1610            8      :   341
##  4284   :   336            23     :   331
##  (Other):  1949            (Other):   202
##  product.manufacturer_country           product.field_description
##  109:   351                   General, Plastic Surgery:   100
##  123:   125                   Immunology              :    71
##  126:551663                   Unknown                 :552251
##  135:   283
##
##
##
##  product.product_report_product_code
##  DHX:    14
##  LKK:551630
##  LTJ:     7
##  MTF:   454
##  MTG:   167
##  PDF:    50
##  PQM:   100
```

```
# ! product.product_report_product_code_LKK
# ! product.issue.type_599
# ! manufacturer_contact_address_1_9476

df2_report <- df2 %>%
  filter( product.product_report_product_code_LKK == "1")
df2_report_reduced <- sample_n(df2_report, size = 60000)

df2_issue <- df2 %>%
  filter( product.issue.type_599 == "1")
df2_issue_reduced <- sample_n(df2_issue, size = 60000)

df2_address <- df2 %>%
  filter( manufacturer_contact_address_1_9476 == "1")
df2_address_reduced <- sample_n(df2_address, size = 60000)
```

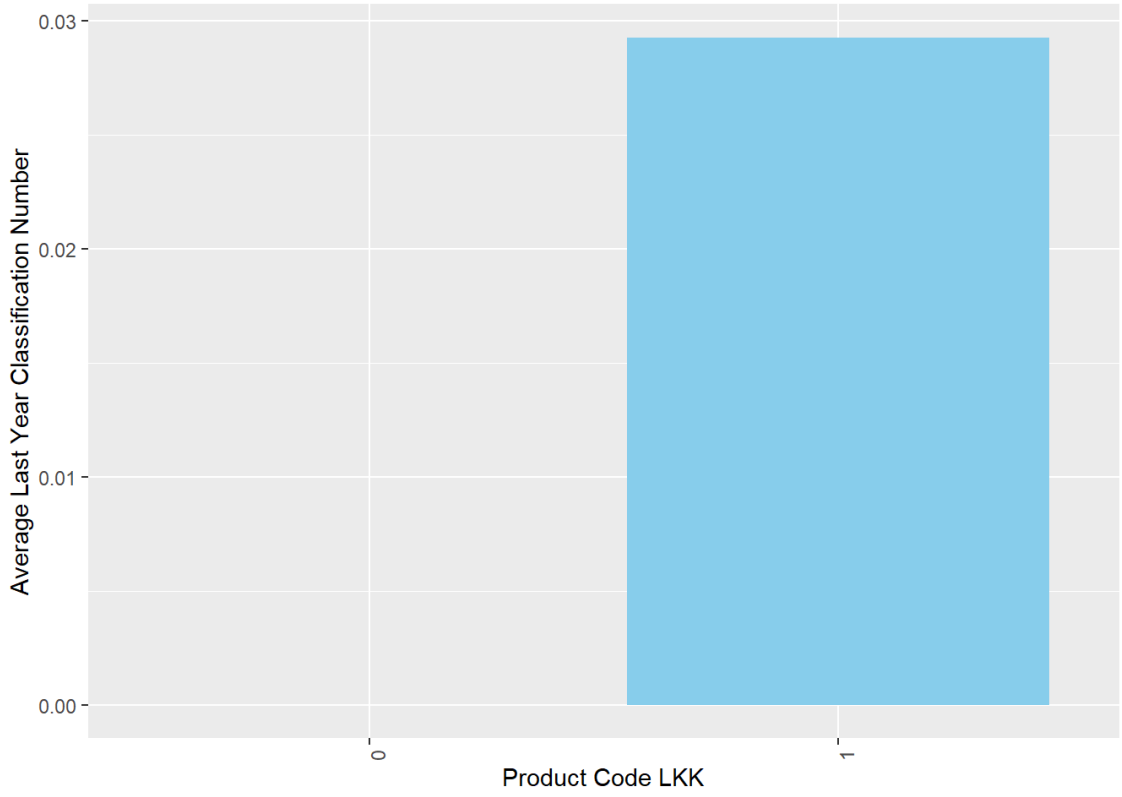
Relations of Product Code LKK 1 or 0 based on averages

For each of the visualizations made the average values of certain variables are different for both levels of Product Code LKK, 1 and 0. This means that Product Code LKK has a correlation to say the least with all of the following variables: Product Code Frequency for last year, Last year classification number.

```
#last_year_all_product_codes_most_freq
avg_data_codes <- df2 %>%
  group_by(product.product_report_product_code_LKK) %>%
  summarise(Average_product_code = mean(last_year_all_product_codes_most_freq, na.rm = TRUE))

ggplot(avg_data_codes, aes(x = as.factor(product.product_report_product_code_LKK), y = Average_product_code)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "Product Code LKK", y = "Average of Product Code Frequency")
```




```
#last_year_classification0_num_uniq
avg <- df2 %>%
  group_by(product.product_report_product_code_LKK) %>%
  summarise(Average = mean(last_year_classification0_num_uniq, na.rm = TRUE))

ggplot(avg, aes(x = as.factor(product.product_report_product_code_LKK), y = Average)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "Product Code LKK", y = "Average Last Year Classification Number")
```



Possible Relations of Other Variables

Last year product quantity average max is very much biased to low frequencies. In totality, majority of the consequences for product issues are Injuries, Death are in the minority. The boxplot displays the distribution of “Last Year Decision Date Average Changes,” with a median close to zero and a fairly symmetrical spread of data. There are several outliers indicating some values significantly lower than the bulk of the data. Product Quantities average max have been consistent throughout the years.

```
df1_short <- df1 %>% sample_n(60000)

summary(df1_short)
```

```
##   ID_non_uniq     date_event        last_year_all_product_codes_num_uniq
##  p860004:59338   Length:60000       Min.   :0.000
##  p080012:  493   Class :character   1st Qu.:1.000
##  p890055:   50   Mode  :character   Median :1.000
##  p930027:   21                      Mean   :1.438
##  p950021:   21                      3rd Qu.:2.000
##  p840001:   20                      Max.   :8.000
##  (Other):   57
##  last_year_all_product_codes_most_freq last_year_brand_name_num_uniq
##  Min.   :   0                          Min.   :0.000
##  1st Qu.:1465                          1st Qu.:1.000
##  Median :1472                          Median :2.000
##  Mean   :1458                          Mean   :1.852
##  3rd Qu.:1480                          3rd Qu.:2.000
##  Max.   :3566                          Max.   :8.000
##
##  last_year_brand_name_most_freq last_year_classification0_num_uniq
##  Min.   :   0                   Min.   :  0.000
##  1st Qu.:2216                   1st Qu.:  0.000
##  Median :4056                   Median :  0.000
##  Mean   :3336                   Mean   :  8.752
##  3rd Qu.:4789                   3rd Qu.:  6.000
##  Max.   :4789                   Max.   :300.000
##
##  last_year_classification1_num_uniq last_year_classification2_num_uniq
##  Min.   :  0.00                     Min.   : 0.0
##  1st Qu.:  4.00                     1st Qu.: 0.0
##  Median : 12.00                     Median : 0.0
##  Mean   : 19.92                     Mean   : 1.9
##  3rd Qu.: 24.00                     3rd Qu.: 0.0
##  Max.   :300.00                     Max.   :85.0
##
##  last_year_company_name_num_uniq last_year_company_name_most_freq
##  Min.   :0.0000                  Min.   :  0.0
##  1st Qu.:1.0000                  1st Qu.:349.0
##  Median :1.0000                  Median :349.0
##  Mean   :0.9886                  Mean   :344.2
##  3rd Qu.:1.0000                  3rd Qu.:349.0
##  Max.   :2.0000                  Max.   :489.0
##
##  last_year_reason_for_legal_announcement_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :2.988
##  3rd Qu.:4.000
##  Max.   :8.000
##
##  last_year_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
##  1st Qu.:  52.0
##  Median : 698.0
##  Mean   : 709.3
##  3rd Qu.:1143.0
##  Max.   :1401.0
##
##  last_year_legal_announcementing_firm_num_uniq
##  Min.   :0.00
##  1st Qu.:1.00
##  Median :2.00
##  Mean   :1.69
##  3rd Qu.:2.00
```

```
##  Max.   :3.00
##
##  last_year_legal_announcementing_firm_most_freq
##  Min.   :  0.0
##  1st Qu.:143.0
##  Median :292.0
##  Mean   :253.6
##  3rd Qu.:292.0
##  Max.   :440.0
##
##  last_year_root_cause_description_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.212
##  3rd Qu.:3.000
##  Max.   :5.000
##
##  last_year_root_cause_description_most_freq
##  Min.   : 0.00
##  1st Qu.: 4.00
##  Median :28.00
##  Mean   :22.66
##  3rd Qu.:28.00
##  Max.   :40.00
##
##  last_year_product_quantity_average_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.496
##  3rd Qu.:3.000
##  Max.   :6.000
##
##  last_year_product_quantity_average_max
##  Min.   :     0
##  1st Qu.:  5463
##  Median : 20286
##  Mean   :104014
##  3rd Qu.: 22298
##  Max.   :636572
##
##  last_year_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  4791
##  Median :  6892
##  Mean   : 43838
##  3rd Qu.: 11154
##  Max.   :333373
##
##  last_year_decision_date_max_changes_in_product
##  Min.   : 0.0
##  1st Qu.:16.0
##  Median :20.0
##  Mean   :20.3
##  3rd Qu.:24.0
##  Max.   :33.0
##
##  last_year_decision_date_average_changes_in_product
##  Min.   : 0.0
##  1st Qu.:16.0
##  Median :20.0
```

```
##  Mean   :20.3
##  3rd Qu.:24.0
##  Max.   :33.0
##
##  last_two_years_all_product_codes_num_uniq
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.454
##  3rd Qu.:2.000
##  Max.   :8.000
##
##  last_two_years_all_product_codes_most_freq last_two_years_brand_name_num_uniq
##  Min.   :   0                               Min.   :0.000
##  1st Qu.:1466                               1st Qu.:1.000
##  Median :1473                               Median :2.000
##  Mean   :1475                               Mean   :1.873
##  3rd Qu.:1480                               3rd Qu.:2.000
##  Max.   :3566                               Max.   :8.000
##
##  last_two_years_brand_name_most_freq last_two_years_classification0_num_uniq
##  Min.   :   0                        Min.   :   0.00
##  1st Qu.:2559                        1st Qu.:   0.00
##  Median :4056                        Median :   6.00
##  Mean   :3374                        Mean   :  17.32
##  3rd Qu.:4789                        3rd Qu.:  20.00
##  Max.   :4789                        Max.   :1449.00
##
##  last_two_years_classification1_num_uniq
##  Min.   :  0.00
##  1st Qu.:  8.00
##  Median : 24.00
##  Mean   : 40.31
##  3rd Qu.: 50.00
##  Max.   :665.00
##
##  last_two_years_classification2_num_uniq last_two_years_company_name_num_uniq
##  Min.   : 0.000                          Min.   :0.0000
##  1st Qu.: 0.000                          1st Qu.:1.0000
##  Median : 0.000                          Median :1.0000
##  Mean   : 2.345                          Mean   :0.9999
##  3rd Qu.: 0.000                          3rd Qu.:1.0000
##  Max.   :85.000                          Max.   :2.0000
##
##  last_two_years_company_name_most_freq
##  Min.   :  0.0
##  1st Qu.:349.0
##  Median :349.0
##  Mean   :348.2
##  3rd Qu.:349.0
##  Max.   :489.0
##
##  last_two_years_reason_for_legal_announcement_num_uniq
##  Min.   : 0.000
##  1st Qu.: 5.000
##  Median : 5.000
##  Mean   : 5.784
##  3rd Qu.: 7.000
##  Max.   :11.000
##
##  last_two_years_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
```

```
##  1st Qu.:  52.0
##  Median :1028.0
##  Mean   : 826.9
##  3rd Qu.:1143.0
##  Max.   :1401.0
##
##  last_two_years_legal_announcementing_firm_num_uniq
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.078
##  3rd Qu.:2.000
##  Max.   :4.000
##
##  last_two_years_legal_announcementing_firm_most_freq
##  Min.   :  0
##  1st Qu.:292
##  Median :292
##  Mean   :283
##  3rd Qu.:292
##  Max.   :440
##
##  last_two_years_root_cause_description_num_uniq
##  Min.   :0.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :3.699
##  3rd Qu.:5.000
##  Max.   :6.000
##
##  last_two_years_root_cause_description_most_freq
##  Min.   : 0.00
##  1st Qu.: 4.00
##  Median :28.00
##  Mean   :24.59
##  3rd Qu.:40.00
##  Max.   :40.00
##
##  last_two_years_product_quantity_average_num_uniq
##  Min.   :0.00
##  1st Qu.:4.00
##  Median :5.00
##  Mean   :4.68
##  3rd Qu.:6.00
##  Max.   :8.00
##
##  last_two_years_product_quantity_average_max
##  Min.   :     0
##  1st Qu.: 13784
##  Median : 22298
##  Mean   :196220
##  3rd Qu.:636572
##  Max.   :636572
##
##  last_two_years_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:  4959
##  Median :  7499
##  Mean   : 65527
##  3rd Qu.:167634
##  Max.   :333373
##
```

```
##  last_two_years_decision_date_max_changes_in_product
##  Min.   : 0.00
##  1st Qu.:34.00
##  Median :40.00
##  Mean   :39.29
##  3rd Qu.:44.00
##  Max.   :55.00
##
##  last_two_years_decision_date_average_changes_in_product
##  Min.   : 0.00
##  1st Qu.:34.00
##  Median :40.00
##  Mean   :39.29
##  3rd Qu.:44.00
##  Max.   :55.00
##
##  last_four_years_all_product_codes_num_uniq
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.454
##  3rd Qu.:2.000
##  Max.   :8.000
##
##  last_four_years_all_product_codes_most_freq
##  Min.   : 302
##  1st Qu.:1466
##  Median :1473
##  Mean   :1476
##  3rd Qu.:1480
##  Max.   :6588
##
##  last_four_years_brand_name_num_uniq last_four_years_brand_name_most_freq
##  Min.   :1.000                       Min.   :  97
##  1st Qu.:1.000                       1st Qu.:2559
##  Median :2.000                       Median :4056
##  Mean   :1.873                       Mean   :3374
##  3rd Qu.:2.000                       3rd Qu.:4789
##  Max.   :8.000                       Max.   :4789
##
##  last_four_years_classification0_num_uniq
##  Min.   :   0.00
##  1st Qu.:  12.00
##  Median :  24.00
##  Mean   :  43.45
##  3rd Qu.:  60.00
##  Max.   :1485.00
##
##  last_four_years_classification1_num_uniq
##  Min.   :   0.00
##  1st Qu.:  14.00
##  Median :  40.00
##  Mean   :  67.45
##  3rd Qu.:  90.00
##  Max.   :2970.00
##
##  last_four_years_classification2_num_uniq last_four_years_company_name_num_uniq
##  Min.   : 0.000                           Min.   :1
##  1st Qu.: 0.000                           1st Qu.:1
##  Median : 0.000                           Median :1
##  Mean   : 2.509                           Mean   :1
##  3rd Qu.: 0.000                           3rd Qu.:1
```

```
##  Max.   :85.000                          Max.   :2
##
##  last_four_years_company_name_most_freq
##  Min.   :111.0
##  1st Qu.:349.0
##  Median :349.0
##  Mean   :348.2
##  3rd Qu.:349.0
##  Max.   :489.0
##
##  last_four_years_reason_for_legal_announcement_num_uniq
##  Min.   : 0.00
##  1st Qu.: 9.00
##  Median :11.00
##  Mean   :10.24
##  3rd Qu.:13.00
##  Max.   :16.00
##
##  last_four_years_reason_for_legal_announcement_most_freq
##  Min.   :   0
##  1st Qu.:1028
##  Median :1028
##  Mean   :1037
##  3rd Qu.:1028
##  Max.   :1361
##
##  last_four_years_legal_announcementing_firm_num_uniq
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.301
##  3rd Qu.:3.000
##  Max.   :5.000
##
##  last_four_years_legal_announcementing_firm_most_freq
##  Min.   : 40
##  1st Qu.:292
##  Median :292
##  Mean   :292
##  3rd Qu.:292
##  Max.   :371
##
##  last_four_years_root_cause_description_num_uniq
##  Min.   :1.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.191
##  3rd Qu.:6.000
##  Max.   :6.000
##
##  last_four_years_root_cause_description_most_freq
##  Min.   : 4.00
##  1st Qu.: 4.00
##  Median : 4.00
##  Mean   :15.19
##  3rd Qu.:40.00
##  Max.   :40.00
##
##  last_four_years_product_quantity_average_num_uniq
##  Min.   : 1.000
##  1st Qu.: 7.000
##  Median : 8.000
```

```
##  Mean   : 7.908
##  3rd Qu.:10.000
##  Max.   :11.000
##
##  last_four_years_product_quantity_average_max
##  Min.   :    55.8
##  1st Qu.: 22298.0
##  Median :636572.0
##  Mean   :435865.2
##  3rd Qu.:636572.0
##  Max.   :636572.0
##
##  last_four_years_product_quantity_average_average
##  Min.   :    55.8
##  1st Qu.:  7499.2
##  Median :123061.1
##  Mean   :111990.6
##  3rd Qu.:168517.2
##  Max.   :333372.5
##
##  last_four_years_decision_date_max_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 68.00
##  Median : 71.00
##  Mean   : 74.32
##  3rd Qu.: 79.00
##  Max.   :107.00
##
##  last_four_years_decision_date_average_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 68.00
##  Median : 71.00
##  Mean   : 74.32
##  3rd Qu.: 79.00
##  Max.   :107.00
##
##  Product.issue.consequence manufacturer_contact_address_1 product.brand_name
##  Death      :  396         9476   :57856                  281286 :56194
##  Injury     :40074         13145  : 1582                  344588 :  726
##  Malfunction:19530         7455   :  418                  43203  :  566
##                            10387  :   33                  154949 :  413
##                            7614   :   20                  153901 :  269
##                            5895   :   17                  238102 :  210
##                            (Other):   74                  (Other): 1622
##  product.generic_name product.issue.type type_of_report.1 reporter_job_code
##  73852  :59399        599    :19349      0:42919          42     :26202
##  44428  :  361        629    : 7342      1:17081          32     :21020
##  44436  :   55        849    : 4262                       17     : 6657
##  45330  :   48        566    : 2951                       28     : 1944
##  44437  :   21        624    : 2425                       52     : 1740
##  89386  :   19        906    : 1816                       1      : 1372
##  (Other):   97        (Other):21855                       (Other): 1065
##   source_type    product.manufacturer_name product.product_operator
##  3      :25439   18383  :31751             15     :59596
##  4      : 8440   19408  :22018             41     :  282
##  11     : 8384   19327  : 5354             0      :   61
##  12     : 8191   12347  :  247             21     :   43
##  6      : 4329   12348  :  185             18     :   15
##  10     : 1660   24392  :  172             24     :    2
##  (Other): 3557   (Other):  273             (Other):    1
##  product.manufacturer_city product.manufacturer_state
##  4513   :52924             48     :52948
```

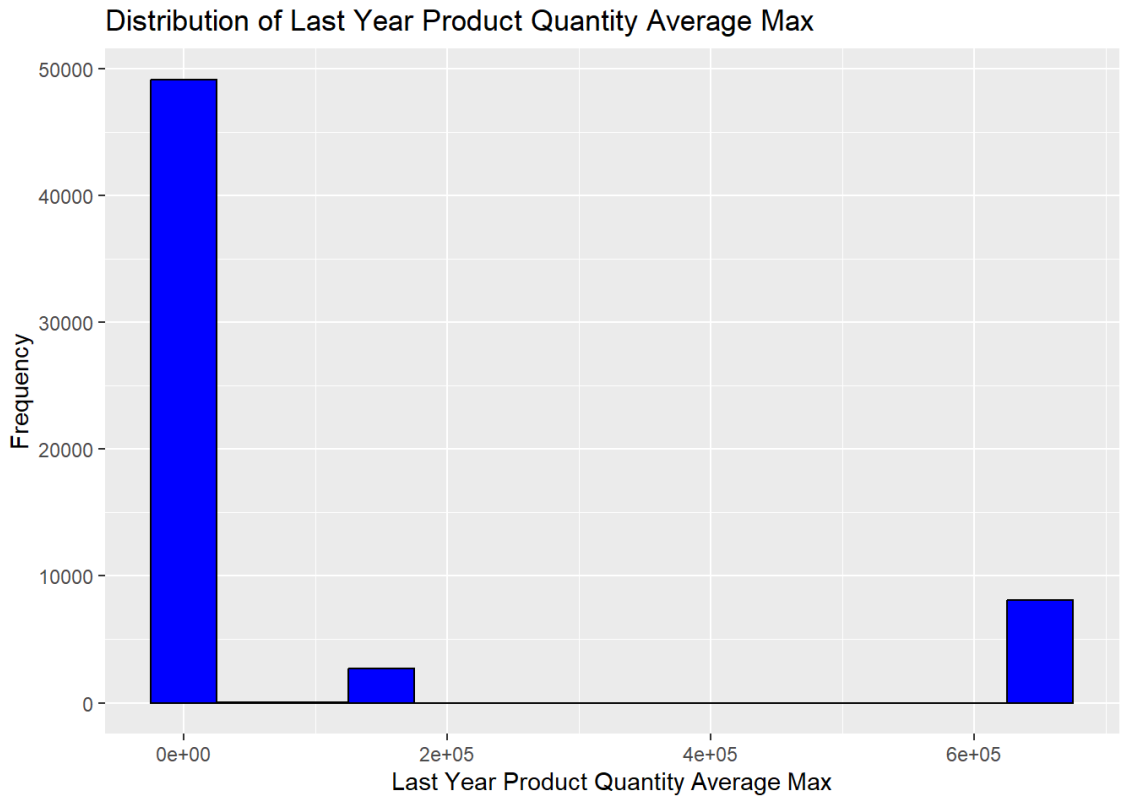
```
##  5990   : 5328             32     : 6379
##  4517   :  859             40     :  481
##  6271   :  473             63     :   99
##  3153   :  172             23     :   36
##  4284   :   38             8      :   34
##  (Other):  206             (Other):   23
##  product.manufacturer_country            product.field_description
##  109:   30                    General, Plastic Surgery:    8
##  123:   20                    Immunology              :    4
##  126:59917                    Unknown                 :59988
##  135:   33
##
##
##
##  product.product_report_product_code
##  DHX:    0
##  LKK:59907
##  LTJ:    0
##  MTF:   62
##  MTG:   19
##  PDF:    4
##  PQM:    8
```

```
str(df1_short)
```

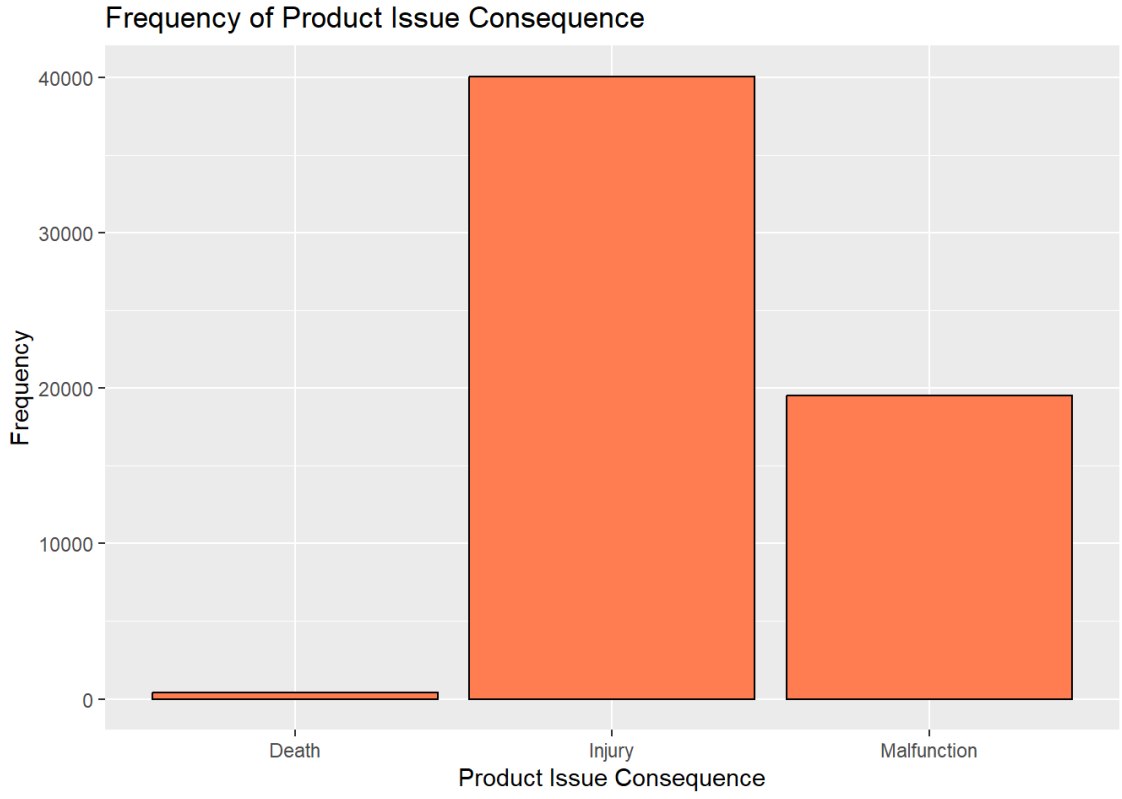
```
## 'data.frame':    60000 obs. of  77 variables:
##  $ ID_non_uniq                                          : Factor w/ 12 levels "p000021","p000027",..: 6 6 3 6
6 6 6 6 6 6 ...
##  $ date_event                                           : chr  "17-05-18" "13-09-17" "02-05-16" "25-06-16"
...
##  $ last_year_all_product_codes_num_uniq                 : int  1 2 1 3 1 1 3 2 3 1 ...
##  $ last_year_all_product_codes_most_freq                : int  1474 1467 1499 1472 1477 1474 1472 1473 1472 1
463 ...
##  $ last_year_brand_name_num_uniq                        : int  2 2 1 2 2 2 2 2 2 2 ...
##  $ last_year_brand_name_most_freq                       : int  166 4057 3409 4789 166 166 4789 2817 4789 4055
...
##  $ last_year_classification0_num_uniq                   : int  0 12 0 0 80 0 0 0 0 4 ...
##  $ last_year_classification1_num_uniq                   : int  8 36 2 32 0 8 44 14 108 4 ...
##  $ last_year_classification2_num_uniq                   : int  4 0 0 0 0 0 44 14 0 0 ...
##  $ last_year_company_name_num_uniq                      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ last_year_company_name_most_freq                     : int  349 349 251 349 349 349 349 349 349 349 ...
##  $ last_year_reason_for_legal_announcement_num_uniq     : int  3 4 2 2 4 4 2 2 3 2 ...
##  $ last_year_reason_for_legal_announcement_most_freq    : int  1361 469 52 52 1143 1028 1361 1361 52 700 ...
##  $ last_year_legal_announcementing_firm_num_uniq        : int  2 2 2 2 2 1 1 1 2 1 ...
##  $ last_year_legal_announcementing_firm_most_freq       : int  292 292 143 143 292 292 292 292 143 292 ...
##  $ last_year_root_cause_description_num_uniq            : int  3 2 2 2 2 4 2 2 2 1 ...
##  $ last_year_root_cause_description_most_freq           : int  20 40 4 4 4 19 20 20 28 40 ...
##  $ last_year_product_quantity_average_num_uniq          : int  3 4 2 2 4 2 2 2 2 2 ...
##  $ last_year_product_quantity_average_max               : num  20286 22298 62 62 636572 ...
##  $ last_year_product_quantity_average_average           : num  7561.3 11197.2 35.5 35.5 266753.1 ...
##  $ last_year_decision_date_max_changes_in_product       : int  28 21 6 21 13 14 31 26 25 19 ...
##  $ last_year_decision_date_average_changes_in_product   : int  28 21 6 21 13 14 31 26 25 19 ...
##  $ last_two_years_all_product_codes_num_uniq            : int  1 2 1 3 1 1 3 2 3 1 ...
##  $ last_two_years_all_product_codes_most_freq           : int  1474 1467 1499 1472 1477 1474 1472 1473 1472 1
463 ...
##  $ last_two_years_brand_name_num_uniq                   : int  2 2 1 2 2 2 2 2 2 2 ...
##  $ last_two_years_brand_name_most_freq                  : int  166 4057 3409 4789 166 166 4789 2817 4789 4055
...
##  $ last_two_years_classification0_num_uniq              : int  4 12 0 0 80 10 44 14 0 4 ...
##  $ last_two_years_classification1_num_uniq              : int  12 60 5 80 0 8 132 42 252 16 ...
##  $ last_two_years_classification2_num_uniq              : int  4 0 0 0 0 0 44 14 0 0 ...
##  $ last_two_years_company_name_num_uniq                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ last_two_years_company_name_most_freq                : int  349 349 251 349 349 349 349 349 349 349 ...
##  $ last_two_years_reason_for_legal_announcement_num_uniq : int  5 6 5 5 4 9 5 5 7 5 ...
##  $ last_two_years_reason_for_legal_announcement_most_freq : int  1361 469 52 52 1143 1028 1361 1361 1028 52 ...
##  $ last_two_years_legal_announcementing_firm_num_uniq   : int  2 3 2 2 2 2 2 2 2 2 ...
##  $ last_two_years_legal_announcementing_firm_most_freq  : int  292 292 292 292 292 292 292 292 292 292 ...
##  $ last_two_years_root_cause_description_num_uniq       : int  3 2 4 4 2 6 3 3 5 3 ...
##  $ last_two_years_root_cause_description_most_freq      : int  40 40 28 28 4 4 40 40 28 40 ...
##  $ last_two_years_product_quantity_average_num_uniq     : int  5 6 3 3 4 6 5 5 4 4 ...
##  $ last_two_years_product_quantity_average_max          : num  22298 22298 5463 5463 636572 ...
##  $ last_two_years_product_quantity_average_average      : num  8997 7477 1845 1845 266753 ...
##  $ last_two_years_decision_date_max_changes_in_product  : int  46 46 11 42 35 30 54 48 40 41 ...
##  $ last_two_years_decision_date_average_changes_in_product : int  46 46 11 42 35 30 54 48 40 41 ...
##  $ last_four_years_all_product_codes_num_uniq           : int  1 2 1 3 1 1 3 2 3 1 ...
##  $ last_four_years_all_product_codes_most_freq          : int  1474 1467 1499 1472 1477 1474 1472 1473 1472 1
463 ...
##  $ last_four_years_brand_name_num_uniq                  : int  2 2 1 2 2 2 2 2 2 2 ...
##  $ last_four_years_brand_name_most_freq                 : int  166 4057 3409 4789 166 166 4789 2817 4789 4055
...
##  $ last_four_years_classification0_num_uniq             : int  4 12 6 96 80 12 44 14 216 24 ...
##  $ last_four_years_classification1_num_uniq             : int  32 132 8 128 0 8 352 112 252 36 ...
##  $ last_four_years_classification2_num_uniq             : int  4 0 0 0 0 0 44 14 0 0 ...
##  $ last_four_years_company_name_num_uniq                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ last_four_years_company_name_most_freq               : int  349 349 251 349 349 349 349 349 349 349 ...
##  $ last_four_years_reason_for_legal_announcement_num_uniq : int  10 12 13 13 4 9 10 10 12 15 ...
##  $ last_four_years_reason_for_legal_announcement_most_freq : int  1361 1028 1028 1028 1143 1028 1361 1361 1028 1
```

```
028 ...
##  $ last_four_years_legal_announcementing_firm_num_uniq     : int  3 3 2 2 2 2 3 3 2 2 ...
##  $ last_four_years_legal_announcementing_firm_most_freq    : int  292 292 292 292 292 292 292 292 292 292 ...
##  $ last_four_years_root_cause_description_num_uniq         : int  5 5 6 6 2 6 5 5 6 6 ...
##  $ last_four_years_root_cause_description_most_freq        : int  40 40 4 4 4 4 40 40 4 4 ...
##  $ last_four_years_product_quantity_average_num_uniq       : int  8 8 10 10 4 7 8 8 9 10 ...
##  $ last_four_years_product_quantity_average_max            : num  22298 22298 636572 636572 636572 ...
##  $ last_four_years_product_quantity_average_average        : num  6315 7123 123061 123061 266753 ...
##  $ last_four_years_decision_date_max_changes_in_product    : int  88 81 24 68 69 64 95 90 72 73 ...
##  $ last_four_years_decision_date_average_changes_in_product: int  88 81 24 68 69 64 95 90 72 73 ...
##  $ Product.issue.consequence                               : Factor w/ 3 levels "Death","Injury",..: 2 2 2 2 2 2
2 3 3 2 ...
##  $ manufacturer_contact_address_1                          : Factor w/ 22 levels "2179","5833",..: 22 16 11 16 1
6 16 22 16 16 16 ...
##  $ product.brand_name                                      : Factor w/ 128 levels "130","22382",..: 103 103 83 1
03 103 15 103 103 120 103 ...
##  $ product.generic_name                                    : Factor w/ 46 levels "77","15128","16785",..: 37 37
17 37 37 37 37 37 37 37 ...
##  $ product.issue.type                                      : Factor w/ 252 levels "2","3","4","8",..: 177 170 18
1 140 175 177 57 170 170 227 ...
##  $ type_of_report.1                                        : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 2 2 1 ...
##  $ reporter_job_code                                       : Factor w/ 22 levels "1","2","3","8",..: 22 19 19 14
19 19 22 14 19 19 ...
##  $ source_type                                             : Factor w/ 17 levels "3","4","5","6",..: 1 1 8 1 9 8
1 2 1 1 ...
##  $ product.manufacturer_name                               : Factor w/ 33 levels "7480","7483",..: 10 10 7 10 10
20 22 22 20 10 ...
##  $ product.product_operator                                : Factor w/ 9 levels "0","15","18",..: 2 2 9 2 2 2 2
2 2 2 ...
##  $ product.manufacturer_city                               : Factor w/ 33 levels "1375","1717",..: 12 12 23 12 1
2 19 12 12 19 12 ...
##  $ product.manufacturer_state                              : Factor w/ 9 levels "8","13","23",..: 8 8 6 8 8 5 8
8 5 8 ...
##  $ product.manufacturer_country                            : Factor w/ 4 levels "109","123","126",..: 3 3 3 3 3
3 3 3 3 3 ...
##  $ product.field_description                               : Factor w/ 3 levels "General, Plastic Surgery",..: 3
3 3 3 3 3 3 3 3 ...
##  $ product.product_report_product_code                     : Factor w/ 7 levels "DHX","LKK","LTJ",..: 2 2 2 2 2
2 2 2 2 2 ...
```

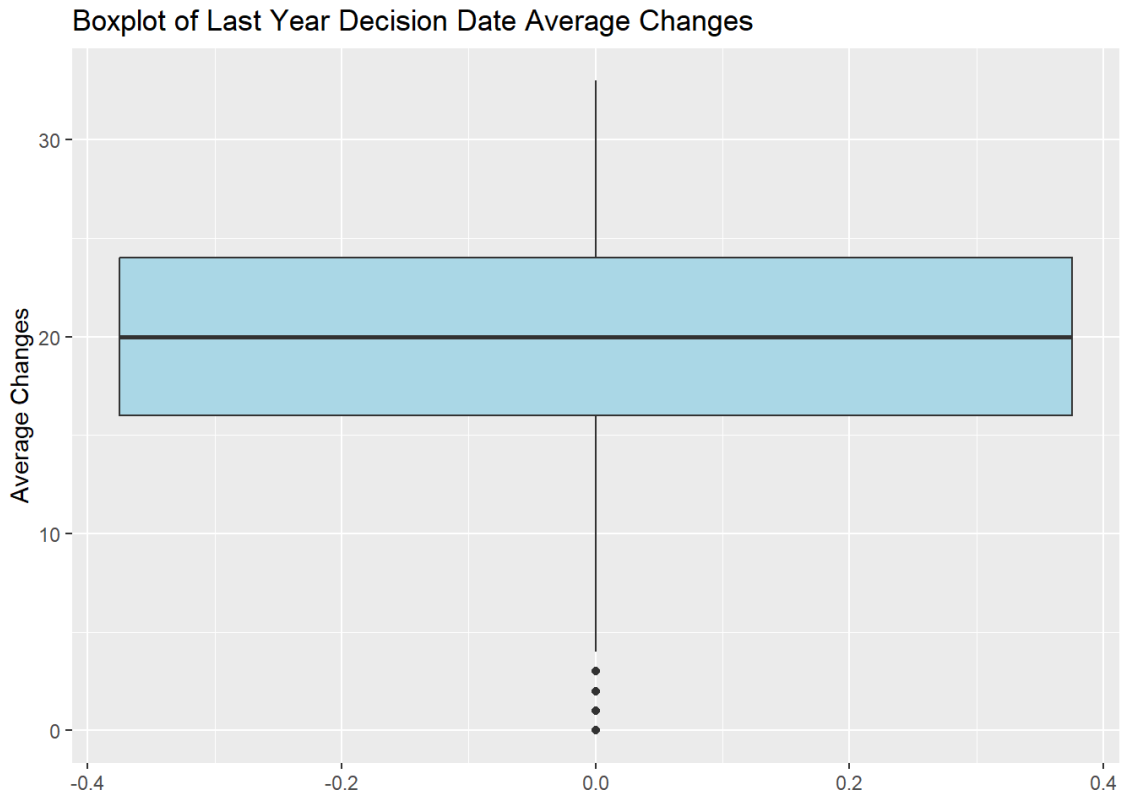
```
##
ggplot(df1_short, aes(x = last_year_product_quantity_average_max)) +
  geom_histogram(binwidth = 50000, fill = "blue", color = "black") +
  labs(title = "Distribution of Last Year Product Quantity Average Max",
       x = "Last Year Product Quantity Average Max",
       y = "Frequency")
```



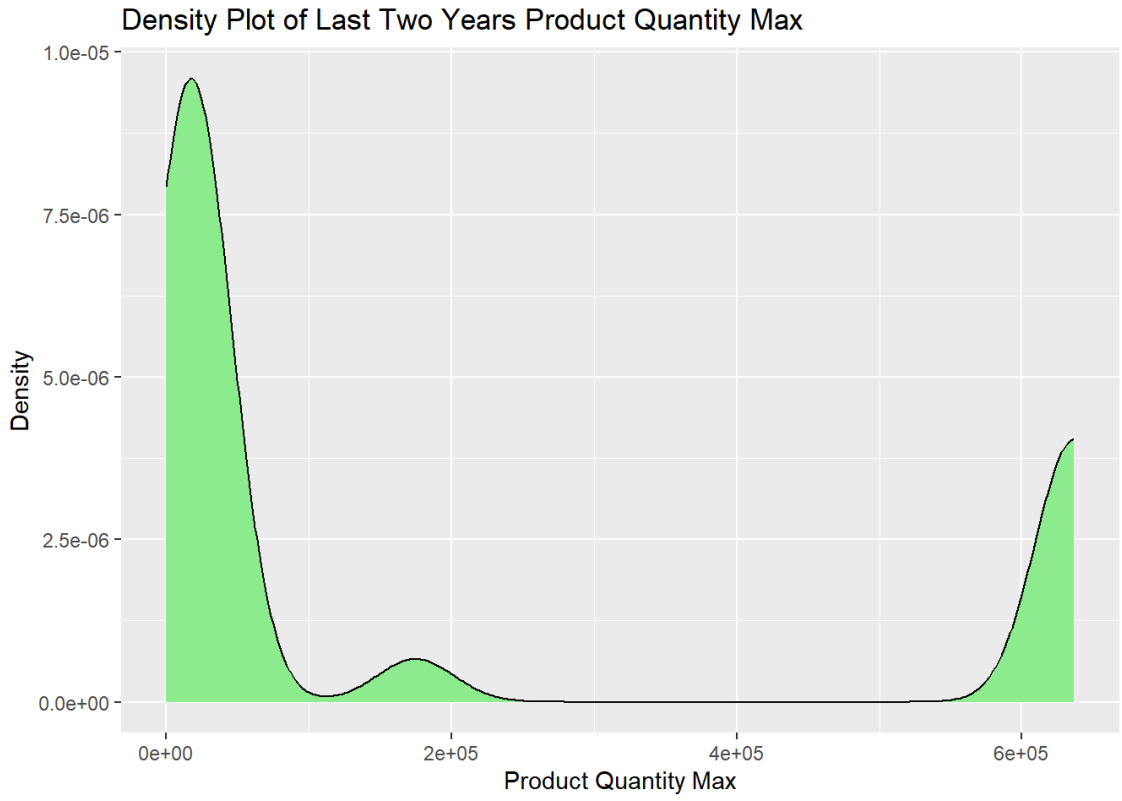

```
ggplot(df1_short, aes(x = Product.issue.consequence)) +
  geom_bar(fill = "coral", color = "black") +
  labs(title = "Frequency of Product Issue Consequence",
       x = "Product Issue Consequence",
       y = "Frequency")
```



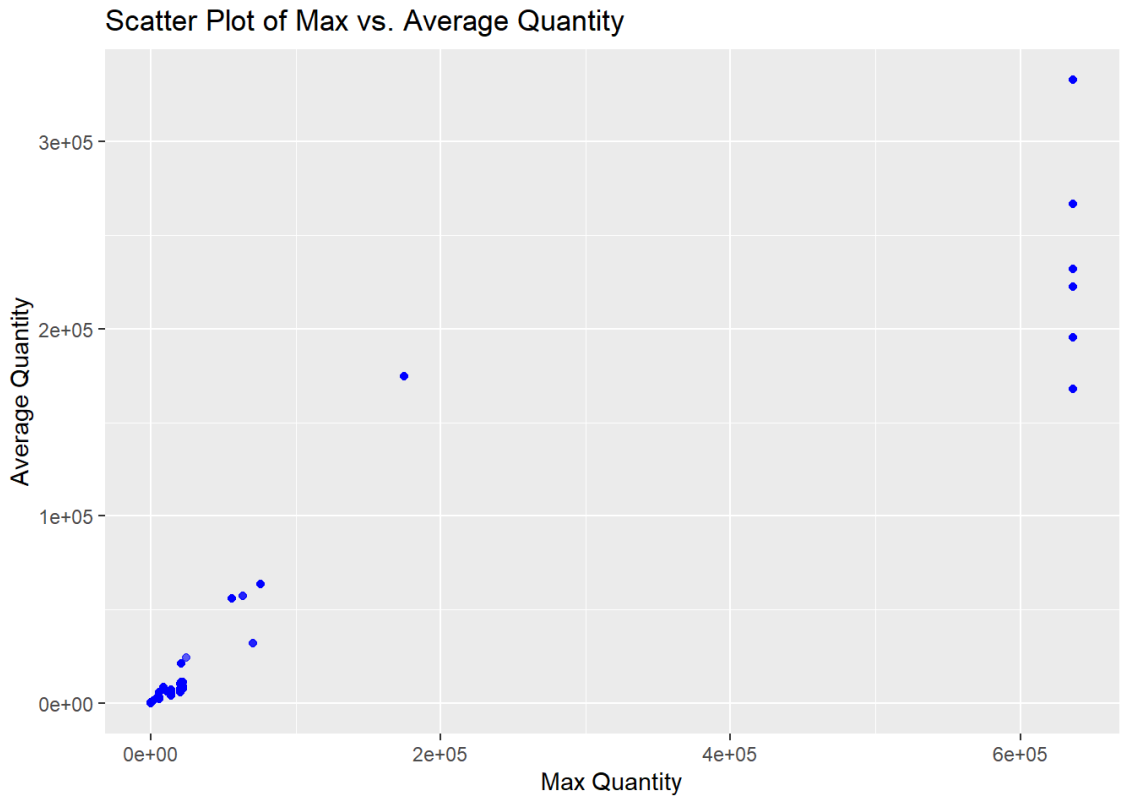
```
ggplot(df1_short, aes(y = last_year_decision_date_average_changes_in_product)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Last Year Decision Date Average Changes", y = "Average Changes")
```




```
ggplot(df1_short, aes(x = last_two_years_product_quantity_average_max)) +
  geom_density(fill = "lightgreen") +
  labs(title = "Density Plot of Last Two Years Product Quantity Max", x = "Product Quantity Max", y = "Density")
```

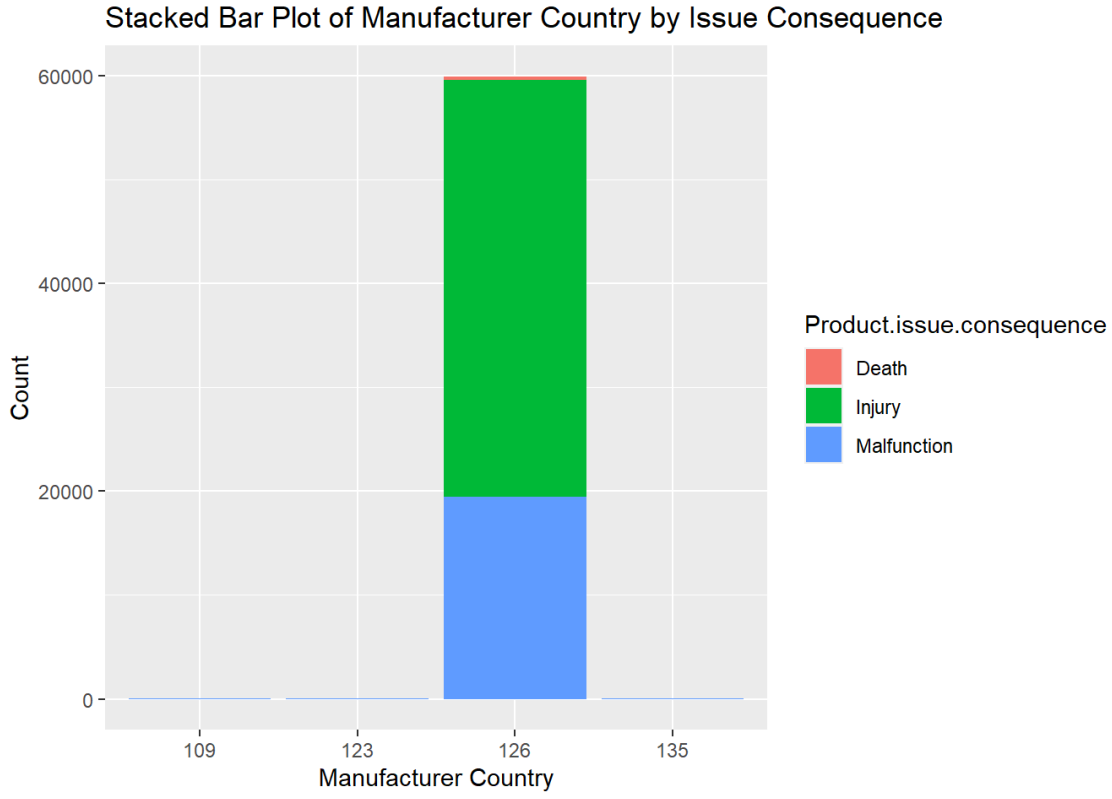


```
ggplot(df1_short, aes(x = last_year_product_quantity_average_max, y = last_year_product_quantity_average_average))
+
  geom_point(alpha = 0.6, color = "blue") +
  labs(title = "Scatter Plot of Max vs. Average Quantity", x = "Max Quantity", y = "Average Quantity")
```




```
ggplot(df1_short, aes(x = Product.issue.consequence, y = last_year_company_name_num_uniq, fill = Product.issue.cons
equence)) +
  geom_violin() +
  labs(title = "Violin Plot of Company Name Uniqueness by Issue Consequence", x = "Issue Consequence", y = "Company
Name Uniqueness")
```


```
ggplot(df1_short, aes(x = product.manufacturer_country, fill = Product.issue.consequence)) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Plot of Manufacturer Country by Issue Consequence", x = "Manufacturer Country", y = "Co
unt")
```



##Zero Analysis 1

This table tells us that there are at most 9 variables that are contributing to majority of the zero counts in the dataset. Top 3 of these variables are related to classification2 throughout the four years where up to a maximum 81% of there observations are zero values. All top 9 are either classification0, 1 or 2 variables.

```r
#Zero Analysis 1
results <- data.frame(
  Variable = character(),
  Number_of_Zero_Rows = integer(),
  Percentage_of_Zero_Rows = numeric(),
  stringsAsFactors = FALSE
)
count_zero_rows <- function(x) {
  sum(x == 0, na.rm = TRUE)
}

for (var in names(df1)) {
  num_zero_rows <- sum(df1[[var]] == 0, na.rm = TRUE)
  percent_zero_rows <- (num_zero_rows / nrow(df1)) * 100
  results <- rbind(results, data.frame(Variable = var, Number_of_Zero_Rows = num_zero_rows, Percentage_of_Zero_Rows
= percent_zero_rows))
}

results_zero <- results[order(-results$Number_of_Zero_Rows),]
print(results_zero)
```

```
##                                                 Variable Number_of_Zero_Rows
## 9                      last_year_classification2_num_uniq               452754
## 29                 last_two_years_classification2_num_uniq               430503
## 49                last_four_years_classification2_num_uniq               423576
## 68                                     type_of_report.1               396032
## 7                      last_year_classification0_num_uniq               340643
## 27                 last_two_years_classification0_num_uniq               158970
## 8                      last_year_classification1_num_uniq                95971
## 28                 last_two_years_classification1_num_uniq                78228
## 48                last_four_years_classification1_num_uniq                78042
## 12         last_year_reason_for_legal_announcement_num_uniq                30669
## 13        last_year_reason_for_legal_announcement_most_freq                30669
## 32    last_two_years_reason_for_legal_announcement_num_uniq                24452
## 33   last_two_years_reason_for_legal_announcement_most_freq                24452
## 52   last_four_years_reason_for_legal_announcement_num_uniq                24366
## 53  last_four_years_reason_for_legal_announcement_most_freq                24366
## 21           last_year_decision_date_max_changes_in_product                 6448
## 22       last_year_decision_date_average_changes_in_product                 6448
## 18              last_year_product_quantity_average_num_uniq                 6304
## 19                   last_year_product_quantity_average_max                 6304
## 20               last_year_product_quantity_average_average                 6304
## 3                      last_year_all_product_codes_num_uniq                 6258
## 4                     last_year_all_product_codes_most_freq                 6258
## 5                             last_year_brand_name_num_uniq                 6258
## 6                            last_year_brand_name_most_freq                 6258
## 10                          last_year_company_name_num_uniq                 6258
## 11                         last_year_company_name_most_freq                 6258
## 14            last_year_legal_announcementing_firm_num_uniq                 6258
## 15           last_year_legal_announcementing_firm_most_freq                 6258
## 16                last_year_root_cause_description_num_uniq                 6258
## 17               last_year_root_cause_description_most_freq                 6258
## 47                 last_four_years_classification0_num_uniq                  869
## 72                                  product.product_operator                  516
## 41     last_two_years_decision_date_max_changes_in_product                  154
## 42  last_two_years_decision_date_average_changes_in_product                  154
## 38         last_two_years_product_quantity_average_num_uniq                  146
## 39              last_two_years_product_quantity_average_max                  146
## 40          last_two_years_product_quantity_average_average                  146
## 23                last_two_years_all_product_codes_num_uniq                   86
## 24               last_two_years_all_product_codes_most_freq                   86
## 25                       last_two_years_brand_name_num_uniq                   86
## 26                      last_two_years_brand_name_most_freq                   86
## 30                     last_two_years_company_name_num_uniq                   86
## 31                    last_two_years_company_name_most_freq                   86
## 34       last_two_years_legal_announcementing_firm_num_uniq                   86
## 35      last_two_years_legal_announcementing_firm_most_freq                   86
## 36           last_two_years_root_cause_description_num_uniq                   86
## 37          last_two_years_root_cause_description_most_freq                   86
## 58        last_four_years_product_quantity_average_num_uniq                   25
## 59             last_four_years_product_quantity_average_max                   25
## 60         last_four_years_product_quantity_average_average                   25
## 61     last_four_years_decision_date_max_changes_in_product                    7
## 62 last_four_years_decision_date_average_changes_in_product                    7
## 1                                            ID_non_uniq                    0
## 2                                             date_event                    0
## 43               last_four_years_all_product_codes_num_uniq                    0
## 44              last_four_years_all_product_codes_most_freq                    0
## 45                      last_four_years_brand_name_num_uniq                    0
## 46                     last_four_years_brand_name_most_freq                    0
## 50                    last_four_years_company_name_num_uniq                    0
## 51                   last_four_years_company_name_most_freq                    0
## 54      last_four_years_legal_announcementing_firm_num_uniq                    0
```

```
## 55     last_four_years_legal_announcementing_firm_most_freq                   0
## 56          last_four_years_root_cause_description_num_uniq                   0
## 57         last_four_years_root_cause_description_most_freq                   0
## 63                                Product.issue.consequence                   0
## 64                           manufacturer_contact_address_1                   0
## 65                                     product.brand_name                     0
## 66                                   product.generic_name                     0
## 67                                     product.issue.type                     0
## 69                                      reporter_job_code                     0
## 70                                            source_type                    0
## 71                                product.manufacturer_name                  0
## 73                                product.manufacturer_city                  0
## 74                               product.manufacturer_state                  0
## 75                             product.manufacturer_country                  0
## 76                                product.field_description                  0
## 77                         product.product_report_product_code                0
##    Percentage_of_Zero_Rows
## 9              81.957995880
## 29             77.930096919
## 49             76.676164237
## 68             71.690120958
## 7              61.663547071
## 27             28.776913302
## 8              17.372769368
## 28             14.160913215
## 48             14.127243303
## 12              5.551734000
## 13              5.551734000
## 32              4.426326251
## 33              4.426326251
## 52              4.410758442
## 53              4.410758442
## 21              1.167223608
## 22              1.167223608
## 18              1.141156580
## 19              1.141156580
## 20              1.141156580
## 3               1.132829612
## 4               1.132829612
## 5               1.132829612
## 6               1.132829612
## 10              1.132829612
## 11              1.132829612
## 14              1.132829612
## 15              1.132829612
## 16              1.132829612
## 17              1.132829612
## 47              0.157307276
## 72              0.093406852
## 41              0.027877239
## 42              0.027877239
## 38              0.026429071
## 39              0.026429071
## 40              0.026429071
## 23              0.015567809
## 24              0.015567809
## 25              0.015567809
## 26              0.015567809
## 30              0.015567809
## 31              0.015567809
## 34              0.015567809
## 35              0.015567809
```

```
## 36            0.015567809
## 37            0.015567809
## 58            0.004525526
## 59            0.004525526
## 60            0.004525526
## 61            0.001267147
## 62            0.001267147
## 1             0.000000000
## 2             0.000000000
## 43            0.000000000
## 44            0.000000000
## 45            0.000000000
## 46            0.000000000
## 50            0.000000000
## 51            0.000000000
## 54            0.000000000
## 55            0.000000000
## 56            0.000000000
## 57            0.000000000
## 63            0.000000000
## 64            0.000000000
## 65            0.000000000
## 66            0.000000000
## 67            0.000000000
## 69            0.000000000
## 70            0.000000000
## 71            0.000000000
## 73            0.000000000
## 74            0.000000000
## 75            0.000000000
## 76            0.000000000
## 77            0.000000000
```

##Zero Analysis 2

The table provides counts of rows with zero occurrences for three categories—death, injury, and malfunction—across four different time spans: the last year, the last 2 years, the last 4 years, and others. In each category, the number of rows with zero occurrences decreases as the time span increases, suggesting there are more instances of non-zero occurrences (events happening) as the time frame expands.

```
#Zero Analysis 2
death_subset <- df[df$Product.issue.consequence == "Death", ]
injury_subset <- df[df$Product.issue.consequence == "Injury", ]
malfunction_subset <- df[df$Product.issue.consequence == "Malfunction", ]

count_zero_rows_any <- function(df, cols) {
  sum(apply(df[, cols, drop = FALSE] == 0, 1, any, na.rm = TRUE))
}
resultsMatrix <- matrix(nrow = 4, ncol = 4, dimnames = list(
  c("Total Rows with 0", "Total Rows with 0 for Death", "Total Rows with 0 for Injury", "Total Rows with 0 for Malf
unction"),
  c("Last Year", "Last 2 Years", "Last 4 Years", "Others")
))
resultsMatrix[1, ] <- c(
  count_zero_rows_any(df, last_year_cols),
  count_zero_rows_any(df, last_two_years_cols),
  count_zero_rows_any(df, last_four_years_cols),
  count_zero_rows_any(df, other_cols)
)
resultsMatrix[2, ] <- c(
  count_zero_rows_any(death_subset, last_year_cols),
  count_zero_rows_any(death_subset, last_two_years_cols),
  count_zero_rows_any(death_subset, last_four_years_cols),
  count_zero_rows_any(death_subset, other_cols)
)
resultsMatrix[3, ] <- c(
  count_zero_rows_any(injury_subset, last_year_cols),
  count_zero_rows_any(injury_subset, last_two_years_cols),
  count_zero_rows_any(injury_subset, last_four_years_cols),
  count_zero_rows_any(injury_subset, other_cols)
)
resultsMatrix[4, ] <- c(
  count_zero_rows_any(malfunction_subset, last_year_cols),
  count_zero_rows_any(malfunction_subset, last_two_years_cols),
  count_zero_rows_any(malfunction_subset, last_four_years_cols),
  count_zero_rows_any(malfunction_subset, other_cols)
)
results_zero_2 <- as.data.frame(resultsMatrix)
print(results_zero_2)
```

```
##                                 Last Year Last 2 Years Last 4 Years Others
## Total Rows with 0                  547880       449267       423576 396224
## Total Rows with 0 for Death          3703         3420         3384   2707
## Total Rows with 0 for Injury       364650       296179       278186 284438
## Total Rows with 0 for Malfunction  179527       149668       142006 109079
```

Statistical Analysis

Now beyond this point, we will perform Forward Selection, Logistic Regression and LDA for the subsetted datasets.

The data that is subsetted at the top level is for

1 - Product Field of Unknown 2 - Report code of LKK 3 - Issue type of 599 4 - Manufacturer Contact Address 1 of 9476

For each of these subsets of data, Forward selection is applied on selecting levels from the following variables

1 - State 3 - Source Type 2 - Operator 3 - Last Year Variables 4 - Last Two Year Variables 5 - Last Four Year Variables 6 - City

For each of the Forward selections applied, multiple variables were selected based on the variables that were significant.

Ultimately after the variable selection, Logistic Regression and LDA was applied for each of the top level subsets. Hence we end up with

1 - 4 Logistic Regression Models 2 - 4 LDA Models

#Code for Forward Selection for Unknown Field

#Forward Selection for unknown field by state

df2_unknown_state <- df2_unknown %>% dplyr::select(death_or_not, starts_with(“product.manufacturer_state”))

summary(df2_unknown_state)

df2_unknown_state <- sample_n(df2_unknown_state, size = 100000)

summary(df2_unknown_state)

df2_unknown_state$death_or_not < -as.numeric(df2_unknown_state$death_or_not)

initial_model_state <- glm(death_or_not ~ 1, data = df2_unknown_state, family = binomial())

full_model_formula_state <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_unknown_state), “death_or_not”), collapse = “+”))) forward_selected_model_state <- stepAIC(initial_model_state, scope = list(lower = initial_model_state$formula, upper = full_model_formula_state), direction = “forward”, trace = FALSE)

summary(forward_selected_model_state)

#! significant variables are: product.manufacturer_state_40,product.manufacturer_state_32

coef(forward_selected_model_state)

par(mfrow = c(2, 2)) plot(forward_selected_model_state)

Forward Selection for unknown field by source type

df2_unknown_source <- df2_unknown %>% dplyr::select(death_or_not, starts_with(“source_type”))

df2_unknown_source <- sample_n(df2_unknown_source, size = 100000)

df2_unknown_source$death_or_not < -as.numeric(df2_unknown_source$death_or_not)

initial_model_source <- glm(death_or_not ~ 1, data = df2_unknown_source, family = binomial())

full_model_formula_source <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_unknown_source), “death_or_not”), collapse = “+”))) forward_selected_model_source <- stepAIC(initial_model_source, scope = list(lower = initial_model_source$formula, upper = full_model_formula_source), direction = “forward”, trace = FALSE)

summary(forward_selected_model_source) # ! significant variables are: source_type_3, source_type_17, source_type_11, source_type_16, source_type_5,source_type_4, source_type_15

coef(forward_selected_model_source)

par(mfrow = c(2, 2)) plot(forward_selected_model_source)

Forward Selection for unknown field by product operator

df2_unknown_operator <- df2_unknown %>% dplyr::select(death_or_not, starts_with(“product.product_operator”))

df2_unknown_operator <- sample_n(df2_unknown_operator, size = 100000)

df2_unknown_operator$death_or_not < -as.numeric(df2_unknown_operator$death_or_not)

initial_model_operator <- glm(death_or_not ~ 1, data = df2_unknown_operator, family = binomial())

full_model_formula_operator <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_unknown_operator), “death_or_not”), collapse = “+”))) forward_selected_model_operator <- stepAIC(initial_model_operator, scope = list(lower = initial_model_operator$formula, upper = full_model_formula_operator), direction = “forward”, trace = FALSE)

summary(forward_selected_model_operator) # ! significant variables are: product.product_operator_41, product.product_operator_18

coef(forward_selected_model_operator)

par(mfrow = c(2, 2)) plot(forward_selected_model_operator)

#Forward selection for Unknown by last year variables

df2_unknown_last_year <- df2_unknown %>% dplyr::select(death_or_not, starts_with("last_year"))

df2_unknown_last_year <- sample_n(df2_unknown_last_year, size = 100000)

df2_unknown_last_year$death_or_not <- -as.numeric(df2_unknown_last_year$death_or_not)

initial_model_last_year <- glm(death_or_not ~ 1, data = df2_unknown_last_year, family = binomial())

full_model_formula_last_year <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_unknown_last_year), "death_or_not"), collapse = "+"))) forward_selected_model_last_year <- stepAIC(initial_model_last_year, scope = list(lower = initial_model_last_year$formula, upper = full_model_formula_last_year), direction = "forward", trace = FALSE)

summary(forward_selected_model_last_year) # ! selected variables that are significant: last_year_decision_date_max_changes_in_product, last_year_company_name_num_uniq, last_year_classification1_num_uniq, last_year_reason_for_legal_announcement_most_freq # ! last_year_product_quantity_average_average, last_year_product_quantity_average_max, last_year_reason_for_legal_announcement_num_uniq, last_year_reason_for_legal_announcement_num_uniq # ! last_year_legal_announcementing_firm_most_freq, last_year_product_quantity_average_num_uniq, last_year_classification2_num_uniq , last_year_brand_name_num_uniq, last_year_brand_name_most_freq, last_year_company_name_most_freq

coef(forward_selected_model_last_year)

par(mfrow = c(2, 2)) plot(forward_selected_model_last_year)

#Forward selection of Unknown for last two years

df2_unknown_last_two_years <- df2_unknown %>% dplyr::select(death_or_not, starts_with("last_two_years"))

df2_unknown_last_two_years <- sample_n(df2_unknown_last_two_years, size = 100000)

df2_unknown_last_two_years$death_or_not <- -as.numeric(df2_unknown_last_two_years$death_or_not)

initial_model_last_two_years <- glm(death_or_not ~ 1, data = df2_unknown_last_two_years, family = binomial())

full_model_formula_last_two_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_unknown_last_two_years), "death_or_not"), collapse = "+"))) forward_selected_model_last_two_years <- stepAIC(initial_model_last_two_years, scope = list(lower = initial_model_last_two_years$formula, upper = full_model_formula_last_two_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_last_two_years)

! Selected variables that are significant are as follows:

! last_two_years_decision_date_max_changes_in_product

#! last_two_years_reason_for_legal_announcement_num_uniq
#! last_two_years_root_cause_description_most_freq
#! last_two_years_classification1_num_uniq
#! last_two_years_product_quantity_average_max
#! last_two_years_product_quantity_average_average
#! last_two_years_product_quantity_average_num_uniq
#! last_two_years_legal_announcementing_firm_num_uniq
#! last_two_years_root_cause_description_num_uniq
#! last_two_years_company_name_most_freq

coef(forward_selected_model_last_two_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_last_two_years)

Forward selection of unknown for last four years

df2_unknown_last_four_years <- df2_unknown %>% dplyr::select(death_or_not, starts_with("last_four_years"))

df2_unknown_last_four_years <- sample_n(df2_unknown_last_four_years, size = 100000)

df2_unknown_last_four_years$death_or_not <- as.numeric(df2_unknown_last_four_years$death_or_not)

initial_model_last_four_years <- glm(death_or_not ~ 1, data = df2_unknown_last_four_years, family = binomial())

full_model_formula_last_four_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_unknown_last_four_years), "death_or_not"), collapse = "+"))) forward_selected_model_last_four_years <- stepAIC(initial_model_last_four_years, scope = list(lower = initial_model_last_four_years$formula, upper = full_model_formula_last_four_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_last_four_years) #! varibales selected that are significant are as follows: #! last_four_years_decision_date_max_changes_in_product #! last_four_years_classification1_num_uniq #! last_four_years_legal_announcementing_firm_num_uniq #! last_four_years_classification2_num_uniq #! last_four_years_product_quantity_average_average #! last_four_years_company_name_most_freq

coef(forward_selected_model_last_four_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_last_four_years)

! Taking the variables that I did also had a reason. It was requested in the guidelines not to take any variables greater than 10 levels

! but that made the present options very small as variables have many large levels.

! and variables needed to be selected so that they made sense together.

! We are trying to observe changes over the years and trying to find the root of the products that might be causing deaths

! Root by manufacturer, source type and city.

#Forward Selection of Unknown by city

df2_unknown_city <- df2_unknown %>% dplyr::select(death_or_not, starts_with("product.manufacturer_city"))

df2_unknown_city <- sample_n(df2_unknown_city, size = 100000)

df2_unknown_city$death_or_not <- as.numeric(df2_unknown_city$death_or_not)

initial_model_city <- glm(death_or_not ~ 1, data = df2_unknown_city, family = binomial())

full_model_formula_city <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_unknown_city), "death_or_not"), collapse = "+"))) forward_selected_model_city <- stepAIC(initial_model_city, scope = list(lower = initial_model_city$formula, upper = full_model_formula_city), direction = "forward", trace = FALSE)

summary(forward_selected_model_city) # ! selected variables that are significant are as follows: # ! product.manufacturer_city_6271 # ! product.manufacturer_city_2085 # ! product.manufacturer_city_4513 # ! product.manufacturer_city_5092 # ! product.manufacturer_city_3153

coef(forward_selected_model_city)

par(mfrow = c(2, 2)) plot(forward_selected_model_city)

summary(df1$product.field_description)

Code for Forward Selection for Prouct Code 599

df2_issue <- df2 %>% filter(product.issue.type_599 == “1”)

#Forward Selection by state

df2_issue_state <- df2_issue %>% dplyr::select(death_or_not, starts_with(“product.manufacturer_state”))

df2_issue_state <- sample_n(df2_issue_state, size = 100000)

summary(df2_issue_state)

df2_issue_state$death_or_not <- as.numeric(df2_issue_state$death_or_not)

initial_model_issue_state <- glm(death_or_not ~ 1, data = df2_issue_state, family = binomial())

full_model_formula_issue_state <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_issue_state), “death_or_not”), collapse = “+”))) forward_selected_model_issue_state <- stepAIC(initial_model_issue_state, scope = list(lower = initial_model_issue_state$formula, upper = full_model_formula_issue_state), direction = “forward”, trace = FALSE)

summary(forward_selected_model_issue_state) # ! No significance

coef(forward_selected_model_issue_state)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_state)

#Forward selection by source type

df2_issue_source <- df2_issue %>% dplyr::select(death_or_not, starts_with(“source_type”))

df2_issue_source <- sample_n(df2_issue_source, size = 100000)

summary(df2_issue_source)

df2_issue_source$death_or_not <- as.numeric(df2_issue_source$death_or_not)

initial_model_issue_source <- glm(death_or_not ~ 1, data = df2_issue_source, family = binomial())

full_model_formula_issue_source <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_issue_source), “death_or_not”), collapse = “+”))) forward_selected_model_issue_source <- stepAIC(initial_model_issue_source, scope = list(lower = initial_model_issue_source$formula, upper = full_model_formula_issue_source), direction = “forward”, trace = FALSE)

summary(forward_selected_model_issue_source) # ! No significance

coef(forward_selected_model_issue_source)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_source)

#Forward selection by operator

df2_issue_operator <- df2_issue %>% dplyr::select(death_or_not, starts_with(“product.product_operator”))

df2_issue_operator <- sample_n(df2_issue_operator, size = 100000)

summary(df2_issue_operator)

df2_issue_operator$death_or_not <- as.numeric(df2_issue_operator$death_or_not)

initial_model_issue_operator <- glm(death_or_not ~ 1, data = df2_issue_operator, family = binomial())

full_model_formula_issue_operator <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_issue_operator), “death_or_not”), collapse = “+”))) forward_selected_model_issue_operator <- stepAIC(initial_model_issue_operator, scope = list(lower = initial_model_issue_operator$formula, upper = full_model_formula_issue_operator), direction = “forward”, trace = FALSE)

summary(forward_selected_model_issue_operator) # !No significance

coef(forward_selected_model_issue_operator)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_operator)

#Forward selection on the basis of last year variables

df2_issue_last_year <- df2_issue %>% dplyr::select(death_or_not, starts_with("last_year"))

df2_issue_last_year <- sample_n(df2_issue_last_year, size = 100000)

summary(df2_issue_last_year)

df2_issue_last_year$death_or_not <- -as.numeric(df2_issue_last_year$death_or_not)

initial_model_issue_last_year <- glm(death_or_not ~ 1, data = df2_issue_last_year, family = binomial())

full_model_formula_issue_last_year <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_issue_last_year), "death_or_not"), collapse = "+"))) forward_selected_model_issue_last_year <- stepAIC(initial_model_issue_last_year, scope = list(lower = initial_model_issue_last_year$formula, upper = full_model_formula_issue_last_year), direction = "forward", trace = FALSE)

summary(forward_selected_model_issue_last_year) #!selected variables: #!last_year_root_cause_description_most_freq
#!last_year_classification0_num_uniq
#!last_year_product_quantity_average_num_uniq
#!last_year_decision_date_max_changes_in_product
#!last_year_reason_for_legal_announcement_num_uniq
#!last_year_reason_for_legal_announcement_most_freq #!last_year_classification2_num_uniq
#!last_year_brand_name_num_uniq
#!last_year_brand_name_most_freq

coef(forward_selected_model_issue_last_year)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_last_year)

#Forward selection on the basis of last two year variables

df2_issue_last_two_years <- df2_issue %>% dplyr::select(death_or_not, starts_with("last_two_years"))

df2_issue_last_two_years <- sample_n(df2_issue_last_two_years, size = 100000)

summary(df2_issue_last_two_years)

df2_issue_last_two_years$death_or_not <- -as.numeric(df2_issue_last_two_years$death_or_not)

initial_model_issue_last_two_years <- glm(death_or_not ~ 1, data = df2_issue_last_two_years, family = binomial())

full_model_formula_issue_last_two_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_issue_last_two_years), "death_or_not"), collapse = "+"))) forward_selected_model_issue_last_two_years <- stepAIC(initial_model_issue_last_two_years, scope = list(lower = initial_model_issue_last_two_years$formula, upper = full_model_formula_issue_last_two_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_issue_last_two_years) #!selected variables
#!last_two_years_reason_for_legal_announcement_num_uniq
#!last_two_years_classification0_num_uniq
#!last_two_years_product_quantity_average_average
#!last_two_years_legal_announcementing_firm_num_uniq
#!last_two_years_product_quantity_average_max
#!last_two_years_reason_for_legal_announcement_most_freq #!last_two_years_root_cause_description_num_uniq

coef(forward_selected_model_issue_last_two_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_last_two_years)

#Forward selection on the basis of last four year variables

df2_issue_last_four_years <- df2_issue %>% dplyr::select(death_or_not, starts_with("last_four_years"))

df2_issue_last_four_years <- sample_n(df2_issue_last_four_years, size = 100000)

summary(df2_issue_last_four_years)

df2_issue_last_four_years$death_or_not <- -as.numeric(df2_issue_last_four_years$death_or_not)

initial_model_issue_last_four_years <- glm(death_or_not ~ 1, data = df2_issue_last_four_years, family = binomial())

full_model_formula_issue_last_four_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_issue_last_four_years), "death_or_not"), collapse = "+"))) forward_selected_model_issue_last_four_years <- stepAIC(initial_model_issue_last_four_years, scope = list(lower = initial_model_issue_last_four_years$formula, upper = full_model_formula_issue_last_four_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_issue_last_four_years) #!variables selected: #!last_four_years_product_quantity_average_average #!last_four_years_classification2_num_uniq
#!last_four_years_legal_announcementing_firm_num_uniq
#!last_four_years_decision_date_average_changes_in_product #!last_four_years_product_quantity_average_num_uniq
#!last_four_years_reason_for_legal_announcement_most_freq

coef(forward_selected_model_issue_last_four_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_last_four_years)

#Forward selection on the basis of city

df2_issue_city <- df2_issue %>% dplyr::select(death_or_not, starts_with("product.manufacturer_city"))

df2_issue_city <- sample_n(df2_issue_city, size = 100000)

summary(df2_issue_city)

df2_issue_city$death_or_not <- -as.numeric(df2_issue_city$death_or_not)

initial_model_issue_city <- glm(death_or_not ~ 1, data = df2_issue_city, family = binomial())

full_model_formula_issue_city <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_issue_city), "death_or_not"), collapse = "+"))) forward_selected_model_issue_city <- stepAIC(initial_model_issue_city, scope = list(lower = initial_model_issue_city$formula, upper = full_model_formula_issue_city), direction = "forward", trace = FALSE)

summary(forward_selected_model_issue_city)

#!No significance

coef(forward_selected_model_issue_city)

par(mfrow = c(2, 2)) plot(forward_selected_model_issue_city)

Code for Forward selection for Contact Address 9476

df2_address <- df2 %>% filter(manufacturer_contact_address_1_9476 == "1")

Forward selection by state

df2_address_state <- df2_address %>% dplyr::select(death_or_not, starts_with("product.manufacturer_state"))

df2_address_state <- sample_n(df2_address_state, size = 100000)

summary(df2_address_state)

df2_address_state$death_or_not <- -as.numeric(df2_address_state$death_or_not)

initial_model_address_state <- glm(death_or_not ~ 1, data = df2_address_state, family = binomial())

full_model_formula_address_state <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_state), "death_or_not"), collapse = "+"))) forward_selected_model_address_state <- stepAIC(initial_model_address_state, scope = list(lower = initial_model_address_state$formula, upper = full_model_formula_address_state), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_state)

#! variable selected: product.manufacturer_state_32

coef(forward_selected_model_address_state)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_state)

#forward selection on the basis of source type

df2_address_source <- df2_address %>% dplyr::select(death_or_not, starts_with("source_type"))

df2_address_source <- sample_n(df2_address_source, size = 100000)

summary(df2_address_source)

df2_address_source$death_or_not <- -as.numeric(df2_address_source$death_or_not)

initial_model_address_source <- glm(death_or_not ~ 1, data = df2_address_source, family = binomial())

full_model_formula_address_source <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_source), "death_or_not"), collapse = "+"))) forward_selected_model_address_source <- stepAIC(initial_model_address_source, scope = list(lower = initial_model_address_source$formula, upper = full_model_formula_address_source), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_source) #! selected variables: #!source_type_17 #!source_type_3 #!source_type_11 #!source_type_5 #!source_type_4 #!source_type_15

coef(forward_selected_model_address_source)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_source)

#forward selection by operator

df2_address_operator <- df2_address %>% dplyr::select(death_or_not, starts_with("product.product_operator"))

df2_address_operator <- sample_n(df2_address_operator, size = 100000)

summary(df2_address_operator)

df2_address_operator$death_or_not <- -as.numeric(df2_address_operator$death_or_not)

initial_model_address_operator <- glm(death_or_not ~ 1, data = df2_address_operator, family = binomial())

full_model_formula_address_operator <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_operator), "death_or_not"), collapse = "+"))) forward_selected_model_address_operator <- stepAIC(initial_model_address_operator, scope = list(lower = initial_model_address_operator$formula, upper = full_model_formula_address_operator), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_operator)

#!nothing statistically significant

coef(forward_selected_model_address_operator)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_operator)

#forward selection on the basis of last year variables

df2_address_last_year <- df2_address %>% dplyr::select(death_or_not, starts_with("last_year"))

df2_address_last_year <- sample_n(df2_address_last_year, size = 100000)

summary(df2_address_last_year)

df2_address_last_year$death_or_not <- -as.numeric(df2_address_last_year$death_or_not)

initial_model_address_last_year <- glm(death_or_not ~ 1, data = df2_address_last_year, family = binomial())

full_model_formula_address_last_year <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_last_year), "death_or_not"), collapse = "+"))) forward_selected_model_address_last_year <- stepAIC(initial_model_address_last_year, scope = list(lower = initial_model_address_last_year$formula, upper = full_model_formula_address_last_year), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_last_year) #!variables selected: #!last_year_product_quantity_average_average
#!last_year_product_quantity_average_max
#!last_year_root_cause_description_num_uniq
#!last_year_legal_announcementing_firm_most_freq
#!last_year_classification1_num_uniq
#!last_year_brand_name_num_uniq
#!last_year_reason_for_legal_announcement_num_uniq
#!last_year_reason_for_legal_announcement_most_freq #!last_year_product_quantity_average_num_uniq

coef(forward_selected_model_address_last_year)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_last_year)

#forward selection on the basis of last two year variables

df2_address_last_two_years <- df2_address %>% dplyr::select(death_or_not, starts_with("last_two_years"))

df2_address_last_two_years <- sample_n(df2_address_last_two_years, size = 100000)

summary(df2_address_last_two_years)

df2_address_last_two_years$death_or_not <- as.numeric(df2_address_last_two_years$death_or_not)

initial_model_address_last_two_years <- glm(death_or_not ~ 1, data = df2_address_last_two_years, family = binomial())

full_model_formula_address_last_two_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_last_two_years), "death_or_not"), collapse = "+"))) forward_selected_model_address_last_two_years <- stepAIC(initial_model_address_last_two_years, scope = list(lower = initial_model_address_last_two_years$formula, upper = full_model_formula_address_last_two_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_last_two_years) #! selected variables: #!last_two_years_product_quantity_average_max #!last_two_years_root_cause_description_num_uniq #!last_two_years_classification1_num_uniq
#!last_two_years_product_quantity_average_num_uniq
#!last_two_years_root_cause_description_most_freq
#!last_two_years_product_quantity_average_average
#!last_two_years_classification0_num_uniq

coef(forward_selected_model_address_last_two_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_last_two_years)

#forward selection on the basis of last four years

df2_address_last_four_years <- df2_address %>% dplyr::select(death_or_not, starts_with("last_four_years"))

df2_address_last_four_years <- sample_n(df2_address_last_four_years, size = 100000)

summary(df2_address_last_four_years)

df2_address_last_four_years$death_or_not <- as.numeric(df2_address_last_four_years$death_or_not)

initial_model_address_last_four_years <- glm(death_or_not ~ 1, data = df2_address_last_four_years, family = binomial())

full_model_formula_address_last_four_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_last_four_years), "death_or_not"), collapse = "+"))) forward_selected_model_address_last_four_years <- stepAIC(initial_model_address_last_four_years, scope = list(lower = initial_model_address_last_four_years$formula, upper = full_model_formula_address_last_four_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_last_four_years)

#!variables selected: #!last_four_years_classification0_num_uniq #!last_four_years_legal_announcementing_firm_num_uniq
#!last_four_years_product_quantity_average_average

coef(forward_selected_model_address_last_four_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_last_four_years)

#forward selection on the basis of city

df2_address_city <- df2_address %>% dplyr::select(death_or_not, starts_with("product.manufacturer_city"))

df2_address_city <- sample_n(df2_address_city, size = 100000)

summary(df2_address_city)

df2_address_city$death_or_not <- as.numeric(df2_address_city$death_or_not)

initial_model_address_city <- glm(death_or_not ~ 1, data = df2_address_city, family = binomial())

full_model_formula_address_city <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_address_city), "death_or_not"), collapse = "+"))) forward_selected_model_address_city <- stepAIC(initial_model_address_city, scope = list(lower = initial_model_address_city$formula, upper = full_model_formula_address_city), direction = "forward", trace = FALSE)

summary(forward_selected_model_address_city) #! variables selected: #!product.manufacturer_city_2085
#!product.manufacturer_city_4513
#!product.manufacturer_city_3153

coef(forward_selected_model_address_city)

par(mfrow = c(2, 2)) plot(forward_selected_model_address_city)

Forward Selection for Product Code LKK

#subseting data for product.product_report_product_code_LKK, running forward selection with the same variables. Then followed by a regression and LDA from the selected results of forward selection.

df2_report <- df2 %>% filter(product.product_report_product_code_LKK == “1”)

#Forward Selection by state

df2_report_state <- df2_report %>% dplyr::select(death_or_not, starts_with(“product.manufacturer_state”))

df2_report_state <- sample_n(df2_report_state, size = 100000)

summary(df2_report_state)

df2_report_state$death_or_not <- as.numeric(df2_report_state$death_or_not)

initial_model_report_state <- glm(death_or_not ~ 1, data = df2_report_state, family = binomial())

full_model_formula_report_state <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_report_state), “death_or_not”), collapse = “+”))) forward_selected_model_report_state <- stepAIC(initial_model_report_state, scope = list(lower = initial_model_report_state$formula, upper = full_model_formula_report_state), direction = “forward”, trace = FALSE)

summary(forward_selected_model_report_state) # ! selected variables: # ! product.manufacturer_state_40 # ! product.manufacturer_state_48 # ! product.manufacturer_state_63

coef(forward_selected_model_report_state)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_state)

#Forward selection by source type

df2_report_source <- df2_report %>% dplyr::select(death_or_not, starts_with(“source_type”))

df2_report_source <- sample_n(df2_report_source, size = 100000)

summary(df2_report_source)

df2_report_source$death_or_not <- as.numeric(df2_report_source$death_or_not)

initial_model_report_source <- glm(death_or_not ~ 1, data = df2_report_source, family = binomial())

full_model_formula_report_source <- as.formula(paste(“death_or_not ~”, paste(setdiff(names(df2_report_source), “death_or_not”), collapse = “+”))) forward_selected_model_report_source <- stepAIC(initial_model_report_source, scope = list(lower = initial_model_report_source$formula, upper = full_model_formula_report_source), direction = “forward”, trace = FALSE)

summary(forward_selected_model_report_source)

! selected variables:

! source_type_17

! source_type_3

! source_type_11

! source_type_16

! source_type_4

! source_type_15

! source_type_18

coef(forward_selected_model_report_source)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_source)

#Forward selection by operator

df2_report_operator <- df2_report %>% dplyr::select(death_or_not, starts_with("product.product_operator"))

df2_report_operator <- sample_n(df2_report_operator, size = 100000)

summary(df2_report_operator)

df2_report_operator$death_or_not <- -as.numeric(df2_report_operator$death_or_not)

initial_model_report_operator <- glm(death_or_not ~ 1, data = df2_report_operator, family = binomial())

full_model_formula_report_operator <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_report_operator), "death_or_not"), collapse = "+"))) forward_selected_model_report_operator <- stepAIC(initial_model_report_operator, scope = list(lower = initial_model_report_operator$formula, upper = full_model_formula_report_operator), direction = "forward", trace = FALSE)

summary(forward_selected_model_report_operator) # ! nothing statistically significant

coef(forward_selected_model_report_operator)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_operator)

#forward selection on the basis of last year variables

df2_report_last_year <- df2_report %>% dplyr::select(death_or_not, starts_with("last_year"))

df2_report_last_year <- sample_n(df2_report_last_year, size = 100000)

summary(df2_report_last_year)

df2_report_last_year$death_or_not <- -as.numeric(df2_report_last_year$death_or_not)

initial_model_report_last_year <- glm(death_or_not ~ 1, data = df2_report_last_year, family = binomial())

full_model_formula_report_last_year <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_report_last_year), "death_or_not"), collapse = "+"))) forward_selected_model_report_last_year <- stepAIC(initial_model_report_last_year, scope = list(lower = initial_model_report_last_year$formula, upper = full_model_formula_report_last_year), direction = "forward", trace = FALSE)

summary(forward_selected_model_report_last_year)

#! selected variables: #!last_year_decision_date_max_changes_in_product
#! last_year_all_product_codes_most_freq
#! last_year_classification1_num_uniq
#! last_year_reason_for_legal_announcement_most_freq #! last_year_reason_for_legal_announcement_num_uniq
#! last_year_product_quantity_average_average
#! last_year_product_quantity_average_max
#! last_year_legal_announcementing_firm_most_freq #! last_year_company_name_most_freq

coef(forward_selected_model_report_last_year)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_last_year)

#Forward selection on the basis of last two years variables

df2_report_last_two_years <- df2_report %>% dplyr::select(death_or_not, starts_with("last_two_years"))

df2_report_last_two_years <- sample_n(df2_report_last_two_years, size = 100000)

summary(df2_report_last_two_years)

df2_report_last_two_years$death_or_not <- -as.numeric(df2_report_last_two_years$death_or_not)

initial_model_report_last_two_years <- glm(death_or_not ~ 1, data = df2_report_last_two_years, family = binomial())

full_model_formula_report_last_two_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_report_last_two_years), "death_or_not"), collapse = "+"))) forward_selected_model_report_last_two_years <- stepAIC(initial_model_report_last_two_years, scope = list(lower = initial_model_report_last_two_years$formula, upper = full_model_formula_report_last_two_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_report_last_two_years) #variables selected: #!
last_two_years_decision_date_max_changes_in_product
#! last_two_years_reason_for_legal_announcement_num_uniq #! last_two_years_classification1_num_uniq
#! last_two_years_legal_announcementing_firm_most_freq
#! last_two_years_all_product_codes_num_uniq

coef(forward_selected_model_report_last_two_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_last_two_years)

#Forward selection on the basis of last four years variables

df2_report_last_four_years <- df2_report %>% dplyr::select(death_or_not, starts_with("last_four_years"))

df2_report_last_four_years <- sample_n(df2_report_last_four_years, size = 100000)

summary(df2_report_last_four_years)

df2_report_last_four_years$death_or_not <- -as.numeric(df2_report_last_four_years$death_or_not)

initial_model_report_last_four_years <- glm(death_or_not ~ 1, data = df2_report_last_four_years, family = binomial())

full_model_formula_report_last_four_years <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_report_last_four_years), "death_or_not"), collapse = "+"))) forward_selected_model_report_last_four_years <- stepAIC(initial_model_report_last_four_years, scope = list(lower = initial_model_report_last_four_years$formula, upper = full_model_formula_report_last_four_years), direction = "forward", trace = FALSE)

summary(forward_selected_model_report_last_four_years) #! selected variables: #!
last_four_years_decision_date_max_changes_in_product #! last_four_years_classification1_num_uniq
#! last_four_years_legal_announcementing_firm_num_uniq

coef(forward_selected_model_report_last_four_years)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_last_four_years)

#Forward selection on the basis of city

df2_report_city <- df2_report %>% dplyr::select(death_or_not, starts_with("product.manufacturer_city"))

df2_report_city <- sample_n(df2_report_city, size = 100000)

summary(df2_report_city)

df2_report_city$death_or_not <- -as.numeric(df2_report_city$death_or_not)

initial_model_report_city <- glm(death_or_not ~ 1, data = df2_report_city, family = binomial())

full_model_formula_report_city <- as.formula(paste("death_or_not ~", paste(setdiff(names(df2_report_city), "death_or_not"), collapse = "+"))) forward_selected_model_report_city <- stepAIC(initial_model_report_city, scope = list(lower = initial_model_report_city$formula, upper = full_model_formula_report_city), direction = "forward", trace = FALSE)

summary(forward_selected_model_report_city) # ! selected variables: # ! product.manufacturer_city_6271
! product.manufacturer_city_2085
! product.manufacturer_city_5092
! product.manufacturer_city_4513

coef(forward_selected_model_report_city)

par(mfrow = c(2, 2)) plot(forward_selected_model_report_city)

Regression for Product Field Unknown

Several variables are significant predictors (indicated by stars next to the Pr(>|z|) values). For instance, product.manufacturer_city_2085, last_four_years_classification1_num_uniq, and source_type_17 show high significance levels with p-values less than 0.001, suggesting a strong association with the outcome variable.

The range of residuals indicates there are outliers or points with high leverage.

The model needed 18 Fisher Scoring iterations to converge, which is within normal limits but on the higher side, indicating a possibly complex model fit.

The model has a high accuracy of 99.34%, but this measure alone can be misleading, especially if the data is imbalanced (which seems to be the case since the prevalence is very high at 99.338%).

The Kappa of 0.0224 is very low, indicating that the model is not adding much predictive power beyond what would be expected by chance.

It has near perfect sensitivity at 99.992%, meaning the model is excellent at predicting the non-event.

True Negative is extremely low at 1.163%, indicating the model is almost always predicting the majority class and is poor at predicting the actual events.

For Mcnemar's Test, The p-value is significant, suggesting a difference in the type of errors made by the model (false positives versus false negatives).

The Area Under the ROC Curve (AUC) is 0.7117, which is fair but not excellent. The AUC measures the model's ability to discriminate between the positive and negative classes.

While the logistic regression model seems to have a high accuracy and fair AUC value, it is crucial to consider the extremely imbalanced nature of the outcome variable. The model's sensitivity is high, but the specificity is very low, meaning it fails to identify the positive cases reliably. The significant predictors in the model do provide some discrimination power, but the overall performance in the context of the actual positive events is not strong. In practice, this model would predict most outcomes as the majority class, missing the critical events you're trying to predict. This highlights the importance of looking beyond accuracy in imbalanced datasets and considering other metrics such as AUC, sensitivity, specificity, and predictive values.

```
# Performing logistic regression for unknown field

df2$death_or_not <- as.numeric(df2$death_or_not)

df_logistic_lda <- df2 %>%
  dplyr::select(
    `product.manufacturer_city_6271`,
    `product.manufacturer_city_2085`,
    `product.manufacturer_city_4513`,
    `product.manufacturer_city_5092`,
    `product.manufacturer_city_3153`,
    `last_four_years_decision_date_max_changes_in_product`,
    `last_four_years_classification1_num_uniq`,
    `last_four_years_legal_announcementing_firm_num_uniq`,
    `last_four_years_classification2_num_uniq`,
    `last_four_years_product_quantity_average_average`,
    `last_four_years_company_name_most_freq`,
    `last_two_years_decision_date_max_changes_in_product`,
    `last_two_years_reason_for_legal_announcement_num_uniq`,
    `last_two_years_root_cause_description_most_freq`,
    `last_two_years_classification1_num_uniq`,
    `last_two_years_product_quantity_average_max`,
    `last_two_years_product_quantity_average_average`,
    `last_two_years_product_quantity_average_num_uniq`,
    `last_two_years_legal_announcementing_firm_num_uniq`,
    `last_two_years_root_cause_description_num_uniq`,
    `last_two_years_company_name_most_freq`,
    `last_year_decision_date_max_changes_in_product`,
    `last_year_company_name_num_uniq`,
    `last_year_classification1_num_uniq`,
    `last_year_reason_for_legal_announcement_most_freq`,
    `last_year_product_quantity_average_average`,
    `last_year_product_quantity_average_max`,
    `last_year_reason_for_legal_announcement_num_uniq`,
    `last_year_legal_announcementing_firm_most_freq`,
    `last_year_product_quantity_average_num_uniq`,
    `last_year_classification2_num_uniq`,
    `last_year_brand_name_num_uniq`,
    `last_year_brand_name_most_freq`,
    `last_year_company_name_most_freq`,
    `product.product_operator_41`,
    `product.product_operator_18`,
    `source_type_3`,
    `source_type_17`,
    `source_type_11`,
    `source_type_16`,
    `source_type_5`,
    `source_type_4`,
    `source_type_15`,
    `product.manufacturer_state_40`,
    `product.manufacturer_state_32`,
    `death_or_not`
  )



df_logistic_lda <- sample_n(df_logistic_lda, size = 65000)



set.seed(123)
index <- createDataPartition(df_logistic_lda$death_or_not, p = 0.80, list = FALSE)
```

```
trainData <- df_logistic_lda[index, ]
testData <- df_logistic_lda[-index, ]

model_logistic <- glm(death_or_not ~ ., data = trainData, family = binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_logistic)
```

```
##
## Call:
## glm(formula = death_or_not ~ ., family = binomial(), data = trainData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2032  -0.1315  -0.0961  -0.0646   3.9847
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       -4.131e+01  2.429e+03
## product.manufacturer_city_6271                     1.755e+01  1.865e+03
## product.manufacturer_city_2085                     4.580e+00  1.545e+00
## product.manufacturer_city_4513                     1.757e+01  6.312e+02
## product.manufacturer_city_5092                    -2.617e+00  2.669e+03
## product.manufacturer_city_3153                     1.264e+00  4.893e-01
## last_four_years_decision_date_max_changes_in_product -7.694e-03  2.410e-02
## last_four_years_classification1_num_uniq          -6.407e-03  4.524e-03
## last_four_years_legal_announcementing_firm_num_uniq -6.973e-01  5.370e-01
## last_four_years_classification2_num_uniq          -1.977e-01  2.227e-01
## last_four_years_product_quantity_average_average   1.579e+00  1.366e+00
## last_four_years_company_name_most_freq             2.794e+01  3.626e+03
## last_two_years_decision_date_max_changes_in_product -2.979e+00  1.781e+00
## last_two_years_reason_for_legal_announcement_num_uniq -4.066e-01  1.771e-01
## last_two_years_root_cause_description_most_freq    2.510e+00  1.295e+00
## last_two_years_classification1_num_uniq            1.000e-02  7.379e-03
## last_two_years_product_quantity_average_max       -3.026e+00  3.581e+00
## last_two_years_product_quantity_average_average    8.458e+00  6.787e+00
## last_two_years_product_quantity_average_num_uniq   8.069e-01  2.262e-01
## last_two_years_legal_announcementing_firm_num_uniq 9.849e-01  5.126e-01
## last_two_years_root_cause_description_num_uniq     2.709e-01  2.143e-01
## last_two_years_company_name_most_freq             -1.762e+01  3.487e+03
## last_year_decision_date_max_changes_in_product     1.076e-01  1.281e+00
## last_year_company_name_num_uniq                    1.259e+01  1.154e+03
## last_year_classification1_num_uniq                -2.100e-02  9.909e-03
## last_year_reason_for_legal_announcement_most_freq -5.748e-04  3.245e-04
## last_year_product_quantity_average_average        -3.836e+00  7.070e+00
## last_year_product_quantity_average_max             1.737e+00  4.033e+00
## last_year_reason_for_legal_announcement_num_uniq  -1.293e-01  2.003e-01
## last_year_legal_announcementing_firm_most_freq    -3.128e+00  9.495e-01
## last_year_product_quantity_average_num_uniq       -5.837e-02  2.026e-01
## last_year_classification2_num_uniq                 2.425e-01  2.230e-01
## last_year_brand_name_num_uniq                      1.219e-01  8.139e-02
## last_year_brand_name_most_freq                    -1.112e-02  1.577e-01
## last_year_company_name_most_freq                   1.176e+00  1.896e+03
## product.product_operator_41                       -6.335e-01  5.149e-01
## product.product_operator_18                       -1.257e+01  1.857e+03
## source_type_3                                     -2.392e-02  2.300e-01
## source_type_17                                     2.013e+00  2.512e-01
## source_type_11                                    -5.025e-01  2.726e-01
## source_type_16                                     2.872e+00  5.625e-01
## source_type_5                                     -9.963e-01  4.577e-01
## source_type_4                                      3.237e-01  2.200e-01
## source_type_15                                     1.643e+00  4.633e-01
## product.manufacturer_state_40                      4.695e+00  2.094e+03
## product.manufacturer_state_32                      1.752e+01  6.312e+02
##                                                   z value Pr(>|z|)
## (Intercept)                                        -0.017 0.986430
## product.manufacturer_city_6271                      0.009 0.992494
## product.manufacturer_city_2085                      2.965 0.003029 **
## product.manufacturer_city_4513                      0.028 0.977789
## product.manufacturer_city_5092                     -0.001 0.999218
```

```
## product.manufacturer_city_3153                              2.584 0.009779 **
## last_four_years_decision_date_max_changes_in_product    -0.319 0.749567
## last_four_years_classification1_num_uniq                -1.416 0.156689
## last_four_years_legal_announcementing_firm_num_uniq     -1.299 0.194115
## last_four_years_classification2_num_uniq                -0.888 0.374658
## last_four_years_product_quantity_average_average         1.156 0.247872
## last_four_years_company_name_most_freq                   0.008 0.993851
## last_two_years_decision_date_max_changes_in_product     -1.673 0.094340 .
## last_two_years_reason_for_legal_announcement_num_uniq   -2.295 0.021710 *
## last_two_years_root_cause_description_most_freq          1.938 0.052633 .
## last_two_years_classification1_num_uniq                  1.356 0.175192
## last_two_years_product_quantity_average_max             -0.845 0.398088
## last_two_years_product_quantity_average_average          1.246 0.212659
## last_two_years_product_quantity_average_num_uniq         3.567 0.000361 ***
## last_two_years_legal_announcementing_firm_num_uniq       1.922 0.054660 .
## last_two_years_root_cause_description_num_uniq           1.264 0.206178
## last_two_years_company_name_most_freq                   -0.005 0.995968
## last_year_decision_date_max_changes_in_product           0.084 0.933050
## last_year_company_name_num_uniq                          0.011 0.991293
## last_year_classification1_num_uniq                      -2.120 0.034040 *
## last_year_reason_for_legal_announcement_most_freq       -1.771 0.076538 .
## last_year_product_quantity_average_average              -0.543 0.587446
## last_year_product_quantity_average_max                   0.431 0.666786
## last_year_reason_for_legal_announcement_num_uniq        -0.646 0.518467
## last_year_legal_announcementing_firm_most_freq          -3.294 0.000987 ***
## last_year_product_quantity_average_num_uniq             -0.288 0.773220
## last_year_classification2_num_uniq                       1.087 0.276896
## last_year_brand_name_num_uniq                            1.498 0.134047
## last_year_brand_name_most_freq                          -0.071 0.943783
## last_year_company_name_most_freq                         0.001 0.999505
## product.product_operator_41                             -1.230 0.218609
## product.product_operator_18                             -0.007 0.994602
## source_type_3                                           -0.104 0.917174
## source_type_17                                           8.015 1.10e-15 ***
## source_type_11                                          -1.843 0.065284 .
## source_type_16                                           5.106 3.28e-07 ***
## source_type_5                                           -2.177 0.029510 *
## source_type_4                                            1.472 0.141116
## source_type_15                                           3.546 0.000391 ***
## product.manufacturer_state_40                            0.002 0.998211
## product.manufacturer_state_32                            0.028 0.977854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4268.2  on 51999  degrees of freedom
## Residual deviance: 3886.0  on 51954  degrees of freedom
## AIC: 3978
##
## Number of Fisher Scoring iterations: 17
```

```
probabilities_logistic <- predict(model_logistic, newdata = testData, type = "response")
predictedClasses_logistic <- ifelse(probabilities_logistic > 0.5, 1, 0)

confusionMatrix(data = as.factor(predictedClasses_logistic), reference = as.factor(testData$death_or_not))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12922    76
##          1     0     2
##
##                Accuracy : 0.9942
##                  95% CI : (0.9927, 0.9954)
##     No Information Rate : 0.994
##     P-Value [Acc > NIR] : 0.4396
##
##                   Kappa : 0.0497
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.00000
##             Specificity : 0.02564
##          Pos Pred Value : 0.99415
##          Neg Pred Value : 1.00000
##              Prevalence : 0.99400
##          Detection Rate : 0.99400
##    Detection Prevalence : 0.99985
##       Balanced Accuracy : 0.51282
##
##        'Positive' Class : 0
##
```

```
rocResult_logistic <- roc(response = testData$death_or_not, predictor = probabilities_logistic)
```

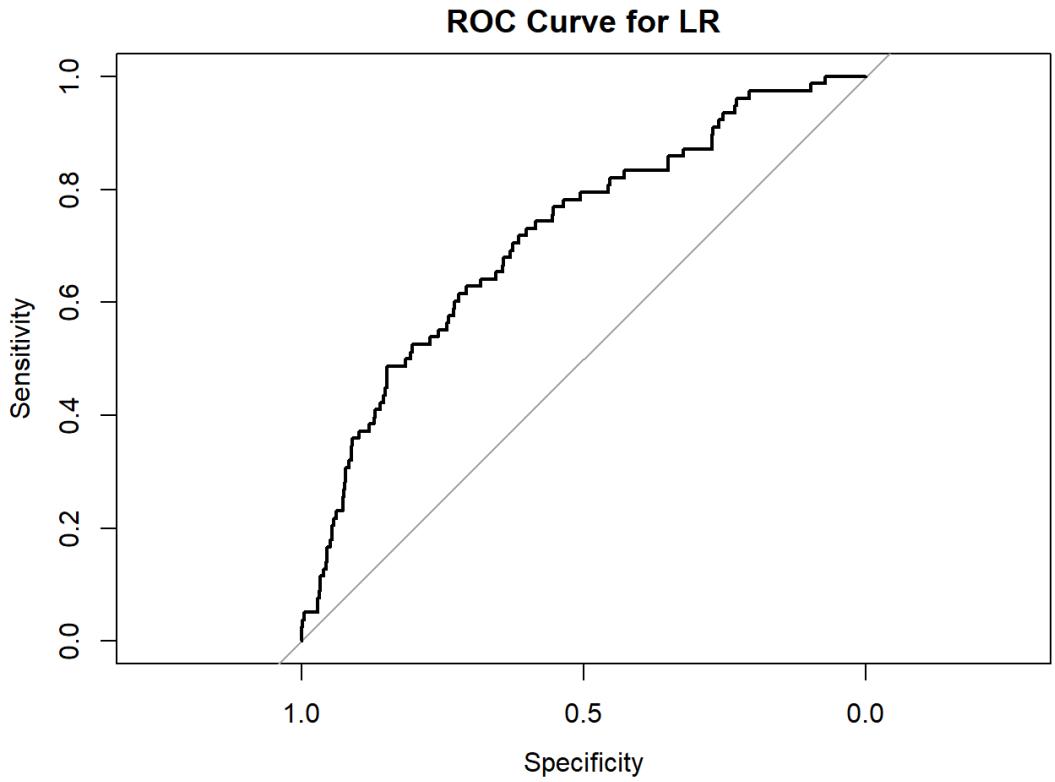
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(rocResult_logistic)
```

```
## Area under the curve: 0.7167
```

```
plot(rocResult_logistic, main = "ROC Curve for LR")
```



Regression for Product Report Code LKK

The AUC for the second model is higher at 0.7447, compared to 0.7117 for the first model. This indicates an improved ability to distinguish between the two classes. Both models have very high accuracy (0.9934 for the first model and 0.9928 for the second), but given the prevalence rate, this could be largely due to the models predicting the majority class well.

Sensitivity remains nearly perfect in both models, but this is likely due to class imbalance.

The Positive Predictive Value (PPV) decreased slightly from the first to the second model, while the Negative Predictive Value (NPV) showed a significant decrease, suggesting the second model may not be as effective in predicting true negatives when the threshold is set at 0.5. The Balanced Accuracy is about the same for both models, indicating that neither model is particularly good when considering both sensitivity and specificity.

Some variables became significant in the second model while they were not in the first, and vice versa. This could suggest differences in how the models are capturing the relationships in the data.

The Residual Deviance is slightly higher in the second model, which might indicate a slightly poorer fit to the data. The AIC is also higher in the second model, suggesting it might not perform as well in terms of the trade-off between the goodness of fit and model complexity.

Although the second model shows a higher AUC, other performance measures like specificity and PPV/NPV suggest that there may not be a practical improvement in model performance, especially in identifying the positive class. Both models exhibit signs of overfitting to the majority class due to the imbalanced dataset. The higher AUC in the second model indicates some improvement, but the real-world applicability of this improvement would depend on the specific context and costs associated with false positives and false negatives. The low specificity and kappa values in both models highlight the need for further model tuning or data sampling strategies to handle the class imbalance before these models could be reliably used for prediction.

```
# Regression

df_logistic_lda_2 <- df2 %>%
  dplyr::select(
    `source_type_17`
    ,`source_type_3`
    ,`source_type_11`
    ,`source_type_16`
    ,`source_type_4`
    ,`source_type_15`
    ,`source_type_18`

    ,`product.manufacturer_state_40`
    ,`product.manufacturer_state_48`
    ,`product.manufacturer_state_63`

    ,`last_year_decision_date_max_changes_in_product`
    ,`last_year_all_product_codes_most_freq`
    ,`last_year_classification1_num_uniq`
    ,`last_year_reason_for_legal_announcement_most_freq`
    ,`last_year_reason_for_legal_announcement_num_uniq`
    ,`last_year_product_quantity_average_average`
    ,`last_year_product_quantity_average_max`
    ,`last_year_legal_announcementing_firm_most_freq`
    ,`last_year_company_name_most_freq`

    ,`last_two_years_decision_date_max_changes_in_product`
    ,`last_two_years_reason_for_legal_announcement_num_uniq`
    ,`last_two_years_classification1_num_uniq`
    ,`last_two_years_legal_announcementing_firm_most_freq`
    ,`last_two_years_all_product_codes_num_uniq`

    ,`last_four_years_decision_date_max_changes_in_product`
    ,`last_four_years_classification1_num_uniq`
    ,`last_four_years_legal_announcementing_firm_num_uniq`

    ,`product.manufacturer_city_6271`
    ,`product.manufacturer_city_2085`
    ,`product.manufacturer_city_5092`
    ,`product.manufacturer_city_4513`
    ,`death_or_not`
  )


df_logistic_lda_2 <- sample_n(df_logistic_lda_2, size = 65000)

index <- createDataPartition(df_logistic_lda_2$death_or_not, p = 0.80, list = FALSE)
trainData <- df_logistic_lda_2[index, ]
testData <- df_logistic_lda_2[-index, ]

model_logistic_2 <- glm(death_or_not ~ ., data = trainData, family = binomial())
summary(model_logistic_2)
```

```
##
## Call:
## glm(formula = death_or_not ~ ., family = binomial(), data = trainData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7752  -0.1227  -0.0994  -0.0779   3.7747
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                     -6.851e+00  2.705e+00
## source_type_17                                   1.436e+00  2.958e-01
## source_type_3                                   -1.079e-01  2.196e-01
## source_type_11                                  -3.668e-01  2.535e-01
## source_type_16                                   2.503e+00  5.959e-01
## source_type_4                                    2.806e-01  2.118e-01
## source_type_15                                   1.654e+00  4.607e-01
## source_type_18                                  -1.442e+01  9.750e+02
## product.manufacturer_state_40                   -1.329e+01  1.171e+03
## product.manufacturer_state_48                   -1.090e+01  3.547e+02
## product.manufacturer_state_63                   -1.961e+01  5.169e+02
## last_year_decision_date_max_changes_in_product  -5.327e-01  1.003e+00
## last_year_all_product_codes_most_freq           -1.353e+00  5.349e+00
## last_year_classification1_num_uniq              -2.013e-02  1.028e-02
## last_year_reason_for_legal_announcement_most_freq -2.290e-04  2.196e-04
## last_year_reason_for_legal_announcement_num_uniq  3.561e-01  1.217e-01
## last_year_product_quantity_average_average        4.218e+00  1.479e+00
## last_year_product_quantity_average_max           -3.409e+00  1.112e+00
## last_year_legal_announcementing_firm_most_freq   -9.941e-01  6.342e-01
## last_year_company_name_most_freq                  4.894e+00  4.137e+00
## last_two_years_decision_date_max_changes_in_product  9.946e-02  1.562e+00
## last_two_years_reason_for_legal_announcement_num_uniq -6.363e-02  6.889e-02
## last_two_years_classification1_num_uniq           8.579e-03  6.320e-03
## last_two_years_legal_announcementing_firm_most_freq  4.147e-01  9.290e-01
## last_two_years_all_product_codes_num_uniq         6.803e-02  9.976e-02
## last_four_years_decision_date_max_changes_in_product -9.111e-03  2.224e-02
## last_four_years_classification1_num_uniq         -3.667e-03  3.259e-03
## last_four_years_legal_announcementing_firm_num_uniq -2.320e-01  2.357e-01
## product.manufacturer_city_6271                    1.610e+01  1.171e+03
## product.manufacturer_city_2085                    4.529e+00  1.098e+00
## product.manufacturer_city_5092                    2.621e+01  5.169e+02
## product.manufacturer_city_4513                    1.096e+01  3.547e+02
##                                                 z value Pr(>|z|)
## (Intercept)                                      -2.532 0.011330 *
## source_type_17                                    4.857 1.19e-06 ***
## source_type_3                                    -0.491 0.623189
## source_type_11                                   -1.447 0.147789
## source_type_16                                    4.200 2.67e-05 ***
## source_type_4                                     1.325 0.185291
## source_type_15                                    3.590 0.000331 ***
## source_type_18                                   -0.015 0.988203
## product.manufacturer_state_40                    -0.011 0.990942
## product.manufacturer_state_48                    -0.031 0.975482
## product.manufacturer_state_63                    -0.038 0.969740
## last_year_decision_date_max_changes_in_product   -0.531 0.595406
## last_year_all_product_codes_most_freq            -0.253 0.800264
## last_year_classification1_num_uniq               -1.958 0.050266 .
## last_year_reason_for_legal_announcement_most_freq -1.043 0.296897
## last_year_reason_for_legal_announcement_num_uniq   2.926 0.003437 **
## last_year_product_quantity_average_average         2.852 0.004342 **
## last_year_product_quantity_average_max           -3.065 0.002173 **
## last_year_legal_announcementing_firm_most_freq   -1.567 0.117025
```

```
## last_year_company_name_most_freq                          1.183 0.236739
## last_two_years_decision_date_max_changes_in_product     0.064 0.949214
## last_two_years_reason_for_legal_announcement_num_uniq  -0.924 0.355655
## last_two_years_classification1_num_uniq                 1.358 0.174598
## last_two_years_legal_announcementing_firm_most_freq     0.446 0.655302
## last_two_years_all_product_codes_num_uniq               0.682 0.495275
## last_four_years_decision_date_max_changes_in_product   -0.410 0.682079
## last_four_years_classification1_num_uniq               -1.125 0.260571
## last_four_years_legal_announcementing_firm_num_uniq    -0.984 0.324920
## product.manufacturer_city_6271                          0.014 0.989027
## product.manufacturer_city_2085                          4.124 3.73e-05 ***
## product.manufacturer_city_5092                          0.051 0.959561
## product.manufacturer_city_4513                          0.031 0.975345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4138.3  on 51999  degrees of freedom
## Residual deviance: 3814.6  on 51968  degrees of freedom
## AIC: 3878.6
##
## Number of Fisher Scoring iterations: 16
```

```
probabilities_logistic_2 <- predict(model_logistic_2, newdata = testData, type = "response")
predictedClasses_logistic_2 <- ifelse(probabilities_logistic_2 > 0.5, 1, 0)

confusionMatrix(data = as.factor(predictedClasses_logistic_2), reference = as.factor(testData$death_or_not))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12900    96
##          1     2     2
##
##                Accuracy : 0.9925
##                  95% CI : (0.9908, 0.9939)
##     No Information Rate : 0.9925
##     P-Value [Acc > NIR] : 0.5268
##
##                   Kappa : 0.0386
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.99984
##             Specificity : 0.02041
##          Pos Pred Value : 0.99261
##          Neg Pred Value : 0.50000
##              Prevalence : 0.99246
##          Detection Rate : 0.99231
##    Detection Prevalence : 0.99969
##       Balanced Accuracy : 0.51013
##
##        'Positive' Class : 0
##
```

```
rocResult_logistic_2 <- roc(response = testData$death_or_not, predictor = probabilities_logistic_2)
```

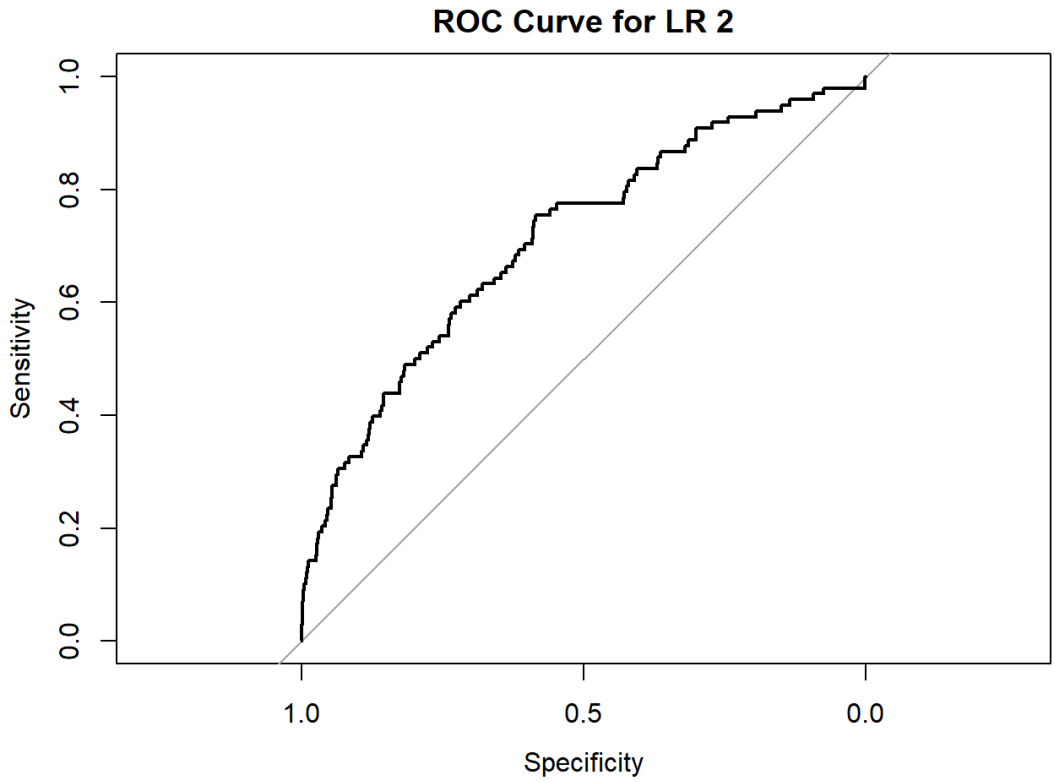
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(rocResult_logistic_2)
```

```
## Area under the curve: 0.7104
```

```
plot(rocResult_logistic_2, main = "ROC Curve for LR 2")
```



Regression For Product Issue Type 599

The AUC is 0.7028, which is lower than the second model (0.7447) and slightly lower than the first model (0.7117). This suggests that LR 3 has a slightly poorer ability to discriminate between the positive and negative classes compared to the second model but is roughly similar to the first.

The accuracy of all three models is very high and similar across the board (ranging from 0.9928 to 0.9934). However, this is likely influenced by the class imbalance and the models' tendency to predict the majority class.

Sensitivity is perfect (1.0000) in the third model, but specificity is 0.0000, indicating that the model predicted all test cases as the negative class and is therefore unable to correctly identify any true positive cases.

The Positive Predictive Value (PPV) is consistent with the accuracy due to the prevalence of the negative class being so high. However, the Negative Predictive Value (NPV) is not calculable (NaN) because there were no true positive predictions.

The Balanced Accuracy is 0.5000, the lowest of all three models, which reflects the model's inability to balance the sensitivity and specificity; it essentially performs as well as random guessing.

The third model shows a different set of significant predictors compared to the previous models. This could be due to changes in the dataset, modeling process, or simply differences in how each model captures the data relationships. The variable last_four_years_decision_date_average_changes_in_product is highly significant in LR 3 ($p < 0.0001$) and appears to be a strong predictor.

Based on the AUC values, LR 3 would have a ROC curve that does not perform as well as the second model but is roughly similar to the first in discriminating between the positive and negative classes.

The Residual Deviance for LR 3 is higher than both previous models, which might indicate a poorer fit to the data.

Comparing all three models, the second model appears to be the strongest in terms of AUC, indicating better discriminative power. However, all models struggle with specificity, as they fail to correctly identify positive cases. This is a common issue in datasets with significant class imbalance. In practice, the models are primarily predicting the majority class. Measures like AUC and Kappa, along with confusion matrix statistics, are crucial in these scenarios to understand the models' true performance beyond just accuracy.

```
# Regression

df_logistic_lda_3 <- df2 %>%
  dplyr::select(
    `last_four_years_product_quantity_average_average`
    ,`last_four_years_classification2_num_uniq`
    ,`last_four_years_legal_announcementing_firm_num_uniq`
    ,`last_four_years_decision_date_average_changes_in_product`
    ,`last_four_years_product_quantity_average_num_uniq`
    ,`last_four_years_reason_for_legal_announcement_most_freq`

    ,`last_two_years_reason_for_legal_announcement_num_uniq`
    ,`last_two_years_classification0_num_uniq`
    ,`last_two_years_product_quantity_average_average`
    ,`last_two_years_legal_announcementing_firm_num_uniq`
    ,`last_two_years_product_quantity_average_max`
    ,`last_two_years_reason_for_legal_announcement_most_freq`
    ,`last_two_years_root_cause_description_num_uniq`

    ,`last_year_root_cause_description_most_freq`
    ,`last_year_classification0_num_uniq`
    ,`last_year_product_quantity_average_num_uniq`
    ,`last_year_decision_date_max_changes_in_product`
    ,`last_year_reason_for_legal_announcement_num_uniq`
    ,`last_year_reason_for_legal_announcement_most_freq`
    ,`last_year_classification2_num_uniq`
    ,`last_year_brand_name_num_uniq`
    ,`last_year_brand_name_most_freq`
    ,`death_or_not`
  )


df_logistic_lda_3 <- sample_n(df_logistic_lda_3, size = 65000)

set.seed(123)  # for reproducibility
index <- createDataPartition(df_logistic_lda_3$death_or_not, p = 0.80, list = FALSE)
trainData <- df_logistic_lda_3[index, ]
testData <- df_logistic_lda_3[-index, ]

model_logistic_3 <- glm(death_or_not ~ ., data = trainData, family = binomial())
summary(model_logistic_3)
```

```
##
## Call:
## glm(formula = death_or_not ~ ., family = binomial(), data = trainData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1969  -0.1255  -0.1030  -0.0766   3.7014
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -4.1209259  1.1307664
## last_four_years_product_quantity_average_average -2.1447921  1.2882219
## last_four_years_classification2_num_uniq         -0.0611830  0.0858222
## last_four_years_legal_announcementing_firm_num_uniq -1.5526808  0.5870118
## last_four_years_decision_date_average_changes_in_product -5.7626858  1.0099834
## last_four_years_product_quantity_average_num_uniq  0.1915129  0.1248532
## last_four_years_reason_for_legal_announcement_most_freq -0.0004279  0.0006986
## last_two_years_reason_for_legal_announcement_num_uniq -0.2289089  0.1613047
## last_two_years_classification0_num_uniq          -7.8510354  6.3746563
## last_two_years_product_quantity_average_average   8.8728974  1.9151622
## last_two_years_legal_announcementing_firm_num_uniq  1.7296019  0.4570619
## last_two_years_product_quantity_average_max      -3.8239362  1.0862541
## last_two_years_reason_for_legal_announcement_most_freq  0.0003412  0.0004032
## last_two_years_root_cause_description_num_uniq    0.5456153  0.1785436
## last_year_root_cause_description_most_freq        0.0873141  0.3586890
## last_year_classification0_num_uniq                1.5982846  1.6249511
## last_year_product_quantity_average_num_uniq       0.0887127  0.1797758
## last_year_decision_date_max_changes_in_product    1.1783900  1.0380540
## last_year_reason_for_legal_announcement_num_uniq -0.1999008  0.1613553
## last_year_reason_for_legal_announcement_most_freq  0.0007604  0.0003154
## last_year_classification2_num_uniq                0.0566859  0.0871654
## last_year_brand_name_num_uniq                     0.0611208  0.0801895
## last_year_brand_name_most_freq                   -0.1901397  0.1549880
##                                                  z value Pr(>|z|)
## (Intercept)                                       -3.644 0.000268 ***
## last_four_years_product_quantity_average_average  -1.665 0.095928 .
## last_four_years_classification2_num_uniq          -0.713 0.475905
## last_four_years_legal_announcementing_firm_num_uniq  -2.645 0.008168 **
## last_four_years_decision_date_average_changes_in_product  -5.706 1.16e-08 ***
## last_four_years_product_quantity_average_num_uniq  1.534 0.125053
## last_four_years_reason_for_legal_announcement_most_freq  -0.613 0.540157
## last_two_years_reason_for_legal_announcement_num_uniq  -1.419 0.155867
## last_two_years_classification0_num_uniq           -1.232 0.218098
## last_two_years_product_quantity_average_average    4.633 3.60e-06 ***
## last_two_years_legal_announcementing_firm_num_uniq  3.784 0.000154 ***
## last_two_years_product_quantity_average_max       -3.520 0.000431 ***
## last_two_years_reason_for_legal_announcement_most_freq  0.846 0.397425
## last_two_years_root_cause_description_num_uniq    3.056 0.002244 **
## last_year_root_cause_description_most_freq        0.243 0.807676
## last_year_classification0_num_uniq                0.984 0.325317
## last_year_product_quantity_average_num_uniq       0.493 0.621686
## last_year_decision_date_max_changes_in_product    1.135 0.256295
## last_year_reason_for_legal_announcement_num_uniq -1.239 0.215388
## last_year_reason_for_legal_announcement_most_freq  2.411 0.015911 *
## last_year_classification2_num_uniq                0.650 0.515482
## last_year_brand_name_num_uniq                     0.762 0.445938
## last_year_brand_name_most_freq                   -1.227 0.219897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 4198.4  on 51999  degrees of freedom
## Residual deviance: 3932.0  on 51977  degrees of freedom
## AIC: 3978
##
## Number of Fisher Scoring iterations: 10
```

```
probabilities_logistic_3 <- predict(model_logistic_3, newdata = testData, type = "response")
predictedClasses_logistic_3 <- ifelse(probabilities_logistic_3 > 0.5, 1, 0)

confusionMatrix(data = as.factor(predictedClasses_logistic_3), reference = as.factor(testData$death_or_not))
```

```
## Warning in confusionMatrix.default(data =
## as.factor(predictedClasses_logistic_3), : Levels are not in the same order for
## reference and data. Refactoring data to match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12914    86
##          1     0     0
##
##                Accuracy : 0.9934
##                  95% CI : (0.9918, 0.9947)
##     No Information Rate : 0.9934
##     P-Value [Acc > NIR] : 0.5286
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9934
##          Neg Pred Value :    NaN
##              Prevalence : 0.9934
##          Detection Rate : 0.9934
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

```
rocResult_logistic_3 <- roc(response = testData$death_or_not, predictor = probabilities_logistic_3)
```

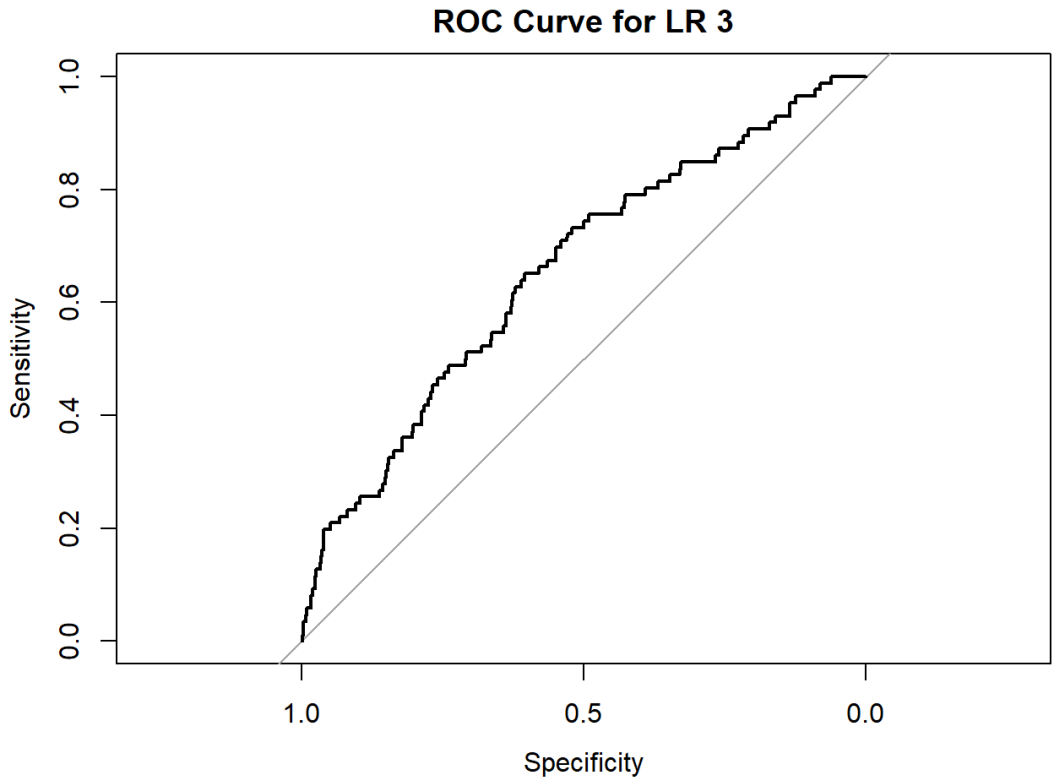
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(rocResult_logistic_3)
```

```
## Area under the curve: 0.6559
```

```
plot(rocResult_logistic_3, main = "ROC Curve for LR 3")
```



Regression for Manufacturer Address 9476

Assuming the ROC curve shown reflects the model's performance, it seems similar to the AUCs of the previous models, which are in the range of 0.7028 to 0.7447. The shape of the ROC curve in the visual provided suggests that LR 4 likely has an AUC that doesn't deviate much from the AUCs of the previous models.

The accuracy (0.993) is very slightly higher than that of the previous models, but given the very high prevalence of the majority class, this measure is likely not the most informative.

The Positive Predictive Value remains high due to the prevalence of the majority class, and the Negative Predictive Value is perfect at 1.00000, but this is misleading since there is only one true positive case.

The Kappa statistic has improved slightly to 0.0214 compared to 0 for LR 3 but is still near zero, suggesting no meaningful agreement between prediction and actuality.

Balanced Accuracy is essentially the same as in LR 3, at 0.50543, indicating the model is not effective in predicting the minority class.

Certain variables like product.manufacturer_city_4513, product.manufacturer_state_32, and source_type_17 are significant predictors in LR 4, reflecting different associations with the outcome compared to the previous models.

Variables like source_type_11 and last_year_classification1_num_uniq show significance in influencing the response variable, which varies from previous models.

The Residual Deviance for LR 4 is slightly lower than the previous models, suggesting a marginally better fit to the training data.

The AIC of LR 4 is lower compared to the previous models, indicating a more favorable balance between the model's fit and complexity.

Fewer iterations are needed for convergence in LR 4 (9 iterations), which could imply a more straightforward fit to the data or possibly a simpler model structure.

The fourth model shows marginal improvements in some fit statistics like AIC, but it doesn't meaningfully advance in predictive performance for the minority class as evidenced by the confusion matrix and the specificity measure. All models struggle with an imbalance in the dataset, and while they can predict the majority class with near-perfect accuracy, their ability to detect the minority class is limited.

The high AUC values relative to the specificity values across all models suggest that the ROC curve is not reflecting the actual practical performance of the models for predicting positive cases. It's worth noting that the AUC can sometimes be an optimistic measure of model performance in highly imbalanced datasets.

```
df_logistic_lda_3 <- df2 %>%
  dplyr::select(
    `product.manufacturer_city_2085`
    ,`product.manufacturer_city_4513`
    ,`product.manufacturer_city_3153`

    ,`last_four_years_classification0_num_uniq`
    ,`last_four_years_legal_announcementing_firm_num_uniq`
    ,`last_four_years_product_quantity_average_average`

    ,`last_two_years_product_quantity_average_max`
    ,`last_two_years_root_cause_description_num_uniq`
    ,`last_two_years_classification1_num_uniq`
    ,`last_two_years_product_quantity_average_num_uniq`
    ,`last_two_years_root_cause_description_most_freq`
    ,`last_two_years_product_quantity_average_average`
    ,`last_two_years_classification0_num_uniq`

    ,`last_year_product_quantity_average_average`
    ,`last_year_product_quantity_average_max`
    ,`last_year_root_cause_description_num_uniq`
    ,`last_year_legal_announcementing_firm_most_freq`
    ,`last_year_classification1_num_uniq`
    ,`last_year_brand_name_num_uniq`
    ,`last_year_reason_for_legal_announcement_num_uniq`
    ,`last_year_reason_for_legal_announcement_most_freq`
    ,`last_year_product_quantity_average_num_uniq`

    ,`source_type_17`
    ,`source_type_3`
    ,`source_type_11`
    ,`source_type_5`
    ,`source_type_4`
    ,`source_type_15`

    ,`product.manufacturer_state_32`
    ,`death_or_not`
  )


df_logistic_lda_3 <- sample_n(df_logistic_lda_3, size = 65000)

set.seed(123)  # for reproducibility
index <- createDataPartition(df_logistic_lda_3$death_or_not, p = 0.80, list = FALSE)
trainData <- df_logistic_lda_3[index, ]
testData <- df_logistic_lda_3[-index, ]

model_logistic_3 <- glm(death_or_not ~ ., data = trainData, family = binomial())
summary(model_logistic_3)
```

```
##
## Call:
## glm(formula = death_or_not ~ ., family = binomial(), data = trainData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9236  -0.1247  -0.0974  -0.0778   3.7377
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                    -4.0053868  1.1486584
## product.manufacturer_city_2085                  4.1480015  1.1319764
## product.manufacturer_city_4513                 -2.3449707  0.2140903
## product.manufacturer_city_3153                 -0.0730922  1.0190815
## last_four_years_classification0_num_uniq       -3.0271908  6.8077034
## last_four_years_legal_announcementing_firm_num_uniq -0.3339265  0.2654104
## last_four_years_product_quantity_average_average  1.9551824  1.0359881
## last_two_years_product_quantity_average_max    -4.5127232  3.2151001
## last_two_years_root_cause_description_num_uniq -0.0598792  0.1707735
## last_two_years_classification1_num_uniq         0.0067829  0.0068886
## last_two_years_product_quantity_average_num_uniq  0.1615033  0.1685556
## last_two_years_root_cause_description_most_freq  1.2043279  0.9832588
## last_two_years_product_quantity_average_average  8.8101615  6.5289193
## last_two_years_classification0_num_uniq         3.0259468  6.6272494
## last_year_product_quantity_average_average     -4.1313247  6.4618526
## last_year_product_quantity_average_max          0.6961678  3.2935839
## last_year_root_cause_description_num_uniq       0.1012032  0.2098555
## last_year_legal_announcementing_firm_most_freq -0.8671232  0.6222198
## last_year_classification1_num_uniq             -0.0235348  0.0107768
## last_year_brand_name_num_uniq                   0.1178463  0.0767657
## last_year_reason_for_legal_announcement_num_uniq  0.2742862  0.1804780
## last_year_reason_for_legal_announcement_most_freq -0.0003163  0.0002227
## last_year_product_quantity_average_num_uniq     0.0127650  0.1961950
## source_type_17                                  1.1977260  0.2977396
## source_type_3                                  -0.5783929  0.2254833
## source_type_11                                 -0.8724257  0.2678719
## source_type_5                                  -1.4809858  0.5852504
## source_type_4                                  -0.1767382  0.2244742
## source_type_15                                  1.5133272  0.4627945
## product.manufacturer_state_32                  -2.3623822  0.2617363
##                                                z value Pr(>|z|)
## (Intercept)                                     -3.487 0.000488 ***
## product.manufacturer_city_2085                   3.664 0.000248 ***
## product.manufacturer_city_4513                 -10.953  < 2e-16 ***
## product.manufacturer_city_3153                  -0.072 0.942822
## last_four_years_classification0_num_uniq        -0.445 0.656557
## last_four_years_legal_announcementing_firm_num_uniq  -1.258 0.208337
## last_four_years_product_quantity_average_average  1.887 0.059125 .
## last_two_years_product_quantity_average_max     -1.404 0.160437
## last_two_years_root_cause_description_num_uniq  -0.351 0.725862
## last_two_years_classification1_num_uniq          0.985 0.324792
## last_two_years_product_quantity_average_num_uniq  0.958 0.337982
## last_two_years_root_cause_description_most_freq  1.225 0.220638
## last_two_years_product_quantity_average_average  1.349 0.177207
## last_two_years_classification0_num_uniq          0.457 0.647965
## last_year_product_quantity_average_average      -0.639 0.522601
## last_year_product_quantity_average_max           0.211 0.832598
## last_year_root_cause_description_num_uniq        0.482 0.629627
## last_year_legal_announcementing_firm_most_freq  -1.394 0.163440
## last_year_classification1_num_uniq              -2.184 0.028973 *
## last_year_brand_name_num_uniq                    1.535 0.124749
## last_year_reason_for_legal_announcement_num_uniq  1.520 0.128567
```

```
## last_year_reason_for_legal_announcement_most_freq    -1.420 0.155480
## last_year_product_quantity_average_num_uniq           0.065 0.948124
## source_type_17                                        4.023 5.75e-05 ***
## source_type_3                                        -2.565 0.010314 *
## source_type_11                                       -3.257 0.001126 **
## source_type_5                                        -2.531 0.011389 *
## source_type_4                                        -0.787 0.431081
## source_type_15                                        3.270 0.001076 **
## product.manufacturer_state_32                        -9.026  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4198.4  on 51999  degrees of freedom
## Residual deviance: 3904.2  on 51970  degrees of freedom
## AIC: 3964.2
##
## Number of Fisher Scoring iterations: 9
```

```
probabilities_logistic_3 <- predict(model_logistic_3, newdata = testData, type = "response")
predictedClasses_logistic_3 <- ifelse(probabilities_logistic_3 > 0.5, 1, 0)

confusionMatrix(data = as.factor(predictedClasses_logistic_3), reference = as.factor(testData$death_or_not))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12908    91
##          1     0     1
##
##                Accuracy : 0.993
##                  95% CI : (0.9914, 0.9944)
##     No Information Rate : 0.9929
##     P-Value [Acc > NIR] : 0.486
##
##                   Kappa : 0.0214
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.00000
##             Specificity : 0.01087
##          Pos Pred Value : 0.99300
##          Neg Pred Value : 1.00000
##              Prevalence : 0.99292
##          Detection Rate : 0.99292
##    Detection Prevalence : 0.99992
##       Balanced Accuracy : 0.50543
##
##        'Positive' Class : 0
##
```

```
rocResult_logistic_3 <- roc(response = testData$death_or_not, predictor = probabilities_logistic_3)
```

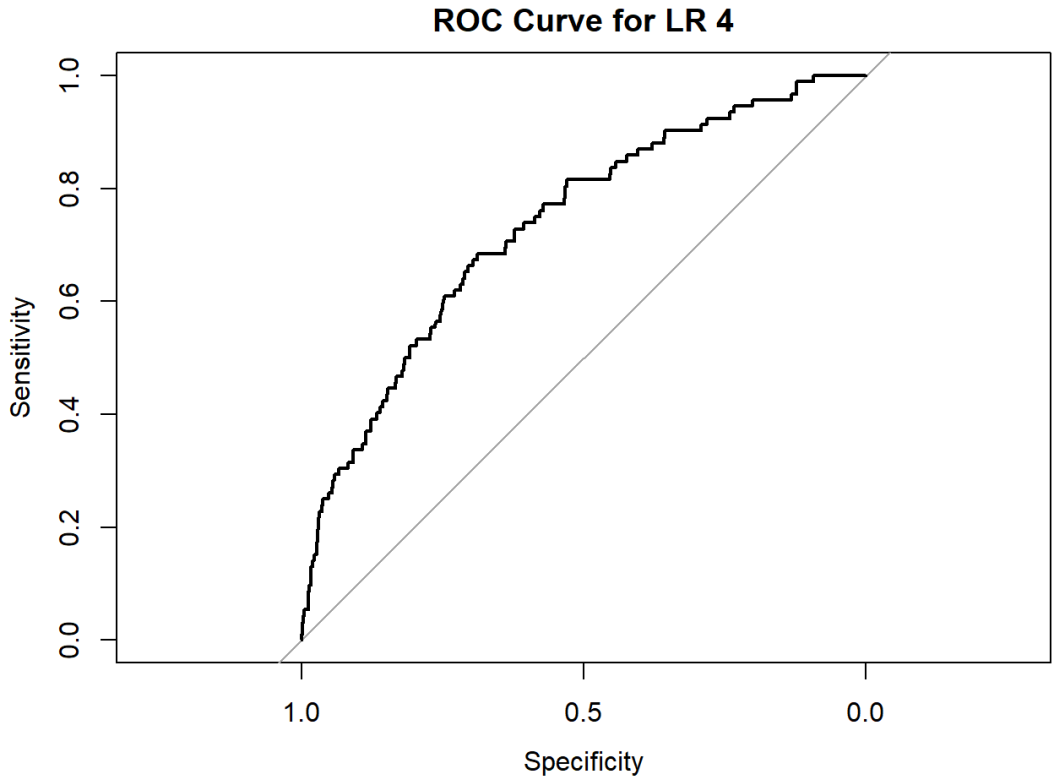
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(rocResult_logistic_3)
```

```
## Area under the curve: 0.7339
```

```
plot(rocResult_logistic_3, main = "ROC Curve for LR 4")
```



##LDA for Product Field Unknown

The ROC curve indicates a fair level of discrimination with an Area Under Curve (AUC) of 0.7308. This is a measure of the model's ability to correctly classify the positive class. The AUC is higher than the third logistic regression model (LR 3) which had an AUC of 0.7028, but lower than the second logistic regression model (LR 2) with an AUC of 0.7447.

The overall accuracy of the LDA model is 98.04%, which is lower than all the logistic regression models (LR 1-4), which had accuracies ranging from 99.28% to 99.34%. Despite the high accuracy, it's important to consider the class imbalance which likely inflates this metric.

Specificity or True Negative Rate is very low at 9.783%, which is consistent with the other models and indicates a continued difficulty in correctly identifying the positive class.

The Positive Predictive Value (PPV) is high at 99.353%, similar to the logistic regression models, reflecting the class imbalance.

The Balanced Accuracy is 54.225%, which reflects the model's limited ability to balance sensitivity and specificity and is the lowest among the models we've seen.

Compared to the logistic regression models, the LDA model has a slightly better AUC than LR 3 but not as good as LR 2. While the accuracy is lower, this may not necessarily be a disadvantage, considering the severe class imbalance in the dataset. The LDA model does seem to recognize a few more true positives, as seen in the lower sensitivity and higher NPV, but this is at the expense of a larger number of false positives, leading to lower specificity and PPV.

This model, like the logistic regression models, shows a high degree of class imbalance influence, as indicated by the high accuracy but low specificity. The higher AUC suggests some potential in the LDA model for discrimination between the classes, but there's still room for improvement, especially in correctly identifying the minority class. Techniques such as resampling the data, using different thresholds for classification, or employing cost-sensitive learning may help improve performance in future iterations.

```
# Performing LDA for unknown field


lda_model <- lda(death_or_not ~ ., data = trainData)

predictions_lda <- predict(lda_model, testData)
predictedClasses_lda <- predictions_lda$class
posteriorProbabilities_lda <- predictions_lda$posterior[,2]  # Probabilities of class 1

confusionMatrix(data = as.factor(predictedClasses_lda), reference = as.factor(testData$death_or_not))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12736    83
##          1   172     9
##
##                Accuracy : 0.9804
##                  95% CI : (0.9779, 0.9827)
##     No Information Rate : 0.9929
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0571
##
##  Mcnemar's Test P-Value : 3.573e-08
##
##             Sensitivity : 0.98667
##             Specificity : 0.09783
##          Pos Pred Value : 0.99353
##          Neg Pred Value : 0.04972
##              Prevalence : 0.99292
##          Detection Rate : 0.97969
##    Detection Prevalence : 0.98608
##       Balanced Accuracy : 0.54225
##
##        'Positive' Class : 0
##
```

```
roc_result_lda <- roc(response = testData$death_or_not, predictor = posteriorProbabilities_lda)
```

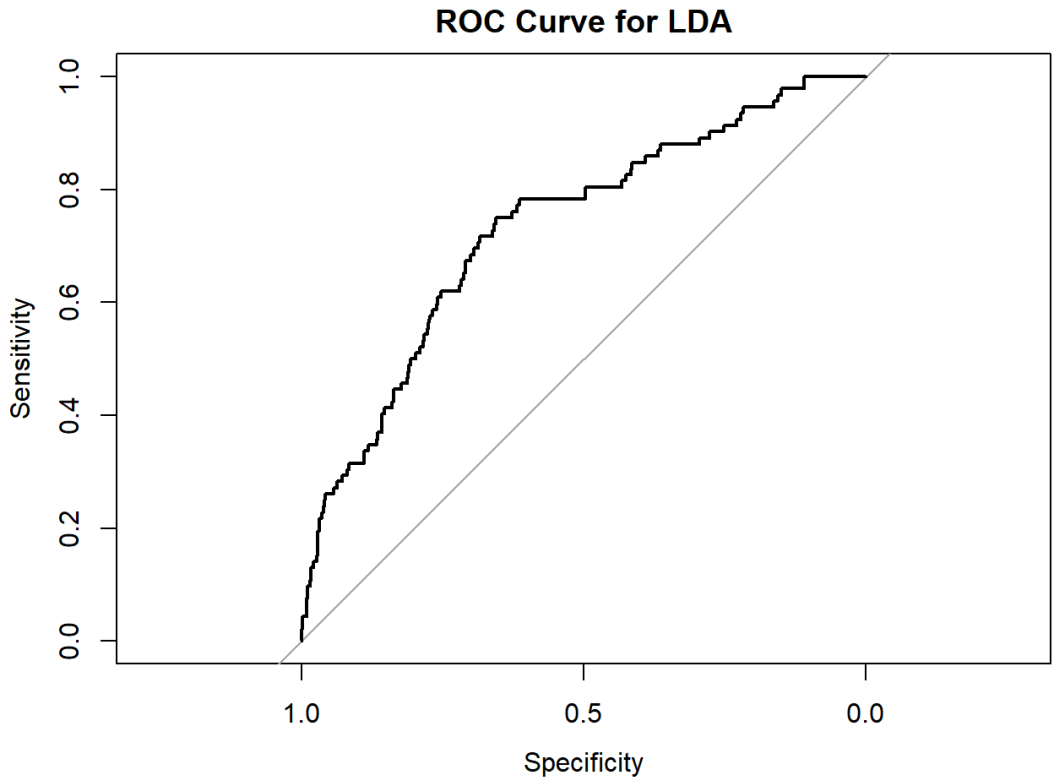
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(roc_result_lda)
```

```
## Area under the curve: 0.7308
```

```
plot(roc_result_lda, main = "ROC Curve for LDA")
```



LDA for Product Report Code LKK

The AUC remains at 0.7308, indicating no change in the model's ability to discriminate between the two classes compared to the first LDA model.

The accuracy is unchanged at 98.04%, and all other metrics from the confusion matrix (sensitivity, specificity, PPV, NPV, prevalence, detection rate, detection prevalence, and balanced accuracy) are the same.

The McNemar's Test p-value remains highly significant, suggesting an imbalance between the number of false negatives and false positives, which persists from the first LDA model.

When compared to the logistic regression models (LR 1-4), the accuracy of both LDA models is lower. However, in terms of AUC, both LDA models are higher than the AUC of LR 3 (0.7028) and close to LR 1 (0.7117), but not as high as LR 2 (0.7447). This indicates that the discriminative ability of LDA is not as strong as the best logistic regression model.

The sensitivity and specificity of both LDA models are lower than those reported in the logistic regression models, which had near-perfect sensitivity but also very low specificity. This suggests that LDA models may be slightly better at detecting the minority class but still suffer from the same issue of low specificity.

The PPV is high in all models due to the class imbalance, which leads to a high number of true negatives.

The NPV is very low in all models, but the NPVs in the logistic regression models are not provided for direct comparison.

Given that the metrics for both LDA models are identical, it raises the question of whether there was a variation in the model inputs or parameters. If the inputs to both LDA models were the same, the identical metrics suggest that the models have converged to the same solution.

In summary, the LDA models are consistent with each other but do not exhibit a marked improvement over the logistic regression models in terms of the ability to predict the minority class correctly. All models indicate high accuracy, but this is a misleading metric due to the class imbalance, and they all suffer from low specificity, indicating challenges in correctly identifying the positive class in an imbalanced dataset.

```
#LDA

lda_model_2 <- lda(death_or_not ~ ., data = trainData)

predictions_lda_2 <- predict(lda_model_2, testData)
predictedClasses_lda_2 <- predictions_lda_2$class
posteriorProbabilities_lda_2 <- predictions_lda_2$posterior[,2]  # Probabilities of class 1

confusionMatrix(data = as.factor(predictedClasses_lda_2), reference = as.factor(testData$death_or_not))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12736    83
##          1   172     9
##
##                Accuracy : 0.9804
##                  95% CI : (0.9779, 0.9827)
##     No Information Rate : 0.9929
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0571
##
##  Mcnemar's Test P-Value : 3.573e-08
##
##             Sensitivity : 0.98667
##             Specificity : 0.09783
##          Pos Pred Value : 0.99353
##          Neg Pred Value : 0.04972
##              Prevalence : 0.99292
##          Detection Rate : 0.97969
##    Detection Prevalence : 0.98608
##       Balanced Accuracy : 0.54225
##
##        'Positive' Class : 0
##
```

```
roc_result_lda_2 <- roc(response = testData$death_or_not, predictor = posteriorProbabilities_lda_2)
```

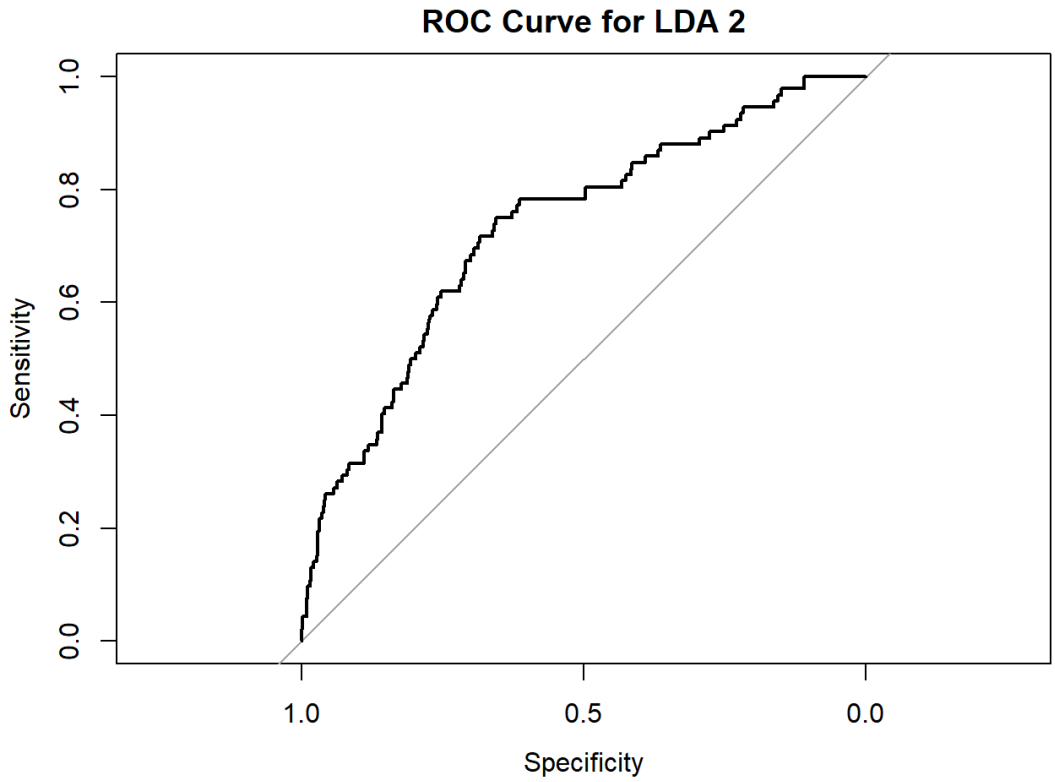
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(roc_result_lda_2)
```

```
## Area under the curve: 0.7308
```

```
plot(roc_result_lda_2, main = "ROC Curve for LDA 2")
```



LDA for Product Issue Type 599

The AUC for LDA 3 is 0.7308, which is exactly the same as both LDA 1 and LDA 2. This suggests no change in the model's ability to discriminate between the two classes.

LDA 3 maintains an accuracy of 98.04%, identical to LDA 1 and LDA 2

At 98.67%, LDA 3 has high sensitivity, the same as the previous two LDA models

The specificity remains low at 9.783%, indicating the model continues to struggle with identifying the positive class.

NPV remains low at 4.972%, indicating a poor performance in predicting true positives.

The Kappa statistic value is 0.0571, indicating the same level of agreement between the predicted and actual values as LDA 1 and LDA 2.

At 54.225%, the balanced accuracy, which averages sensitivity and specificity, suggests limited performance

Compared to the logistic regression models (LR 1-4), the accuracies of the LDA models are slightly lower, but this difference is minimal.

The AUC of LDA 1, LDA 2, and LDA 3 is higher than that of LR 3 (0.7028), similar to LR 1 (0.7117), but not as high as LR 2 (0.7447), indicating that while the LDA models don't have the highest AUC, they are relatively consistent.

All the LDA models show near-perfect sensitivity but very low specificity, similar to the logistic regression models, suggesting a common difficulty in positively identifying the minority class across all models.

In summary, across all LDA models and compared to the logistic regression models, we see high sensitivity and PPV but low specificity and NPV. This pattern indicates a strong influence of class imbalance and highlights the need for techniques that address the imbalance or alter the decision threshold to improve minority class detection.

```
#LDA

lda_model_3 <- lda(death_or_not ~ ., data = trainData)

predictions_lda_3 <- predict(lda_model_3, testData)
predictedClasses_lda_3 <- predictions_lda_3$class
posteriorProbabilities_lda_3 <- predictions_lda_3$posterior[,2]  # Probabilities of class 1

confusionMatrix(data = as.factor(predictedClasses_lda_3), reference = as.factor(testData$death_or_not))

roc_result_lda_3 <- roc(response = testData$death_or_not, predictor = posteriorProbabilities_lda_3)
auc(roc_result_lda_3)
plot(roc_result_lda_3, main = "ROC Curve for LDA 3")

summary(lda_model_3)
```

LDA for Manufacturer 1 Contact Address 9476

The AUC for LDA 4 is 0.7308, consistent with the AUC values reported for LDA 1, LDA 2, and LDA 3. This suggests that all LDA models are equal in terms of discrimination ability between the classes.

All four LDA models have the same accuracy (0.9804), which is slightly lower than the accuracies reported for the logistic regression models (LR 1-4), which were all above 0.9928.

Sensitivity is consistent at 0.98667 across all LDA models, indicating a high true positive rate for the majority class.

Specificity remains low at 0.09783 for all LDA models, indicating poor performance in correctly identifying the minority class.

The PPV is high at 0.99353, as is common in datasets with a significant class imbalance where the model correctly predicts the majority class.

The NPV is very low at 0.04972, underscoring the models' difficulty in accurately predicting positive cases in the minority class.

The balanced accuracy of 0.54225 for all LDA models indicates that the models do not perform well in a balanced manner for both classes.

The logistic regression models (LR 1-4) showed slightly better accuracy but also struggled with low specificity. This pattern suggests that while logistic regression models are better at predicting the majority class without error, they are not necessarily more effective at identifying the minority class.

The AUC for the logistic regression models varied, with LR 2 showing the highest AUC at 0.7447 and LR 3 the lowest at 0.7028. The LDAs' AUC values are in the lower range of these results, suggesting that while the LDAs have some ability to discriminate between classes, they are not the top performers.

The identical results across the four LDA models suggest that either the models are being trained on the same features and data and thus arriving at the same statistical conclusions, or there may be an issue with the analysis or reporting process. It is unusual for different models or iterations to yield exactly the same metrics, especially in different runs, unless the underlying data and model structure are identical.

```
#LDA

lda_model_4 <- lda(death_or_not ~ ., data = trainData)

predictions_lda_4 <- predict(lda_model_4, testData)
predictedClasses_lda_4 <- predictions_lda_4$class
posteriorProbabilities_lda_4 <- predictions_lda_4$posterior[,2]  # Probabilities of class 1

confusionMatrix(data = as.factor(predictedClasses_lda_4), reference = as.factor(testData$death_or_not))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 12736    83
##          1   172     9
##
##                Accuracy : 0.9804
##                  95% CI : (0.9779, 0.9827)
##     No Information Rate : 0.9929
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0571
##
##  Mcnemar's Test P-Value : 3.573e-08
##
##             Sensitivity : 0.98667
##             Specificity : 0.09783
##          Pos Pred Value : 0.99353
##          Neg Pred Value : 0.04972
##              Prevalence : 0.99292
##          Detection Rate : 0.97969
##    Detection Prevalence : 0.98608
##       Balanced Accuracy : 0.54225
##
##        'Positive' Class : 0
##
```

```
roc_result_lda_4 <- roc(response = testData$death_or_not, predictor = posteriorProbabilities_lda_4)
```

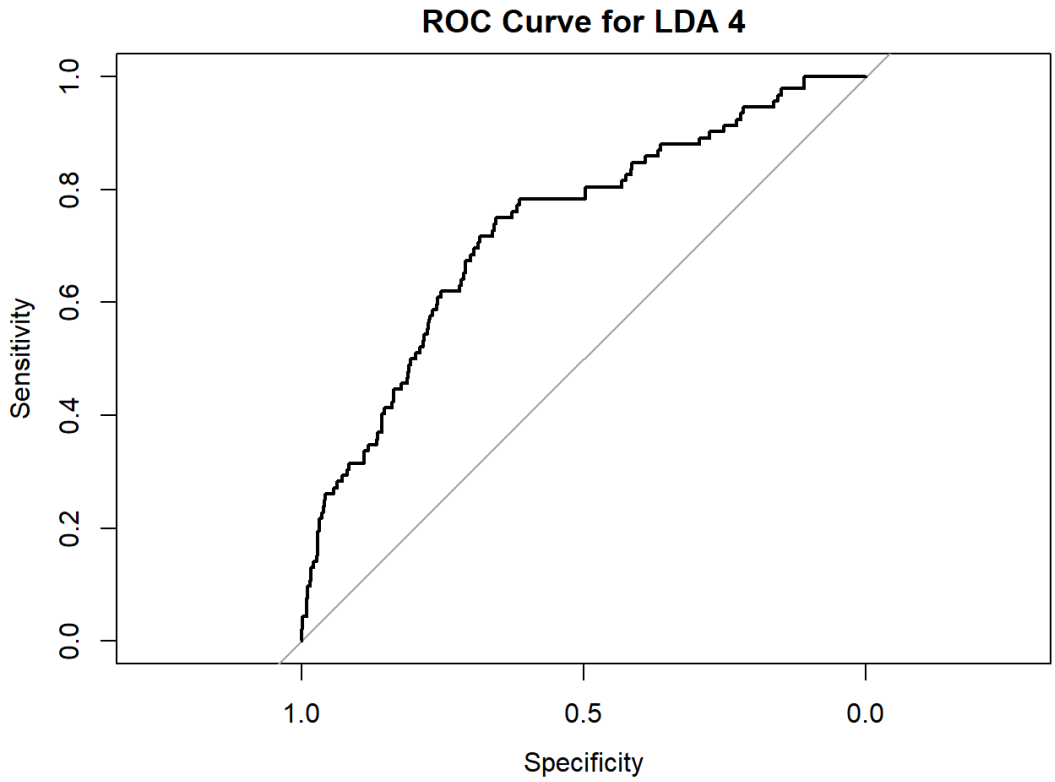
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(roc_result_lda_4)
```

```
## Area under the curve: 0.7308
```

```
plot(roc_result_lda_4, main = "ROC Curve for LDA 4")
```



Final Remarks

All models, both LDA and logistic regression, show the impact of class imbalance with high accuracy and PPV, but poor specificity and NPV. Despite high accuracies, the practical usefulness of these models is limited without addressing the imbalance or employing techniques to improve minority class prediction.