

# Base model fitting for RTC Data

Matthew Knowles

2023-03-15

## 1: Data Cleaning

We begin by converting to a binary variable by severity. Accidents with a severity of 3 become a 0, and others are converted to a 1.

```
accidents <- accidents %>%
  mutate(bin_severity = ifelse(.data$accident_severity == 3, 0, 1))
```

We now reduce the size of the dataset to contain only variables we care about.

```
sub_accidents <- accidents %>%
  select(
    "accident_index",
    "bin_severity",
    "day_of_week",
    "road_type",
    "speed_limit",
    "time",
    "light_conditions",
    "weather_conditions",
    "local_authority_ons_district",
    "road_surface_conditions"
  )

sub_vehicles <- vehicles %>%
  select(
    "accident_index",
    "sex_of_driver",
    "age_band_of_driver"
  )

sub_casualties <- casualty %>%
  select(
    "accident_index",
    "sex_of_casualty",
    "age_band_of_casualty"
  )

df_sub <- merge.data.frame(sub_vehicles, sub_casualties, by = "accident_index")
df_sub_unique <- unique(df_sub)
df <- merge.data.frame(sub_accidents, df_sub_unique, by = "accident_index") %>%
  filter(sex_of_casualty != 9)
```

Before dealing with other variables in the dataset, we deal with time on its own. The times provided in the

dataset are characters, for example of the form “12:14”. There are many unique times in the dataset, which would cause issues when fitting models as these would all be treated as individual factors. We therefore split the 24 hour day into 6 chunks of 4 hours, as this dramatically reduces the number of factors, but also gives good information about the relationship between time of day and accident severity.

The first thing to do is build a function to parse the times by splitting the character before the colon, and converting that to an integer. A series of if-statements are then checked to put the time into the correct category. This isn’t the most computationally efficient method, but it does the job. We then apply this function to every time observation to create a new column of data called “time\_group”, and drop the original time variable from the data so it doesn’t cause problems later on.

```
# Split time into 6 chunks of four hours
parse_times <- function(time){
  split_time <- stringr::str_split(time, ":", simplify = TRUE)
  hour <- as.integer(split_time[1])
  if (hour < 4) return(1)
  if (hour < 8 ) return(2)
  if (hour < 12) return(3)
  if (hour < 16) return(4)
  ifelse(hour < 20, return(5), return(6))
}

df <- df %>%
  mutate(time_group = purrr::map_int(time, parse_times)) %>%
  select(-time)
```

Location is to be included, but in the current state, the *local\_authority\_ons\_district* variable has too many unique values. To account for this, the region codes (those beginning with E0, N, S or W) are matched to a wider local area, such as the West Midlands, Yorkshire and The Humber etc. Across England, Wales, Scotland & Northern Ireland, there are 12 unique regions. Each region is assigned a numerical value. This assignment is done alphabetically, so East Midlands is 1, all the way through to Yorkshire and The Humber at 12. With this in mind, we can add region code to the data.

```
# Obtains the numerical value of a region from a given ONS coded
obtain_region <- function(region_code){
  return(regions$Region.Code[which(regions$Code == region_code)])
}

df <- df %>%
  mutate(region = purrr::map_int(local_authority_ons_district, obtain_region)) %>%
  select(-local_authority_ons_district)
```

We want all variables, except speed limit, to be factors. To achieve this we mutate across all integer columns and change the type to factor. However, we need speed limit to stay as an integer to make the model adaptable to non-standard speed limits. The “bin\_severity” column needs to be a factor, so we ensure this by calling the as.factor function once more just to ensure it is in fact a factor. By viewing the head of the data as a tibble we can check that the columns are of expected type.

```
df <- df %>%
  mutate(across(where(is.integer), as.factor))
df$speed_limit <- as.integer(df$speed_limit) #Speed limit goes back to integer
df$bin_severity <- as.factor(df$bin_severity)
head(tibble(df))

## # A tibble: 6 x 14
##   accident_index bin_s~1 day_o~2 road_~3 speed~4 light~5 weath~6 road_~7 sex_o~8
##   <chr>          <fct>    <fct>    <fct>    <int> <fct>    <fct>    <fct>    <fct>
```

```

## 1 2020010219808 0      3      6      2 1      9      9      2
## 2 2020010220496 0      2      6      2 1      1      1      1
## 3 2020010228005 0      4      6      3 4      1      2      3
## 4 2020010228006 1      4      6      3 4      1      1      1
## 5 2020010228011 0      4      6      3 4      1      1      1
## 6 2020010228012 0      4      2      2 4      1      1      1
## # ... with 5 more variables: age_band_of_driver <fct>, sex_of_casualty <fct>,
## #   age_band_of_casualty <fct>, time_group <fct>, region <fct>, and abbreviated
## #   variable names 1: bin_severity, 2: day_of_week, 3: road_type,
## #   4: speed_limit, 5: light_conditions, 6: weather_conditions,
## #   7: road_surface_conditions, 8: sex_of_driver

```

Some observations of the data are -1. We don't want these in the data, so we trim the data down to remove any rows that contain -1 in any of the columns. This helps with run-time of the model as well due to the size of the data before this is done. At this stage we also remove the accident index variable, as it was only needed for combining the three datasets, and isn't needed in fitting the GLM itself.

```

has.neg <- apply(df, 1, function(row) any(row == -1))
df <- df[-which(has.neg), ] %>%
  select(-accident_index)

```

## 2: Fitting

We fit a general glm to the training data. We specify that the response variable is binomial, and that we wish to use a logit link function. No other interaction terms are included, as this will be built upon later.

```

fit <- glm(
  data = df,
  formula = bin_severity ~ .,
  family = binomial(link = "logit")
)

```

Let us take a look at the summary and plots of this model.

```

summary(fit)

##
## Call:
## glm(formula = bin_severity ~ ., family = binomial(link = "logit"),
##      data = df)
##
## Deviance Residuals:
##    Min      1Q      Median      3Q      Max
## -1.5739 -0.7530 -0.6218 -0.4157  2.8225
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.644006  0.431747 -3.808 0.000140 ***
## day_of_week2                -0.178770  0.023107 -7.737 1.02e-14 ***
## day_of_week3                -0.187139  0.023000 -8.136 4.07e-16 ***
## day_of_week4                -0.192339  0.022858 -8.415 < 2e-16 ***
## day_of_week5                -0.118646  0.022515 -5.270 1.37e-07 ***
## day_of_week6                -0.141966  0.022166 -6.405 1.51e-10 ***
## day_of_week7                -0.038302  0.022692 -1.688 0.091439 .
## road_type2                  0.323376  0.060938  5.307 1.12e-07 ***
## road_type3                  0.061049  0.032112  1.901 0.057282 .

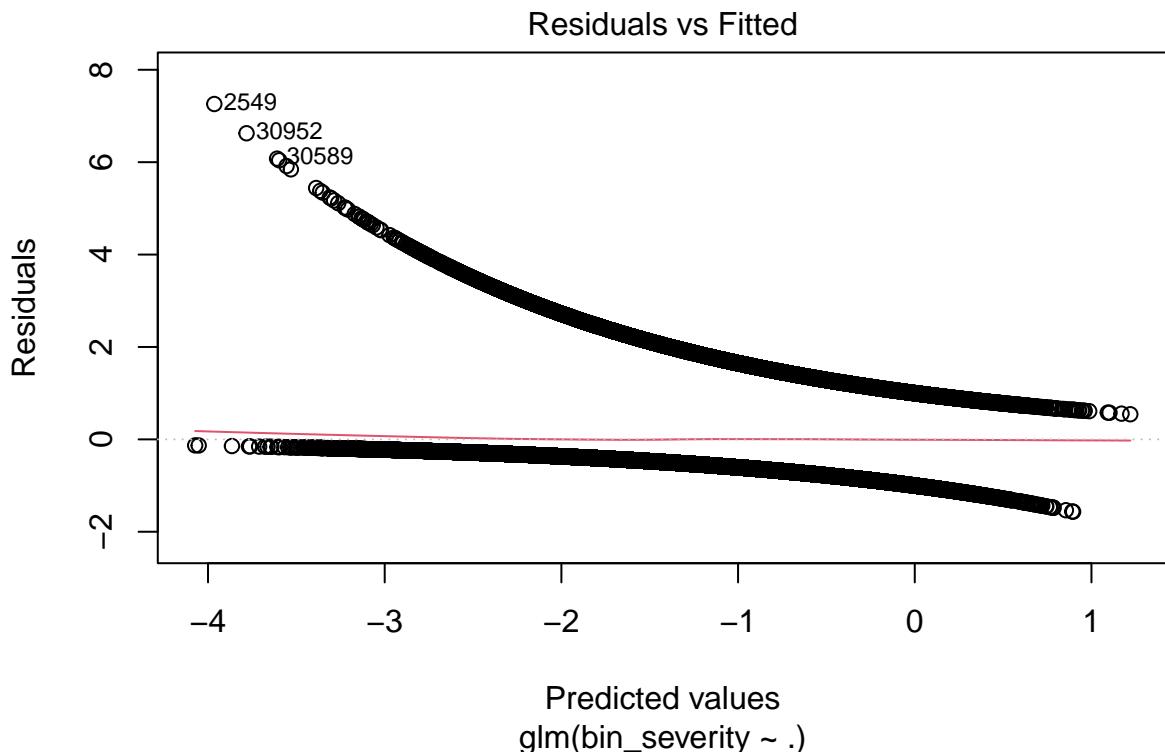
```

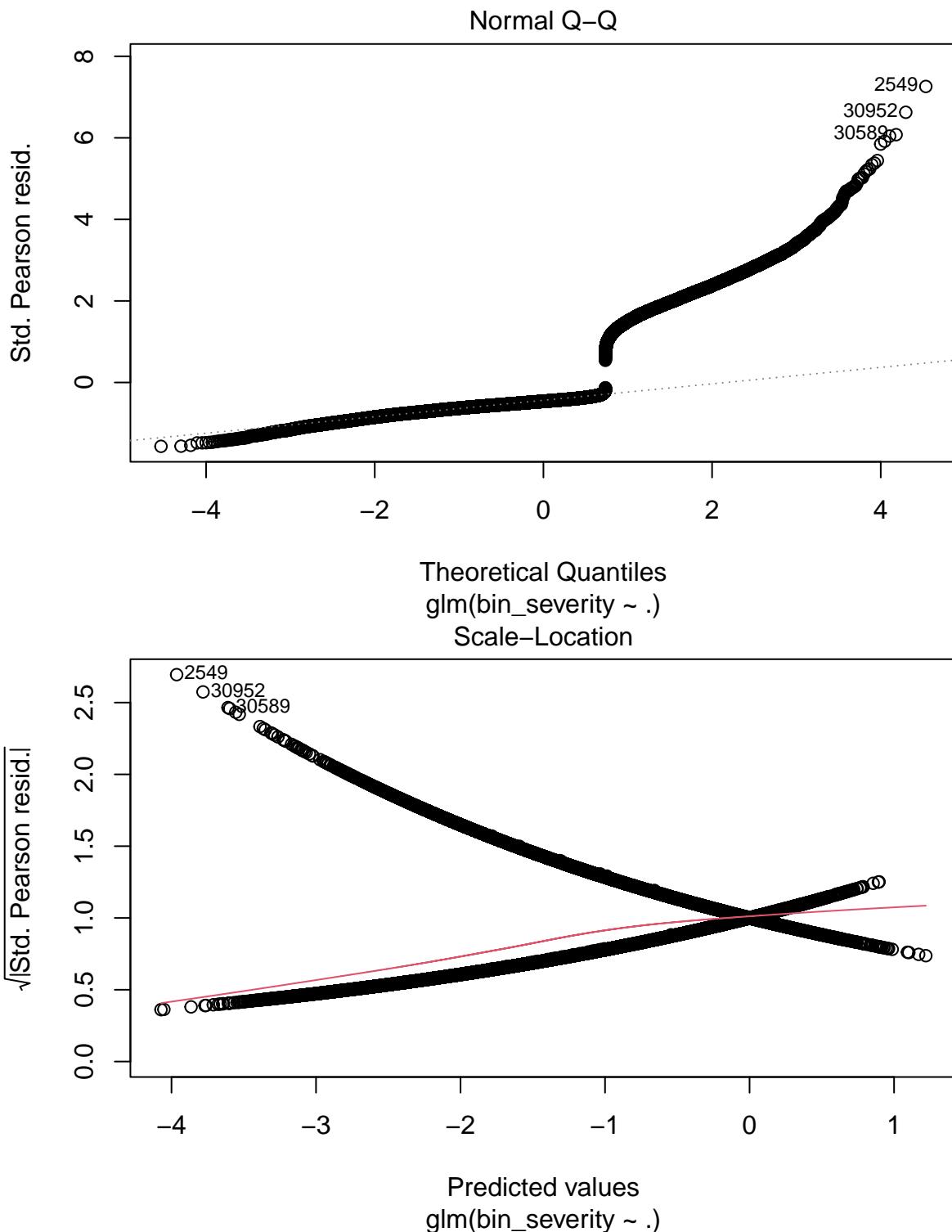
## road_type6	0.488735	0.028699	17.030	< 2e-16	***
## road_type7	0.018231	0.055495	0.329	0.742525	
## road_type9	-0.312001	0.084522	-3.691	0.000223	***
## speed_limit	0.188456	0.004715	39.972	< 2e-16	***
## light_conditions4	0.040551	0.019996	2.028	0.042565	*
## light_conditions5	0.053542	0.068794	0.778	0.436401	
## light_conditions6	0.240118	0.026899	8.927	< 2e-16	***
## light_conditions7	-0.473057	0.056234	-8.412	< 2e-16	***
## weather_conditions2	-0.124128	0.022273	-5.573	2.50e-08	***
## weather_conditions3	0.093784	0.145102	0.646	0.518062	
## weather_conditions4	0.183752	0.044103	4.166	3.09e-05	***
## weather_conditions5	-0.036287	0.044353	-0.818	0.413276	
## weather_conditions6	0.539645	0.209884	2.571	0.010136	*
## weather_conditions7	-0.010957	0.073136	-0.150	0.880910	
## weather_conditions8	-0.068316	0.040012	-1.707	0.087748	.
## weather_conditions9	-0.365490	0.055126	-6.630	3.35e-11	***
## road_surface_conditions2	-0.027202	0.016692	-1.630	0.103172	
## road_surface_conditions3	-1.078139	0.195372	-5.518	3.42e-08	***
## road_surface_conditions4	-0.276089	0.067228	-4.107	4.01e-05	***
## road_surface_conditions5	-0.220648	0.132876	-1.661	0.096801	.
## road_surface_conditions9	-0.086495	0.137311	-0.630	0.528749	
## sex_of_driver2	-0.166366	0.014417	-11.540	< 2e-16	***
## sex_of_driver3	-0.479985	0.060223	-7.970	1.59e-15	***
## age_band_of_driver2	-0.153648	0.460146	-0.334	0.738446	
## age_band_of_driver3	0.107394	0.436205	0.246	0.805527	
## age_band_of_driver4	0.149958	0.431535	0.347	0.728216	
## age_band_of_driver5	0.161412	0.431280	0.374	0.708208	
## age_band_of_driver6	0.144557	0.431023	0.335	0.737337	
## age_band_of_driver7	0.141425	0.431086	0.328	0.742862	
## age_band_of_driver8	0.201195	0.431147	0.467	0.640750	
## age_band_of_driver9	0.282638	0.431266	0.655	0.512230	
## age_band_of_driver10	0.268665	0.431707	0.622	0.533724	
## age_band_of_driver11	0.172723	0.432274	0.400	0.689475	
## sex_of_casualty2	-0.349442	0.013490	-25.905	< 2e-16	***
## age_band_of_casualty2	-0.093993	0.064138	-1.465	0.142788	
## age_band_of_casualty3	0.082312	0.058835	1.399	0.161807	
## age_band_of_casualty4	-0.050879	0.052633	-0.967	0.333709	
## age_band_of_casualty5	-0.184877	0.051735	-3.574	0.000352	***
## age_band_of_casualty6	-0.215375	0.049903	-4.316	1.59e-05	***
## age_band_of_casualty7	-0.165505	0.050590	-3.271	0.001070	**
## age_band_of_casualty8	-0.082391	0.050767	-1.623	0.104607	
## age_band_of_casualty9	0.089703	0.051712	1.735	0.082798	.
## age_band_of_casualty10	0.312644	0.054737	5.712	1.12e-08	***
## age_band_of_casualty11	0.469714	0.057027	8.237	< 2e-16	***
## time_group2	-0.328354	0.039882	-8.233	< 2e-16	***
## time_group3	-0.544970	0.038565	-14.131	< 2e-16	***
## time_group4	-0.491883	0.037625	-13.073	< 2e-16	***
## time_group5	-0.450272	0.034891	-12.905	< 2e-16	***
## time_group6	-0.271676	0.035784	-7.592	3.15e-14	***
## region2	-0.036145	0.027132	-1.332	0.182797	
## region3	-0.291325	0.027280	-10.679	< 2e-16	***
## region4	-0.039928	0.037877	-1.054	0.291810	
## region5	0.003611	0.027774	0.130	0.896552	
## region7	0.615285	0.031638	19.448	< 2e-16	***

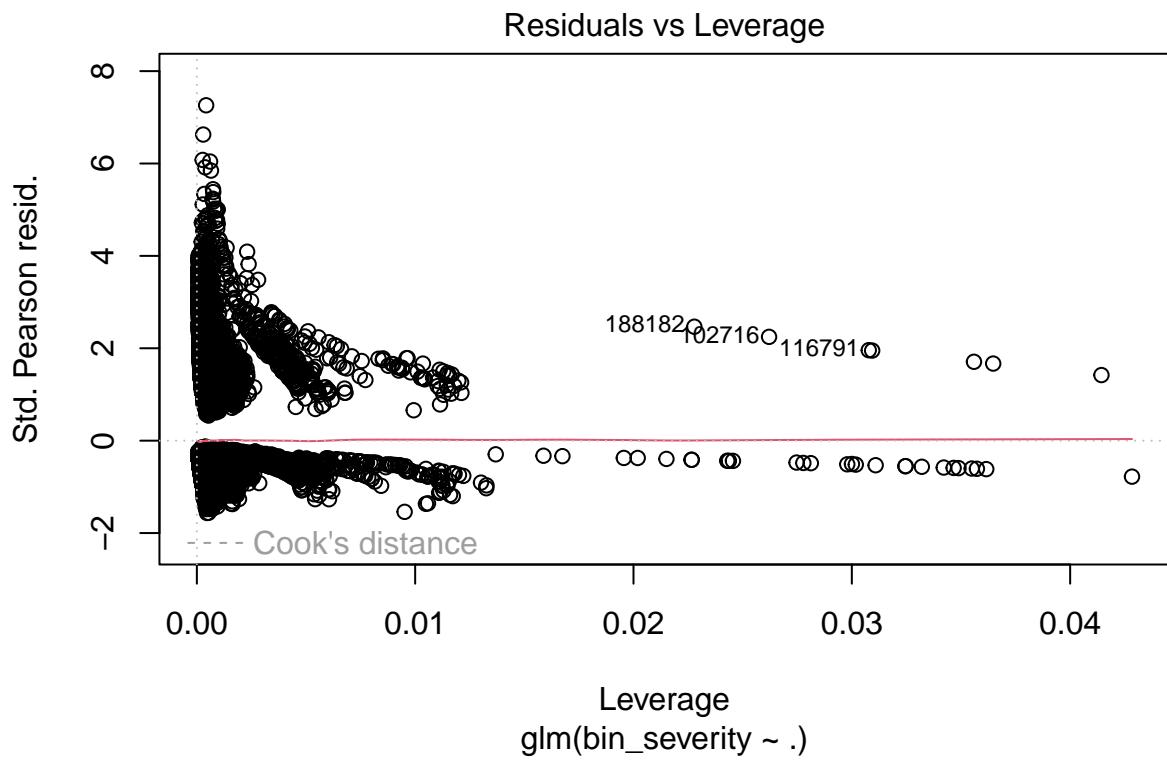
```

## region8          0.044661  0.024867  1.796 0.072501 .
## region9         -0.267143  0.028760 -9.289 < 2e-16 ***
## region10        -0.066668  0.037540 -1.776 0.075746 .
## region11        -0.160294  0.029819 -5.376 7.63e-08 ***
## region12         0.053446  0.028424  1.880 0.060061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 187526  on 173397  degrees of freedom
## Residual deviance: 178926  on 173330  degrees of freedom
## AIC: 179062
##
## Number of Fisher Scoring iterations: 4
plot(fit)

```

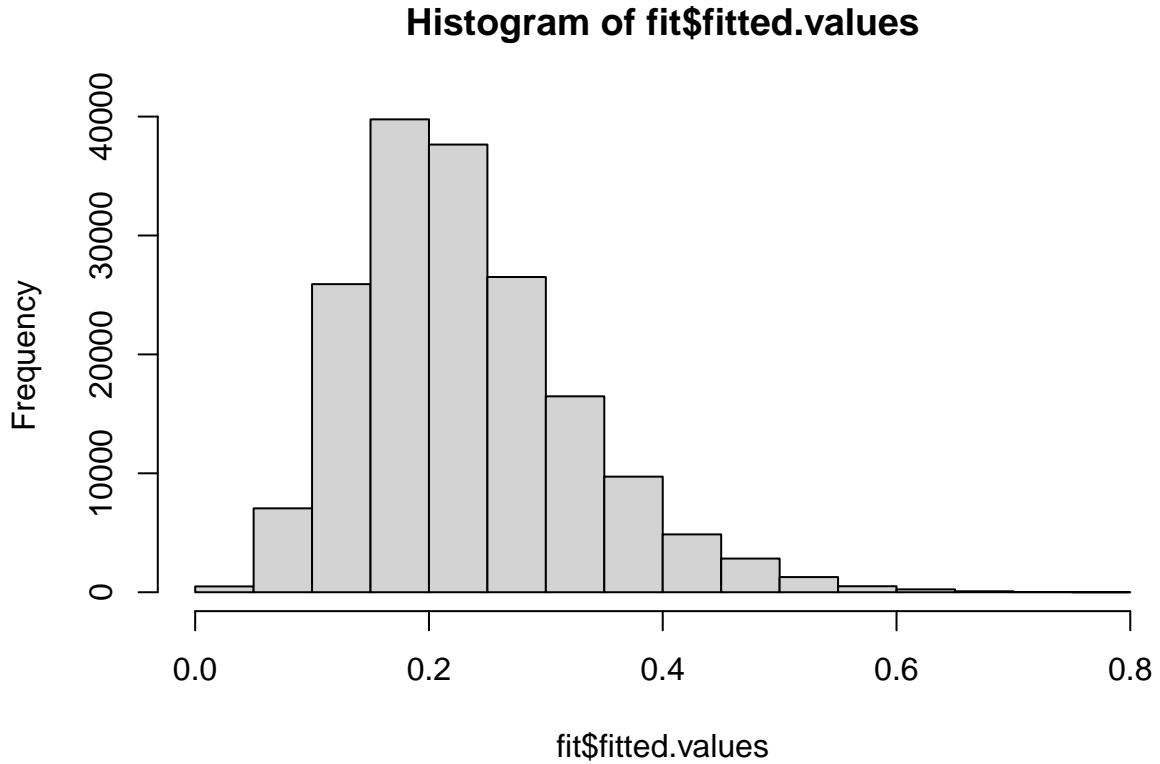






This histogram shows the distribution of fitted values from the training data.

```
hist(fit$fitted.values)
```

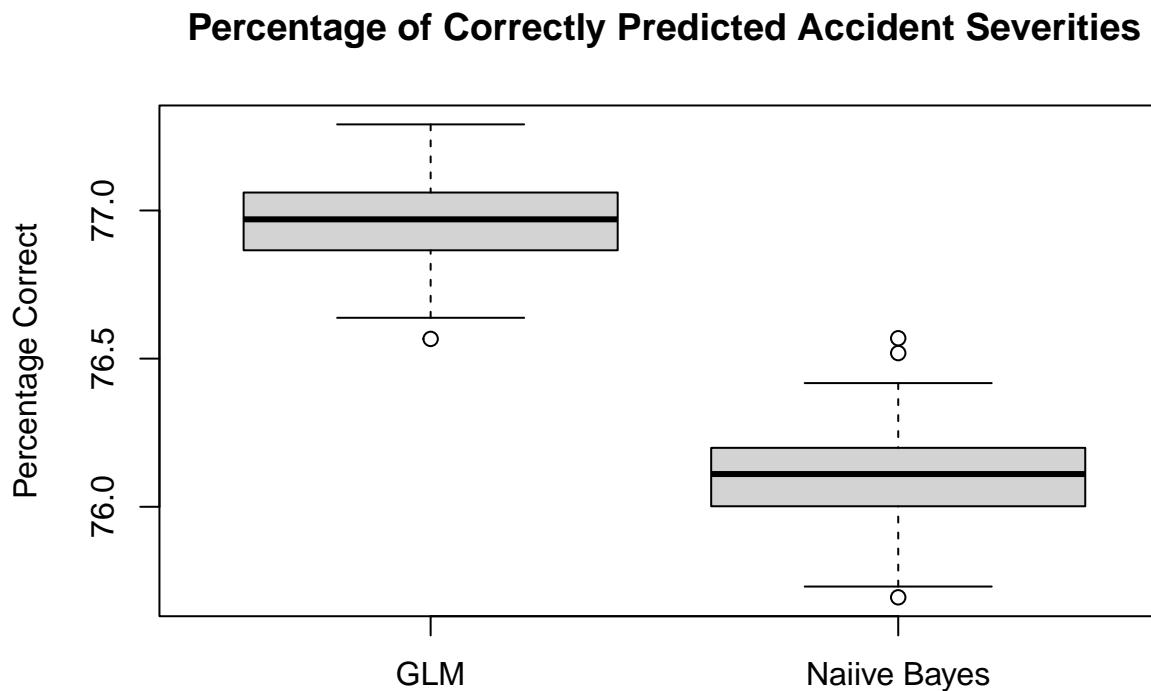


### 3: Cross Validation & Model Refining

We can perform some bootstrapping on the model

We can look at several useful plots to see how our model compares against a Naive-Bayes classifier.

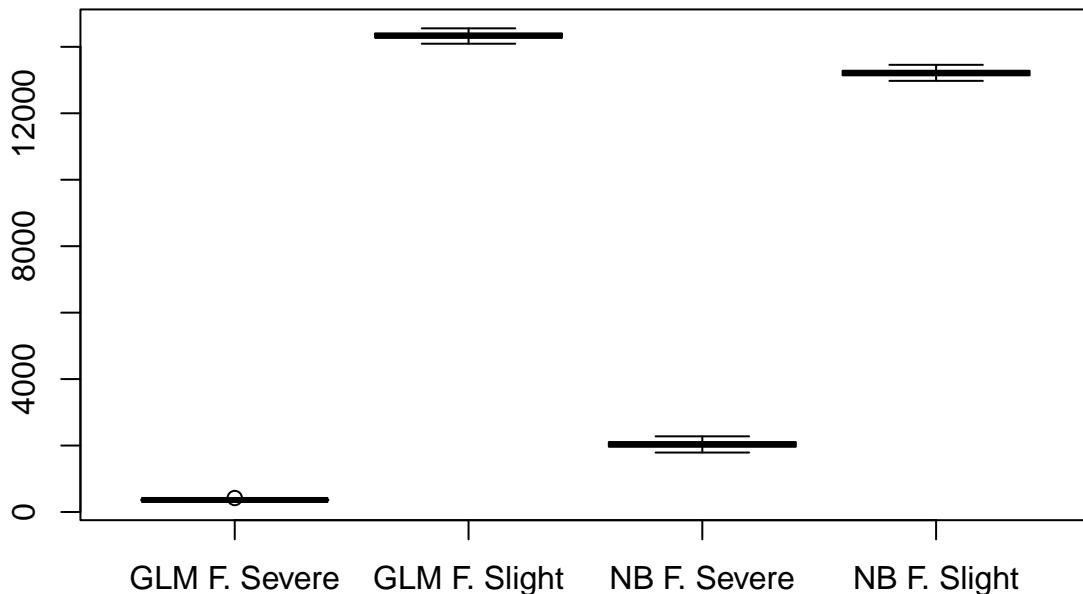
```
boxplot(accuracy$GLMPercCorrect, accuracy$NBPercCorrect,
        names = c("GLM", "Naïve Bayes"),
        ylab = "Percentage Correct",
        main = "Percentage of Correctly Predicted Accident Severities")
```



The plot below presents the proportion false predictions in each category. So a case when the model predicted severe but the accident was slight or visa-versa.

```
boxplot(accuracy$GLM.False.Severe, accuracy$GLM.False.Slight,
        accuracy$NB.False.Severe, accuracy$NB.False.Slight,
        names = c("GLM F. Severe", "GLM F. Slight",
                 "NB F. Severe", "NB F. Slight"),
        main = "Proportion of misclassifications"
)
```

## Proportion of misclassifications



Interestingly, the GLM model had a lower proportion of false severes than Naive-Bayes, but a higher proportion of false slight accidents than Naive-Bayes.

## 4: Speed Limit and Road Type analysis

We now present a slightly more specific model than the one we have just seen. We include an interaction term between speed limit and road type, allowing us to assess whether or not speed limits on a given road type are appropriate based on the contribution they make towards accident severity.

```
fit_rt_sl <- glm(
  data = df %>% select(-region),
  formula = bin_severity ~ . + speed_limit*road_type,
  family = binomial(link = "logit")
)
```

We can view the summary of this model as before.

```
summary(fit_rt_sl)

##
## Call:
## glm(formula = bin_severity ~ . + speed_limit * road_type, family = binomial(link = "logit"),
##      data = df %>% select(-region))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.5904   -0.7457   -0.6276   -0.4384    2.8480
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.493e+00  4.360e-01 -3.425 0.000615 ***
## day_of_week2             -1.633e-01  2.306e-02 -7.083 1.41e-12 ***
## day_of_week3             -1.792e-01  2.296e-02 -7.806 5.88e-15 ***
##
```

```

## day_of_week4 -1.799e-01 2.282e-02 -7.882 3.21e-15 ***
## day_of_week5 -1.125e-01 2.248e-02 -5.007 5.52e-07 ***
## day_of_week6 -1.282e-01 2.212e-02 -5.796 6.81e-09 ***
## day_of_week7 -2.979e-02 2.265e-02 -1.315 0.188365
## road_type2 -2.030e-01 2.398e-01 -0.846 0.397282
## road_type3 2.122e-01 9.644e-02 2.201 0.027768 *
## road_type6 2.073e-03 8.803e-02 0.024 0.981213
## road_type7 1.408e-01 1.604e-01 0.877 0.380238
## road_type9 -7.140e-01 2.263e-01 -3.155 0.001605 **
## speed_limit 1.075e-01 2.152e-02 4.996 5.85e-07 ***
## light_conditions4 6.281e-02 1.997e-02 3.144 0.001664 **
## light_conditions5 6.588e-02 6.876e-02 0.958 0.337998
## light_conditions6 2.124e-01 2.690e-02 7.898 2.83e-15 ***
## light_conditions7 -4.848e-01 5.613e-02 -8.637 < 2e-16 ***
## weather_conditions2 -1.300e-01 2.221e-02 -5.856 4.75e-09 ***
## weather_conditions3 2.333e-01 1.445e-01 1.615 0.106386
## weather_conditions4 1.879e-01 4.398e-02 4.272 1.94e-05 ***
## weather_conditions5 -1.629e-02 4.413e-02 -0.369 0.712115
## weather_conditions6 6.118e-01 2.090e-01 2.927 0.003417 **
## weather_conditions7 -3.561e-02 7.313e-02 -0.487 0.626306
## weather_conditions8 -9.837e-02 3.984e-02 -2.469 0.013538 *
## weather_conditions9 -3.495e-01 5.480e-02 -6.379 1.79e-10 ***
## road_surface_conditions2 -1.842e-02 1.663e-02 -1.107 0.268098
## road_surface_conditions3 -1.023e+00 1.953e-01 -5.236 1.64e-07 ***
## road_surface_conditions4 -2.752e-01 6.688e-02 -4.114 3.88e-05 ***
## road_surface_conditions5 -1.597e-01 1.324e-01 -1.207 0.227480
## road_surface_conditions9 -1.591e-01 1.369e-01 -1.162 0.245108
## sex_of_driver2 -1.670e-01 1.440e-02 -11.602 < 2e-16 ***
## sex_of_driver3 -4.869e-01 5.994e-02 -8.122 4.57e-16 ***
## age_band_of_driver2 -1.723e-01 4.576e-01 -0.377 0.706536
## age_band_of_driver3 6.973e-02 4.336e-01 0.161 0.872231
## age_band_of_driver4 9.137e-02 4.289e-01 0.213 0.831304
## age_band_of_driver5 1.050e-01 4.287e-01 0.245 0.806491
## age_band_of_driver6 9.090e-02 4.284e-01 0.212 0.831963
## age_band_of_driver7 8.727e-02 4.285e-01 0.204 0.838613
## age_band_of_driver8 1.460e-01 4.285e-01 0.341 0.733254
## age_band_of_driver9 2.270e-01 4.287e-01 0.530 0.596376
## age_band_of_driver10 2.132e-01 4.291e-01 0.497 0.619321
## age_band_of_driver11 1.107e-01 4.297e-01 0.258 0.796635
## sex_of_casualty2 -3.474e-01 1.347e-02 -25.792 < 2e-16 ***
## age_band_of_casualty2 -1.009e-01 6.402e-02 -1.575 0.115186
## age_band_of_casualty3 7.540e-02 5.872e-02 1.284 0.199104
## age_band_of_casualty4 -8.238e-02 5.253e-02 -1.568 0.116773
## age_band_of_casualty5 -2.196e-01 5.162e-02 -4.255 2.09e-05 ***
## age_band_of_casualty6 -2.496e-01 4.977e-02 -5.015 5.30e-07 ***
## age_band_of_casualty7 -1.982e-01 5.047e-02 -3.927 8.58e-05 ***
## age_band_of_casualty8 -1.143e-01 5.065e-02 -2.257 0.024035 *
## age_band_of_casualty9 6.083e-02 5.160e-02 1.179 0.238477
## age_band_of_casualty10 2.826e-01 5.463e-02 5.173 2.30e-07 ***
## age_band_of_casualty11 4.361e-01 5.692e-02 7.661 1.84e-14 ***
## time_group2 -3.299e-01 3.973e-02 -8.303 < 2e-16 ***
## time_group3 -5.454e-01 3.841e-02 -14.199 < 2e-16 ***
## time_group4 -4.858e-01 3.747e-02 -12.964 < 2e-16 ***
## time_group5 -4.459e-01 3.473e-02 -12.836 < 2e-16 ***

```

```

## time_group6      -2.757e-01  3.563e-02  -7.736 1.02e-14 ***
## id              1.881e-06  1.270e-07  14.812 < 2e-16 ***
## road_type2:speed_limit 1.450e-01  8.378e-02   1.731 0.083528 .
## road_type3:speed_limit -2.503e-03  2.300e-02  -0.109 0.913345
## road_type6:speed_limit 1.316e-01  2.212e-02   5.949 2.70e-09 ***
## road_type7:speed_limit 1.876e-03  3.307e-02   0.057 0.954770
## road_type9:speed_limit 8.912e-02  6.806e-02   1.309 0.190391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 187526  on 173397  degrees of freedom
## Residual deviance: 179636  on 173334  degrees of freedom
## AIC: 179764
##
## Number of Fisher Scoring iterations: 4

```

Notice the coefficients for road type 2 and 6 (one way street, single carriageway respectively) are on the same order as speed limit overall. This would suggest that perhaps a lower speed limit would be appropriate on these particular road types.