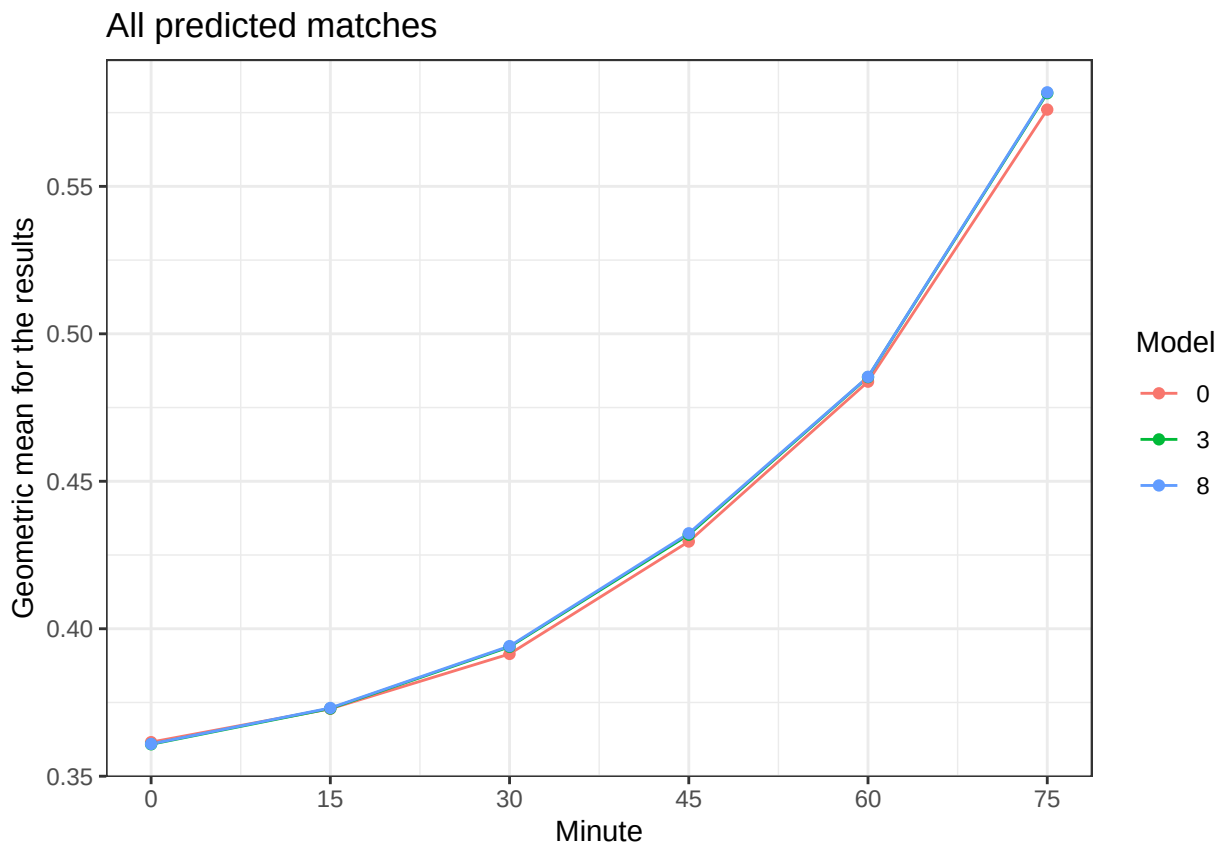
Geometric mean

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(knitr)

load("data/HDA_dc.RData")
load("data/first_matches.RData")

HDA = HDA_dc %>%
  anti_join(first_matches)

nrow(HDA)
```

```
## [1] 1858
```

```
HDA[,c(9:98)][which(HDA[,c(9:98)] == 0, arr.ind = TRUE)] = 10^-5
```

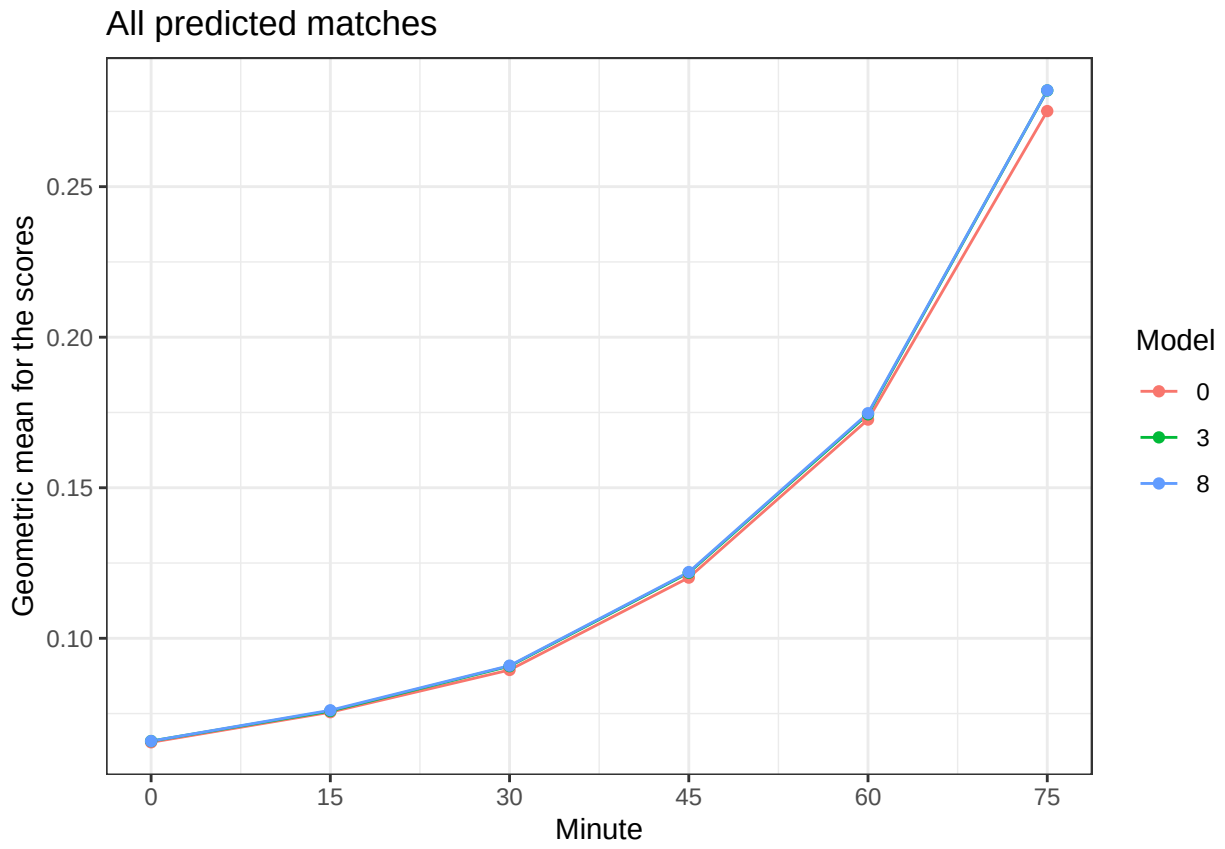
```
results = tibble(GeoMean = apply(HDA[,c(63:80)], 2, EnvStats::geoMean),
                 Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                 Model = factor(c(rep("0", 6),
                                  rep("3", 6),
                                  rep("8", 6))))
results %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches") +
  ylab("Geometric mean for the results")
```




```
results %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3615542	0.3728595	0.3914571	0.4295694	0.4837697	0.5760223
3	0.3608178	0.3729549	0.3938743	0.4318778	0.4853132	0.5815989
8	0.3609456	0.3731334	0.3941103	0.4323616	0.4854227	0.5818806

```
scores = tibble(GeoMean = apply(HDA[,c(81:98)], 2, EnvStats::geoMean),
                Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                Model = factor(c(rep("0", 6),
                                 rep("3", 6),
                                 rep("8", 6))))
scores %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches") +
  ylab("Geometric mean for the scores")
```




```
scores %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.0654583	0.0754529	0.0894830	0.1201492	0.1726362	0.2751038
3	0.0658623	0.0758537	0.0907471	0.1217557	0.1743915	0.2819345
8	0.0658657	0.0760945	0.0909170	0.1219899	0.1747439	0.2819894

```
load("~/GitHub/soccer-live-predictions/soccer-live-predictions/scrape/data/reds.RData")

matches = reds %>%
  filter(Season > 2015, Half == 1) %>%
  select(Season, Match)

HDA_reds = HDA %>%
  inner_join(matches)
```

```
## Joining, by = c("Season", "Match")
```

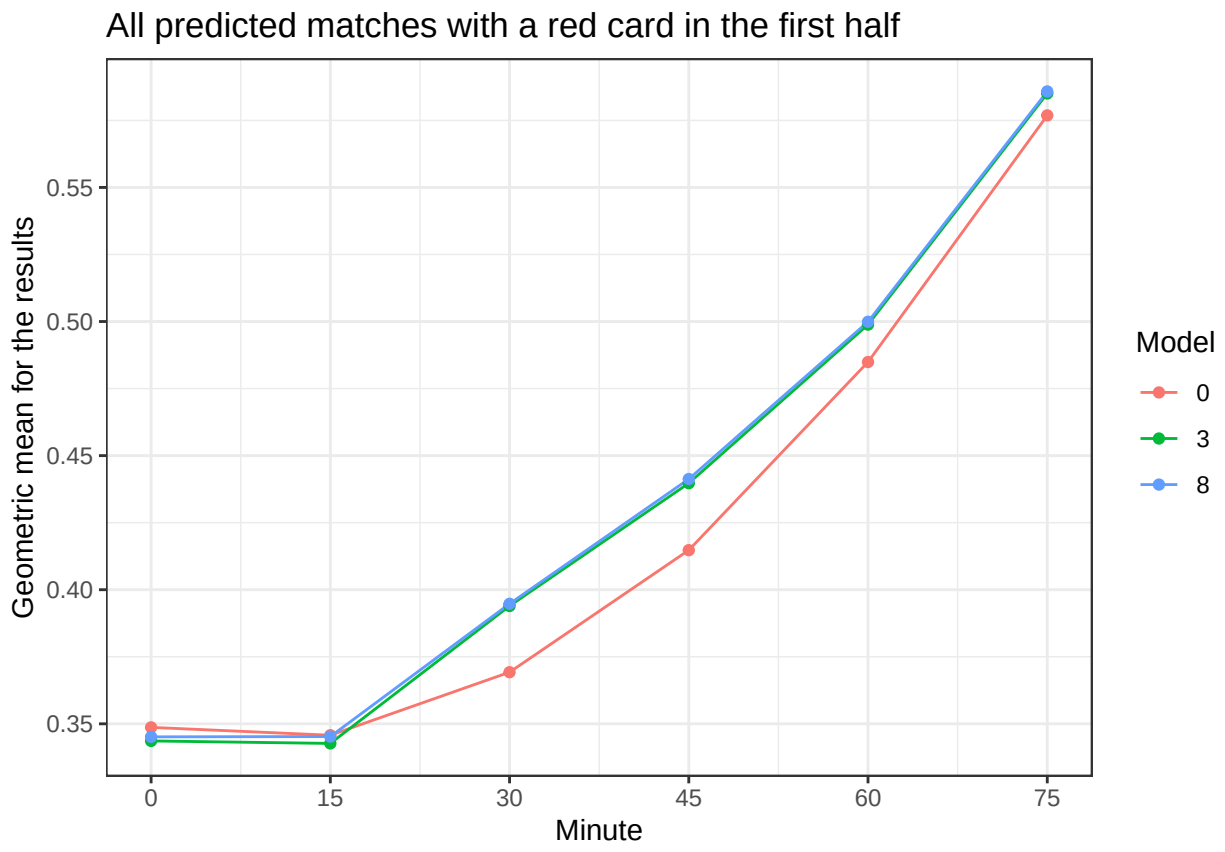
```
HDA_no_reds = HDA %>%
  anti_join(matches)
```

```
## Joining, by = c("Season", "Match")
```

```
nrow(HDA_reds)
```

```
## [1] 82
```

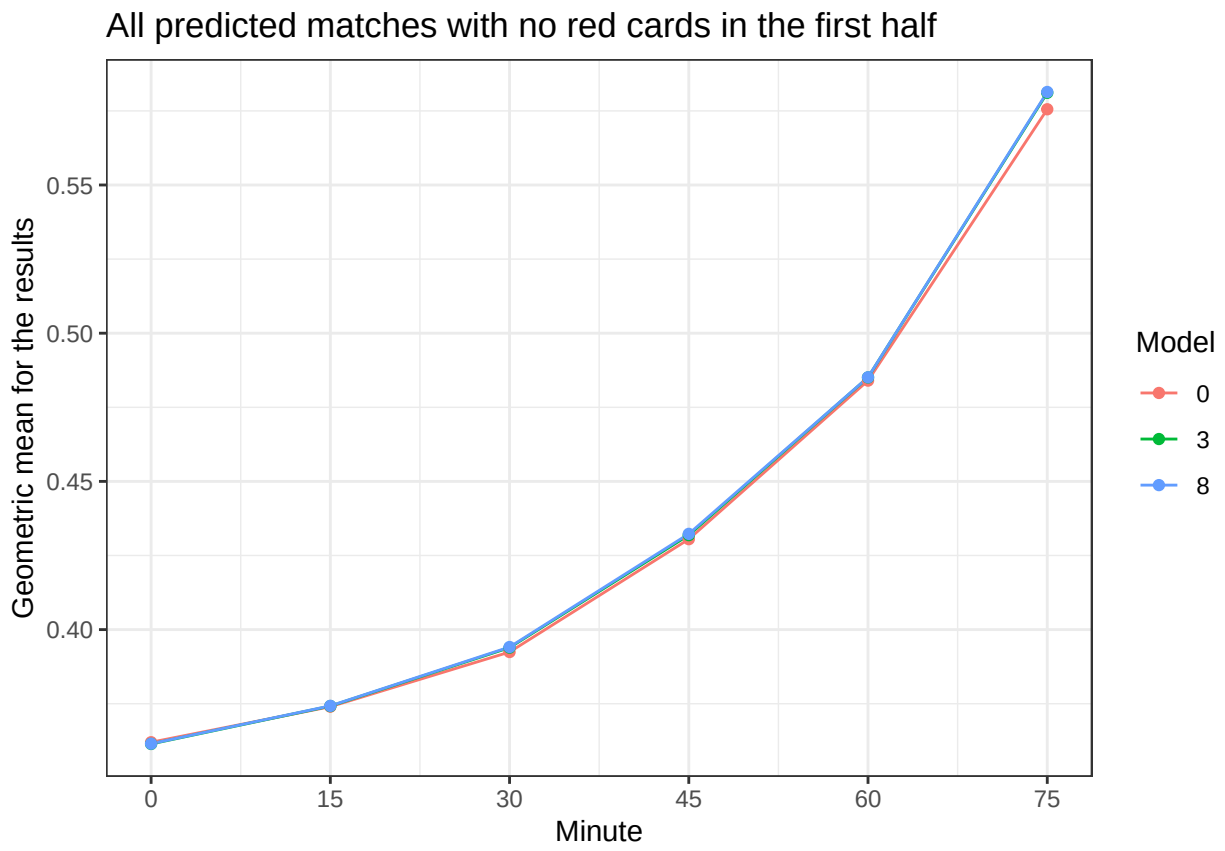
```
results_reds = tibble(GeoMean = apply(HDA_reds[,c(63:80)], 2, EnvStats::geoMean),
                      Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                      Model = factor(c(rep("0", 6),
                                       rep("3", 6),
                                       rep("8", 6))))
results_reds %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with a red card in the first half") +
  ylab("Geometric mean for the results")
```



```
results_reds %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3486645	0.3457368	0.3692382	0.4147450	0.4848982	0.5768597
3	0.3436144	0.3426827	0.3939457	0.4397161	0.4988566	0.5850398
8	0.3451358	0.3452399	0.3947515	0.4412706	0.4999056	0.5857873

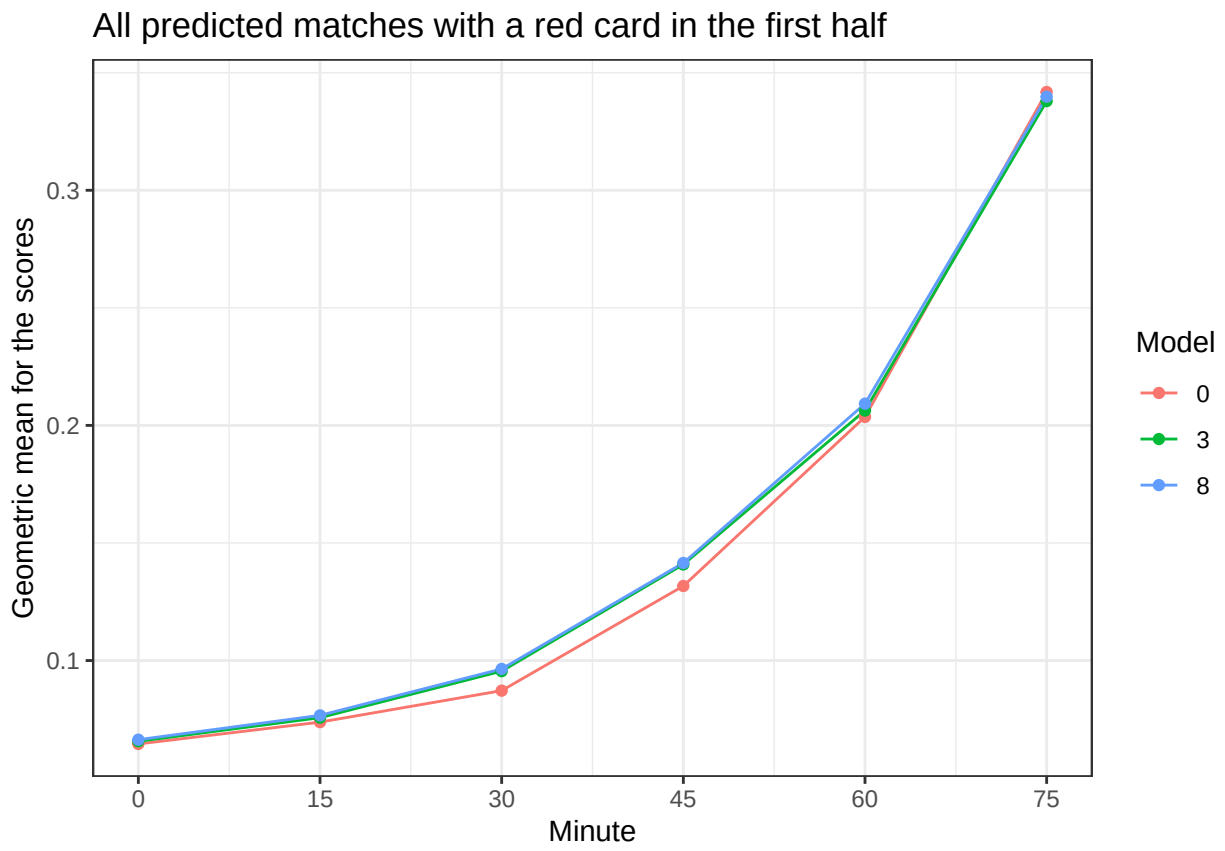
```
results_no_reds = tibble(GeoMean = apply(HDA_no_reds[,c(63:80)], 2, EnvStats::geoMean),
                         Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                         Model = factor(c(rep("0", 6),
                                          rep("3", 6),
                                          rep("8", 6))))
results_no_reds %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with no red cards in the first half") +
  ylab("Geometric mean for the results")
```



```
results_no_reds %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3620003	0.3739905	0.3924370	0.4305123	0.4840060	0.5755491
3	0.3614173	0.3741974	0.3939049	0.4318678	0.4850609	0.5811073
8	0.3615027	0.3742770	0.3941192	0.4322963	0.4851196	0.5813781

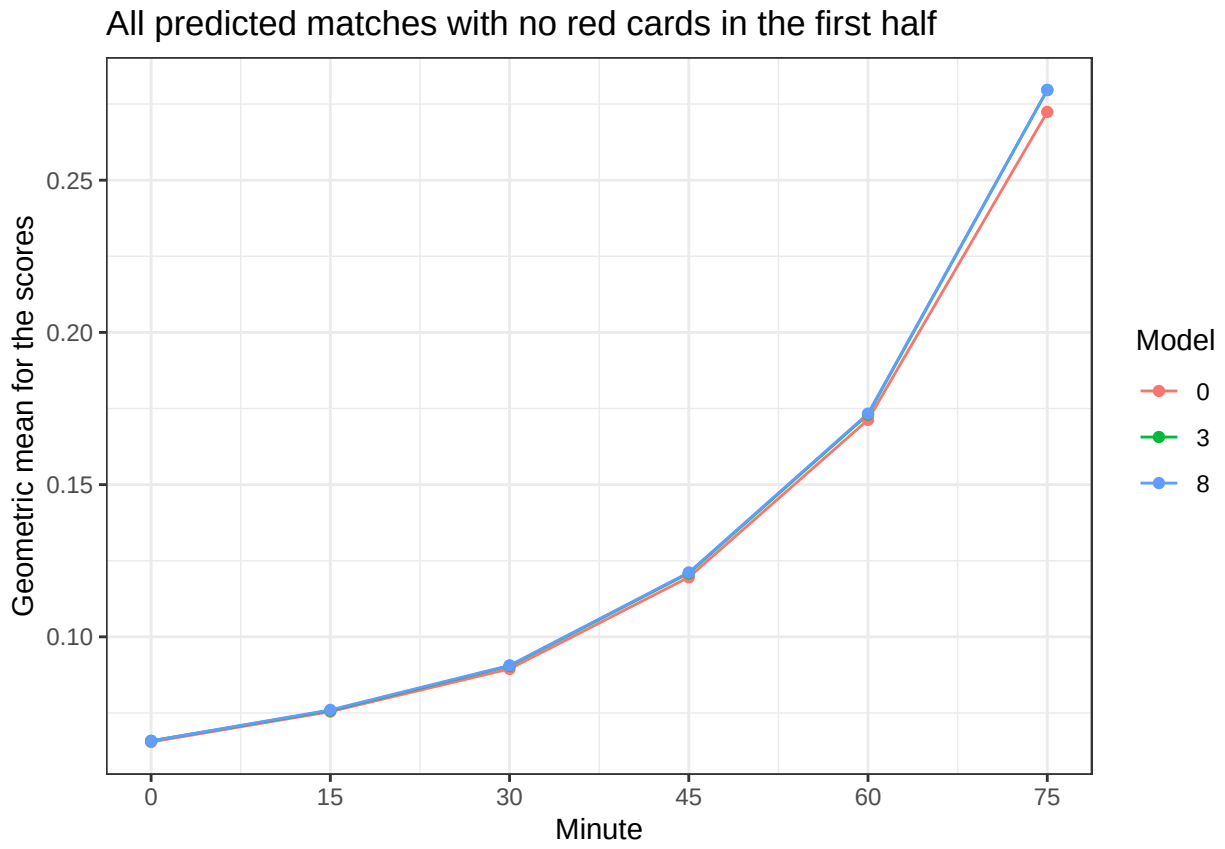
```
scores_reds = tibble(GeoMean = apply(HDA_reds[,c(81:98)], 2, EnvStats::geoMean),
                     Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                     Model = factor(c(rep("0", 6),
                                      rep("3", 6),
                                      rep("8", 6))))
scores_reds %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with a red card in the first half") +
  ylab("Geometric mean for the scores")
```



```
scores_reds %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.0645187	0.0738059	0.0871930	0.1316973	0.2036072	0.3417896
3	0.0656550	0.0756983	0.0954845	0.1408456	0.2063144	0.3379100
8	0.0663443	0.0766767	0.0964263	0.1414901	0.2092313	0.3397818

```
scores_no_reds = tibble(GeoMean = apply(HDA_no_reds[,c(81:98)], 2, EnvStats::geoMean),
                        Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                        Model = factor(c(rep("0", 6),
                                         rep("3", 6),
                                         rep("8", 6))))
scores_no_reds %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with no red cards in the first half") +
  ylab("Geometric mean for the scores")
```



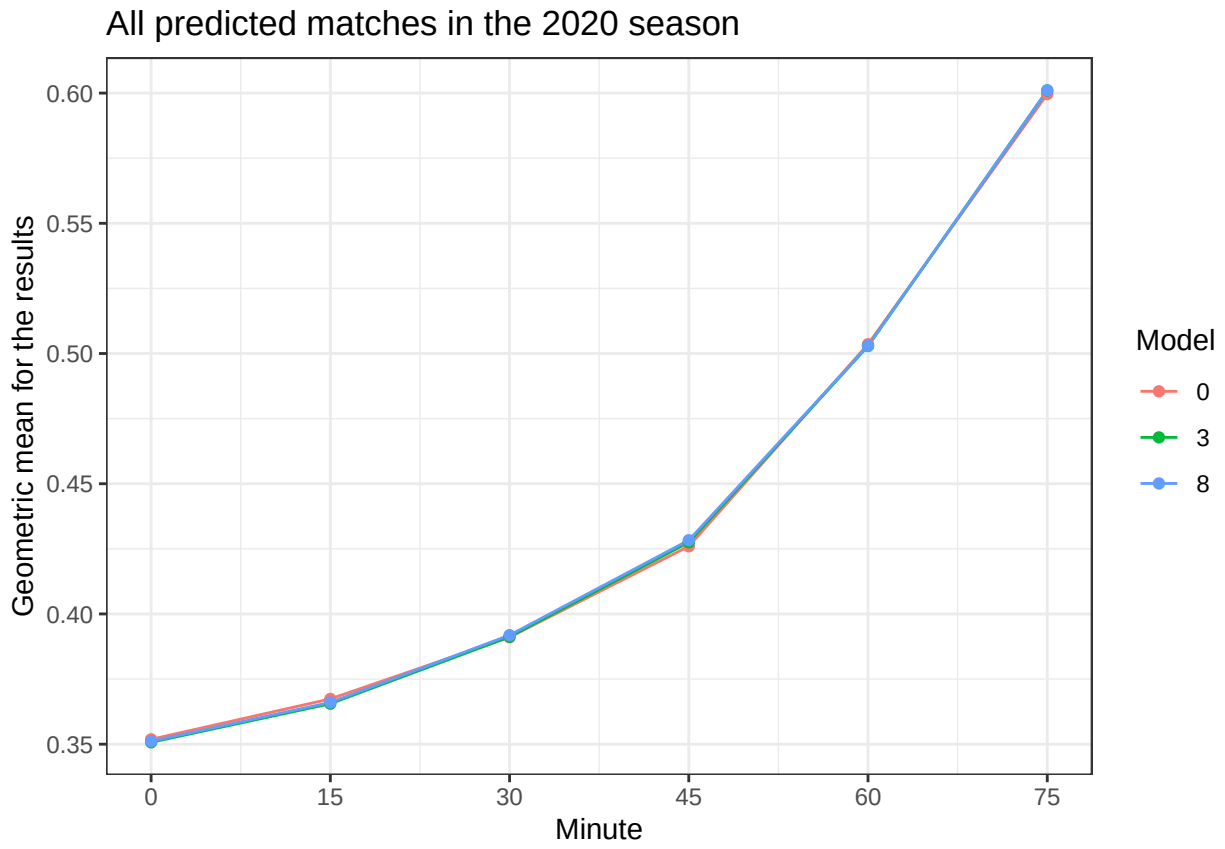
```r
scores_no_reds %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.0654788	0.0754585	0.0894830	0.1195775	0.1712945	0.2724196
3	0.0658462	0.0757965	0.0904839	0.1209245	0.1730690	0.2796281
8	0.0658215	0.0760032	0.0906152	0.1211394	0.1733257	0.2796183

```r
HDA_2020 = HDA %>%
  filter(Season == 2020)

nrow(HDA_2020)
```

```
## [1] 376
```

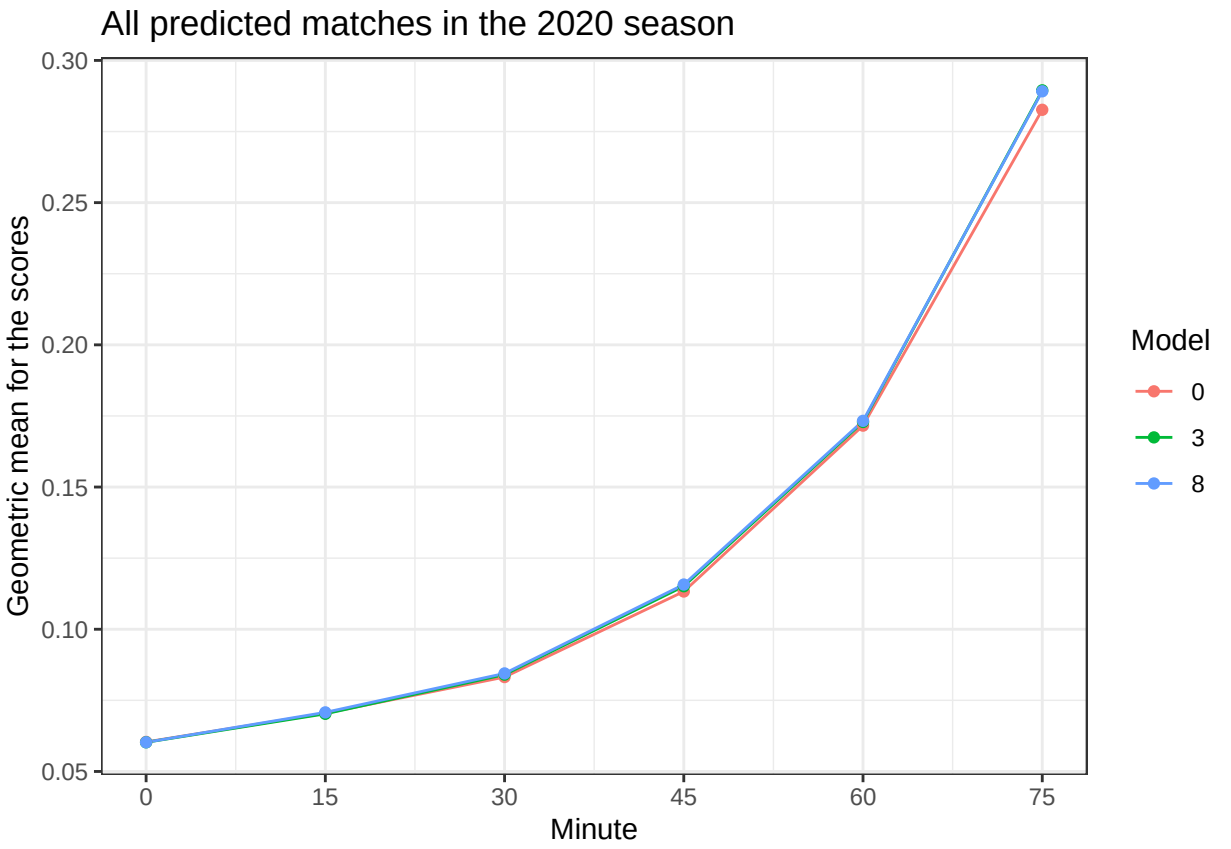
```r
results_2020 = tibble(GeoMean = apply(HDA_2020[,c(63:80)], 2, EnvStats::geoMean),
                      Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                      Model = factor(c(rep("0", 6),
                                       rep("3", 6),
                                       rep("8", 6))))
results_2020 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches in the 2020 season") +
  ylab("Geometric mean for the results")
```




```
results_2020 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3518546	0.3673683	0.3910833	0.4259717	0.5035654	0.5996244
3	0.3507179	0.3655351	0.3912212	0.4275058	0.5028590	0.6010051
8	0.3511148	0.3660320	0.3918488	0.4283029	0.5029385	0.6010540

```
scores_2020 = tibble(GeoMean = apply(HDA_2020[,c(81:98)], 2, EnvStats::geoMean),
                     Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                     Model = factor(c(rep("0", 6),
                                      rep("3", 6),
                                      rep("8", 6))))
scores_2020 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches in the 2020 season") +
  ylab("Geometric mean for the scores")
```




```
scores_2020 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.0603763	0.0704758	0.0831918	0.1132685	0.1715947	0.2826407
3	0.0602509	0.0702607	0.0840340	0.1151168	0.1728721	0.2894863
8	0.0602494	0.0707715	0.0844645	0.1157232	0.1732687	0.2891548

```
load("~/GitHub/soccer-live-predictions/soccer-live-predictions/scrape/data/results.RData")
load("~/GitHub/soccer-live-predictions/soccer-live-predictions/scrape/data/goals.RData")
```

```
at_45 = results %>%
  select(Season, Match) %>%
  filter(Season > 2015)
```

```
home_score_at_45 <- function(season, match) {
  goals %>%
    filter(Season == season,
           Match == match,
           Team == 1,
           Half == 1) %>%
```

```
    nrow()
}

away_score_at_45 <- function(season, match) {
  goals %>%
    filter(Season == season,
           Match == match,
           Team == 2,
           Half == 1) %>%
    nrow()
}
```

```
at_45 = at_45 %>%
  rowwise() %>%
  mutate(Home_Score = home_score_at_45(Season, Match),
         Away_Score = away_score_at_45(Season, Match),
         abs_dif = abs(Home_Score - Away_Score))
```

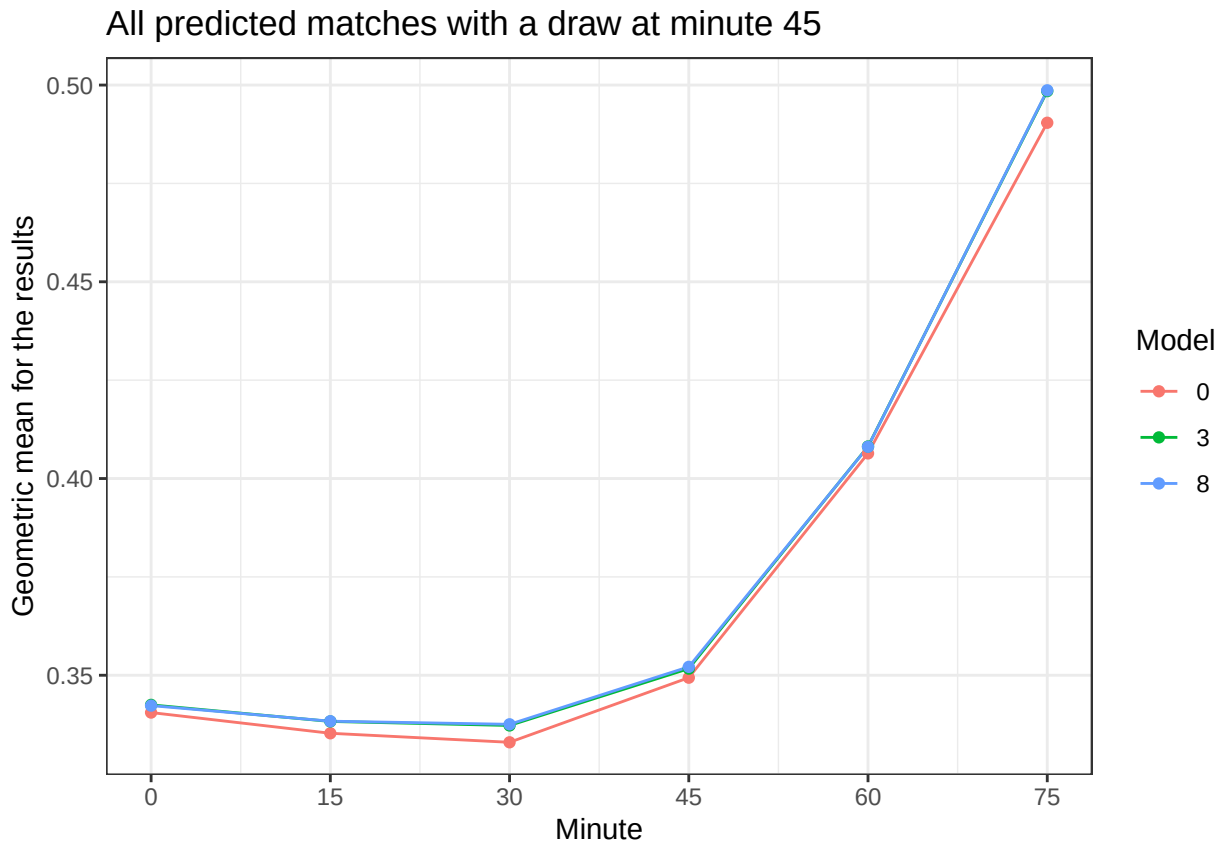
```
tmp_00 = at_45 %>%
  filter(abs_dif == 0) %>%
  select(Season, Match)

HDA_00 = HDA %>%
  inner_join(tmp_00)
```

```
## Joining, by = c("Season", "Match")
```

```
nrow(HDA_00)
```

```
## [1] 838
```

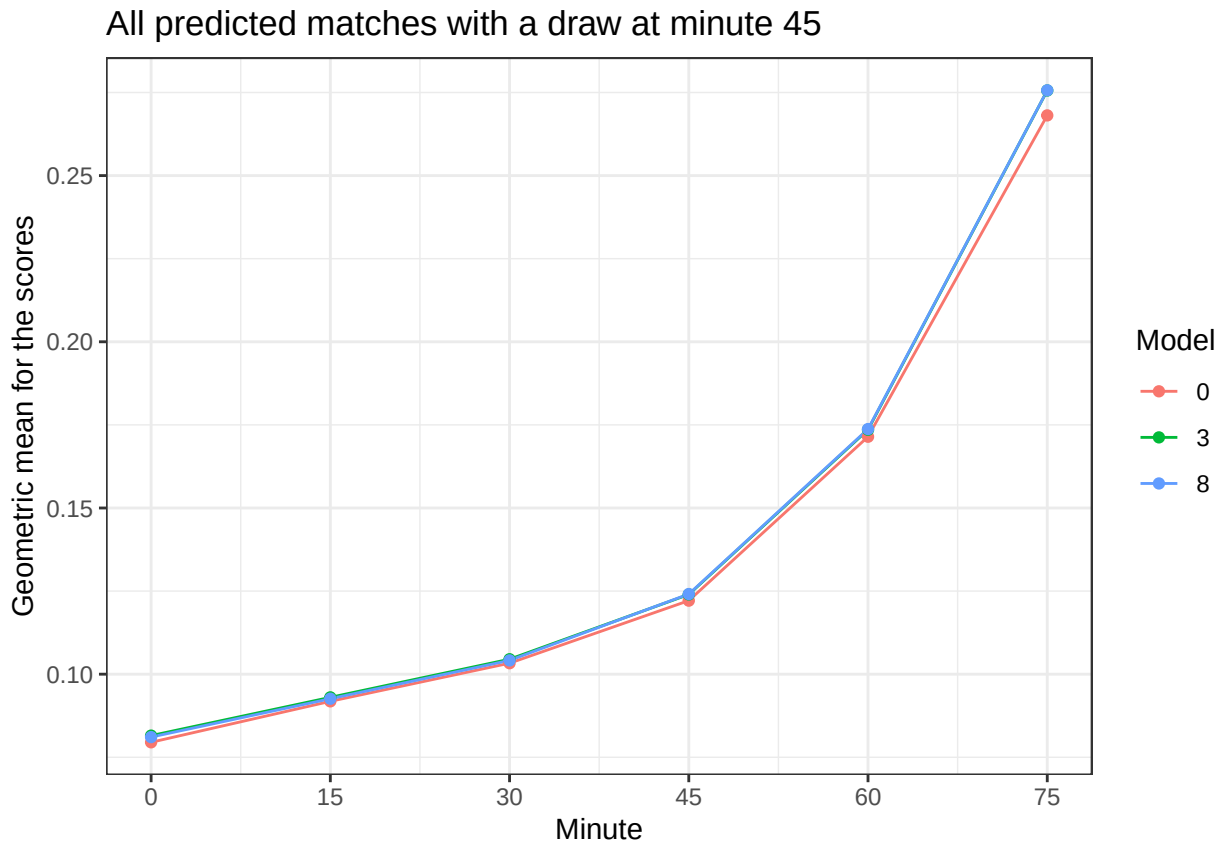
```
results_00 = tibble(GeoMean = apply(HDA_00[,c(63:80)], 2, EnvStats::geoMean),
                    Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                    Model = factor(c(rep("0", 6),
                                     rep("3", 6),
                                     rep("8", 6))))
results_00 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with a draw at minute 45") +
  ylab("Geometric mean for the results")
```




```
results_00 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3405184	0.3352532	0.3329615	0.3493841	0.4063878	0.4904210
3	0.3424718	0.3382645	0.3372412	0.3516865	0.4081422	0.4984660
8	0.3422894	0.3383228	0.3375331	0.3521376	0.4080746	0.4986579

```
scores_00 = tibble(GeoMean = apply(HDA_00[,c(81:98)], 2, EnvStats::geoMean),
                   Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                   Model = factor(c(rep("0", 6),
                                    rep("3", 6),
                                    rep("8", 6))))
scores_00 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with a draw at minute 45") +
  ylab("Geometric mean for the scores")
```




```
scores_00 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.0795046	0.0918301	0.1033027	0.1221252	0.1714369	0.2681104
3	0.0815104	0.0930254	0.1044979	0.1239824	0.1735399	0.2756050
8	0.0810795	0.0926119	0.1041446	0.1241021	0.1737373	0.2756520

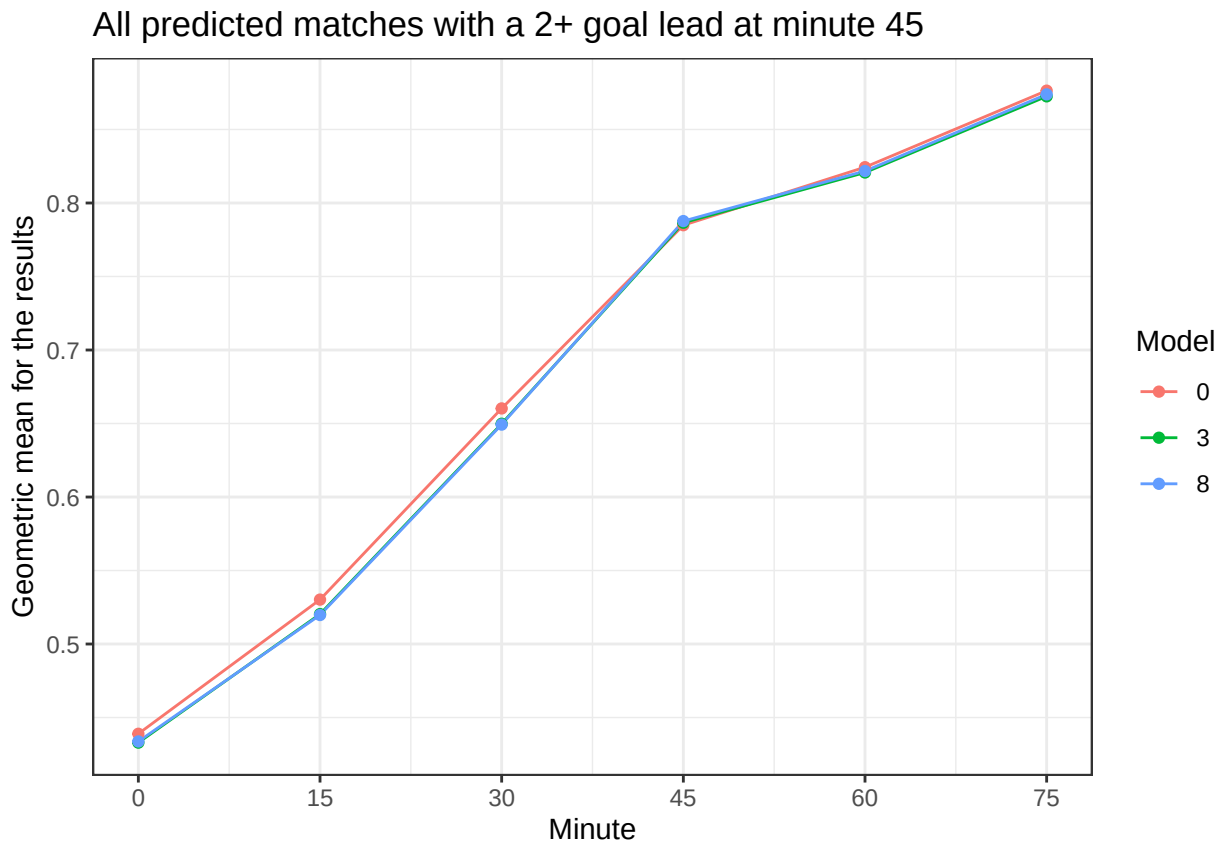
```
tmp_20 = at_45 %>%
  filter(abs_dif >= 2) %>%
  select(Season, Match)

HDA_20 = HDA %>%
  inner_join(tmp_20)
```

```
## Joining, by = c("Season", "Match")
```

```
nrow(HDA_20)
```

```
## [1] 211
```

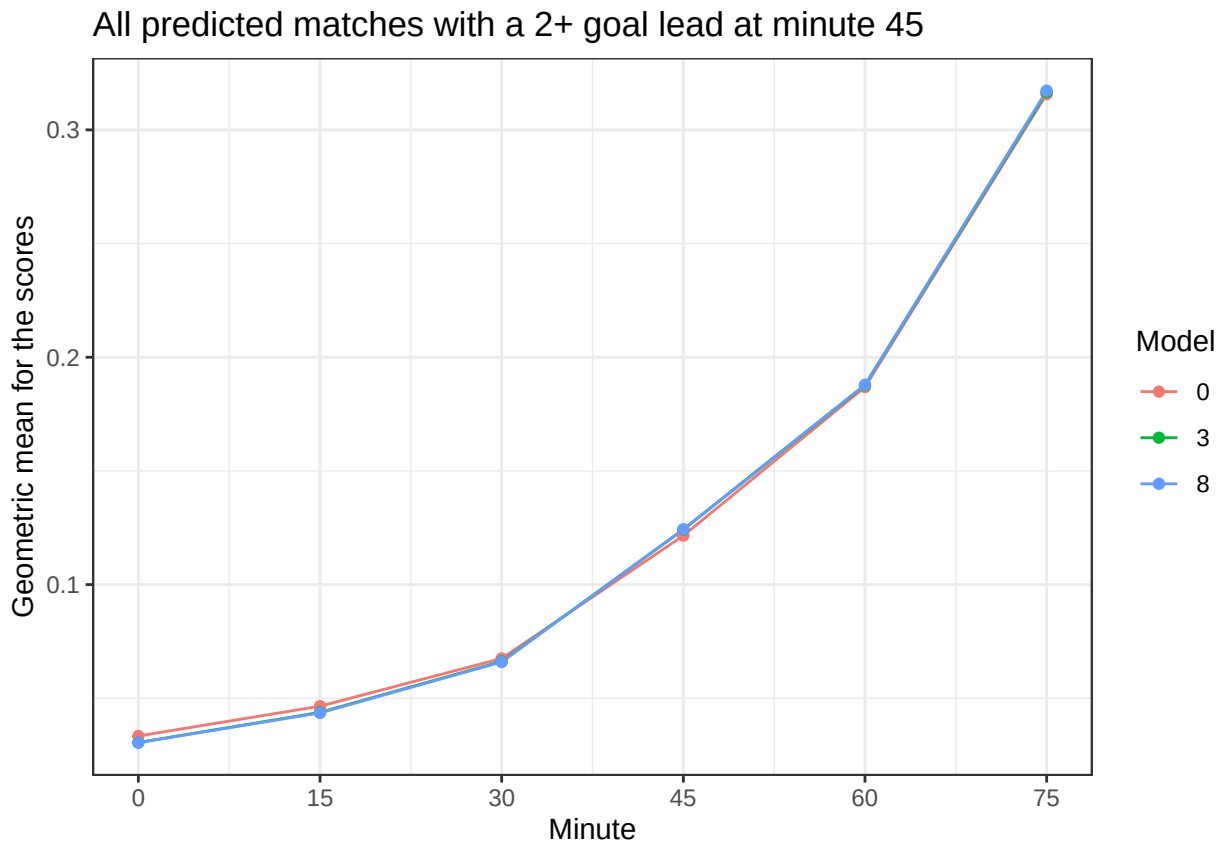
```
results_20 = tibble(GeoMean = apply(HDA_20[,c(63:80)], 2, EnvStats::geoMean),
                    Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                    Model = factor(c(rep("0", 6),
                                     rep("3", 6),
                                     rep("8", 6))))
results_20 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with a 2+ goal lead at minute 45") +
  ylab("Geometric mean for the results")
```




```
results_20 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.4389412	0.5302146	0.6602535	0.7848906	0.8243508	0.8764061
3	0.4329183	0.5203485	0.6498265	0.7865919	0.8207394	0.8725380
8	0.4337269	0.5197316	0.6492730	0.7876867	0.8217482	0.8738818

```
scores_20 = tibble(GeoMean = apply(HDA_20[,c(81:98)], 2, EnvStats::geoMean),
                   Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                   Model = factor(c(rep("0", 6),
                                    rep("3", 6),
                                    rep("8", 6))))
scores_20 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches with a 2+ goal lead at minute 45") +
  ylab("Geometric mean for the scores")
```




```
scores_20 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.0334097	0.0465290	0.0675163	0.1215685	0.1868990	0.3156954
3	0.0305472	0.0438189	0.0662227	0.1241918	0.1876758	0.3165560
8	0.0304989	0.0435676	0.0659792	0.1243353	0.1878899	0.3172397