Project 3 - London Marathon Predictions

William Sorg.74

2025-04-27

Question Formulation

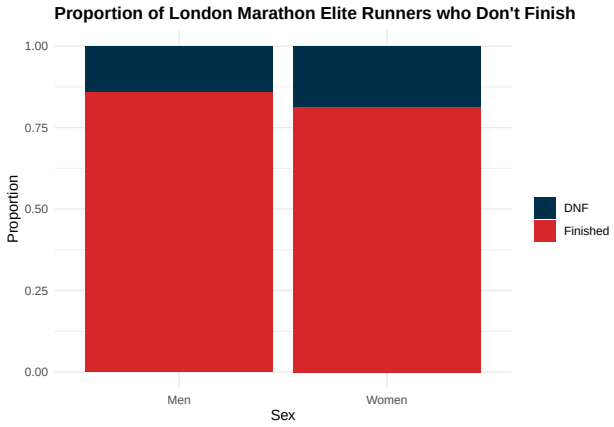
The London Marathon is one the most prestigious races in the world. As one of the Abbott World Marathon Majors, the London Marathon attracts over 50,000 runners to compete each year, along with approximately 750,000 spectators. Among the runners are some of the greatest elite distance runners in the world. Eliud Kipchoge, Mo Farah, Sifan Hassan, and the late Kelvin Kiptum have all raced the streets of London.
Despite the marathons's significance, it's difficult to find statistically driven predictions for the race. Nearly every other sport has some sort of win probability model, but there is a void for runnning. In a sport that can witness extreme unpredictability, I believe it would be interesting and informative to build a model that predicts the winner and top finishers of the London Marathon. Given that the 2025 London Marathon will begin at 4:00 a.m. EST on April 27th, this project will focus on predicting the results of the 2025 race. I will then briefly evaluate the accuracy of the predictions after the race concludes.

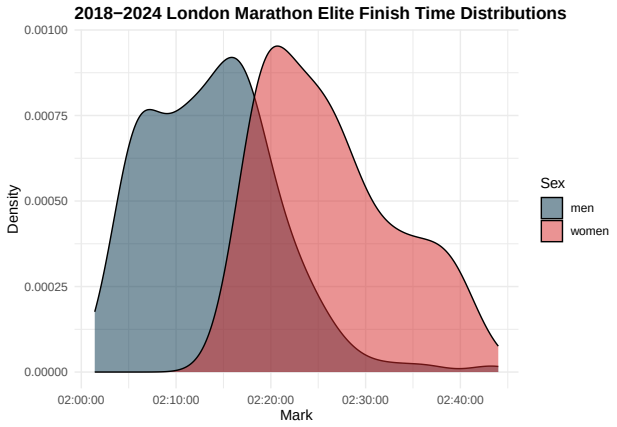
Data Selection

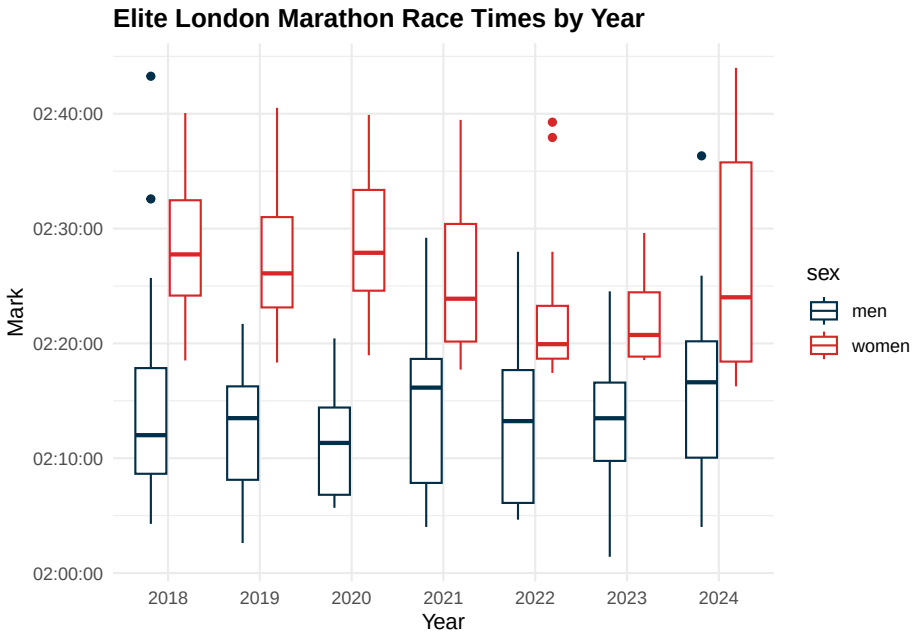
This project will use the 2018-2024 London Marathons to train and test different models with the goal of developing predictions for the 2025 London Marathon on April 27th. The London Marathon Elite runner results were scraped from each year's London Marathon results website. Then, for each athlete, every past race result was scraped off of the World Athletics website. From this data, several different statistics were calculated. These include number of races, number of finishes, PR, worst time, average time, time standard deviation, and most recent result for marathons, half marathons, and 10Ks. In total, this dataset contains 383 athletes when the 2025 elite field is included. All code for scraping can be found on my GitHub. (Note, scraping the data was the biggest time requirement of this project.)

```
## # A tibble: 6 x 36
##   raceDate   Name             Nat   continent sex   age   Mark  Place dob
##   <date>     <chr>            <chr> <chr>     <chr> <drt> <Per> <int> <date>
## 1 2018-04-22 Guye Idemo ADOLA ETH   Africa    men   1004~ 9155S    17 1990-10-20
## 2 2018-04-22 Tracy BARLOW     GBR   Europe    women    N~ 9129S     9 NA
## 3 2018-04-22 Kenenisa BEKELE  ETH   Africa    men   1309~ 7733S     6 1982-06-13
## 4 2018-04-22 Tadelech BEKELE  ETH   Africa    women  987~ 8500S     3 1991-04-11
## 5 2018-04-22 Stephanie BRUCE  USA   Americas  women    N~ 9148S    10 NA
## 6 2018-04-22 Fernando CABADA  USA   Americas  men   1314~ 8259S    13 1982-04-22
## # i 27 more variables: lastRaceDate <date>, numMarathons <int>,
## #   finishesMarathons <int>, finishpctMarathons <dbl>, prMarathon <dbl>,
## #   worstMarathon <dbl>, avgMarathon <dbl>, sdMarathon <dbl>,
## #   recentMarathon <Period>, marathons <list>, numHalfMarathons <int>,
## #   finishesHalfMarathons <int>, finishpctHalfMarathons <dbl>,
## #   prHalfMarathon <dbl>, worstHalfMarathon <dbl>, avgHalfMarathon <dbl>,
## #   sdHalfMarathon <dbl>, recentHalfMarathon <Period>, ...
```

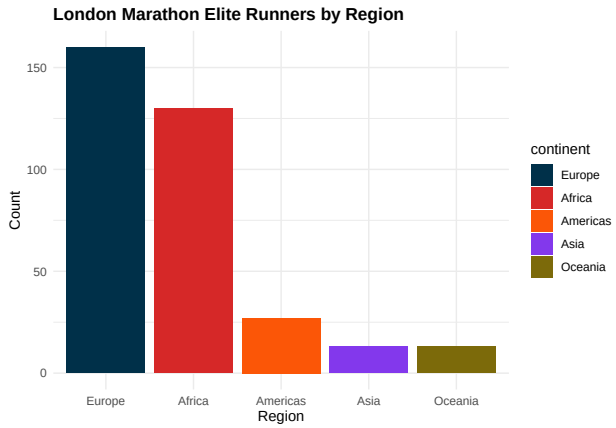
Exploratory Analysis

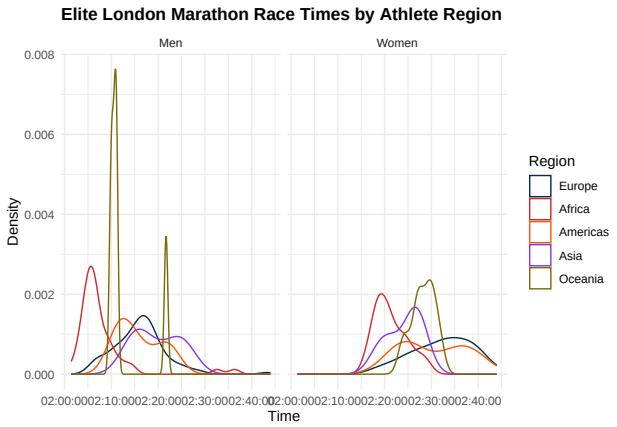




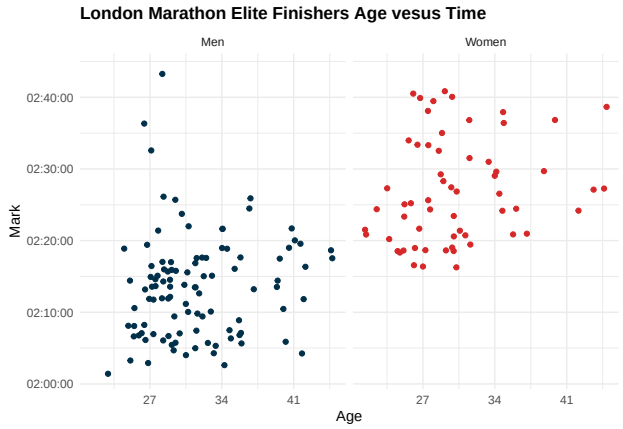


It appears that around 85% of elite runners don't finish the London Marathon, with women finishing at a marginally lower rate. Of the men who finish, the distribution of finish times is concentrated between 2:04 and 2:20. The distribution also appears to be somewhat bimodal with spikes at 2:05 and 2:15. This could be showcasing the difference between the typical lead pack and the rest of the elite field who may not be truly challenging for the win. For the women, most times are between 2:18 and 2:30. The women's times also tail off slower, with a noticeable number of finishers coming in a little under 2:40. The year by year races depict the variation between races. There isn't a noticeable trend over time outside of each year being unique. Some of this is likely caused by weather and elite field quality, but I believe it mostly comes down to race tactics. The London Marathon is able to consistently draw elite talent, so I don't believe one of these years would have a much lower or higher quality elite field. Also, 2022 for example, saw a fast women's race, but an average or below average men's race. If it were weather, I would expect both the men's and women's races to have the same trend. Overall, these plots show that some elite runners don't finish, that finish times vary, and that each race injects its own uniqueness.

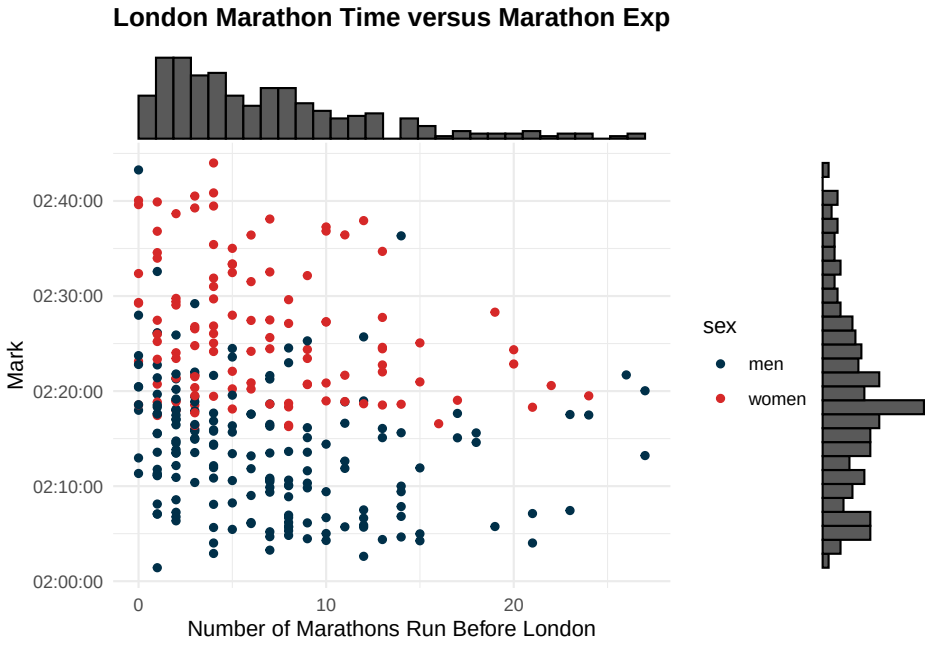


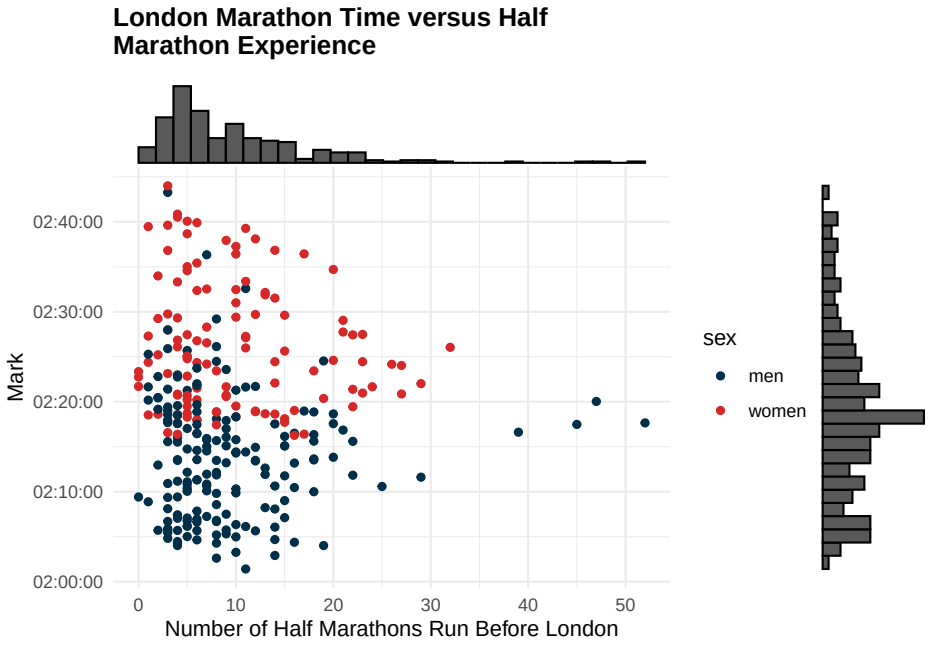


The London Marathon elite field is made up of mostly European and African athletes. Since 2018, the African athletes clearly tend to finish faster than everyone else. Outside of Africa, I don't think there is a noticeable trend by region, nor is there a large enough sample size for some of the regions to draw a conclusion. Because of this, I think it makes sense to evaluate by an athlete's region as Africa or rest of the world.

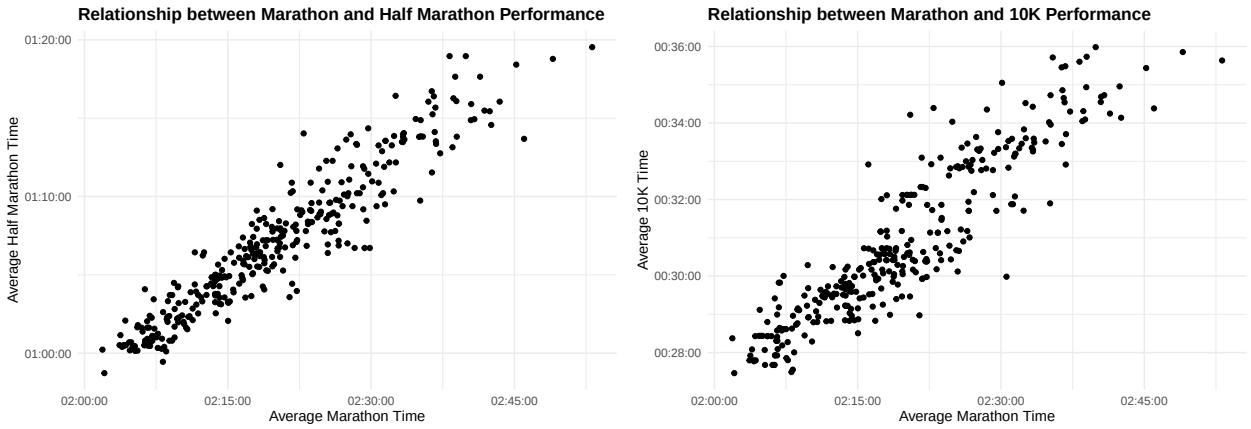


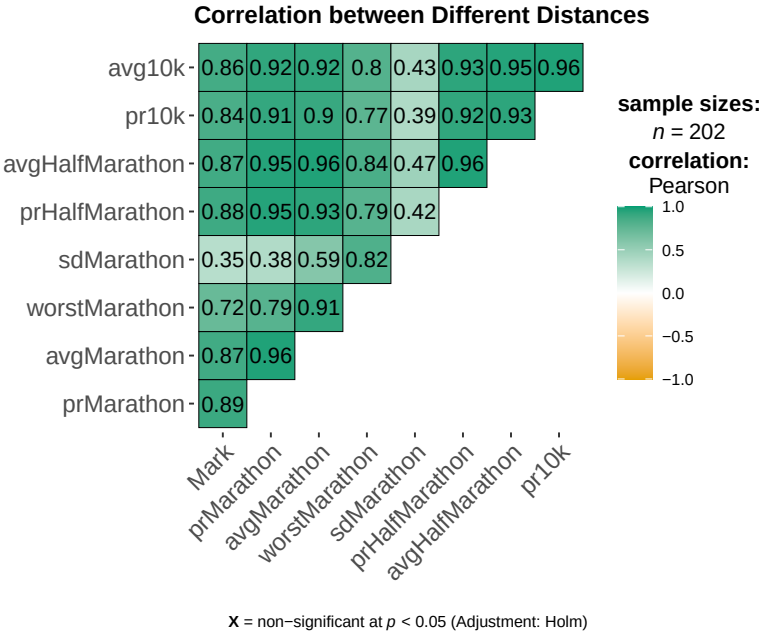
Based on these two plots, I don't see any discernible relationship between age and finish time. I still believe it may be interesting to cluster with age as an input, but I don't think age will be an important factor in determining who will win the London Marathon.



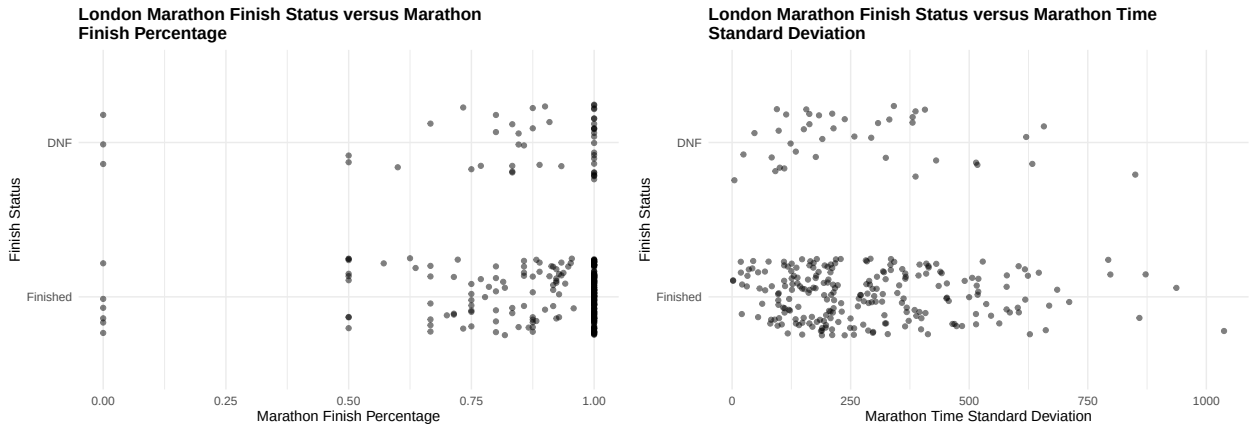


Similar to age, the number of either marathons or half marathons run doesn't seem to have a large impact on race success. Again, I think it will still be interesting to use race experience to help cluster athletes so we can see which athletes are similar, but I don't think there is a clear takeaway from the above plots.

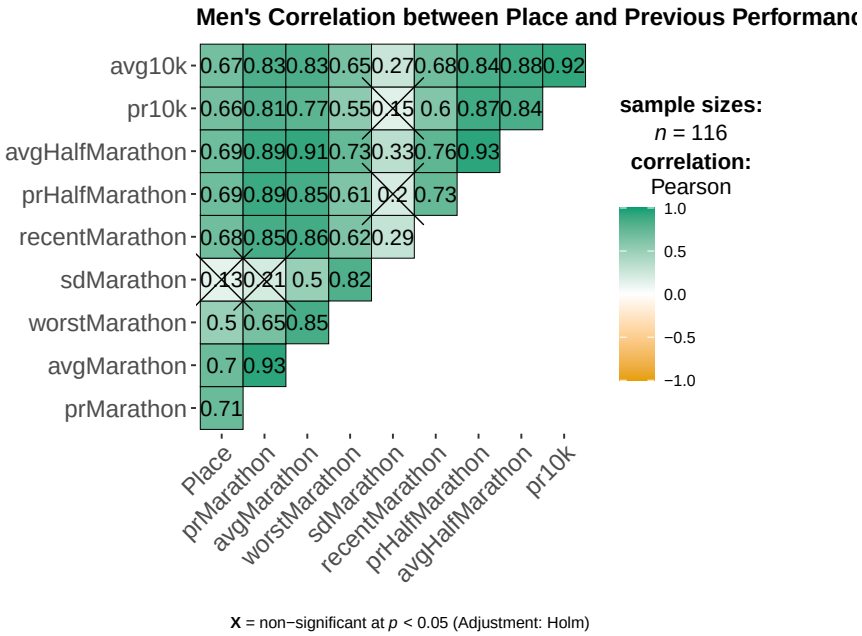


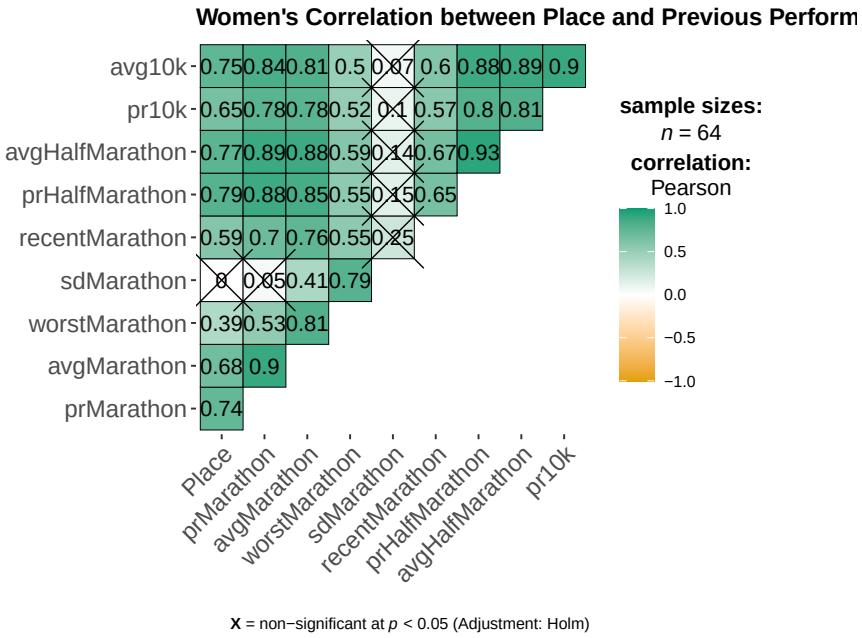


As one would expect, full and half marathon race times are closely related. 10k times are also closely related to marathon times, but there is a little more noise in the relationship. These relationships will be helpful creating variables with very few missing values. Every year at London, there are typically a few athletes making their marathon debut. Because of this, they don't have a marathon PR or any time for reference. Some years, these athletes could probably be removed from consideration. However, one of the favorites in the 2025 race is Jacob Kiplimo, who will be making his marathon debut. Kiplimo set the half marathon world record in February with a time of 56:42, so he definitely will be a factor in the race. To include Kiplimo and other debut runners, I will use a linear model to convert times between distances.

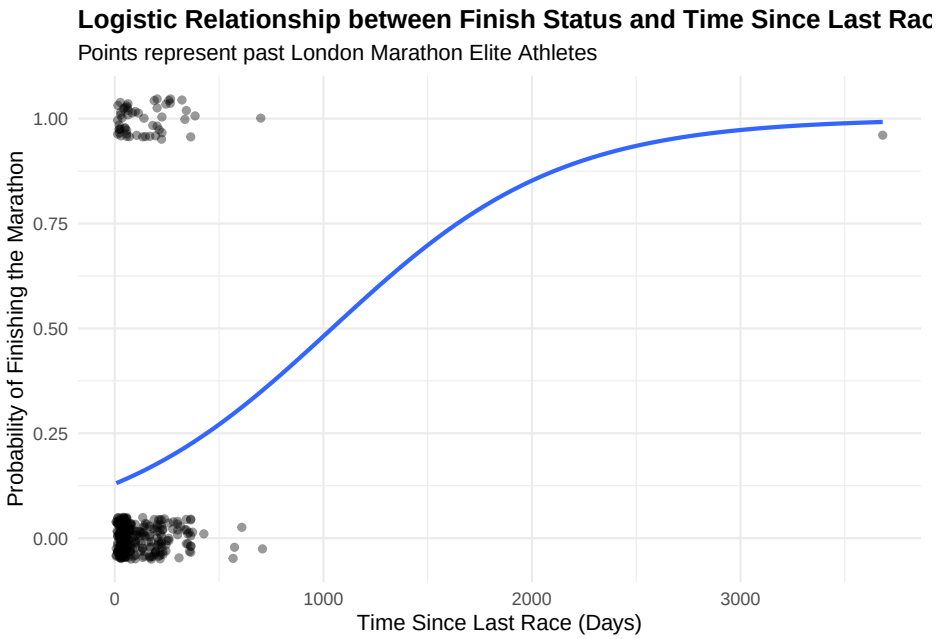


These plots show no significant evidence of a relationship between past DNFs or marathon consistency and not finishing the London Marathon. Based on this and previous plots, I think it will be difficult to correctly predict who won't finish the race. I also believe a probability prediction will also be more informative because I don't think many runners will have a discrete prediction of not finishing.





These correlation plots mainly focus on the correlation between previous times and finish place. It looks like PR times are possibly the best predictor of success. More noticeable is the weak relationship between worst marathon time and place. This makes sense though because some athletes probably haven't raced enough to have had a truly bad marathon time. Also, some athletes would rather drop out than finish the race slowly. Also, the insignificance of the marathon time standard deviation stands out.



Time since the last race doesn't seem to be a very helpful variable beyond highlighting athletes that haven't raced in an extremely long time. For this, though, I do think it could help inform predictions, especially for whether or not an athlete will finish the race.

Analysis

For the race predictions, I want to look at the probability of finishing and then at the finish time and placement. First, though, I want to create a subset of the original dataset and fill in some missing values.

Subsetting Data

```
HalftoFull <- lm(prMarathon ~ prHalfMarathon, data = LMAnalysisData %>%
                   filter(is.finite(prMarathon) & is.finite(prHalfMarathon)))
HalftoFull
```

```
##
## Call:
## lm(formula = prMarathon ~ prHalfMarathon, data = LMAnalysisData %>%
##     filter(is.finite(prMarathon) & is.finite(prHalfMarathon)))
##
## Coefficients:
##    (Intercept)  prHalfMarathon
##        437.450           1.983
```

```
tenktoFull <- lm(prMarathon ~ pr10k, data = LMAnalysisData %>%
                   filter(is.finite(prMarathon) & is.finite(pr10k)))
tenktoFull
```

```
##
## Call:
## lm(formula = prMarathon ~ pr10k, data = LMAnalysisData %>% filter(is.finite(prMarathon) &
##     is.finite(pr10k)))
##
## Coefficients:
## (Intercept)        pr10k
##    1034.362        3.981
```

```
## # A tibble: 1 x 4
##   Name          Nat   prHalfMarathon marathonEffort
##   <chr>         <chr>          <dbl>          <dbl>
## 1 Jacob KIPLIMO UGA             3402          7183.
```

After applying the conversions, Jacob Kiplimo's half marathon world record is converted into a 1:59:43. This would be a world record in its own right and seems a little ambitious, but I don't think it's that different from a 56:42 half, which he ran.

```
## # A tibble: 6 x 11
##   raceDate   Name  Africa sex   Mark  Place finished marathonEffort numMarathons
##   <date>     <chr> <lgl>  <chr> <Per> <int> <lgl>              <dbl>        <int>
## 1 2018-04-22 Guye~ TRUE   men   9155S    17 TRUE                7426            1
## 2 2018-04-22 Trac~ FALSE  women 9129S     9 TRUE                9042            9
## 3 2018-04-22 Kene~ TRUE   men   7733S     6 TRUE                7383            8
## 4 2018-04-22 Tade~ TRUE   women 8500S     3 TRUE                8514           11
## 5 2018-04-22 Step~ FALSE  women 9148S    10 TRUE                8975            5
## 6 2018-04-22 Fern~ FALSE  men   8259S    13 TRUE                7896           17
## # i 2 more variables: numHalfMarathons <int>, timesince <drtn>
```

Will they finish?

These models will attempt to predict whether or not an elite athlete will finish the London Marathon.

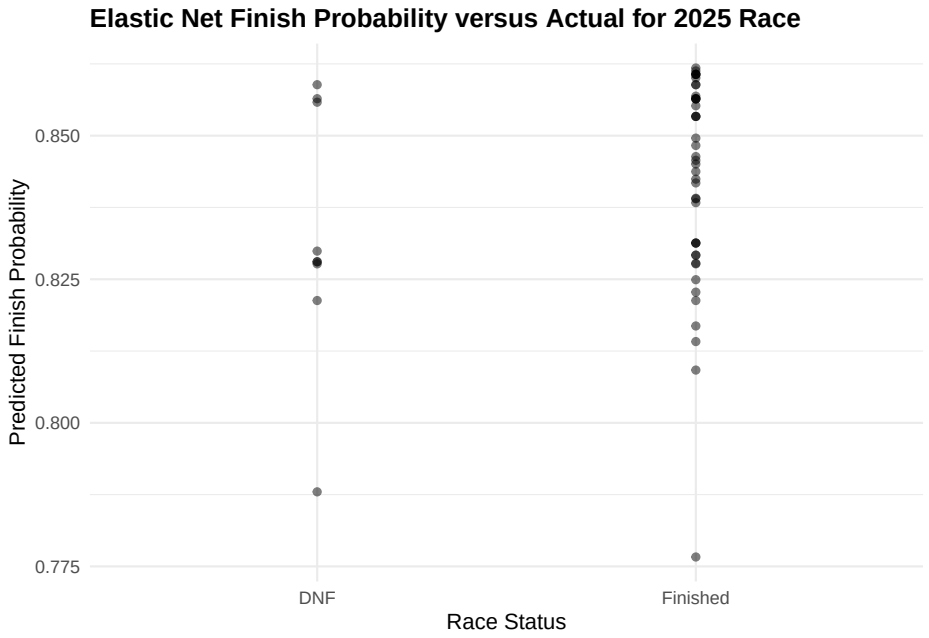
Elastic Net Logistic Regression

```
x <- data.matrix(trainData[,c(3:4, 8:11)])
y <- trainData$finished


ridge_lambda_out <- cv.glmnet(x, y, nfolds = 10, type.measure = "class",
                              alpha = 0.5, family = 'binomial')

lambda <- ridge_lambda_out$lambda.min

ridge_out <- glmnet(x, y, alpha = 0.5,
                    lambda = lambda,
                    family = "binomial")
```

```
ridge_out$beta
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## Africa           -0.2060971112
## sex               .
## marathonEffort    .
## numMarathons      .
## numHalfMarathons  .
## timesince        -0.0007116192
```

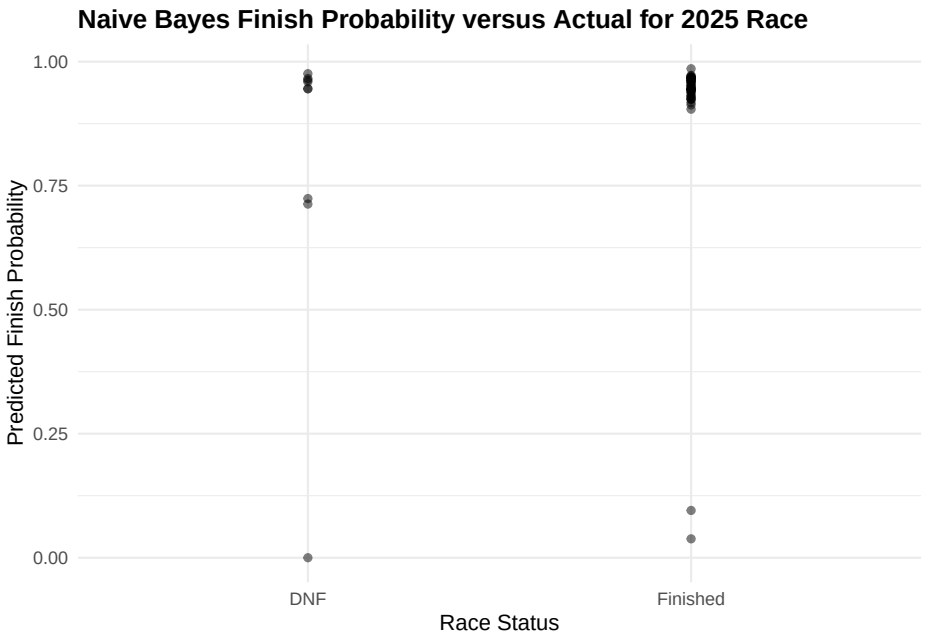
These low beta values suggest that there isn't a very strong relationship between these variables and finishing the race. This is also seen in the predictions since the probability of finishing doesn't vary much.



Naive Bayes

```
nbout <- naiveBayes(finished ~ sex + Africa + numMarathons + timesince,
                    data = trainData, type = 'raw')
```

The naive Bayes predictions are different from the prior model, but they don't appear any more accurate. Actually, they seem to be less realistic since I don't think it makes sense to have a less than 12.5% chance of finishing.

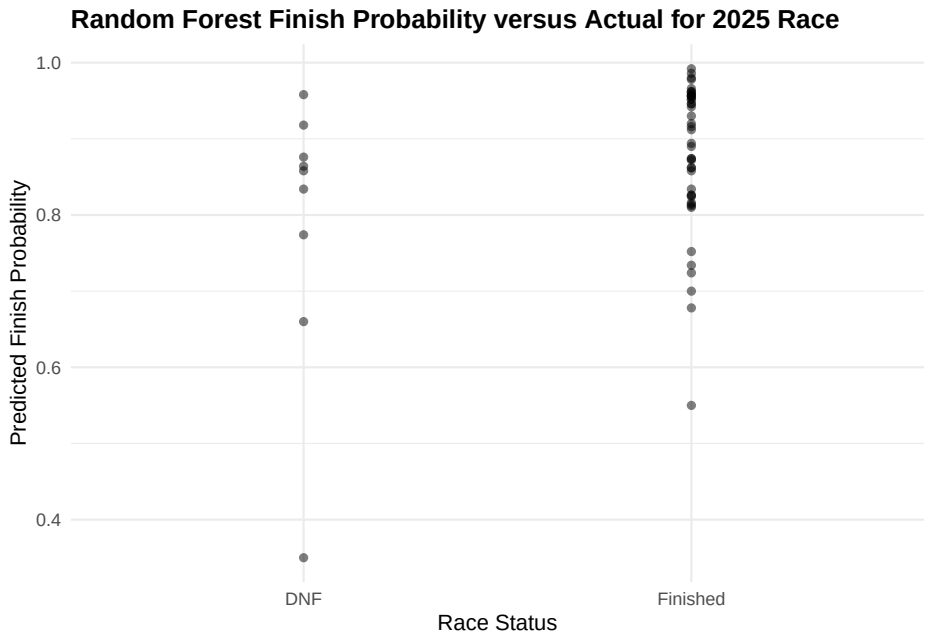


Random Forest

```
rf1 <- randomForest(finished ~ sex + Africa + numMarathons + timesince,
             data = trainData)
rf1
```

```
##
## Call:
##  randomForest(formula = finished ~ sex + Africa + numMarathons +      timesince, data = trainData)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 16.91%
## Confusion matrix:
##       FALSE TRUE class.error
## FALSE     2   53  0.96363636
## TRUE      5  283  0.01736111
```

The random forest predictions seem more realistic than naive Bayes and have more range than the elastic net model. However, like the other two, the model does not perform particularly well on the 2025 race.



Overall, the poor accuracy of these models isn't surprising given the exploratory analysis. It didn't seem like there were any variables with a strong correlation with not finishing. Because of this, it's nearly impossible to construct a model that performs well. I think it's probably more accurate to assume that all elite runners have the same probability of not finishing.

Finish Time Predictions

Elastic Net Regression

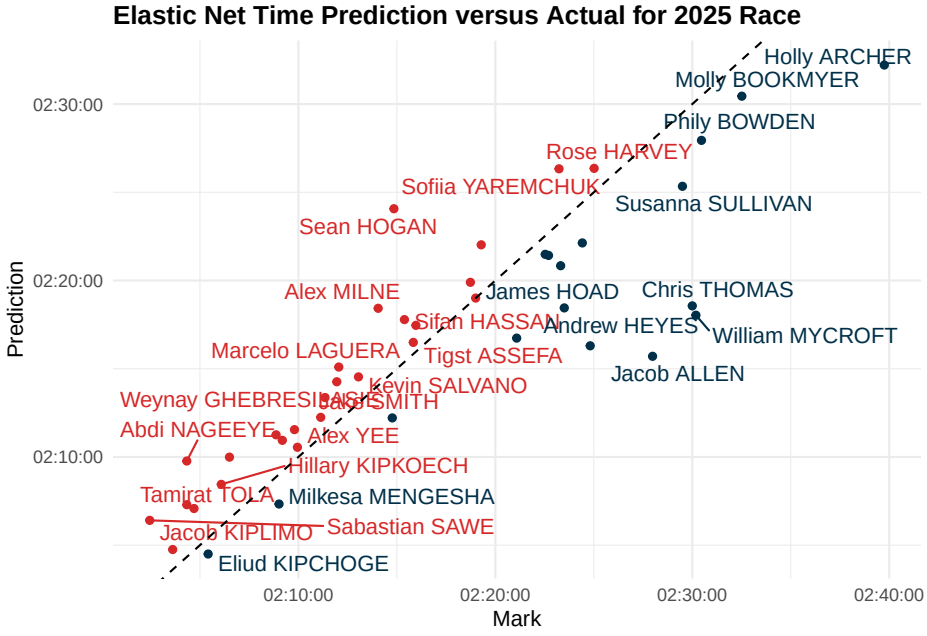
```
x <- data.matrix(trainData %>%
                   filter(!is.na(Mark)) %>%
                          .[,c(3:4, 8:11)])
y <- trainData %>%
  filter(!is.na(Mark)) %>%
  pull(Mark)

ridge_lambda_out <- cv.glmnet(x, y,  nfolds = 10,
                              type.measure = 'mse', alpha = 0.5, family="gaussian")
lambda <- ridge_lambda_out$lambda.min

ridgeMark <- glmnet(x, y, alpha = 0.5,
                    lambda = lambda,
                    family = "gaussian")
ridgeMark$beta
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## Africa           -78.23912589
```

```
## sex              170.98879231
## marathonEffort     0.73671875
## numMarathons       .
## numHalfMarathons   .
## timesince         -0.07382418
```

As one would expect, sex is an important predictor of finish time. I also think their best converted marathon is probably undervalued because the Africa variable captures a similar thing. It's a little disappointing that the number of marathons or half marathons isn't that important. I think it would be interesting if experience impacted performance on this stage, but it doesn't seem to. Overall, the predictions seem to be pretty evenly split between over- and underestimating. A RMSE of approximately 275 seconds (4:35) isn't great, but it doesn't seem like there were many outlier performances.

```
## [1] "RMSE: 274.662068465032"
```

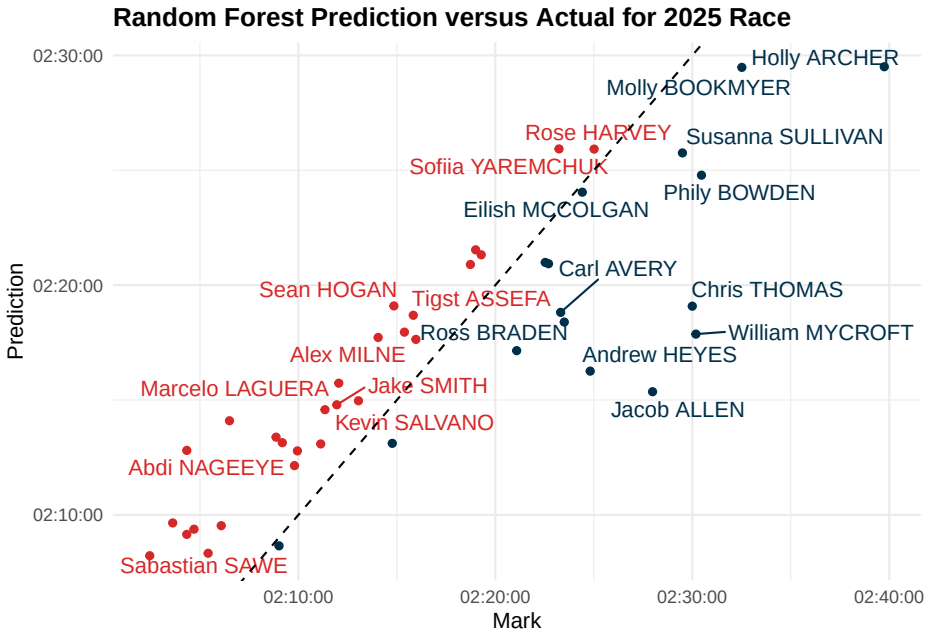


Random Forest

```
rfMark <- randomForest(Mark ~ sex + marathonEffort + Africa + timesince, data = trainData %>% filter(!is
```

The accuracy of the random forest model is not great and is a little concerning. It appears that most of the fast times were considered overperformers while low times were considered underperformers. I believe this is happening because of the small dataset and the fact that both men and women are included together. Because of this, previous women's times are pulling men's predictions back while previous men's times are pushing women's predictions forward.

```
## [1] "RMSE: 311.696127404391"
```



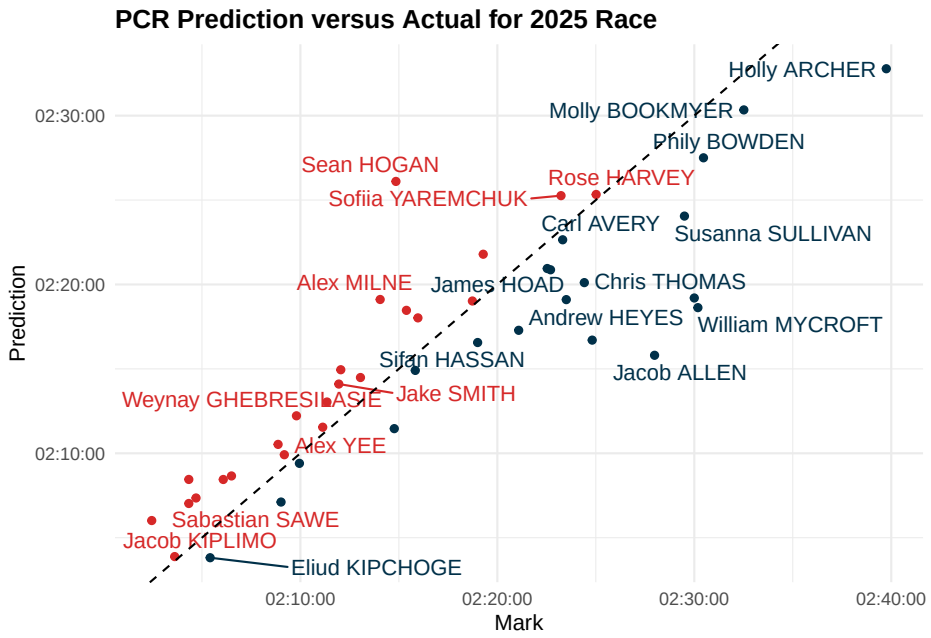
Principal Component Regression

```
pcrMark <- pcr(Mark ~ sex + marathonEffort + Africa + timesince + numMarathons, data = trainData %>% fi
```

The PCR accuracy is similar to the linear regression. Prediction error appears to be pretty consistent and individual predictions seem similar to before. Oddly enough, the model did very well on predicting Jacob Kiplimo and Alex Yee, two men who debuted in the marathon.

```
## [1] "RMSE: 274.06648397162"
```

```
## [1] 274.0665
```



All in all, in think there is more for predicting finishing time than there is predicting if someone will finish. Multiple linear regression and PCR both perform fairly well across the board.

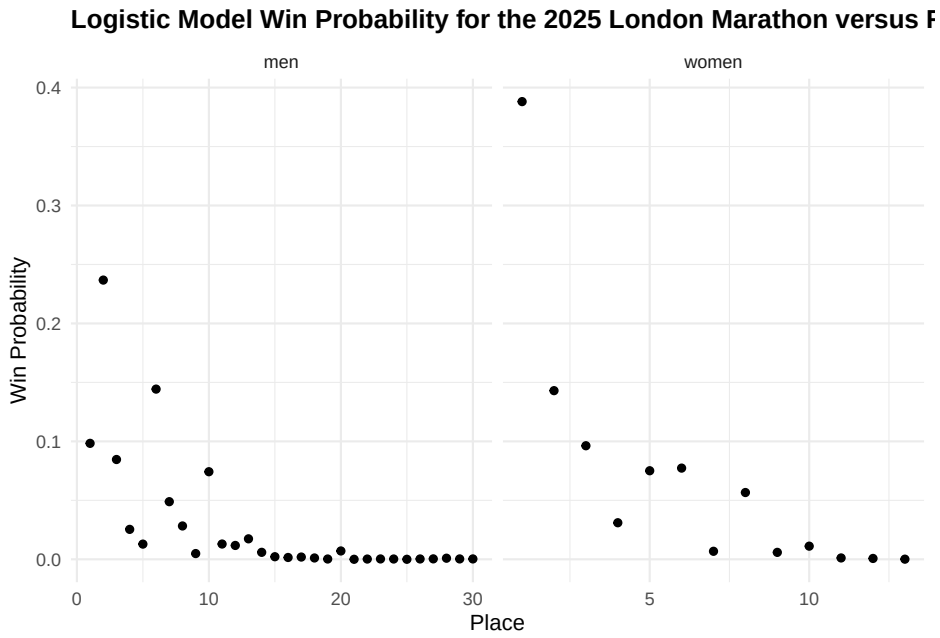
Win Probability

Logistic Regression

```
lmWin <- glm(win ~ sex + marathonEffort + Africa + timesince, data = trainData, family = 'binomial')
lmWin
```

```
##
## Call:  glm(formula = win ~ sex + marathonEffort + Africa + timesince,
##     family = "binomial", data = trainData)
##
## Coefficients:
##    (Intercept)        sexwomen  marathonEffort      AfricaTRUE       timesince
##      46.650711        5.446605       -0.006598        0.761732       -0.004050
##
## Degrees of Freedom: 342 Total (i.e. Null);  338 Residual
## Null Deviance:       117
## Residual Deviance: 85.23     AIC: 95.23
```

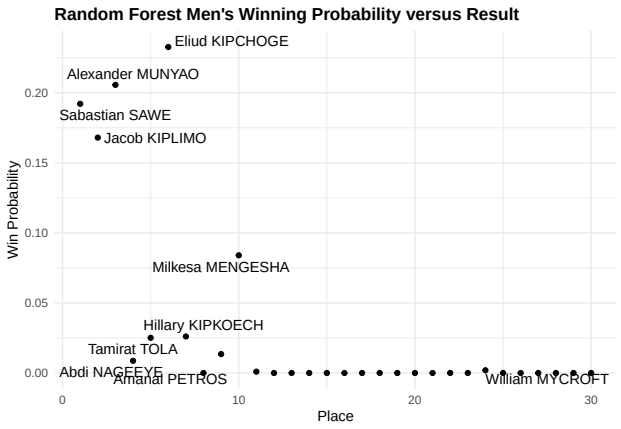
This logistic model performed much better on the 2025 London Marathon than I anticipated. There is a clear decrease in pre-race win probability as a runner's finish place increases. In the men's race, the runner-up, Kiplimo, had the best odds of winning. The winner, Sebastian Sawe, had the third best odds of winning. In the women's race, Tigst Assefa was the clear favorite and won.

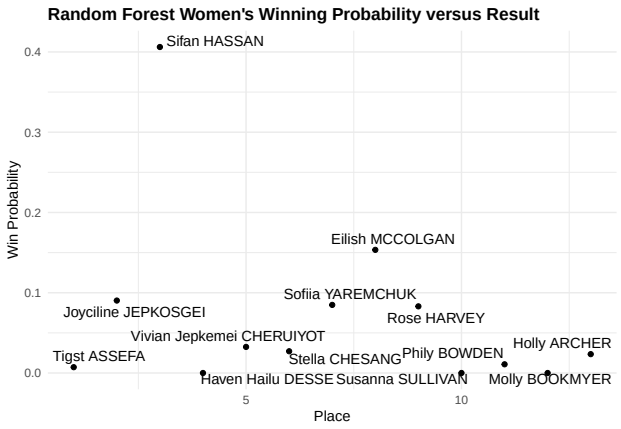


Random Forest

```
menWin <- randomForest(win ~ sex + marathonEffort + Africa + timesince, data = trainData %>% filter(sex
womenWin <- randomForest(win ~ sex + marathonEffort + Africa + timesince, data = trainData %>% filter(s
```

The random forest predictions aren't as accurate at logistic regression. In the men's race, the model gives Kipchoge a much higher probability of winning than before. This makes sense because he has the best PR in the field, but as a fan, it is clear that his career is winding down. After his poor showing at the Olympics, picking Kipchoge as the favorite doesn't seem very logical. The women's race, however, makes even less sense. Tigst Assefa having that low of a win probability is a red flag. The former world record holder was a clear contender going into the race. Sifan Hassan having a high probability of winning isn't that crazy and actually is closer to the betting odds via Bovada, which implied a 63.6% percent chance of winning.}





Between these two models, I think it's clear that the logistic regression performed better on the 2025 race. It gave the eventual winners good pre-race odds, and other top finishers had high probabilities as well.

Recommendations

Considering everything in this project, I believe the biggest takeaway is the difficulty in predicting elite marathon performance. First, trying to determine who will DNF produced very few valuable insights. It seemed like everyone in the elite field had a similar probability of finishing. In the future, I would like to look at other marathons to see if this is true for marathons in general. Since London is flat, I don't think there is as big of a DNF factor. For courses like New York and Boston, with a lot of elevation change, I could see DNFs being more prominent. I also think they may be easier to predict by looking at past performances on difficult courses or cross country. But for London and other flat marathons, I think the past finish percentage may be the best estimate of everyone's probability of finishing.

As for time predictions, the models seemed to perform all right in general as the RMSE was mostly around five minutes. This isn't overly precise, but it does give some concept of expected performance of an athlete. If someone was new to the sport, these projections would give them enough information to understand the race. However, for those who follow the sport, I don't think these projections offer enough accuracy to be that informative.

Lastly, the win percentage models gave the most encouraging results. They performed pretty well on the 2025 London Marathon, especially in the women's race. The logistic model gave eventual winner Tigst Assefa the highest win probability. It also gave Jepkosgei, the runner-up, the second-highest win probability despite the public heavily favoring Sifan Hassan. In the men's race, it gave the second-place finisher Kiplimo the best odds and the eventual champion, Sawe, the fourth-best odds.

Men's 2025 London Marathon Results

Place	Name	Time	Time Prediction	Win Probability
1	**Sabastian SAWE**	2:02:27	2:06:24	9.83%
2	**Jacob KIPLIMO**	2:03:37	2:04:45	23.68%
3	**Alexander MUNYAO**	2:04:20	2:07:17	8.46%
4	**Abdi NAGEEYE**	2:04:20	2:09:46	2.54%
5	**Tamirat TOLA**	2:04:42	2:07:04	1.29%
6	**Eliud KIPCHOGE**	2:05:25	2:04:29	14.43%
7	**Hillary KIPKOECH**	2:06:05	2:08:26	4.89%
8	**Amanal PETROS**	2:06:30	2:09:59	2.83%
9	**Mahamed MAHAMED**	2:08:52	2:11:15	0.49%
10	**Milkesa MENGESHA**	2:09:01	2:07:20	7.43%

Time prediction from elastic net model and win probability from logistic regression.

Women's 2025 London Marathon Results

Place	Name	Time	Time Prediction	Win Probability
1	**Tigst ASSEFA**	2:15:50	2:16:29	38.81%
2	**Joyciline JEPKOSGEI**	2:18:44	2:19:53	14.30%
3	**Sifan HASSAN**	2:19:00	2:19:00	9.63%
4	**Haven Hailu DESSE**	2:19:17	2:22:01	3.10%
5	**Vivian Jepkemei CHERUIYOT**	2:22:32	2:21:29	7.51%
6	**Stella CHESANG**	2:22:42	2:21:25	7.73%
7	**Sofiia YAREMCHUK**	2:23:14	2:26:19	0.68%
8	**Eilish MCCOLGAN**	2:24:25	2:22:07	5.66%
9	**Rose HARVEY**	2:25:01	2:26:21	0.59%
10	**Susanna SULLIVAN**	2:29:30	2:25:20	1.11%

Overall, I believe this project is a good starting point for future analysis in the sport of running. One of the main limitations of this project was the small sample size. I think finding ways to use more data, whether it's finding similar races or conversions between races, will help improve accuracy and highlight trends in the data. In addition to that, I think factoring in weather could be beneficial. London isn't exactly known for harsh conditions, but adjusting times on slightly hotter years may give us a better idea of what times to expect. Also, I think it would be interesting to factor in race styles to predict the race outcome. For example, if the elite field is full of runners who don't like to lead, the race may be slower than if a runner like Connor Mantz is up front keeping the pace honest. Along the same lines, if the elite field is full of a lot of 2:04 guys who can't kick, the slowest one probably doesn't have a great shot at winning. However, I would argue a 2:06 guy with a fast 10k time is more likely to win because they are stylistically different. It's these complexities that I would like to focus on because ultimately, the runners' times are not independent of each other. Another element that I believe should be incorporated is the runner's age and performance curves. Essentially, projecting when athletes will peak and when their performances will begin to worsen. I think this would've helped with predicting Kipchoge's performance this year because he has the fastest PR, but is out of his prime.

The goal of this project was to begin to address the void in running statistics for improving the viewing experience. Win probability, point totals, and other projections are commonplace in most major sports, but elite running hasn't adopted this yet. While data is used extensively in training, very little data is used to improve fan experience and knowledge. For these reasons, I will continue to work on models that predict athlete performance in the marathon and, eventually, other distances. The biggest challenge with this goal, and with this project specifically, is the data collection. Most athletes have their race history on the World Athletics website. However, they offer no API access, so researchers have to resort to scraping. This is particularly difficult given their website layout and the fact that they change their website fairly often (they changed while I was scraping data for this project).

The other, and larger challenge, is the lack of data on how athletes are training. For most athletes, we don't see their weekly mileage or the times they are hitting in workouts. On top of that, there isn't an injury report like in the NBA, so spectators have no idea if someone is showing up to the starting line injured. This is fairly common since most athletes receive appearance fees for just showing up to these races. This lack of transparency in training makes it difficult to predict who's going to race well.

In conclusion, this project only scratched the surface of what I believe is possible in running analytics. With some of the trends identified in this project, future analysis will be able to build and add complexity for better results. These improved predictions will allow fans of the sport to be more knowledgeable and will hopefully help grow interest in running.