ISyE 6740 - Fall 2022
Final Report

Predicting Marathon Race Time

Elise Yang

Problem Statement

Earlier this year in July, I embarked on a 17-week journey to train for my first marathon and race. Having never run more than 3 miles before, I doubted whether finishing 26.2 miles (or for this project, 42.195 kilometers) was even feasible. Pacing yourself is extremely important in long-distance running, amongst many other factors. Starting out too fast can result in injury, fatigue, cramping, and add much more time to your overall finishing time, and may even result in the dreaded DNF (did not finish).

Many running coaches advise first-timers to run below their goal pace for the first 10 miles, at their goal pace for the next 10, and faster than their goal pace in the last 6 to avoid burnout and achieve negative splits (where the pace improves throughout the race with each mile/kilometer). In October, I not only finished the entire 26.2 miles but in 4.5 hours- a whole 30 minutes faster than a current online race time calculator predicted, as well as achieved negative splits for the first time. However, I believe I could have achieved a faster time had I known to pace myself better as I still had some juice left by the end of the race. I would like to build a predictor model to predict a more accurate marathon time based on previous runs (5 km's, 10 km's, etc.).

Typically, a runner's overall race time slows the longer the run distance is. However, because I had never run a race before, I did not factor in race day adrenaline and support from the cheering crowds, along with proper rest and fueling the days leading up to the race. I will build a model based on runners' previous training data to predict my race time.

Data Source

The dataset contains 10,703,690 records of running training during 2019 and 2020, from 36,412 athletes from around the world. The records were obtained through web scraping of a large social network for athletes on the internet. The data with the athletes' activities are contained in dataframe objects (tabular data) and saved in the Parquet file format using the Pandas library, part of the Python ecosystem for data science.

- datetime: date of the running activity
- athlete: a computer-generated ID for the athlete (integer)
- distance: distance of running (floating-point number, in kilometers)
- duration: duration of running (floating-point number, in minutes)
- gender: gender (string 'M' of 'F')

- age_group: age interval (one of the strings '18 - 34', '35 - 54', or '55 +')
- country: country of origin of the athlete (string)
- major: marathon(s) and year(s) the athlete ran (comma-separated list of strings)

Files with the athletes' activities data are sampled at different frequencies: day 'd', week 'w', month 'm', and quarter 'q' for each year, 2019 and 2020. I will use the dataset on daily activities from 2020. The dataset also contains data with different government's stringency indexes for the COVID-19 pandemic.

Methodology

The first step is exploratory data analysis, which is the process of performing initial investigations on the data so as to discover patterns, spot anomalies, test the hypothesis, and check any assumptions with the help of summary statistics and graphical representations. Next is data preprocessing, which involves transforming the data to fit the model. This will take up the majority of the project time. This step involves feature engineering and selection, which will be an iterative process to determine the best features for the model. Feature engineering transforms raw data into features that can be used in supervised learning while also enhancing model accuracy. I will build six regression models using SKlearn packages in Python and choose the best regression model based on their predictive results after assigning accuracy scores. The goal is to ultimately be able to predict marathon times accurately after accounting for gender, age group, and previous run distances and respective times.

Six Regression Models:
Compare the results of the following six regression models and select the best predictor model.

- Linear Regression: Fits linear equation to the data
- Ridge Regression: Analysis of multicollinearity in multiple regression data
- Lasso Regression: Variable selection and regularization technique
- SVM: Supervised learning for data groups
- Decision Tree: Builds a tree to map possible outcomes
- Random Forest: Combination of decision trees

Data Preprocessing

First, I cleaned the dataset by converting values to the correct types. The gender and age group column values needed to be converted to integers so that the model can ingest the fitted data. Age group '18-34' becomes 0, '35-54' becomes 1, and '55+' becomes 2. Next, the dataset is subsetted into only athletes who have completed a marathon distance, which I determined to be between 42-42.5 km, and NaN values are dropped (runners that did not complete all new distance features, 0-5 km, etc., as described below in the feature engineering section). This resulted in only 1,299 athletes' data being used in building the models. Finally, I split the dataset into 70/30 to train and test.

Feature Selection:
I determined the most important predictors of marathon time by visual inspection of graphs as there are not many features provided in the original dataset. Average training distances, run pace, and duration are the best predictors of results on race day. Distinguishing athletes' gender and age group before building models is also essential to generating accurate marathon race time predictions.

Feature Engineering:
New features are created for certain ranges of distances (0-5 km, 5-10 km, etc.). Predicted marathon times vary depending on training run distance so these new distance features will make this distinction. A new feature for run paces is also created by finding the average duration/distance for each run.

Exploratory Data Analysis:

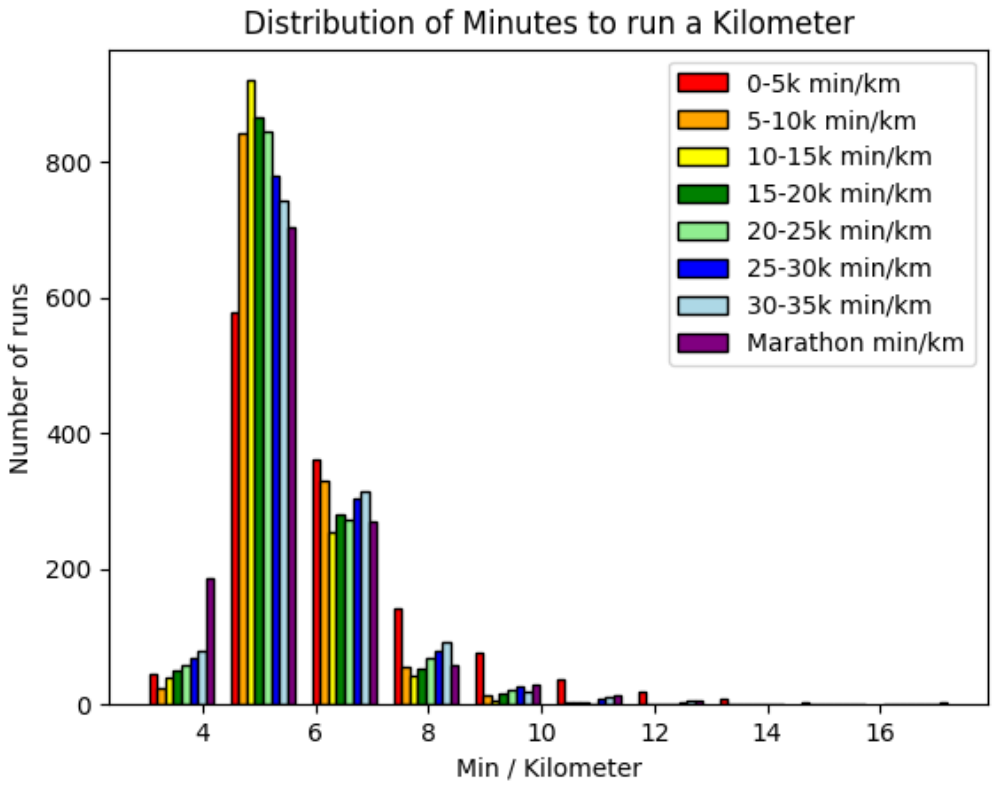

Figure 1: *Evaluating the number of runs for each range of distances and their respective paces*

This graph shows the running pace at each distance, indicated by the bar color, in min/km, on the x axis and the number of occurrences on the y axis. Most of the runs were completed at around 5 minutes per kilometer with the distance 10-15 km being the most common. However, most runs at the slower paces were at the 0-5 km distance. The marathon distance had the fastest pace.

This is notable as in order to build strength and increase speed, runners train with 80 percent easy runs and 20 percent hard runs in a given week. This means most of the mileage is run with a effort of 60-70 percent of max heart rate (easy runs) and a minimal amount of mileage at the 77-93 percent (hard runs).

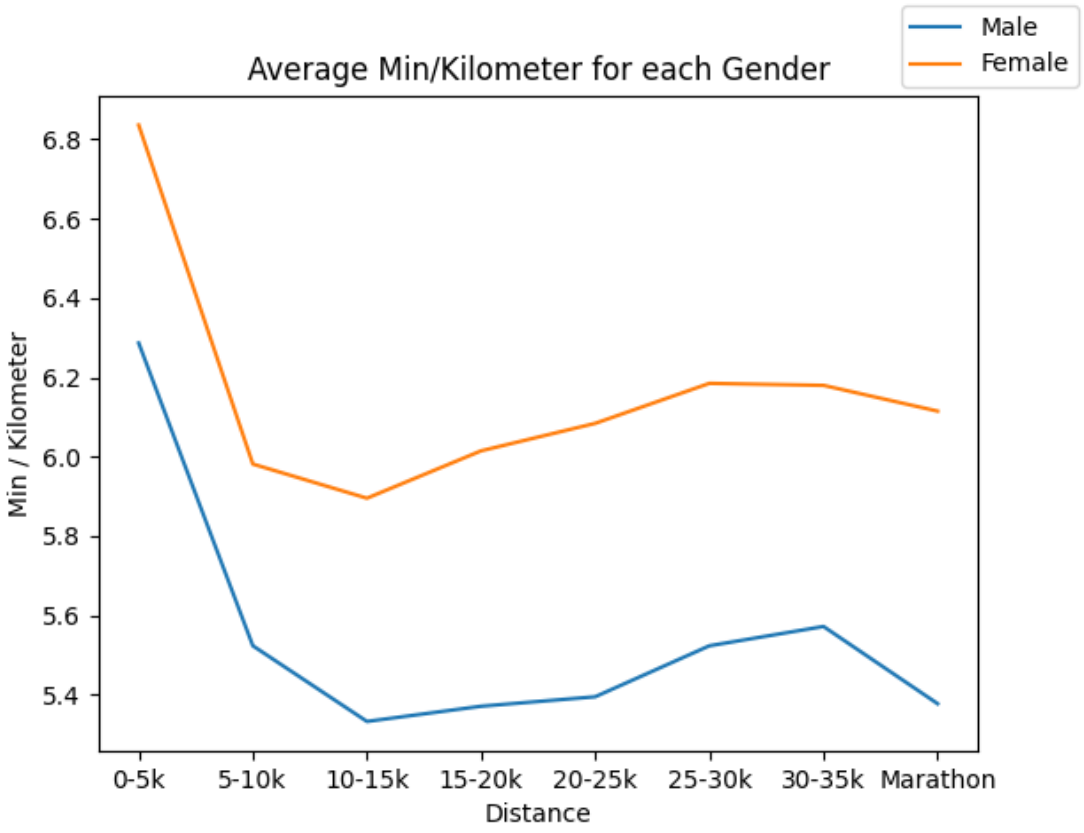

Figure 2: *Analyzing the differences in paces between males and females*

This visual shows the changes in athletes' running paces (min/km) given an increase in mileage. We can see that gender is an important feature to account for. Paces for both genders also change at about the same rate given the run distance.

From looking at the graph, the average paces at 15-25 km is similar to marathon pace. These training distance paces may be good indicators of marathon race pace.

Long run paces peak at 30-35 km. This may be due to most training plans having the longest long run at this distance. Training at distances anywhere above this increases the risk of injury and generally does not improve race day performance. This may also be the reason run paces fall between 35-42 km as many inexperienced runners burn out in the last few kilometers by not pacing themselves correctly.

Evaluation and Final Results

After splitting the data into testing and training sets, I fitted each model and generated an accuracy score. This was done by predicting average race pace then calculating the marathon race time by multiplying by 42.195. Changing the margin of error to one minute above/below pace time led to a 42 minute range above/below the predicted marathon time. This resulted in an approximately 62 percent accuracy rate. Subsequently, a 30 second margin of error led to a 21 minute range and an estimated 84 percent accuracy score.

The table below shows the accuracy rates for each of six regression models for both a 30 second and 1 minute margin of error.

Regression Model	Accuracy Score +/- 30 Seconds	Accuracy Score +/- 1 Minute
Linear Regression	61.94%	82.15%
Lasso CV	62.73%	81.89%
Ridge CV	61.94%	82.15%
SVM	51.71%	90.29%
Decision Tree	49.61%	84.51%
Random Forest	55.38%	90.03%

The SVM, Decision Tree, and Random Forest packages did not allow floats as parameters but rather required integers, so I rounded the paces and thus produced the abnormally high accuracy rates for a generous 1 minute error margin. The model struggles with predicting run pace within 30 seconds but is fairly accurate within a 1 minute error.

These results indicate that Lasso Regression would be the best predictor at the 30 second margin of error and SVM or Random Forest to be the optimal predictor model at the 1 minute error margin.

After fitting each model, I wanted to see how these models would perform predicting my actual marathon race time. I used an app called Strava to record my training runs and marathon race. Below are my average paces for each distance.

Distance (KM)	Pace (Min/KM)
0-5	7.199
5-10	7.137
10-15	7.318
15-20	7.293
20-25	8.673
25-30	7.804
30-35	7.220
42	6.329

The six models produced the following results for my predicted marathon race time.

Regression Model	Pace (Min/KM)	Predicted Marathon Race Time
Linear Regression	7.629	5 H 22 M
Lasso CV	7.485	5 H 16 M
Ridge CV	7.627	5 H 22 M
SVM	7	4 H 55 M
Decision Tree	5	3 H 31 M
Random Forest	6	4 H 13 M

Surprisingly, the Random Forest Classifier predicted closest to my actual marathon race time of 4 hours and 25 minutes.

Project Constraints

I ran into several problems during data preprocessing. Ideally, the data would have the athletes' marathon race times, but this dataset only included the name and year of the

marathon they ran in the past. I would need both training and race data that took place in the same year. Additionally, there is not a comprehensive list of features- in a runner's world, three age groups and gender is sufficient to determine possible race times, but to create a predictor model, more specific features such as height, weight, and smaller age groups would be needed. Third, I averaged all of the run times instead of looking at each individual run leading up to a marathon. In a perfect world, I would only look at training runs leading up to a specific marathon they were training for and that marathon.

Future Work

In the future, tuning the models by finding the optimal parameters and evaluating cross validation results would improve model accuracy. This project was limited by the constraints of the dataset and course schedule. This dataset can also be used to predict other races such as 5 km's, 10 km's, or 50 km's. Another dataset combined with this one can be used to analyze whether different levels of COVID-19 quarantine policies had an affect on training and race times. I am ultimately looking forward to using the Random Forest Classifier model as I train for my next marathon in Berlin 2023.

References

https://figshare.com/articles/dataset/A_public_dataset_on_long-distance_running_training_in_2019_and_2020/16620238