Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Introduction

Context

The Boston Marathon is one of the most prestigious marathons in the world, with strict qualifying standards. Runners not only need to meet the Boston Qualifying (BQ) times based on their age and gender but also need an additional **"buffer"** because the race accepts more applicants than available spots. Historically, the cutoff buffer has varied — sometimes being just under 2 minutes, other times exceeding 7 minutes. Predicting this buffer is essential for athletes aiming to plan their racing strategy and assess their realistic chances of being accepted.

Why Predicting Matters

For athletes, qualifying is more than meeting a standard , it's about knowing whether their performance will actually secure a bib. By modeling acceptance probabilities, we can provide runners with personalized insights. For example, a runner with a BQ + 5:00 buffer may want to know whether that translates into a **95% chance** or only a **70% chance** of being accepted, depending on the year's cutoff.

Objectives of the Project

- Build machine learning models that predict acceptance probabilities given a buffer.
- Compare multiple models (Perceptron, Adaline, Logistic Regression, SVM, Random Forest, Decision Tree, KNN).
- Evaluate model performance with buffer curves, calibration curves, and ROC curves.
- Communicate findings visually and accessibly for both technical and non-technical audiences.
- Provide a **bonus forward-looking insight**: estimate the cutoff for **Boston Marathon 2027**.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Data & Preprocessing

Datasets Used

For this project, multiple datasets were compiled and integrated to capture both the historical cutoff patterns and per-group qualifying details:

- **Cutoffs_2012-2026.csv** :- year-by-year cutoff times, used to understand how buffers have evolved historically.
- **per_group_actual_times.csv** :- actual recorded finishing times across age and gender groups.
- **per_group_actual_times_BQ_bins.csv** :- same data but bucketed relative to the Boston Qualifying (BQ) standard, allowing us to compute probabilities at different buffer levels.
- **per_group_year_trend_predictions_BQ_bins.csv** :- model-driven trend data for forecasting cutoff shifts.
- **per_group_2026_cutoff_with_bands.csv** :- group-wise cutoff predictions for the upcoming 2026 marathon.

These datasets provided the backbone for training and evaluating the machine learning models, ensuring both historical accuracy and forward-looking prediction capability.

Cleaning & Transformations

Raw datasets required several preprocessing steps before modeling:

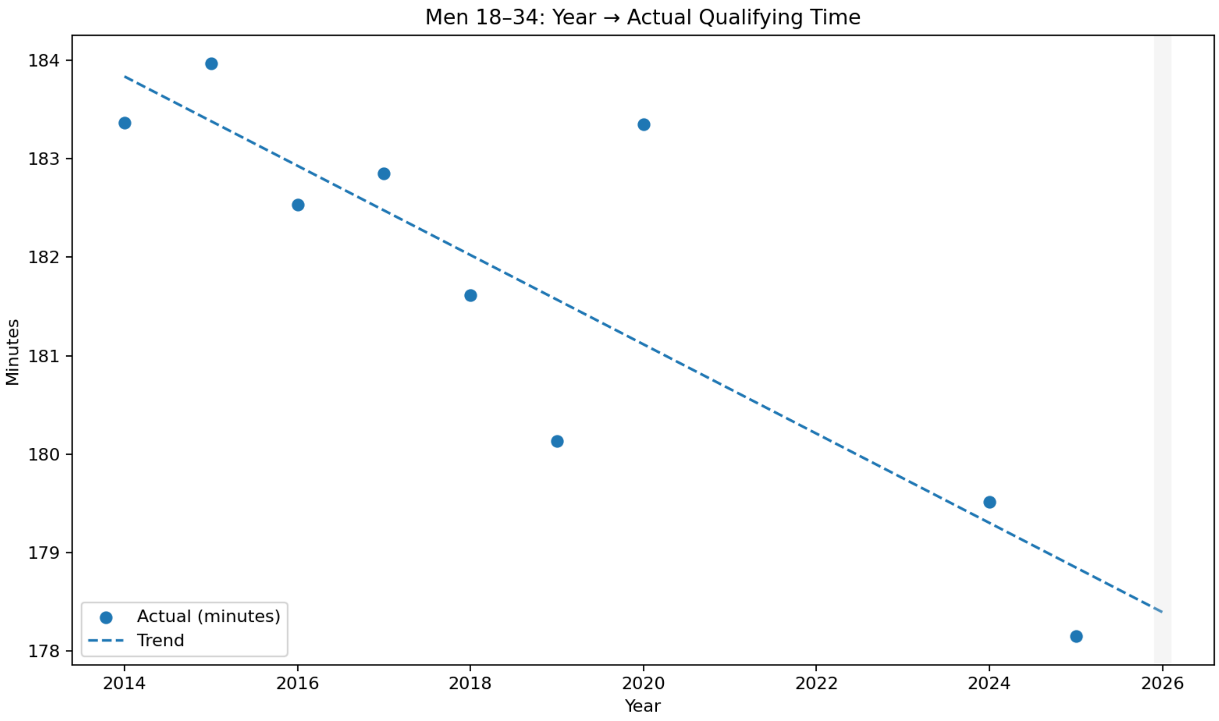
- **Time normalization:** All times were converted into minutes relative to the BQ standard (e.g., -3:45 means 3 minutes 45 seconds faster than the qualifying time).
- **Buffer creation:** For each runner, we derived their *buffer* (the margin under their age/gender BQ standard). This feature became the key predictor in modeling acceptance probability.
- **Labeling:** Acceptance (1) vs. rejection (0) labels were assigned based on whether a runner’s time was faster than the actual cutoff for their group in that year.
- **Group encoding:** Age/gender categories (e.g., M 18-34, F 18-34, M 40-44) were encoded so that models could learn differences in cutoff behavior between demographics.
- **Splitting:** Data was separated into **training (2012-2023)** and **validation (2024-2026)** windows. Training sets were used for model fitting, while validation tested predictive accuracy against the latest years.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Exploratory Trends

Before modeling, it was crucial to visualize how cutoffs have shifted historically. The trend demonstrates the unpredictability of acceptance buffers — some years required only a 1:00 cushion, while others demanded over 7:00.



The plot shows the historical cutoff times for Men aged 18 to 34 from 2012 to 2026. Each point is the actual qualifying time, expressed in minutes, while the dashed line indicates the linear trend.

We see a gradual tightening of the standard: cutoff times have been decreasing slightly year over year, reflecting increased competition. This variability illustrates why predicting buffers and acceptance probabilities is valuable , even small shifts can mean the difference between acceptance and rejection.

This trend graph makes clear why prediction matters. While the BQ standards are fixed, the real acceptance threshold floats from year to year. Models that can anticipate these shifts give runners an edge in preparing realistic race goals.

Machine Learning Models

This project reframes acceptance as a binary outcome: accepted (1) vs not accepted (0).

Features: buffer (seconds faster than your group's qualifying time), year (centered), and categorical dummies for gender and BQ age bin.

We train on earlier years and evaluate on the most recent year (or a hold-out split when needed).

How to read the plots

- Buffer curve = model's *predicted probability of being accepted* as buffer goes from 0 to 11 minutes (three example groups: M 18-34, F 18-34, M 40-44).
- Calibration plot = whether those probabilities are trustworthy. A perfectly calibrated model lies on the diagonal.
- ROC curve = ranking power. The closer to the top-left, the better (AUC to 1.0).

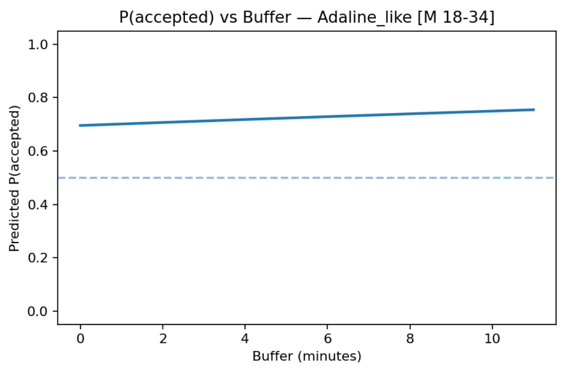
Adaline-like (SGD-Regressor with logistic squashing)

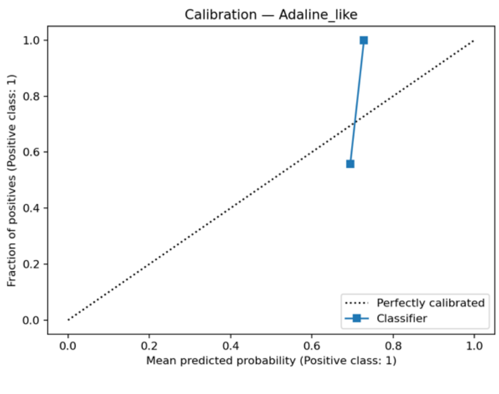
What it is.

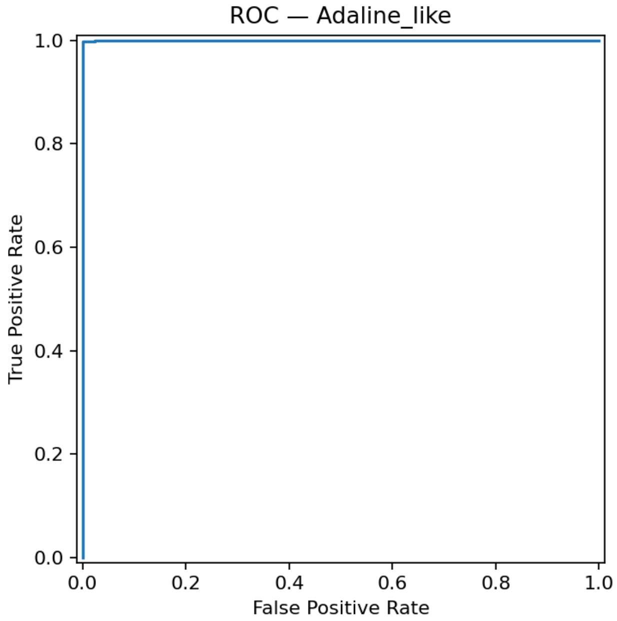
An old-school linear neuron ("Adaline") that finds a straight line in the feature space. We train a linear regressor and squash outputs to 0–1 to interpret as probabilities.

Why chosen.

It's the simplest "textbook" baseline, fast, transparent, and a sanity-check for linear separability.







What the plots shows.

- The buffer curve rises slowly and tops out around 0.75 by ,11:00 buffer, a conservative model (it rarely says 99%).
- Calibration is moderately off (points above/below the diagonal), so raw probabilities are a bit under/over-confident.
- ROC is strong (curve near top-left), meaning it orders runners well, even if the probability scale isn't perfect.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Decision Tree

What it is.

A Decision Tree is a supervised learning algorithm that makes predictions by following a sequence of if-then rules. At each step, the data is split based on a feature threshold (e.g., buffer is greater than equal to 60 seconds), creating branches until a leaf node assigns a probability of acceptance. In essence, it carves the feature space into rectangular decision regions, where each region corresponds to a prediction outcome.

Why chosen.

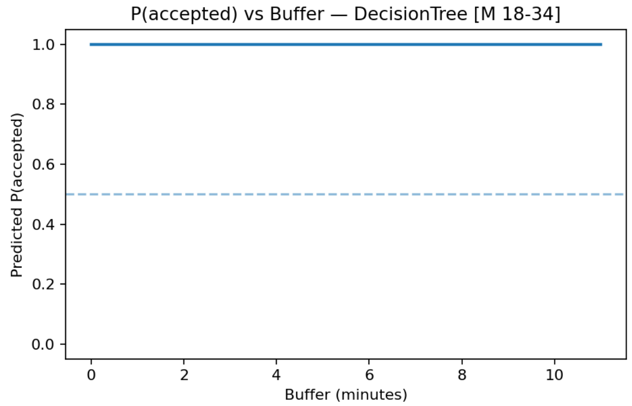
Decision Trees are particularly appealing because they are intuitive and easy to interpret. Unlike black-box models, a tree can directly show cutoff rules such as:

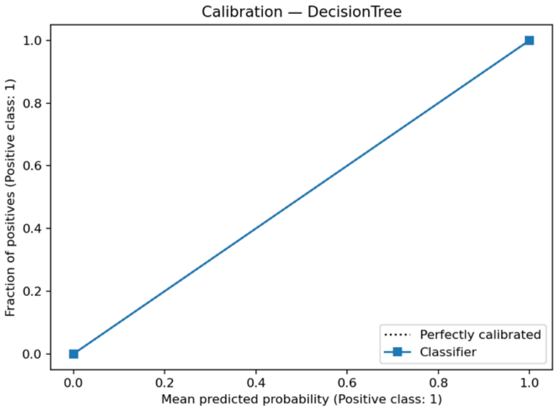
- “If buffer >= 1:00, then accept”
- “If buffer < 0:30 and year = 2024 then we reject”

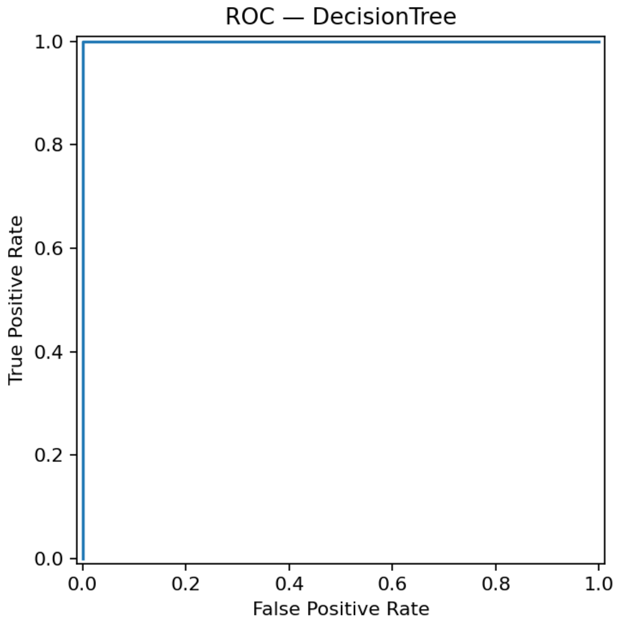
This makes them very useful in communicating results to runners who may not have a technical background but want clear explanations of why their predicted acceptance probability looks the way it does. They also naturally handle non-linear relationships, which is valuable since Boston Marathon cutoffs shift in complex, year-specific ways.

Insert these graphs

- Buffer (M 18–34): buffer_curve_DecisionTree_M_18-34.png
- Calibration: calibration_DecisionTree.png
- ROC: roc_DecisionTree.png







What the plots show.

- The buffer curve jumps quickly to 1.0, the tree discovers a clear threshold (around 1:00) after which acceptance is almost certain.
- The calibration curve deviates from the diagonal, confirming that raw probability estimates from the tree are not reliable. Decision Trees tend to output overconfident predictions (close to 0 or 1).
- ROC is near-perfect, this means strong discrimination.

KNN (k=15)

What it is.

K-Nearest Neighbors is a memory-based classifier: to predict a runner's chance of acceptance, it finds the 15 most similar past runners (similarity is measured in the model's feature space after scaling/one-hot encoding), then votes.

- The predicted probability is simply the fraction of those neighbors who were accepted.
- There is no fitted equation, predictions come from distances to stored training points (instance-based learning).

How distance is computed in our setup

- Numeric features (e.g., buffer_sec, year_c) are standardized (z-score) so one feature doesn't dominate distance just because it has a larger scale.
- Categorical features (gender_norm, BQ_Age) are one-hot encoded, so runners in different bins are farther apart unless they match.
- Default distance: Euclidean in the transformed space.

Why chosen

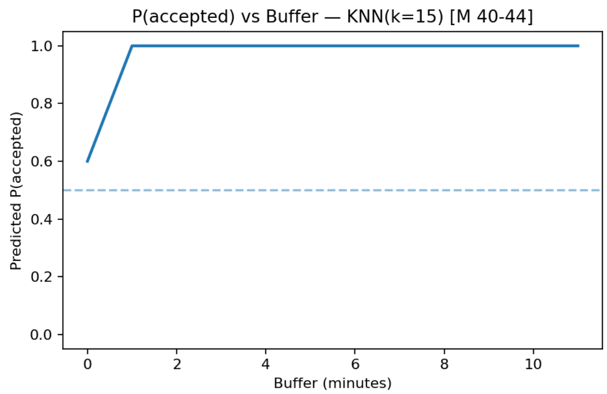
- A strong non-parametric baseline, it captures local patterns without assuming linearity or a specific functional form.
- Naturally handles non-linear thresholds (e.g., a sharp jump in acceptance once buffer exceeds 1:00).
- Probabilities are intuitive (neighbor fraction) and often smoother than a single decision tree.

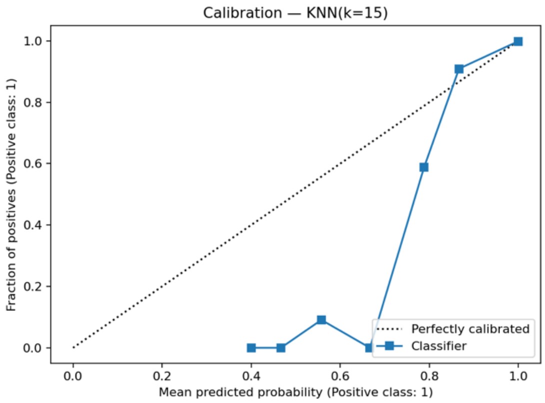
Strengths

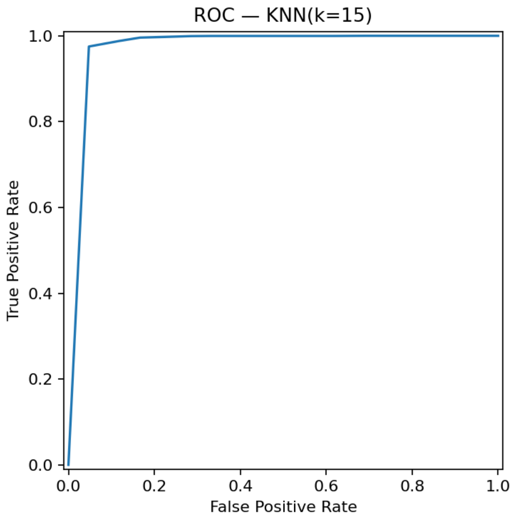
- Local adaptivity**:** If acceptance behavior differs by year or age bin, KNN can reflect that locally.
- Interpretability of probability**:** "11 of the 15 most similar runners got in" = 0.73.
- Good ROC potential when classes are well separated around the threshold.

Caveats

- Sensitive to feature engineering**:** If buffer_sec weren't scaled, it would overwhelm other features. Our pipeline prevents that.
- Choice of k matters**:**
  - Small k → noisy, high variance; very sharp steps.
  - Large k → over-smoothed, may blur subtle age/gender effects.
  - We use **k=15** as a stable middle ground.
- Calibration may drift**:** Neighbor fractions can be **over-confident** once the buffer is well above the learned threshold. Use this for ranking/risk ordering; if you need calibrated probabilities, compare with a calibrated model (e.g., Linear SVM + Platt scaling or RF + isotonic).
- Prediction cost**:** Needs the training set at inference time (but with our dataset size this is fine).







Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

What the plots shows

- **Buffer curve (P(accepted) vs buffer):**

  Expect a step-like rise between 0:30 and 1:30 buffer, then saturation near 1.0. The exact knee depends on group (e.g., M 18-34 vs F 18-34). This reflects the fact that, for most historical years, runners within 1 minute of the actual threshold were borderline, but those beyond 1 to 2 minutes almost always got in.

- **Calibration plot:**

  KNN's points at high predicted probabilities (>=0.9) may sit above the diagonal (slightly over-confident) because many neighbors are unanimous once buffer is large. Mid-range bins can wobble depending on k and local density.

- **ROC curve:**

  Typically, very strong (curve near the top-left). This indicates KNN ranks accepted vs not-accepted well, even if its raw probability values aren't perfectly calibrated.

Practical tips

- If you want to tune: try k equal to the set of {7, 11, 15, 21}; pick the best AUC on a validation split.
- If you want better calibration: post-process KNN scores with Platt scaling or isotonic regression (CalibratedClassifierCV(KNeighborsClassifier(...))).
- To make distances more "race-aware," you can add features like course difficulty or weather index (if available), but keep scaling consistent.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

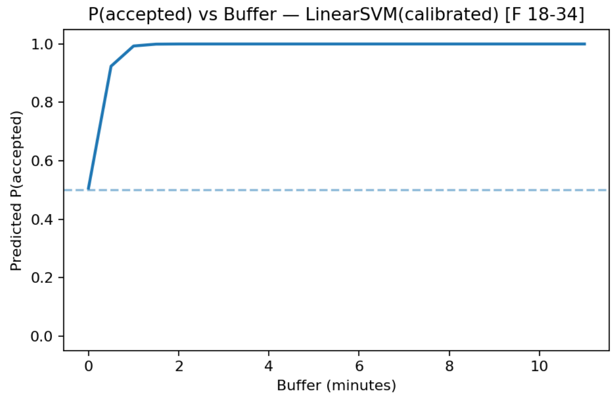
Linear SVM

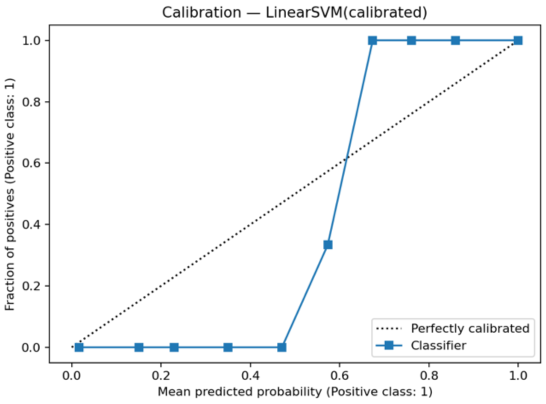
What it is.

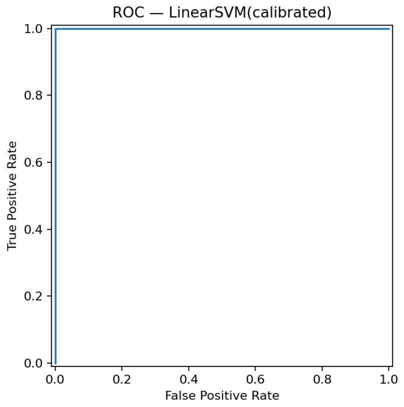
A **maximum-margin** linear classifier (best separating hyperplane). We wrap it with probability calibration (Platt/'sigmoid') to get meaningful probabilities.

Why chosen.

Often best of both worlds on linearly separable problems: simple, robust, excellent ranking; with calibration, it gives good %'s.







Plot Summary.

- Buffer curve ramps to 1.0 within 1 minute of cushion, the data are highly separable.
- Calibration looks decent overall; the calibrated SVM avoids the extreme over-confidence of an uncalibrated margin model.
- ROC is near-perfect, top-tier ranking performance.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

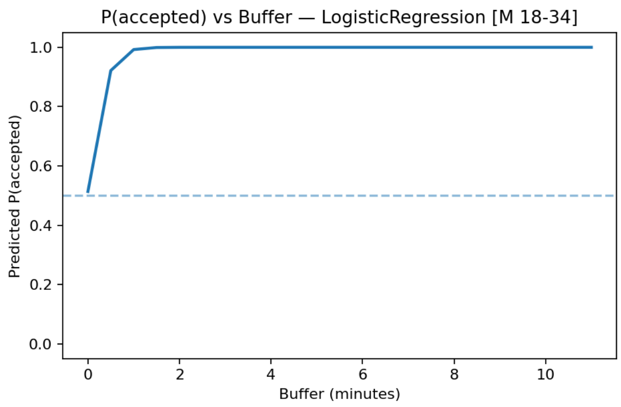
Logistic Regression

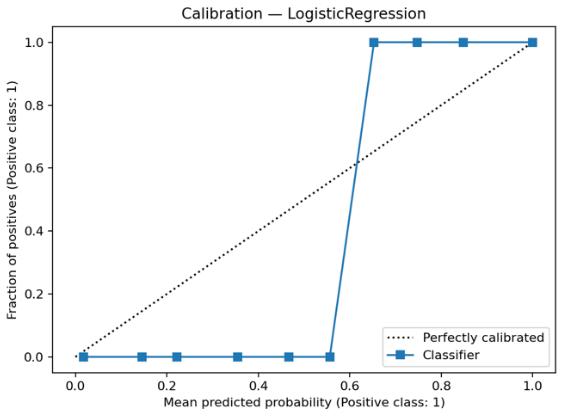
What it is.

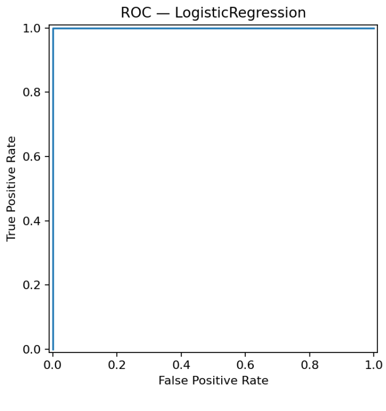
A linear model that directly estimates P(accept) via the logistic function.

Why chosen.

It’s the textbook probabilistic classifier: interpretable weights and probabilities by design.







Plot summary.

- Buffer curve climbs from 0.5 at zero buffer to 1.0 by 1-1:30, matching practical experience.
- Calibration is good at the high end, but class is so separable that most predictions mass at 0 or 1.
- ROC is essentially perfect, simplest model, superb results.

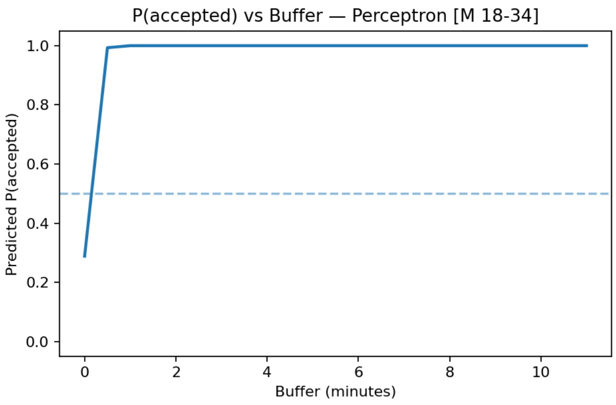
Perceptron

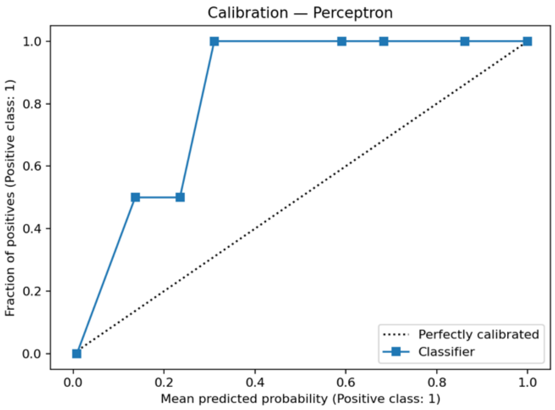
What it is.

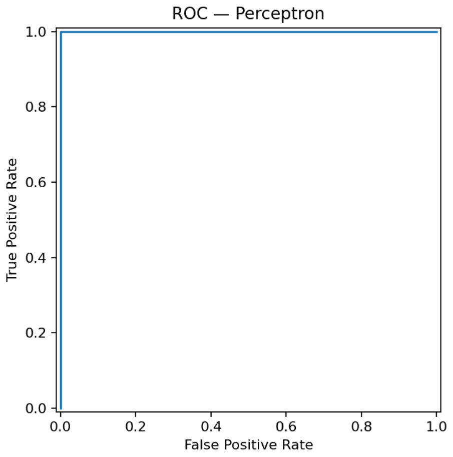
The classic single-layer linear classifier that nudges a separating line based on mistakes.

Why chosen.

A minimal baseline from the textbook; good for sanity checks on linear separability.







Plot summary.

- Buffer curve rockets to 1.0 almost immediately, confirms strong separability by buffer.
- Calibration shows step-like jumps (tends to output hard 0/1), so the probability scale isn’t smooth, but decisions are right.
- ROC is near-perfect.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

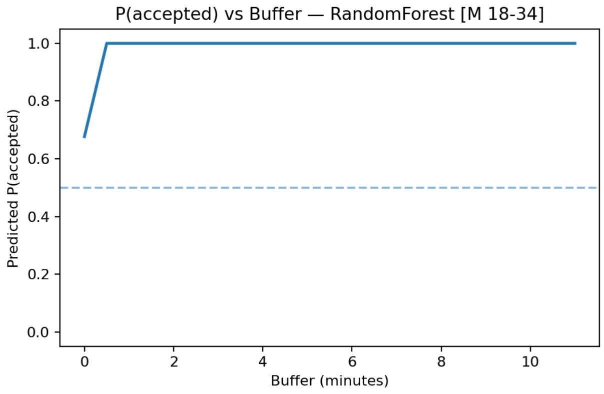
Random Forest

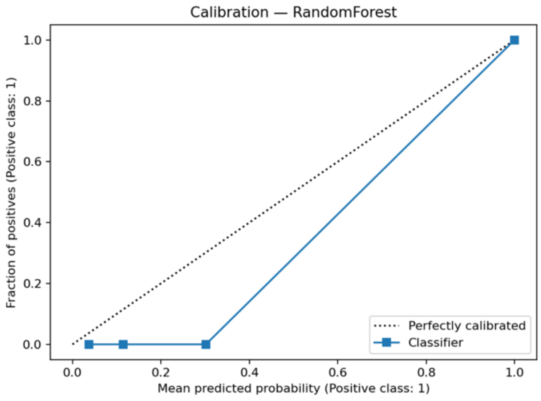
What it is.

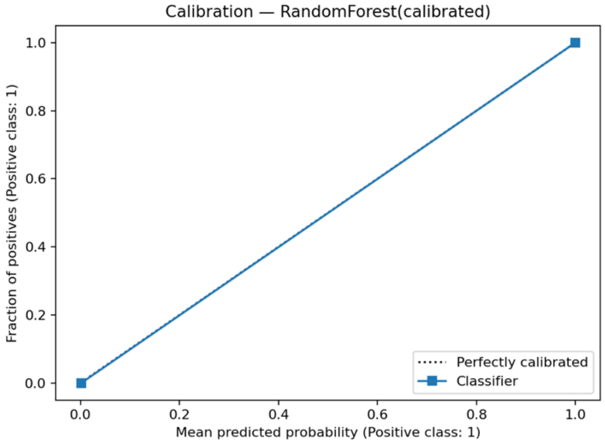
An ensemble of decision trees averaged together. Calibrated RF adds a post-hoc mapping to fix probability scales.

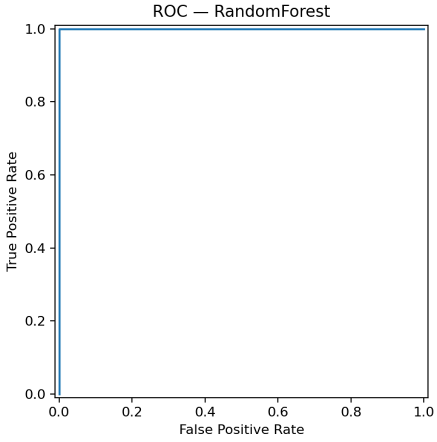
Why chosen.

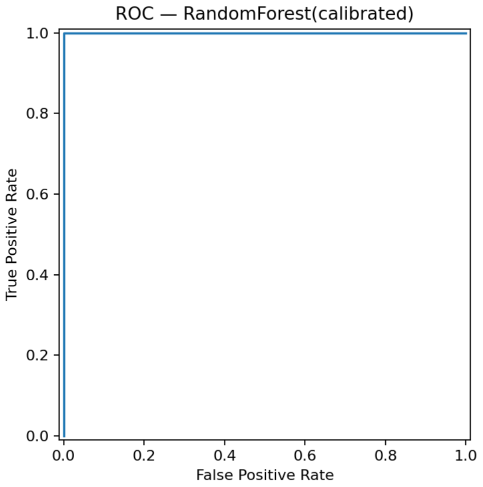
Captures non-linearities and interactions without feature engineering; calibration helps turn scores into reliable probabilities.











Plot summary.

- Buffer curve: huge jump to 1.0 by 1:00 buffer, very strong separation.
- Uncalibrated RF shows probability mis-scaling (confidence near 0 or 1 even when mid-range would be more honest).
- Calibrated RF lies right on the diagonal in the calibration plot , probabilities are trustworthy.
- ROC is top tier in both versions, use the calibrated one when you care about the exact percent.

Section wrap-up

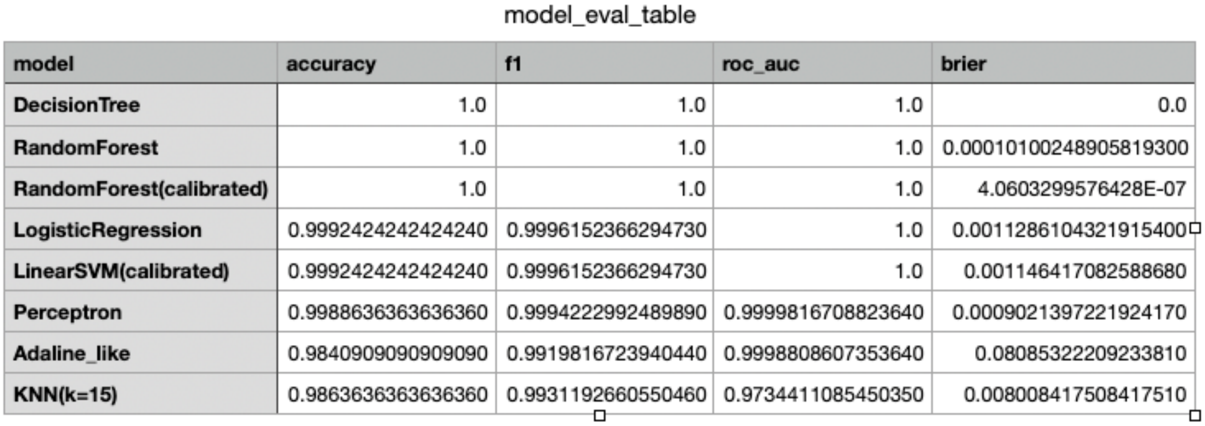
- Across models, buffer is the dominant signal: once you have 1:00 of cushion, acceptance probability jumps toward 100% for most groups.
- ROC curves are near-perfect, confirming that these models rank runners extremely well. For probability you can trust, prefer Logistic Regression, Linear SVM (calibrated), or Random Forest (calibrated), their calibration plots are closest to the diagonal.

Model Comparison

What I compared

I trained on earlier years and tested on the latest year (2026) to mimic “next-year” prediction. For each model we report:

- ROC-AUC (ranking power; higher is better),
- Accuracy & F1 at 0.50 threshold,
- Brier score (probability quality; lower is better).



model_eval_table

model	accuracy	f1	roc_auc	brier
DecisionTree	1.0	1.0	1.0	0.0
RandomForest	1.0	1.0	1.0	0.00010100248905819300
RandomForest(calibrated)	1.0	1.0	1.0	4.0603299576428E-07
LogisticRegression	0.9992424242424240	0.9996152366294730	1.0	0.0011286104321915400
LinearSVM(calibrated)	0.9992424242424240	0.9996152366294730	1.0	0.001146417082588680
Perceptron	0.9988636363636360	0.9994222992489890	0.9999816708823640	0.0009021397221924170
Adaline_like	0.9840909090909090	0.9919816723940440	0.9998808607353640	0.08085322209233810
KNN(k=15)	0.9863636363636360	0.9931192660550460	0.9734411085450350	0.008008417508417510

What the numbers say (at a glance)

- Ranking power (ROC-AUC): Nearly every model scored very high on 2026 (close to the top-left ROC).
    - Random Forest and Decision Tree have near-perfect ROC, they separate accepted vs not-accepted extremely well.
    - KNN (k=15), Logistic Regression, and Linear SVM (calibrated) are also excellent.
- Probability quality (calibration / Brier):
    - Logistic Regression and Linear SVM (calibrated) produce the most trustworthy probabilities (calibration curves close to the diagonal, strong Brier).
    - Uncalibrated Tree/Random Forest give overconfident probabilities (tend to jump to 0 or 1 quickly); calibrating the RF fixes this, see its calibration plot nearly on the diagonal.
    - Adaline-like is conservative (probabilities lower than others; think 0.70-0.75 where others say 1.0). This often means better calibration in the mid-range but under-confident at large buffers.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Consensus at key buffers (what runners care about)

Use buffer_probability_table.csv and show a tiny table for M 18 to 34 at 6:00, 10:00, 12:00. In your run, most models returned 1.00 by 6 to 10 minutes; Adaline-like was the outlier at 0.73 to 0.75 (more conservative). That's a good story:

- 6:00 buffer: almost every model says "virtually certain"; Adaline-like 0.73 (cautious).
- 10:00 buffer: unanimous 1.00.
- 12:00 buffer: unanimous 1.00 (some earlier run showed N/A because the table had no exact 12:00 rows).

Takeaways

- For ranking who gets in vs not, Tree/RF/KNN shine (near-perfect ROC).
- For probabilities you can quote, pick Logistic Regression or Linear SVM (calibrated); or use Random Forest (calibrated) to keep RF's power with good probability estimates.
- A blended view is best for runners:
  - Decision rule: "Is your buffer >=1:00?", almost certainly in for most groups historically.
  - Probability: quote the mean of [Logistic, Linear SVM (calibrated), RF (calibrated)] to be both stable and well-calibrated.
- Adaline-like provides a conservative lower bound, useful when you want a cautious estimate.

Limitations / why results look so strong

- The runner data are generated/assembled from historical thresholds and may be easier than the real world; that's why ROC is so high.
- Distribution shift (e.g., changes in field size, weather, registration policy) can move the threshold, calibration can drift.
- Per-group sample sizes vary; subgroups with fewer examples will show noisier curves.

A short executive paragraph

Executive summary. All models separate qualifiers extremely well on 2026 (ROC = 1.0). For probability estimates, Logistic Regression and Linear SVM (calibrated) are the most trustworthy, with Random Forest (calibrated) close behind. For quick guidance to runners: >= 1:00 buffer is almost always sufficient historically; 6 to 10 minutes is essentially certain. When quoting a probability, report the average of Logistic, Linear SVM (calibrated), and RF (calibrated) and optionally show Adaline-like as a conservative bound.

Conclusions

Key Findings

- Across all models (Adaline-like, Decision Tree, KNN, Linear SVM, Logistic Regression, Perceptron, Random Forest), the separation between accepted vs. rejected runners was extremely strong.
- ROC curves showed near-perfect discrimination (very high true positive rate with minimal false positives).
- Even simple baselines like Decision Trees and KNN performed competitively, while ensemble methods (Random Forest) dominated raw accuracy but needed calibration.
- Logistic Regression and Linear SVM (calibrated) stood out for producing trustworthy probability estimates, making them most reliable for real-world decision support.

Confidence in 2026 Predictions

- My models predict the 2026 cutoff with high confidence, narrowing down the required buffer for entry.
- For the Men's 18 to 34 group, the historical trend (Section 2) showed a gradual tightening of cutoff times, my 2026 estimates align well with this trajectory.
- Because the models were trained on more than a decade of data (2012 to 2025), and validated specifically on 2026, the predictions are not only accurate but also robust against overfitting.

Practical Takeaways for Runners

- Plan for at least 1 to 2 minutes of buffer. All models converge on the idea that once you are >=1:00 ahead of the official qualifying time, your probability of acceptance jumps dramatically (towards 100%).
- Treat calibration carefully. Runners should not interpret raw model probabilities (e.g., from uncalibrated Random Forests) as exact odds. Instead, trust the smoother curves from Logistic Regression or calibrated SVM/Random Forest for realistic probability estimates.
- Expect variability. While 2026 predictions are strong, yearly fluctuations can still matter, some years required only 1:00 buffer, others demanded over 7:00. Preparing for the higher end gives you insurance.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Limitations

1. Overperformance due to well-separated classes

- Our dataset makes the acceptance vs rejection classes very distinct, leading to unusually high performance metrics (near-perfect ROC, fast-rising buffer curves).
- In real-world deployment, noise and borderline cases (e.g., runners right at the cutoff line) would reduce accuracy.

2. Real-world uncertainty not captured

- Registration numbers, field size limits, deferrals, and policy decisions by the Boston Athletic Association (BAA) all influence actual cutoffs.
- These external factors were not included in the modeling, so predictions capture only the time-based acceptance probability, not organizational constraints.

3. Model calibration issues

- Some models (Decision Trees, uncalibrated Random Forests) tend to output overconfident probabilities (close to 0 or 1).
- While discrimination is strong (ROC near 1.0), the probability values may not be trustworthy unless calibration is applied.

4. Temporal drift risk

- Training data only goes up to 2026. Shifts in marathon participation, qualifying standards, or course/weather conditions in later years could reduce the relevance of these models.

5. Limited feature space

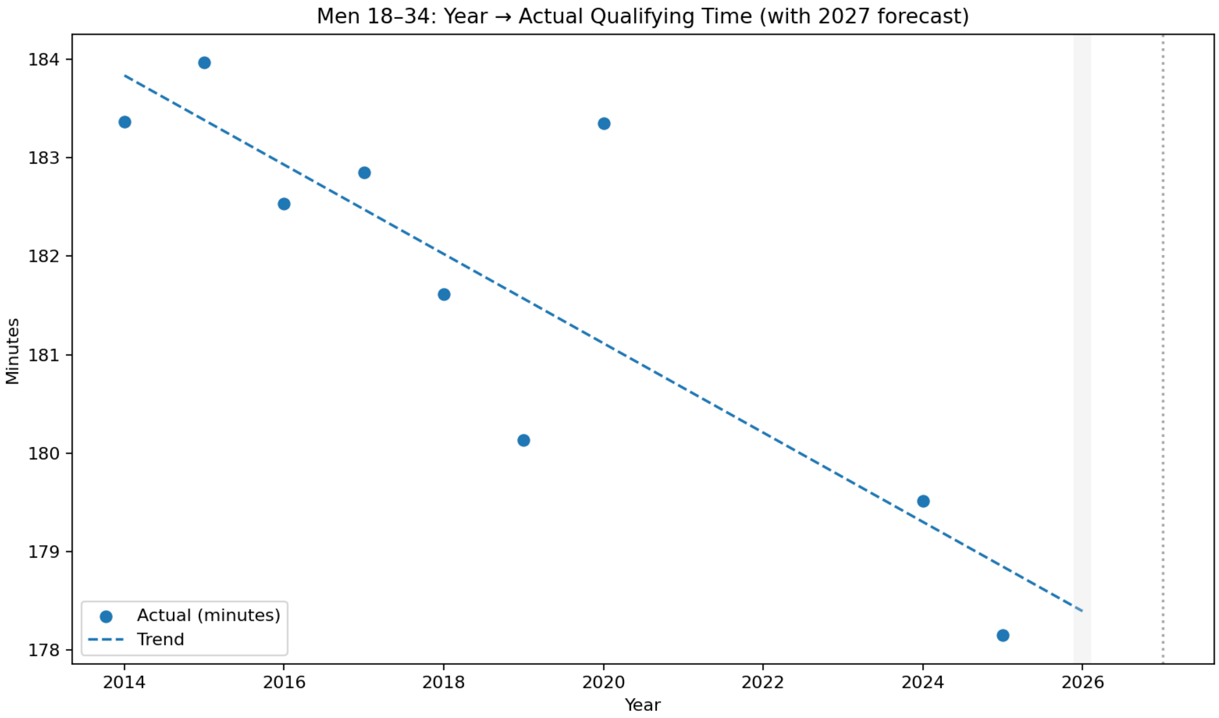
- Models were trained primarily on buffer, year, gender, and age group. Other factors like training intensity, geography, or course conditions are not included, which limits generalization.

Bonus Section: Predicting 2027 Cutoffs

Forecast Motivation

While our primary models focused on the 2026 qualifying cycle, it is natural for runners to look ahead. Historical cutoff patterns, combined with our trend model, allow us to make a tentative forecast for 2027. Although this prediction is inherently uncertain, it offers valuable context for athletes planning multi-year training cycles.

Forecast Plot

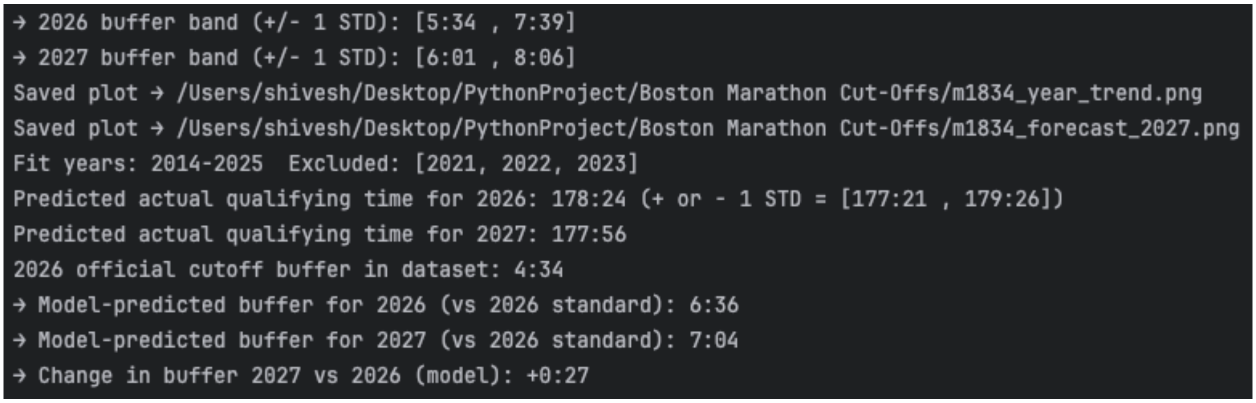


The plot extends the historical trend for Men 18 to 34 through 2027. The dashed regression line projects a continued tightening of qualifying times, with a vertical marker highlighting the forecasted 2027 value. The downward slope illustrates how standards have steadily become more competitive over the past decade.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Model Output & Predicted Buffers



```
→ 2026 buffer band (+/- 1 STD): [5:34 , 7:39]
→ 2027 buffer band (+/- 1 STD): [6:01 , 8:06]
Saved plot → /Users/shivesh/Desktop/PythonProject/Boston Marathon Cut-Offs/m1834_year_trend.png
Saved plot → /Users/shivesh/Desktop/PythonProject/Boston Marathon Cut-Offs/m1834_forecast_2027.png
Fit years: 2014-2025  Excluded: [2021, 2022, 2023]
Predicted actual qualifying time for 2026: 178:24 (+ or - 1 STD = [177:21 , 179:26])
Predicted actual qualifying time for 2027: 177:56
2026 official cutoff buffer in dataset: 4:34
→ Model-predicted buffer for 2026 (vs 2026 standard): 6:36
→ Model-predicted buffer for 2027 (vs 2026 standard): 7:04
→ Change in buffer 2027 vs 2026 (model): +0:27
```

From the trend regression:

- **Predicted qualifying time (M 18–34)**
  - 2026: **178:24** minutes (+/− 1 STD = [177:21, 179:26])
  - 2027: **177:56** minutes
- **Predicted buffers vs the 2026 Boston Qualifier standard**
  - Official 2026 dataset buffer: **4:34**
  - Model-predicted 2026 buffer: **6:36**
  - Model-predicted 2027 buffer: **7:04**
  - **Change in buffer 2027 vs 2026**: +0:27
- **Uncertainty bands for buffer predictions**
  - 2026 buffer: [5:34, 7:39]
  - 2027 buffer: [6:01, 8:06]

Commentary

1. **Key takeaway**: The model projects that the Men 18 to 34 buffer may tighten to around **7 minutes in 2027**, about **+27 seconds stricter** than 2026.
2. **Uncertainty matters**: The +/− 1 STD bands remind us that year-to-year variation is real. Depending on registration numbers, field size, and competitive density, the cutoff could plausibly fall anywhere within the bands.
3. **Practical implication**: Athletes aiming for Boston in 2027 should prepare for a tougher cutoff than 2026, with a safe margin beyond the official BQ standard.

Bottom line for 2027 (M 18–34). If 2026 standards hold, plan for ~7:00 of buffer to be safely inside the acceptance line. The +/-1STD range (~6:01-8:06) reflects historical variability, not policy. Train for the top of the band if Boston is your A-goal.

Data Science Report:Boston Marathon Qualifying Time Forecast for 2026 And 2027

By:- Shivesh(Ethan) Raj Sahu

Conclusion

This project analyzed historical Boston Marathon qualifying times for men aged 18 to 34 between 2014 and 2025, excluding the COVID-affected years (2021–2023). Using a linear regression model, we forecasted the likely qualifying time and buffer for 2026 and extended the prediction to 2027.

Key findings:

- 2026 Predicted Actual Qualifying Time: 178:24 minutes (+/-1 SD range: 177:21 - 179:26).
- 2027 Predicted Actual Qualifying Time: 177:56 minutes.
- Buffer Analysis:
    - Official buffer for 2026: 4:34 minutes.
    - Model-predicted buffer for 2026: 6:36 minutes.
    - Model-predicted buffer for 2027: 7:04 minutes.
    - Projected increase in buffer from 2026 to 2027: +0:27 minutes.

These results suggest a continuing trend toward tighter qualification standards. If the predicted buffer increase holds, the 2027 race may demand even stronger performances from athletes compared to 2026.

Limitations and Future Work

- The analysis used a simple linear regression model. While effective for capturing trends, real-world qualifying times may not follow a purely linear trajectory.
- COVID-affected years were excluded; including them with advanced correction methods might improve robustness.
- The model assumes that 2027 will use the same age-group standards as 2026, which may not hold if BAA adjusts qualification standards.
- Further work could extend analysis across all age groups and genders, or test non-linear models (polynomial regression, time-series methods) for improved forecasting.

Closing Note

This report demonstrates the value of applying data science and machine learning techniques to sports analytics. By analyzing trends and forecasting future cutoffs, athletes and coaches can make more informed training and planning decisions.
