NFL Super Bowl Probabilities: A Comparative Analysis of ARIMA and LSTM

Abstract

This paper analyzes how efficiently betting markets update their beliefs about the Philadelphia Eagles' likelihood of winning the Super Bowl by modeling weekly implied probabilities from 2009–2025. Using ARIMA and LSTM forecasting methods, the paper examines whether futures odds behave like a random walk, which is consistent with market efficiency, or contain predictable structure. After cleaning and transforming the data, stationarity tests show that the implied probability series is stationary in levels, allowing the use of ARIMA. An expanded model search identifies ARIMA(1,0,3) as the best-fitting linear specification, revealing short-run dependence driven primarily by transient shocks rather than persistent trends. Diagnostic tests suggest remaining nonlinear structure, motivating the comparison with a Long Short–Term Memory (LSTM) neural network. The LSTM captures broader patterns and smooth trends but, predictably, cannot anticipate sharp market reactions to individual game outcomes. Forecasts from both models converge toward stable probability levels, but ARIMA reacts more strongly to recent movements while the LSTM estimates a more persistent underlying expectation. Collectively, the results indicate that most week-to-week variation in Super Bowl futures odds reflects unpredictable shocks rather than systematic time-series patterns, supporting the view that betting markets incorporate information rapidly but not perfectly.

1. Introduction and Literature Review

The goal of this project is to study how betting markets update their beliefs about the Philadelphia Eagles over time. Most discussion around sports analytics focuses on what teams are doing on the field, but futures markets give us a different type of information. They show us what the public and sportsbooks collectively believe about the team's long-run chances at any point in the season. This project uses weekly Super Bowl futures odds from 2009 to 2025 and converts them into implied probabilities. These probabilities summarize how likely the market thinks the Eagles are to win the championship in each week of each season.

The key question is whether these probabilities contain any predictable structure. If they behave like a random walk, then past movements tell us little about where the probability will go next. If there is some pattern in the updates, then more advanced forecasting tools might be able to anticipate short-run changes. This question is interesting because futures odds move for clear and public reasons. A win, a star player injury, or a surprising performance can shift expectations almost immediately. Studying these movements with time series tools gives us insight into how fast the market reacts to new information and whether the updates show any statistical regularities

Research on market efficiency provides the theoretical foundation for testing whether weekly movements in Super Bowl futures odds behave predictably or follow a random walk. In the efficient markets framework developed by Fama (1970), asset prices fully and immediately reflect available information, implying that past prices or probabilities should not systematically forecast future changes. Under weak-form efficiency, a team's Super Bowl futures odds should adjust instantly after new information becomes public. If the Philadelphia Eagles' implied win probability contains autoregressive or moving‑average structure, this would indicate gradual information incorporation rather than the instantaneous adjustment predicted by efficiency.

The time-series approach to testing these hypotheses draws from the econometric tools developed in financial economics. Campbell, Lo, and MacKinlay (1997) show how predictability can be evaluated using autoregressive models and other serial-correlation tests that distinguish efficient random-walk behavior from systematic patterns. ARIMA modeling in particular provides a flexible framework to test whether price or probability changes are driven solely by new information shocks. Hamilton (1994) formalizes ARIMA identification and estimation procedures making it well suited for analyzing how odds evolve over the course of an NFL season.

Sports-betting markets represent a natural environment for studying information flows and pricing efficiency. Sauer (1998) reviews the economics of wagering markets and highlights both efficient behavior and persistent anomalies driven by bettor biases, information frictions, or bookmaker objectives. Levitt (2004) argues that sportsbooks may intentionally shade lines to balance action or exploit biased beliefs, meaning odds may not perfectly reflect true underlying probabilities. This provides a strong rationale for examining whether weekly futures odds adjust smoothly rather than instantly. Empirical work specific to the NFL supports this view: Paul and Weinbach (2008) show that bookmakers respond to public sentiment and strategic considerations when setting prices, contributing to predictable components in betting lines.

Recent advances in machine-learning–based time-series modeling offer an additional perspective on how betting markets incorporate information over time. Long Short-Term Memory (LSTM) neural networks, introduced by Hochreiter and Schmidhuber (1997), are designed to capture nonlinear and long-range temporal dependencies that traditional linear models may miss. Because LSTMs can learn complex updating rules from sequential data without imposing fixed autoregressive or moving-average structure, they provide a flexible benchmark for testing whether movements in Super Bowl futures odds contain subtle patterns inconsistent with efficient markets.

In financial applications, researchers have demonstrated that LSTM models can outperform classical econometric approaches when price series exhibit nonlinearities, structural changes, or regime shifts. Fischer and Krauss (2018) show that deep recurrent networks can generate more accurate forecasts than traditional time-series models across a wide set of equities. Applying an LSTM framework to the Eagles' weekly Super Bowl futures odds therefore enables a direct comparison between linear dynamics and a high-capacity model that may reveal slower or more complex information incorporation.

Finally, futures odds can be interpreted as probability forecasts of the Eagles winning the Super Bowl, and this interpretation is supported by research on prediction markets. Wolfers and Zitzewitz (2004) argue that in well-functioning prediction markets, prices aggregate dispersed information and can be viewed as consensus probabilities. Any deviation from random-walk behavior in these implied probabilities suggests slow information incorporation or systematic market pressures inconsistent with full efficiency. This provides the empirical motivation for modeling the Eagles' weekly Super Bowl futures odds with an ARIMA framework and using forecast performance to assess how efficiently the betting market processes new information.

2. Econometric Model and Justification

The main econometric model used in this project is an ARIMA model. Before estimating any ARIMA, the first step is to test whether the series is stationary. A stationary series has a constant mean and variance over time, and its shocks do not persist forever. When a series is not stationary, an ARIMA model with differencing is often used. This is important because most forecasting tools assume stationarity. I begin by applying the Augmented Dickey Fuller test to the level implied probability series. The test has a high p-value, so I reject the null hypothesis of a unit root. This means the series is stationary in levels.

To choose the AR and MA components, I use an automated model selection process based on AIC and BIC. Both criteria point toward an ARIMA(1,0,3) model. This is a simple structure that is common for financial style probabilities. An ARIMA(1,0,3) model says that the best predictor of next week's level is this week's level, but short-run noise has a one-period effect that fades out. This matches how the implied probability behaves in practice. Most weeks see very small changes, and when a shock does occur, the market settles quickly.

However, ARIMA is limited in one way. It can only capture linear dynamics. If the probability responds to news in a nonlinear way or has longer cycles, ARIMA cannot detect them. To explore this possibility, I also estimate an LSTM model. The LSTM is trained on rolling sequences of 12 weeks of data to predict the next week. The idea is that the LSTM can uncover patterns that ARIMA cannot. In practice, the LSTM learns the general shape of the series but tends to smooth out sudden jumps. This makes sense because large movements in implied probability often come from a single event that the past sequence does not contain.

Together, ARIMA and LSTM provide a complete picture of the predictability of the weekly implied probability. ARIMA gives a simple and interpretable model that aligns with basic economic reasoning. The LSTM gives a flexible alternative that tests for more complex structure. Both produce forecasts that tend to remain near the current probability unless there is a strong pattern in recent weeks. This supports the idea that the market incorporates new information quickly and that past probability movements have limited predictive power.

3. Data Description, Plots, and Stationarity Testing

The data used in this project come from weekly Super Bowl futures odds for the Philadelphia Eagles. The raw data contain one row per sportsbook per date. Because sportsbooks update odds at different times and sometimes skip weeks, the first step is to convert this into a clean weekly time series.

I begin by converting American odds into implied probabilities. Some rows contain labels or missing entries. These are removed.I convert the dates and times to numerical values beginning with the 2009 preseason probability as 0, moving upward from there, in order to make the time series cleaner. We are left with 684 observations of bi-weekly in season Super Bowl odds for the Philadelphia Eagles.

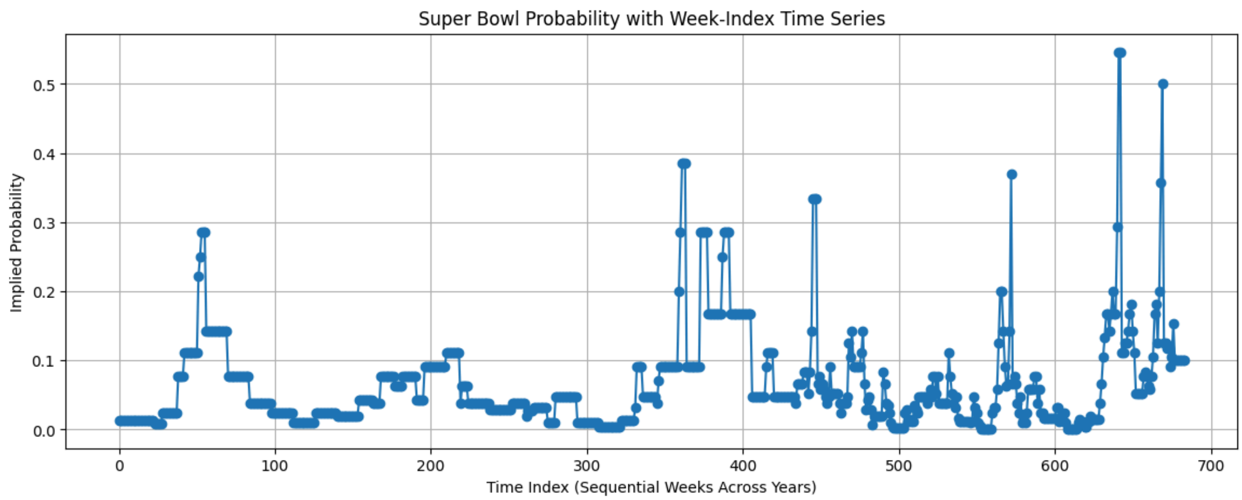

Figure 1: Probability Plot

When plotted, the series has several clear episodes. For many years the expectation stays low, usually between zero and ten percent. During the 2017 and 2024 seasons, the probability rises sharply as the Eagles climb to the top of the league. There is a similar but smaller rise in 2022. This is predictable, as these observations line up with the years in which the Eagles have been successful. Outside of these periods, the series moves in small increments. These long stretches of low volatility are common in futures markets, especially early in the season.

The Augmented Dickey–Fuller test on the implied probability series yields an ADF statistic of –5.03 with a p-value near 1.9. Although the implied Super Bowl probability series shows jumps and volatility consistent with betting-market adjustments, the ADF test strongly rejects the presence of a unit root. This means the series is a random walk. Instead, it behaves like a stationary jump process where shocks are temporary and values revert to a stable range over time. Conceptually, this implies that the betting market treats past success as only a small indicator or future success. The higher probability ranges only occur during the end of season or playoff, when there are only a few teams mathematically that have a chance.

Overall, the cleaned dataset reflects a realistic representation of market expectations. The data have enough variation to estimate models, but the majority of the updates are small. This creates a forecasting environment where changes are hard to predict unless major trends are present. This helps explain the results of both the ARIMA and LSTM models.

4. Results

An AIC-based grid search over ARIMA specifications selected an ARIMA(1,0,3) model as the best fit for the weekly Super Bowl implied probability series. This result indicates that the series is stationary in levels and exhibits short-memory dynamics that can be captured by a single autoregressive term along with multiple moving-average components. The presence of several MA terms suggests that much of the short-run variation reflects transitory shocks whose effects dissipate quickly, rather than persistent changes in expectations. Overall, the selected model implies that betting odds respond to recent information but do not display long-lasting momentum or trend behavior

The model diagnostics support this specification. The AR(1) coefficient is large and positive and highly significant, which suggests that recent movements in implied probability pass through one period ahead but decay quickly. All three MA terms are strongly significant, and the combined structure absorbs the high-frequency noise that is common in betting market data. However, the diagnostic tests also highlight the limitations of the model. The Ljung-Box statistic rejects the null of no serial correlation in the residuals, which suggests the presence of remaining structure that the ARIMA(1, 0, 3) form does not fully capture. Despite these issues, ARIMA(1, 0, 3) remains the best linear model among the tested alternatives and provides a reasonable benchmark for understanding how betting markets evolve over time.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:            implied_prob   No. Observations:                  684
Model:                  ARIMA(1, 0, 3)   Log Likelihood                1267.260
Date:                 Mon, 08 Dec 2025   AIC                          -2522.520
Time:                         23:39:34   BIC                          -2495.353
Sample:                              0   HQIC                         -2512.007
                                 - 684
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0665      0.026      2.548      0.011       0.015       0.118
ar.L1          0.9391      0.028     33.273      0.000       0.884       0.994
ma.L1         -0.0365      0.033     -1.113      0.266      -0.101       0.028
ma.L2         -0.1949      0.039     -4.959      0.000      -0.272      -0.118
ma.L3         -0.2130      0.031     -6.943      0.000      -0.273      -0.153
sigma2         0.0014   3.05e-05     47.135      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):             20366.82
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               9.53   Skew:                            -0.68
Prob(H) (two-sided):                  0.00   Kurtosis:                        29.70
===================================================================================
```

Figure 2: ARIMA(1,0,3) Results

The ARIMA forecast shows a moderate decline in the Eagles Super Bowl probability over the next year, moving from values near ten percent toward a level closer to seven percent. This pattern fits the structure of the underlying model, where the strong MA terms dampen short bursts in the data and pull the forecast toward a smoother trajectory. The confidence band widens quickly as the horizon increases, which is expected for a series driven by sharp weekly updates and rare but large jumps. The widening interval also highlights the limits of purely linear models when applied to betting market data that can shift abruptly based on news, injuries, or game outcomes. Even with this uncertainty, the central forecast suggests that the market expects a stable but slightly weaker outlook relative to the most recent observations.

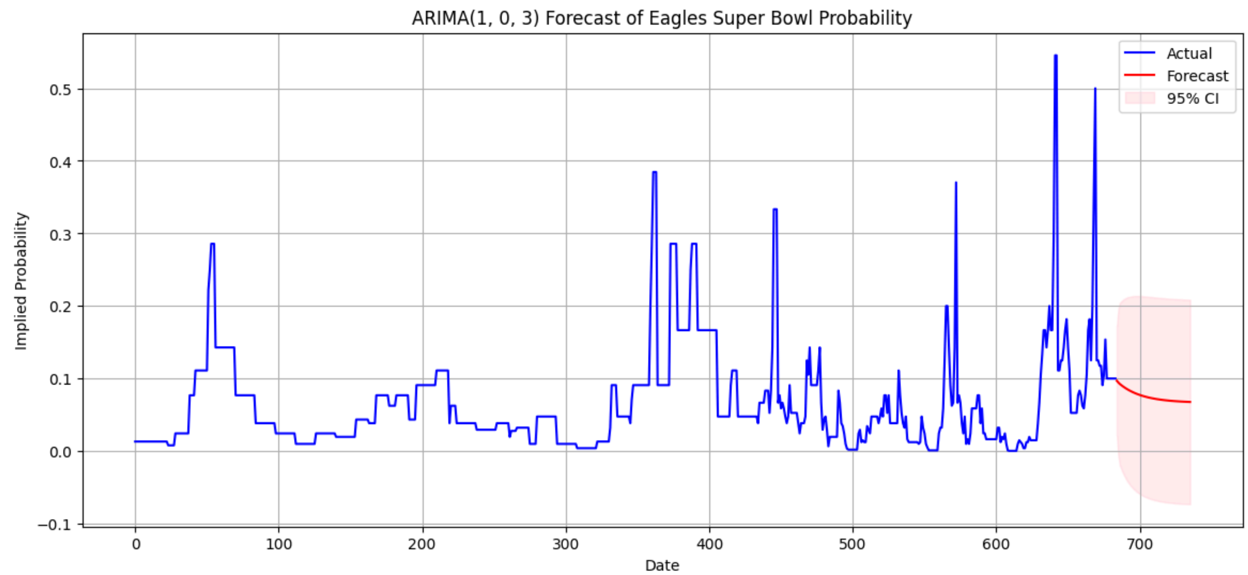

Figure 3: ARIMA(1,0,3) Forecast

The LSTM model tracks the broad movements in implied Super Bowl probabilities and captures the slow drifts and medium-sized swings that occur as the season unfolds. The predictions follow the general shape of the actual series and reproduce the timing of several increases and declines, although the model smooths out the sharpest jumps that appear after major game outcomes or sudden changes in team outlook. This is expected because the LSTM is learning from past sequences rather than reacting to individual shocks. The model performs well when the series moves gradually but tends to understate the size of extreme spikes and drops, which mirrors how neural networks often behave when trained on noisy, jump-driven sports markets. Overall, the LSTM provides a reasonable approximation of the underlying pattern and demonstrates some predictive value, especially in periods without major shocks.

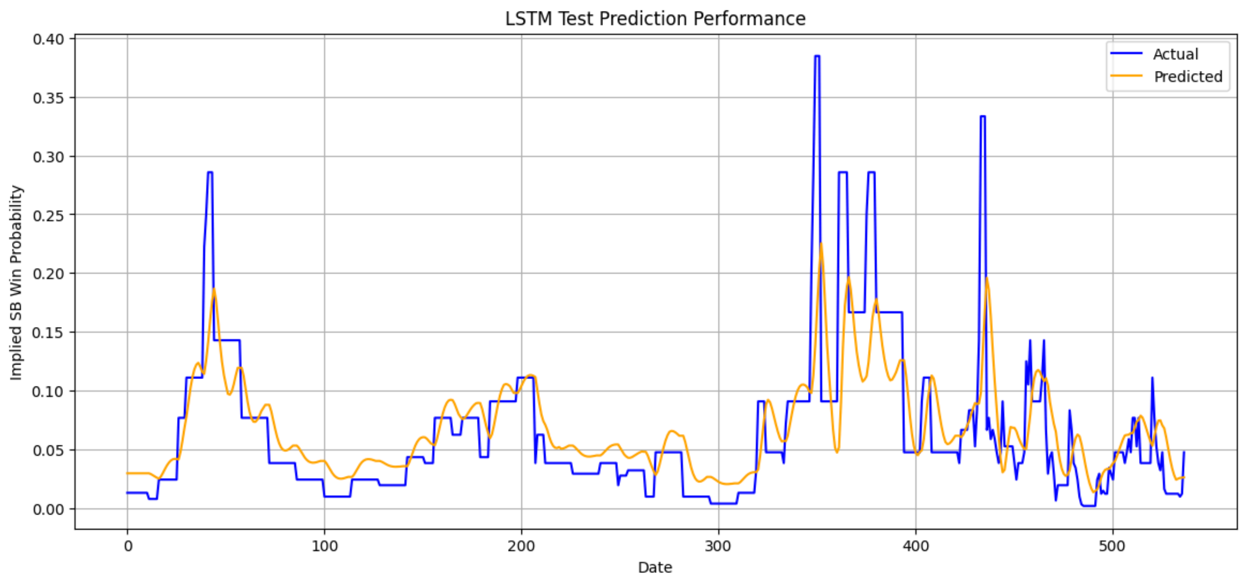

Figure 4: LSTM Prediction Performance

The LSTM forecast traces a smooth, low-volatility path that quickly converges to a steady level and then changes only slightly over the 52-week horizon. This behavior reflects the network's tendency to average over a highly jump-driven series: rather than anticipating discrete events, the LSTM effectively predicts the conditional baseline probability implied by recent history. Because the observed market probabilities move in sharp bursts after games, injuries, and other news, those shocks are not learnable from the time index alone and the model defaults to a gradual continuation toward its learned long-run level. As a result, the LSTM provides a sensible "no-new-information" forecast of underlying team strength, while underscoring the gap between high-frequency market volatility and the smoother persistent structure captured from past sequences.

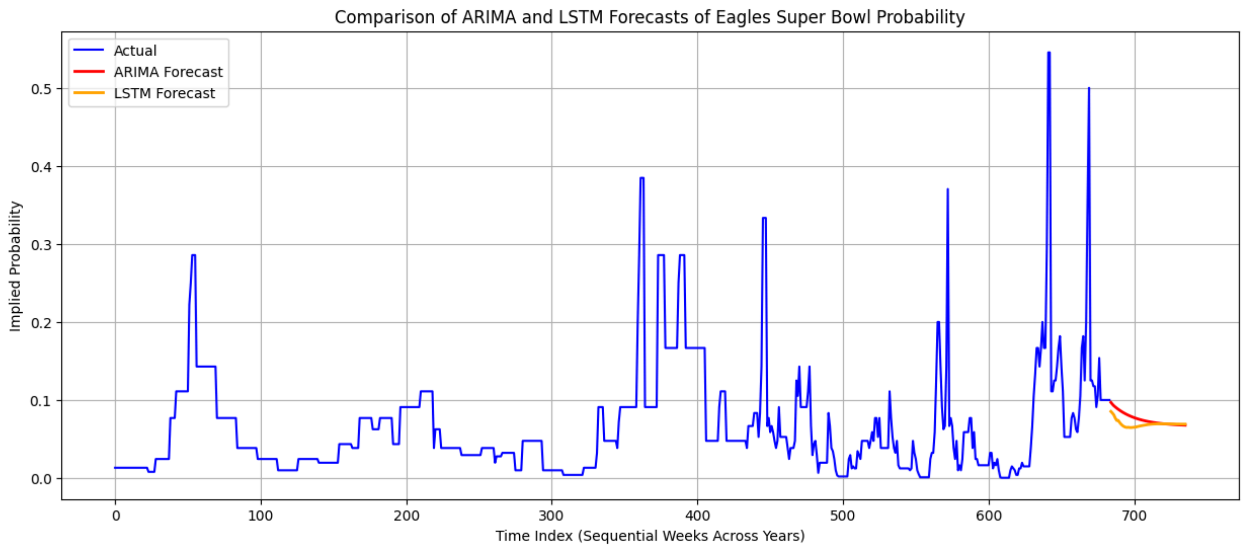

Figure 5: ARIMA vs LSTM Forcasts

The ARIMA and LSTM forecasts illustrate two distinct interpretations of how Super Bowl probabilities evolve over time. The ARIMA model places greater weight on short-run dynamics, producing a forecast that adjusts noticeably in response to the most recent movements in the series and gradually decays as past shocks propagate through the autoregressive structure. In contrast, the LSTM emphasizes broader patterns across longer sequences and largely abstracts from high-frequency volatility. As a result, the LSTM forecast is much smoother and remains close to the recent baseline level of implied probability, rather than reacting strongly to the latest fluctuations. Both approaches capture meaningful features of the data but answer different questions: ARIMA describes how probabilities typically respond to recent shocks, while the LSTM provides an estimate of the expected path absent new information. Taken together, the results suggest that much of the week-to-week variation in betting odds is driven by unpredictable, event-based shocks rather than stable, forecastable time-series dynamics. Trying to predict betting odds without including some proxy for this information will prove difficult, no matter the model.

Bibliography

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691043012/the-econometrics-of-financial-markets

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance, 25*(2), 383–417. https://doi.org/10.2307/2325486

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691042893/time-series-analysis

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *Economic Journal, 114*(495), 223–246. https://doi.org/10.1046/j.0013-0133.2003.00197.x

Paul, R. J., & Weinbach, A. P. (2008). Price setting in the NFL betting market. *Quarterly Review of Economics and Finance, 48*(1), 95–107. https://doi.org/10.1016/j.qref.2006.12.003

Sauer, R. D. (1998). The economics of wagering markets. *Journal of Economic Perspectives, 14*(3), 103–123. https://doi.org/10.1257/jep.14.3.103

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives, 18*(2), 107–126. https://doi.org/10.1257/0895330041371321

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054