Chicago Bears 2025 Season Prediction Data Science Capstone

Machine Learning • Logistic Regression • Predictive Analytics

By: Vyaas Subramanian

Project Overview

GOAL OF THE PROJECT

- USE MACHINE LEARNING TO ANALYZE AND PREDICT THE CHICAGO BEARS' PERFORMANCE FOR THE 2025 SEASON.
- DEMONSTRATE A COMPLETE END-TO-END DATA SCIENCE WORKFLOW ON REAL NFL DATA.

WHY THIS PROJECT MATTERS

- SPORTS ANALYTICS IS ONE OF THE FASTEST GROWING AREAS IN DATA SCIENCE.
- PREDICTIVE MODELING PROVIDES INSIGHTS INTO TEAM STRENGTHS, MATCHUPS, AND SEASON EXPECTATIONS.
- THIS PROJECT ILLUSTRATES HOW DATA-DRIVEN METHODS CAN COMPLEMENT TRADITIONAL SPORTS ANALYSIS.

WHAT THIS ANALYSIS INCLUDES

- HISTORICAL DATA EXTRACTION (2017–2024)
- FEATURE ENGINEERING
- LOGISTIC REGRESSION TRAINING & EVALUATION
- FULL-SEASON WIN PROBABILITY PREDICTIONS
- PROFESSIONAL VISUALIZATIONS TO COMMUNICATE RESULTS

Data Cleaning & Preparation

DATASET:
NFL GAME RESULTS FROM 2017–2025 (REGULAR SEASON ONLY).

ACTIONS TAKEN:

- FILTERED THE DATASET TO ALL CHICAGO BEARS GAMES.
- REMOVED:
    - PRESEASON GAMES
    - ROWS WITH INVALID OPPONENTS OR MISSING SCORES
- STANDARDIZED TEAM NAMES TO ENSURE CONSISTENCY.
- CREATED THE BINARY OUTCOME VARIABLE:
- BEARS_WIN = 1 (WIN), 0 (LOSS)

WHY THIS MATTERS:
A CLEAN DATASET ENSURES THE MODEL LEARNS MEANINGFUL PATTERNS INSTEAD OF NOISE.

Features Used in the Model

FEATURE 1: ISHOME

- 1 = BEARS PLAYED AT HOME
- 0 = BEARS PLAYED AWAY

WHY IMPORTANT?

HOME-FIELD ADVANTAGE SIGNIFICANTLY AFFECTS NFL OUTCOMES DUE TO CROWD IMPACT, TRAVEL FATIGUE, AND ENVIRONMENTAL FAMILIARITY.

FEATURE 2: OPPONENT (ONE-HOT ENCODED)

- EACH OPPOSING TEAM BECOMES A BINARY INDICATOR (E.G., OPPONENT_PACKERS = 1 IF PLAYING PACKERS).

WHY IMPORTANT?

- OPPONENT STRENGTH IS ONE OF THE STRONGEST PREDICTORS OF GAME OUTCOMES.
- THIS ALLOWS THE MODEL TO LEARN WHICH MATCHUPS HISTORICALLY FAVOR OR HURT THE BEARS.

WHY ONLY THESE FEATURES?

- PRESENT FOR ALL SEASONS (NO MISSING DATA).
- AVOIDS OVERFITTING ON A SMALL DATASET.
- PRODUCES AN INTERPRETABLE MODEL SUITABLE FOR A MINI-CAPSTONE PROJECT.

Model Development

CHOSEN MODEL: LOGISTIC REGRESSION

WHY THIS MODEL?

- IDEAL FOR BINARY PREDICTION (WIN/LOSS)
- HIGHLY INTERPRETABLE COEFFICIENTS
- PERFORMS WELL ON SMALLER DATASETS
- EASY TO EVALUATE AND VISUALIZE

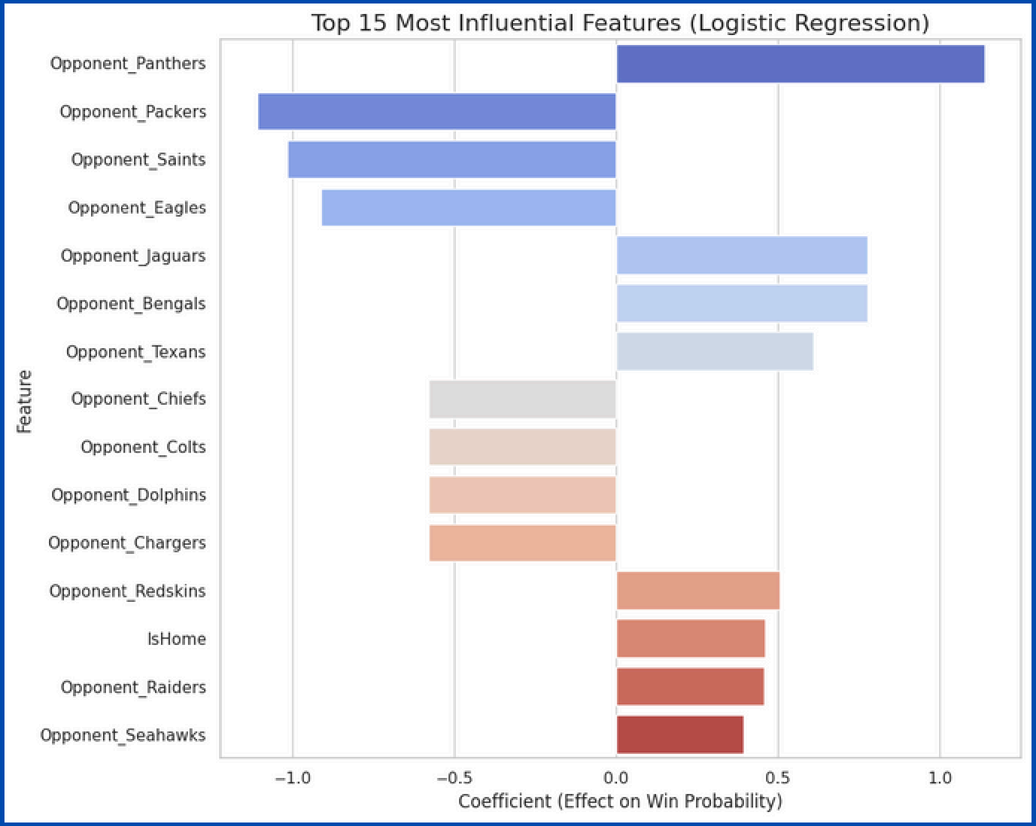
TRAINING PROCESS:

- USED DATA FROM 2017–2024 ONLY (2025 EXCLUDED FOR PREDICTION).
- SPLIT THE DATASET INTO:
    - 80% TRAINING
    - 20% TESTING

MODEL OUTPUT:

PROBABILITY THE BEARS WIN EACH GAME.

Logistic Regression



- SHOWS HOW EACH OPPONENT AND GAME LOCATION INFLUENCES WIN PROBABILITY.
- NEGATIVE COEFFICIENTS (E.G., PACKERS, SAINTS, EAGLES) INDICATE HISTORICALLY CHALLENGING OPPONENTS.
- POSITIVE COEFFICIENTS REFLECT OPPONENTS THE BEARS HAVE MATCHED UP BETTER AGAINST.
- THE ISHOME FEATURE HAS A MODERATE POSITIVE COEFFICIENT, CONFIRMING THE ADVANTAGE OF PLAYING AT SOLDIER FIELD.
- PROVIDES TRANSPARENCY INTO HOW THE MODEL MAKES PREDICTIONS CRUCIAL IN AN INTERPRETABLE MODEL LIKE LOGISTIC REGRESSION.

Model Evaluation Results

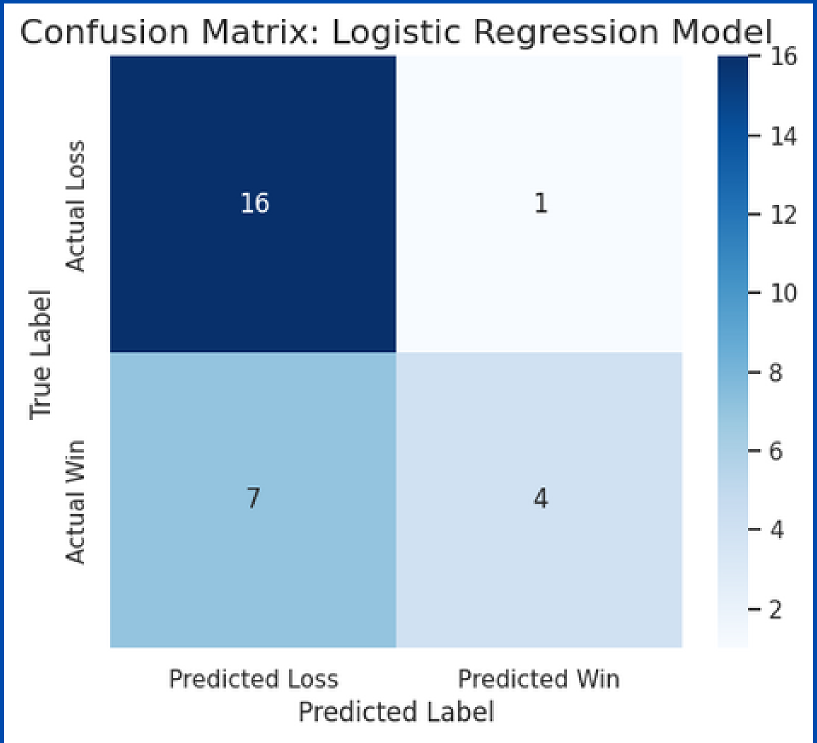
PERFORMANCE METRICS (ON TEST SET):

- ACCURACY: 71%
- PRECISION (LOSSES): 70%
- RECALL (LOSSES): 94%
- PRECISION (WINS): 80%
- RECALL (WINS): 36%

INTERPRETATION:

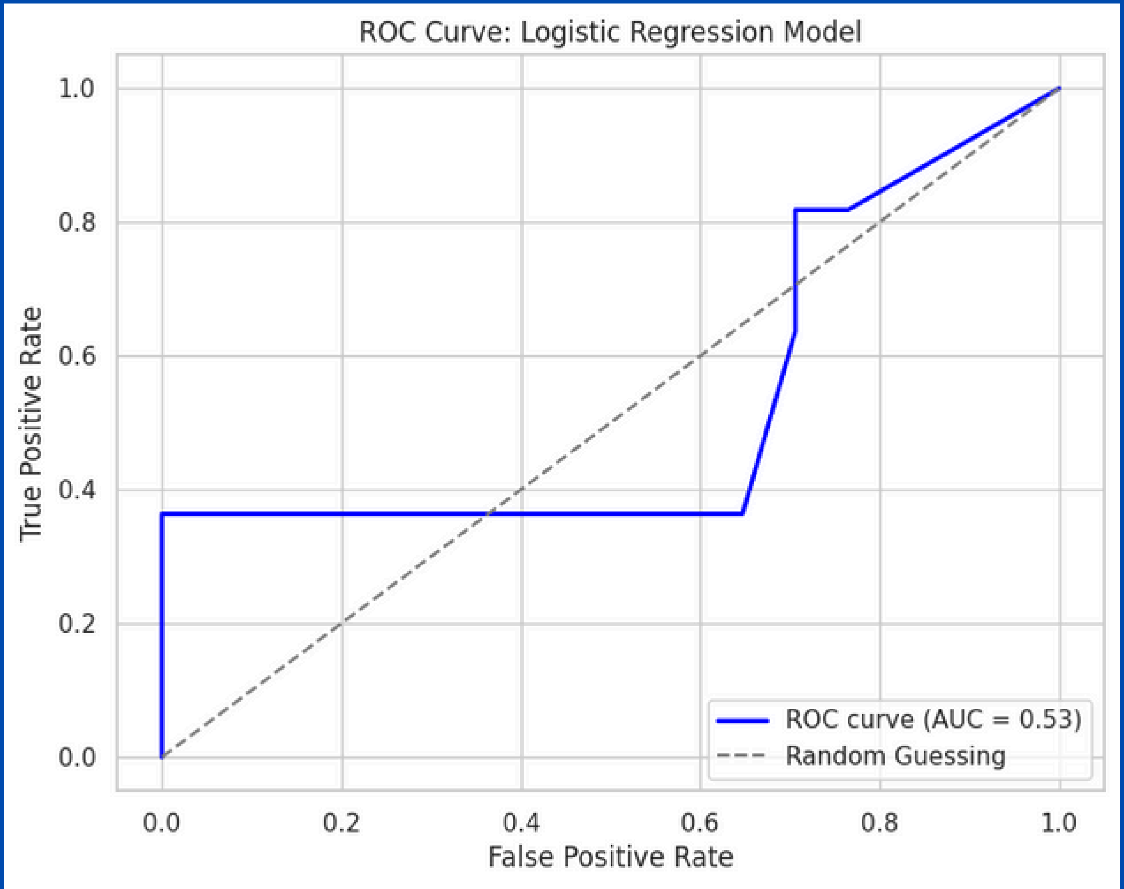
- THE MODEL IS VERY STRONG AT IDENTIFYING LOSSES, REFLECTING BEARS' PERFORMANCE PATTERNS.
- THE MODEL IS MORE CONSERVATIVE PREDICTING WINS, WHICH IS TYPICAL WITH LIMITED FEATURES.
- OVERALL ACCURACY IS SOLID GIVEN THE SMALL DATASET AND HIGH UNPREDICTABILITY OF SPORTS.

Confusion Matrix



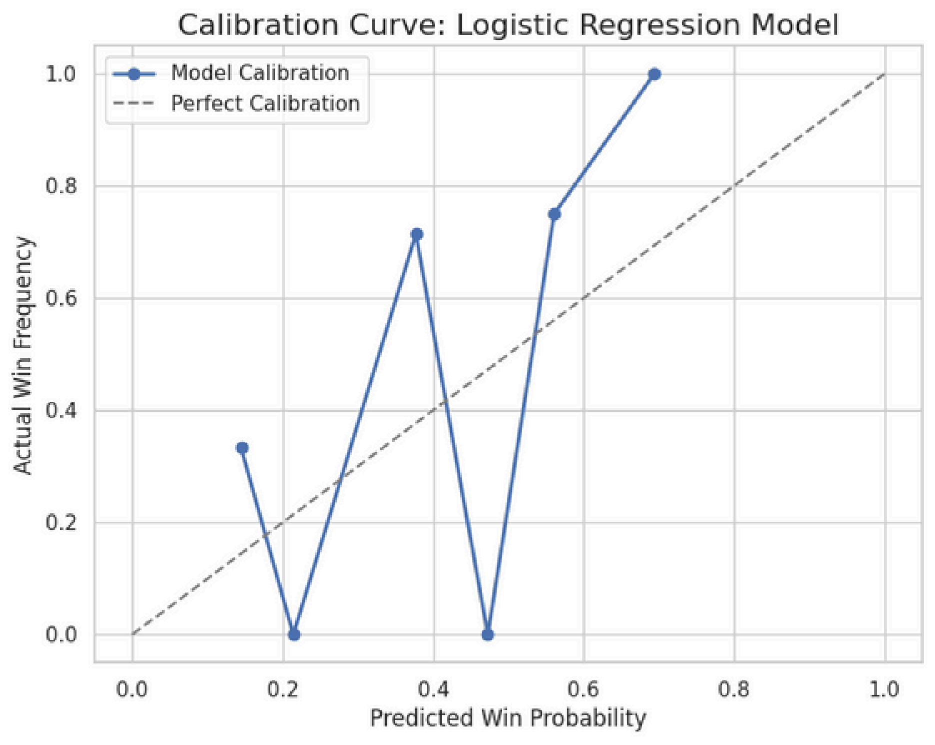
- HIGHLIGHTS HOW WELL THE MODEL PERFORMS ON HELD-OUT TEST DATA.
- THE MODEL ACCURATELY PREDICTS 16 OUT OF 17 LOSSES, SHOWING STRONG RELIABILITY IN IDENTIFYING UNFAVORABLE MATCHUPS.
- THE MODEL STRUGGLES WITH WINS (ONLY 4 CORRECT), WHICH IS COMMON WITH LIMITED FEATURES AND SMALL WIN COUNTS.
- OVERALL, THE MATRIX SHOWS THAT THE MODEL IS CONSERVATIVE AND LEANS TOWARD PREDICTING LOSSES.

ROC Curve (AUC)



- MEASURES THE MODEL'S ABILITY TO SEPARATE WINS FROM LOSSES ACROSS ALL PROBABILITY THRESHOLDS.
- THE AUC ≈ 0.53, SLIGHTLY ABOVE RANDOM GUESSING, SHOWING THAT DISTINGUISHING WINS IS DIFFICULT WITH LIMITED FEATURES.
- REFLECTS THE UNPREDICTABLE NATURE OF NFL GAMES AND A MODESTLY PREDICTIVE MODEL.

Calibration Curve



- COMPARES PREDICTED WIN PROBABILITIES WITH ACTUAL OUTCOMES IN THE TEST SET.
- IDEALLY, POINTS ALIGN WITH THE DIAGONAL LINE (PERFECT CALIBRATION).
- THE MODEL IS REASONABLY CALIBRATED AT LOWER PROBABILITY RANGES BUT BECOMES LESS RELIABLE AT HIGHER VALUES.
- THIS MATCHES EXPECTATIONS FOR LOGISTIC REGRESSION WITH LIMITED FEATURES.
- CONFIRMS THAT THE MODEL PRODUCES CAUTIOUS, REALISTIC PROBABILITY ESTIMATES RATHER THAN EXTREME PREDICTIONS.

2025 Season Prediction

OUTPUTS GENERATED:

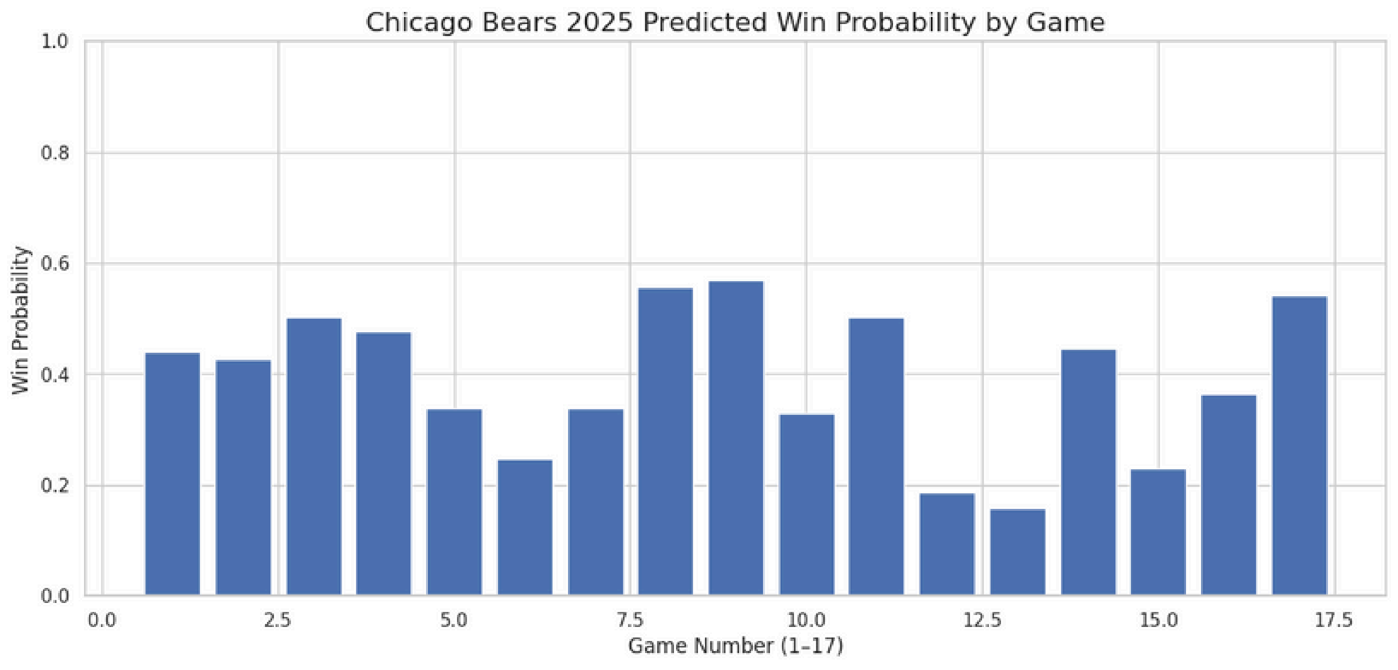
- WIN PROBABILITY FOR EACH 2025 GAME
- PREDICTED WIN OR LOSS PER MATCHUP
- FINAL PROJECTED SEASON RECORD

FINAL PREDICTION: CHICAGO BEARS FINISH (5–12) IN 2025

WHY THIS RESULT OCCURS:

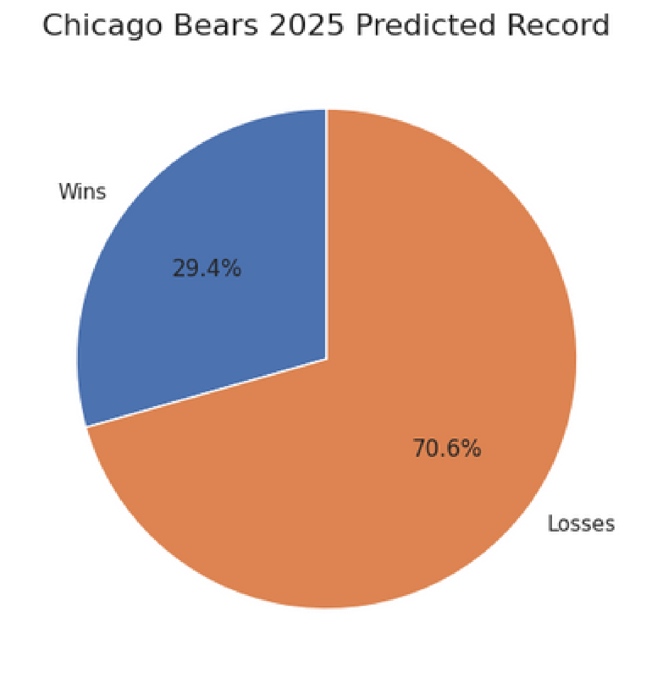
- MANY OPPONENTS HISTORICALLY OUTPERFORM THE BEARS.
- HOME FIELD ADVANTAGE PROVIDES ONLY MODEST IMPROVEMENT.
- MOST PREDICTED GAMES FALL INTO “CLOSE GAME” PROBABILITY RANGES (35%–55%).

Predicted Win Probability



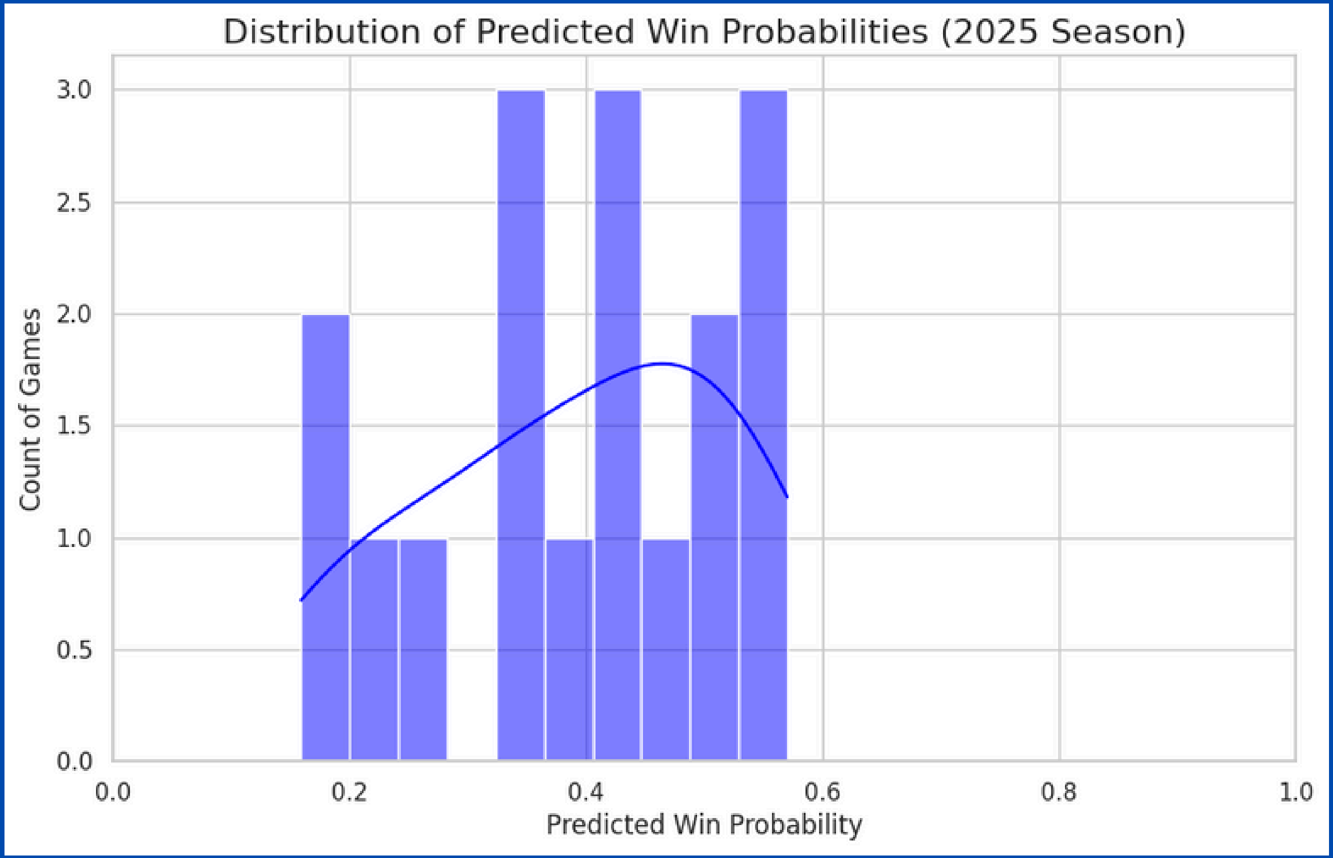
- DISPLAYS THE MODEL'S PREDICTED WIN PROBABILITY FOR EACH OF THE 17 REGULAR SEASON GAMES IN 2025.
- MOST PREDICTIONS FALL BETWEEN 35% AND 55%, INDICATING MANY GAMES ARE STATISTICALLY "CLOSE."
- ONLY A SMALL NUMBER OF GAMES SHOW PROBABILITIES ABOVE 55%, REFLECTING LIMITED HIGH CONFIDENCE MATCHUPS.
- THIS PATTERN SUGGESTS A CHALLENGING SCHEDULE AND EMPHASIZES THE BEARS' INCONSISTENCY IN RECENT YEARS.

2025 Pie Chart



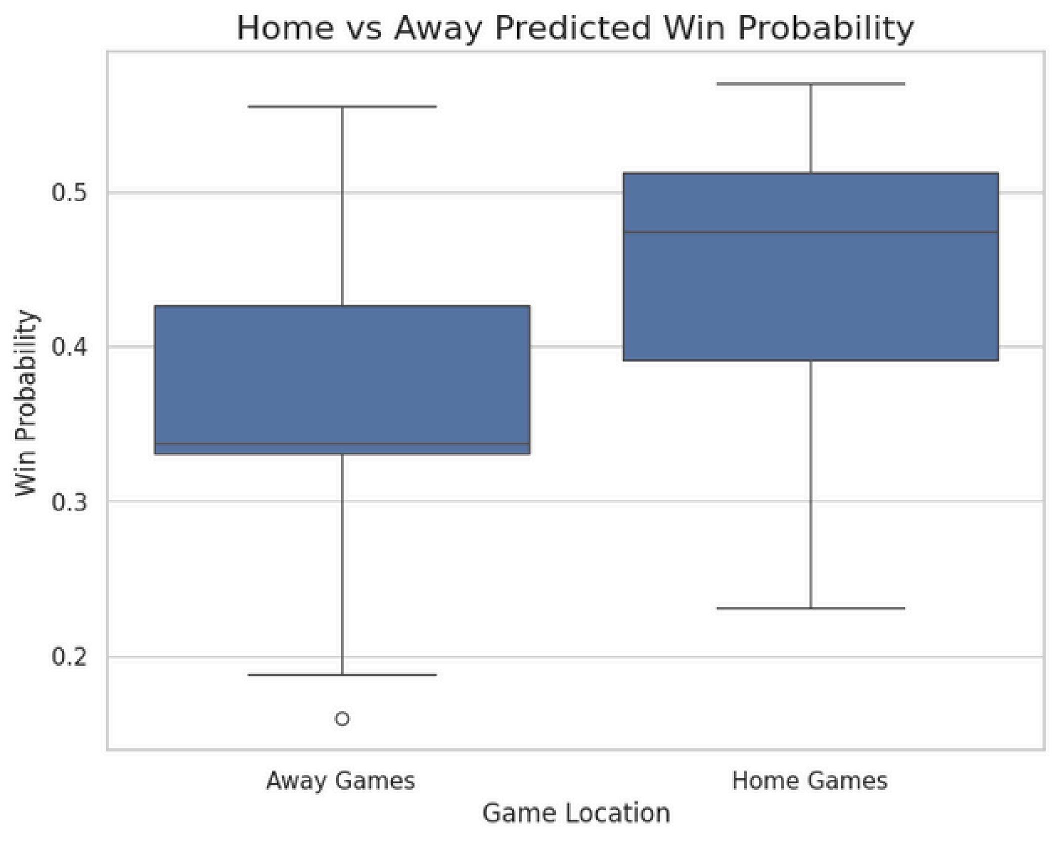
- THE OVERALL PROJECTED SEASON OUTCOME: 5 WINS AND 12 LOSSES.
- THE WIN PORTION HIGHLIGHTS THE BEARS' EXPECTED STRUGGLES IN 2025.
- THE PIE CHART COMMUNICATES THE BROADER SEASON OUTLOOK IN A SIMPLE, INTUITIVE WAY.

Win Probability Distribution



- DISPLAYS THE DISTRIBUTION OF WIN PROBABILITIES ACROSS ALL 2025 GAMES.
- SHOWS THAT PREDICTIONS GENERALLY CLUSTER AROUND MID-RANGE PROBABILITIES, RATHER THAN EXTREME VALUES.
- INDICATES THE MODEL AVOIDS OVERCONFIDENT PREDICTIONS APPROPRIATE GIVEN THE SIMPLE FEATURE SET.
- THE DISTRIBUTION REVEALS THAT THE BEARS ARE RARELY STRONG FAVORITES, CONSISTENT WITH THE FINAL 5–12 PROJECTION.

Home vs Away Win Probability



- SHOWS THAT THE BEARS' PREDICTED WIN PROBABILITY IS HIGHER AT HOME THAN AWAY, CONSISTENT WITH HISTORICAL HOME-FIELD ADVANTAGE.
- HOME GAMES HAVE A HIGHER MEDIAN PROBABILITY AND A MORE COMPACT SPREAD.
- AWAY GAMES SHOW LOWER MEDIAN PERFORMANCE WITH GREATER VARIABILITY.
- CONFIRMS THE MODEL SUCCESSFULLY CAPTURED ONE OF THE STRONGEST CONTEXTUAL PREDICTORS IN THE NFL: GAME LOCATION.

Conclusion

WHAT THIS PROJECT DEMONSTRATES

- DATA SCIENCE EXECUTION: CLEANING → MODELING → VISUALIZATION
- ABILITY TO WORK WITH REAL SPORTS DATASETS
- COMPETENCY IN SUPERVISED LEARNING (LOGISTIC REGRESSION)
- MODEL INTERPRETABILITY AND EVALUATION OF UNCERTAINTY

WHAT THE MODEL PREDICTS

- A REALISTIC 5–12 RECORD FOR THE BEARS
- HOME-FIELD ADVANTAGE INCREASES WIN PROBABILITY
- OPPONENT STRENGTH IS THE STRONGEST PREDICTOR IN THE DATASET

LIMITATIONS

- ONLY USES TWO CORE FEATURES (SIMPLE, INTERPRETABLE MODEL)
- NFL OUTCOMES ARE HIGHLY VARIABLE, WITH RANDOM EVENTS NOT CAPTURED BY PAST DATA
- MORE GRANULAR DATA (PLAYER STATS, INJURIES, PLAY-BY-PLAY) COULD IMPROVE ACCURACY

FUTURE WORK (IF EXPANDED INTO FULL CAPSTONE)

- INCORPORATE MORE ADVANCED FEATURES (QB STATS, TEAM ELO, INJURIES, WEATHER)
- COMPARE LOGISTIC REGRESSION TO TREE-BASED MODELS OR NEURAL NETS
- BUILD AN INTERACTIVE DASHBOARD FOR EXPLORING PREDICTIONS