Introduction to Machine Learning Project 3 - Competition Football pass prediction

DOMINIKA PIECHOTA and PAWEŁ DOROSZ

This report documents the end-to-end development of a machine learning system for predicting pass receivers in football matches. The problem is formulated as a *Learning to Rank* (LTR) task, where the model ranks 22 potential candidates for receivers for each pass. Our investigation followed a rigorous chronological evolution: starting from baseline neural networks (MLP) using raw coordinates and some basic features, progressing to heuristic-based features, and using advanced geometric computations. Through feature importance analysis we achieved a final Top-1 Accuracy of 42.15% on a chronologically split validation set using a mix (50:50) of LightGBM and XGBoost Rankers.

1 Introduction and Problem Definition

The goal of this project is to identify the actual receiver of a pass while having given a snapshot of player positions at the moment of the pass.
Our first model was treating it as a simple classification problem, but it ignores the relative nature of football decision-making. A player is chosen not just because they are "good," but because they are "better option" than others. Therefore, we defined the problem as *Learning to Rank (LTR).*

2 Investigation of Approaches and Feature Evolution

The research was conducted in distinct phases. Each phase introduced new modeling techniques and refined the feature set based on accuracy and brier score analysis, as well as feature importance scores.

2.1 Phase 1: The Baseline (Naive Physics)

Model: Multi-Layer Perceptron (MLP) Classifier, Simple LightGBM Classifier.
Objective: Binary Classification (Is Receiver? 0/1).
In the initial phase, we attempted to model the problem using raw Euclidean physics without domain knowledge.
Feature Set A (Basic):

- `x_s, y_s, x_r, y_r`: $(x_{sender}, y_{sender}), (x_{candidate}, y_{candidate})$.
- `dx, dy`:, Differences $x_{sender} - x_{candidate}, y_{sender} - y_{candidate}$.
- `distance`: $\sqrt{dx^2 + dy^2}$.
- `angle`: $\arctan 2(dy, dx)$.
- `same_team`: teammates or opponents?

Result Analysis: Accuracy: 30.2%, Brier-score: 0.952. The MLP and LightGBM Classifier struggled to capture relationships, achieving an accuracy of only $\approx$ 30.2%. The model heavily biased predictions towards the nearest player, ignoring opponents standing directly in the passing path. This confirmed that raw distance is an insufficient predictor. The MLP model gave a high Brier score because it treated each player as an independent binary event, resulting in poorly calibrated probabilities.

2.2 Phase 2: Tactical Heuristics (Modeling Pressure)

Model: LightGBM Classifier.
To make the model more aware of the actual game scheme, we introduced features representing opponents pressure. We hypothesized that players under heavy pressure are unlikely targets.
Feature Set B (Pressure): Added to Set A.

- `receiver_closest_opponent_dist`: Distance between receiver and the closest opponent to him.
- `sender_closest_opponent_dist`: Distance between sender and the closest opponent to him.
- `receiver_closest_3_opponents_dist`: average distance of three nearest opponents to the receiver
- `receiver_closest_3_teammates_dist`: average distance of three nearest friendly players to the receiver
- `receiver_closest_opp_to_sender_dist`: distance from the sender to the reciever's closest opponent

Result Analysis: Accuracy improved to ≈ 36.66%. Brier-score - still very high.

2.3 Phase 3: Tactical Heuristics

Model: LightGBM Classifier.
Feature Set C (Pressure/Manual penalties): Added to Set B.

- `opponents_in_radius`: Count of opponents within a 6-meter radius.
- `direction`: Penalizes backward play and rewards forward play
- `local_overload`: Difference in the number of teammates and opponents in a radius of 6m
- `pressure_diff`: Difference between distance from the receiver to his closest opponent and from the sender to his closest opponent
- `Manual Penalties`: We experimented with hard-coded logic, adding arbitrary penalty values (e.g., +600) to the ranking score if an opponent was close(<6m).
- `closest_opponent_to_pass_line`: Computed min distance to the pass line.

Result Analysis: Accuracy improved to ≈ 38.2%. The model successfully learned to avoid players having many opponents nearby. However, the manual penalty system was brittle and did not generalize well to long passes where the "danger zone" is larger. The LGBMClassifier achieved high Brier scores because its binary logloss objective focused on global score classification rather than on the distribution of relative probabilities among the 22 candidates within a specific event.

2.4 Phase 4: The Geometric Features and LTR introduction

Model: LightGBM Ranker.

Replacing the classifier model with a ranker significantly improved the results, as the model finally fit the task. In addition, a new feature appeared: an ellipse containing both the sender (at one end) and the receiver (a short distance before the other end of the ellipse).

2.4.1 Ellipse Mathematical Formulation

. **Ellipse creation (size adjustments)**
Let $S(x_s, y_s)$ be the position of the sender and $R(x_r, y_r)$ be the position of a potential receiver.
The pass vector is defined as $\vec{V} = (dx, dy)$, where $dx = x_r - x_s$ and $dy = y_r - y_s$.

The geometry of the ellipse is defined by the following steps:

(1) **Center Placement:** The center of the ellipse $C(x_c, y_c)$ is placed along the pass trajectory at a specific ratio of the total distance:

$$C = S + \text{ratio} \cdot \vec{V} \tag{1}$$

In our implementation, a ratio of 0.85 was found to be optimal through iterative testing.

(2) **Semi-Major Axis (a):** Defined as the distance from the center to the sender, ensuring the sender always lies on the ellipse contour:

$$a^2 = \|C - S\|^2 \tag{2}$$

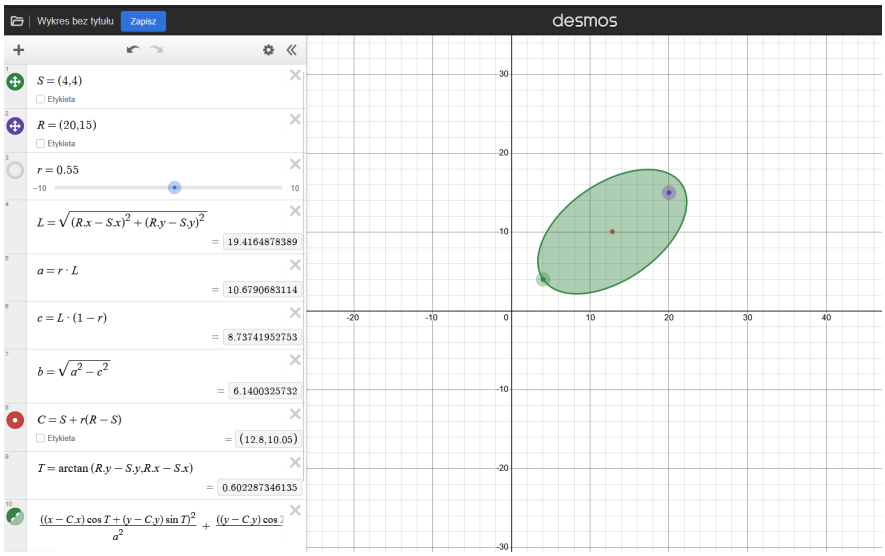

Fig. 1. elipse

(3) **Linear Eccentricity (c):** The distance from the center to the receiver is defined as c. In this configuration, the receiver acts as the focal point (focus) of the ellipse:

$$c^2 = \|R - C\|^2 \tag{3}$$

(4) **Semi-Minor Axis (b):** Using the standard elliptical identity, the semi-minor axis b (representing the lateral "width" of the interception zone) is:

$$b^2 = a^2 - c^2 \tag{4}$$

Coordinate and Rotation adjustments To determine if an opponent $P(x_p, y_p)$ is inside the ellipse, we transform the global coordinates into the local coordinate system of the ellipse. We align the major axis with the x-axis by rotating the points by $-\theta$, where $\theta = \text{atan2}(dy, dx)$. The relative coordinates (x', y') are calculated as:

$$\begin{aligned} x' &= (x_p - x_c)\cos(-\theta) - (y_p - y_c)\sin(-\theta) \\ y' &= (x_p - x_c)\sin(-\theta) + (y_p - y_c)\cos(-\theta) \end{aligned} \tag{5}$$

An opponent is counted as a "potential interceptor" if they satisfy the elliptical inequality:

$$\frac{(x')^2}{a^2} + \frac{(y')^2}{b^2} < 1 \tag{6}$$

The ellipse ratio was set by successively inserting values and selecting the one with the highest accuracy and is equal to 0.85

2.4.2 Tactical Rationale

. The decision to place the sender at the vertex and the receiver at the focus is tactically significant:

- **Cone of Uncertainty:** This configuration naturally creates a shape that is narrowest near the sender and wider near the receiver, accurately modeling how a defender's ability to intercept increases as the ball travels. However, the widest point is not where the receiver is, because we take into account the players ability to "defend" the ball.
- **Dynamic Scaling:** As the distance of the pass increases, the area of the ellipse scales quadratically, which reflects the increased risk associated with long-distance passes.
- **Lateral Coverage:** Unlike a simple radius-based check, the semi-minor axis b captures defenders who are not directly on the pass line but are close enough to react and move into the ball's path.

Feature Set D (Geometry): Added to set C.

- `opponents_in_ellipse`: The count of opponents inside this dynamic shape.

Result Analysis: This approach pushed accuracy to 39 – 40%. The feature importance analysis showed `opponents_in_ellipse` as the top predictor, rendering the manual "penalties" from Phase 2 obsolete.

2.5 Phase 5: XGBoost introduction

Model: LightGBM Ranker and XGBoost.
We decided to add another model (XGBoost) and obtain results as the average of the predictions of these two models. Combining models allows us to use two different tree-building architectures (leaf-wise and level-wise), which results in better capture of both deep interactions and general trends in the data. By averaging the results, we reduce variance and minimise the risk of overfitting, making the final predictions more stable and resistant to noise. Each model is based on a slightly different approach to feature engineering, allowing us to combine different analytical perspectives on pitch geometry and player behaviour.
Feature Set E (Zones):

- `sender_zone_x`: defense, midfield, attack
- `sender_zone_y`: left, middle, right
- `receiver_zone_x`: defense, midfield, attack
- `receiver_zone_y`: left, middle, right
- `zone_progression`: forward or backward pass

Result Analysis: This approach pushed accuracy to 42.15%. The brier score was lower due to the combination of different approaches.

3 Feature Selection and Pruning

As the feature set grew, we performed a rigorous selection process to prevent overfitting and reduce noise. We used the **Gain** metric from LightGBM and XGBoost to assess contribution.

Table 1. Feature Importance Analysis and Pruning Decisions for Most Important Decisions

Feature	Importance	Action Taken
`receiver_closest_opponent_dist`	High (Top 3)	**Kept.** Critical for measuring immediate marking.
`receiver_to_sender_dist`	High	**Kept.** Fundamental constraint (long passes are more risky).
`sender_zone, reciver_zone`	Moderate	**Kept.** Contextual feature (defensive vs attacking pass).
`opponents_in_ellipse`	Moderate	**Kept.** The predictive feature for interception risk.
`opponents_in_radius`	Moderate	**Dropped.** Replaced by an ellipse feature giving better results.
`closest_opponent_to_pass_line`	Moderate	**Dropped.** Replaced by an ellipse feature giving better results.
`manual_penalty` (+700)	Low	**Dropped.** The model learned these patterns automatically via tree splits.
`angle`	Low	**Dropped.** Found to be redundant with vector components.
`zone_progression`	Low	**Dropped.** Not needed while having features such as direction, dx.
`direction` (Hardcoded)	Moderate	**Modified.** Scaling values - arbitrary choice -to improve accuracy.

4 Model Architecture and Assessment

4.1 Validation Strategy

- **Chronological Split:**
  - *Training Set:* First 90% of snapshots.
  - *Validation Set:* Last 10% of snapshots.

4.2 Model A: LightGBM Ranker (Geometry Specialist)

- **Objective:** `lambdarank` (Optimizes NDCG).
- **Configuration:** Deeper trees (`num_leaves=64`) to capture complex geometric dependencies.
- **Top features:**
  - receiver_to_sender_dist
  - receiver_closest_opponent_dist
  - receiver_closest_3_teammates_dist
  - receiver_closest_3_opponents_dist
  - abs_y_diff
  - opponents_in_elipse

4.3 Model B: XGBoost Ranker

- **Objective:** `rank:NDCG (Normalized Discounted Cumulative Gain)`.
- **Configuration:** Histogram-based method (`tree_method='hist'`) with heavy regularization (`max_depth=7`).
- **Role:** Better at utilizing the precomputed pressure statistics matrices (Density, Min Distances).

- **Top features:**
  - receiver_to_sender_dist
  - receiver_closest_opponent_dist
  - receiver_closest_3_teammates_dist
  - average_three_receiver_opponents
  - receiver_zone_y
  - direction

4.4 The softmax function

Softmax with a temperature parameter was used to precisely scale the probability distribution of the recipient selection, which allowed for optimal control of the model's prediction confidence level. The best temperature coefficients and weights of individual models in the ensemble were selected using the *RandomSearch* procedure, which identified the most effective parameters for the test data.

4.5 Hyperparameters

All internal hyperparameters for both models were determined through a systematic Grid Search process to identify the optimal configuration for maximizing Top-1 accuracy.

5 Final Results and Analysis

Table 2 presents the performance on the held-out validation set.

Table 2. Performance Comparison of Investigated Approaches (results from validation test)

Approach	Key Innovation	Top-1 Accuracy	Brier Score
Baseline MLP	Raw Coordinates	30.20%	0.952
LGBM Classifier	+ Pressure features	36.66%	0.980
LGBM Classifier	+ Manual Penalties / Pressure	38.20%	0.970
LGBM Ranker	+ Basic Ellipse	37%	0.760
Hybrid (model A+B)	**Vertex-Focus Ellipse + Blended features**	**42.15%**	**0.712**

5.1 Analysis

The transition from Classification to Ranking provided the structural framework for success (decreasing brier-score). The mix of the models achieved a balance effect: LightGBM handled the "physics" of the pass, while XGBoost handled the tactical density, resulting in a robust 42.15% accuracy on our validation set.

6 Conclusion

This investigation demonstrates that in sports analytics, domain-specific feature engineering is crucial. We successfully moved from a naive distance-based model to a sophisticated geometric rankers. The final system effectively "sees" the pitch not just as points, but as zones of control and interception risks.

7 Sources

Ellipse - Wikipedia
About LTR models
XGBoost documentation
LGBM ranker documentation
LGBM ranker article
NDCG article