Counter-Pressing Effect Value (CPEV)
Technical Report

Liverpool FC – Data Science Department

Summer 2018

1. Introduction and Objective

The 2017–18 Liverpool side under Jürgen Klopp built its identity on high pressing and rapid attacking transitions. As part of our group's analytical study of the team, we conducted a descriptive analysis of match data to assess performance across positions. This analysis highlighted that, while Liverpool's front line was already among Europe's most productive, the midfield contributed relatively less to sustaining attacking pressure immediately after ball losses. Consequently, I decided to focus my technical investigation on identifying and evaluating midfielders capable of reinforcing this aspect of play.

To support this decision analytically, I developed the **Counter-Pressing Effect Value (CPEV)** metric, a machine-learning model estimating the probability that a quick regain leads to a shot **inside the penalty area within 10 seconds**. The model uses event-level Wyscout data and is trained using a logistic regression classifier. Formally, this can be expressed as:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta^T X)}},$$

where $y = 1$ indicates that a shot inside the penalty area occurred within 10 seconds of the regain, and X represents spatial and contextual features of the regain moment.

The goal is to identify players who contribute most to dangerous counter-pressing situations—those whose defensive recoveries translate efficiently into attacking threat.

2. Data Description and Preprocessing

Dataset Overview

We use the Wyscout event dataset for the 2017–18 Premier League season. Each record includes: `matchId`, `playerId`, `teamId`, match period, event type/sub-type, `x, y` coordinates (0–100 scale), event time in seconds, and a unique event `id`. The dataset captures all on-ball actions and their spatial context, forming the foundation for modeling counter-pressing behavior.

Event Filtering and Linking

To model counter-pressing chains, events were categorized as follows:

- **Turnovers:** failed passes, lost duels, mis-controls, or clearances conceded.
- **Regains:** recoveries, interceptions, or successful duels by the opposite team within **7 seconds** of a turnover.
- **Shots:** attempts inside the penalty area occurring within **10 seconds** of a regain.

All coordinates were normalized to a 105 × 68 m pitch. A linking algorithm associated each turnover with the nearest subsequent regain and possible shot event, constructing a single record per sequence:

$$(\text{Turnover}_t, \text{Regain}_{t+\Delta t_1}, \text{Shot}_{t+\Delta t_2}).$$

After cleaning and filtering, the final dataset contained **36,913** valid counter-pressing sequences, referred to as the *CPEV DataFrame*.

Choice of Temporal Windows

The temporal windows of **7 seconds for regains** and **10 seconds for shots** were selected based on both tactical reasoning and empirical validation.

The 7-second window for regains was chosen as a middle ground between the well-known "five-second rule" popularized by Pep Guardiola and the broader 10-second interval sometimes used in pressing studies. While a 5-second limit captures the purest form of counter-pressing intensity, it yielded too few samples for robust model training. Conversely, a 10-second cutoff risked including slower defensive recoveries that no longer reflected immediate pressing intent. Thus, a 7-second interval was deemed an effective compromise—tight enough to reflect authentic counter-pressing behavior while maintaining sufficient sample size for statistical power.

For the subsequent attacking phase, a 10-second window after the regain was used to identify shots. This duration was empirically supported: when testing alternative windows (10s vs. 15s), the 10-second specification produced more *stable results* in model performance, meaning lower variance and more consistent AUC scores across cross-validation folds. Conceptually, it also aligns with football intuition—beyond 10 seconds, attacking sequences are often influenced by additional buildup actions rather than the initial counter-pressing success.

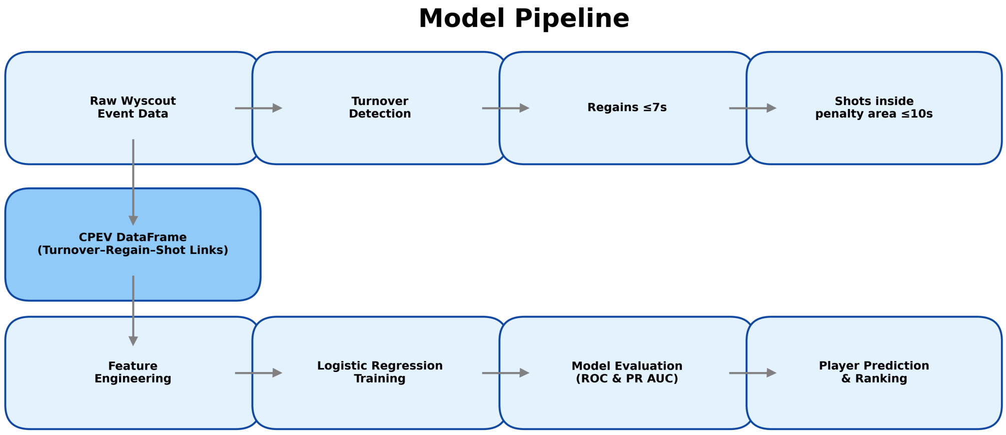
Model Pipeline
Raw Wyscout Event Data
Turnover Detection
Regains ≤7s
Shots inside penalty area ≤10s
CPEV DataFrame (Turnover-Regain-Shot Links)
Feature Engineering
Logistic Regression Training
Model Evaluation (ROC & PR AUC)
Player Prediction & Ranking

Figure 1: Data pipeline: from turnover detection to regain–shot linkage.

3. Feature Engineering

Each regain event was enriched with features capturing its tactical context:

- **Regain X (r_x):** normalized x-coordinate of the regain. Higher values indicate regains closer to the opponent's goal.
- **Centrality (c_y):** $c_y = 1 - \frac{|y-34|}{34}$, representing how central the regain occurs.
- **Out-of-Shape Index (OSI):** a custom indicator quantifying opponent disorganization prior to the ball regain. OSI combines two components: (1) a *temporal decay factor*, modeled as e^{-kt} with $k = 0.2$, which penalizes late recoveries since faster regains imply stronger counterpressure; and (2) a *spatial sparsity measure*, computed from the density of opponent actions (both defensive and offensive) within a 25-m radius and 10-s window before the regain. For each regain, sparsity was defined as:

$$\text{sparsity} = \frac{1}{1 + \sum e^{-d/25}},$$

where d represents the distance of opponent events from the regain location, yielding higher values when few opponent actions occur nearby. The final OSI is given by the product of the two components:

$$\text{OSI} = e^{-0.2t} \times \text{sparsity},$$

and subsequently normalized to range between 0 and 1, so that higher values indicate quicker regains achieved while the opponent was spatially stretched and disorganized.

- **Progressive Next Action:** binary variable indicating whether the next action (pass or carry) of the team advanced play. Progressive passes were identified following Wyscout's definition—requiring a minimum forward movement threshold depending on field zone (30 m in own half, 15 m when crossing halfway, and 10 m in the opponent half). Progressive carries combined two sources: (1) explicit *Acceleration* events provided by Wyscout, where a player advanced the ball at least 10 m, and (2) reconstructed carries inferred from consecutive same-player events within 5 seconds, indicating continuous ball retention and forward movement of at least 10 m.
- **Interaction Term:** $r_x \times c_y$ to capture the non-linear value of central high regains.

All numeric features were normalized to the $[0, 1]$ range to ensure comparable feature scales and prevent variables with larger numeric ranges (e.g., pitch coordinates) from dominating the model's optimization process. This scaling improves numerical stability and speeds up convergence during logistic regression training. Binary variables (e.g., *progressive next action*, *shot in box within 10s*) were converted from boolean to integer format (0/1) to ensure compatibility with the regression model.

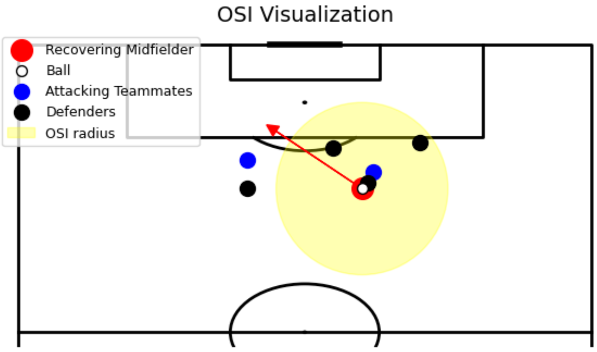

Figure 2: Illustration of the Out-of-Shape Index (OSI). The metric combines time since turnover with local opponent density to quantify structural disorganization.

4. Model Specification and Training

A logistic regression model was trained using `statsmodels` to estimate the probability of producing a shot inside the box within 10 seconds after a ball regain. The input feature vector was defined as:

$$X = [r_x, c_y, \text{OSI}, \text{progressive}, r_x \times c_y],$$

where r_x is the normalized regain location (x-coordinate), c_y is the centrality of the regain, OSI captures opponent disorganization, and progressive is a binary indicator of whether the next team action advanced play. An interaction term $r_x \times c_y$ was included to capture the non-linear effect of regaining the ball in central and advanced areas.

To assess model generalization and robustness, the dataset was evaluated using **five-fold stratified cross-validation**, ensuring class balance across folds. In each iteration, `regain_x` was rescaled via Min–Max normalization, interaction features were re-generated, and a logistic regression was fitted via maximum likelihood estimation. Model performance was assessed using two complementary metrics: the **ROC-AUC** (Receiver Operating Characteristic Area Under Curve) and the **PR-AUC** (Precision–Recall Area Under Curve).

While ROC-AUC measures overall discriminative ability, PR-AUC is particularly informative under strong class imbalance—since "shot within 10 seconds" is a rare outcome. PR-AUC emphasizes the model's precision when identifying these scarce positive cases. A *shuffled-label baseline* was also trained to confirm that the observed predictive signal was not due to random correlations.

Table 1: Fold-wise ROC-AUC and PR-AUC performance.

Fold	**True ROC-AUC**	**True PR-AUC**	**Shuffled ROC-AUC**	**Shuffled PR-AUC**
1	0.886	0.124	0.495	0.015
2	0.874	0.106	0.474	0.015
3	0.874	0.114	0.473	0.017
4	0.883	0.115	0.523	0.017
5	0.873	0.101	0.497	0.015
Mean (±SD)	**0.878 ± 0.005**	**0.112 ± 0.008**	**0.492 ± 0.019**	**0.016 ± 0.001**

The model achieved an average ROC-AUC of **0.878**, indicating excellent discriminative power—nearly a 90% probability of ranking a true shot-inducing regain higher than a random one. The corresponding mean PR-AUC of **0.112** may appear modest in absolute terms, but represents a roughly **sevenfold improvement** over the shuffled baseline (0.016), confirming that the model captures meaningful structure rather than noise. Moreover, the low standard deviation across folds (±0.005 in ROC-AUC) indicates strong **stability and generalization**, with consistent performance across match segments.

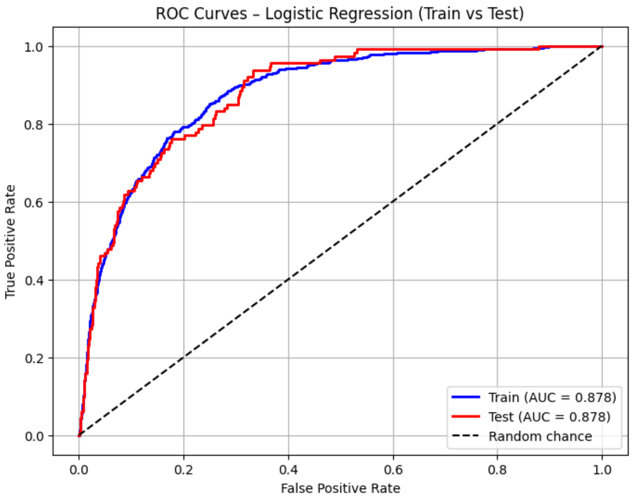

Figure 3: ROC curves for train and test folds. The true model (red) consistently separates positive and negative outcomes, confirming strong generalization and predictive validity.

Overall, the model demonstrates high discriminative performance, resilience to overfitting, and interpretable feature effects linking the spatial and contextual dimensions of ball regains to subsequent shot creation.

5. Results and Model Interpretation

Final Model Summary

The final model achieved a pseudo-R^2 of **0.21**, indicating that the features explain a meaningful share of the variation in shot creation probability after ball regains.

Coefficient Interpretation and Explainability

- **Regain X (r_x) — significant and positive:** The coefficient of +2.54 shows that ball recoveries closer to the opponent's goal dramatically increase the likelihood of producing a shot. Each

Table 2: Final logistic regression results (trained on full dataset).

Feature	Coef.	Std. Err.	z-score	p-value
Intercept	–7.15	0.33	–21.95	0.000
Regain X	2.54	0.43	5.98	0.000
Centrality	–0.61	0.52	–1.18	0.238
OSI	0.60	0.24	2.50	0.012
Progressive Next Action	1.06	0.11	9.68	0.000
Regain X × Centrality	4.37	0.69	6.33	0.000

0.1 increase in normalized regain position raises the log-odds of creating a shot by approximately 25%.

- **Progressive Next Action — highly significant:** The coefficient of +1.06 indicates that immediately playing forward after regaining possession substantially boosts shot probability. This validates the tactical value of vertical, high-tempo transitions.
- **OSI (Opponent Structural Instability) — positive and significant:** A +0.60 effect suggests that pressing against disorganized opponents leads to more successful attacking outcomes. The result empirically supports the intuitive principle that regains during opponent disorganization are more valuable.
- **Interaction ($r_x \times c_y$) — strong positive effect:** The +4.37 coefficient confirms that regaining the ball in **central, advanced zones** creates disproportionately higher threat levels, reflecting high counterpressing leverage.
- **Centrality (c_y) — not significant in isolation:** Although centrality alone is not statistically significant (p = 0.238), its contribution becomes meaningful when combined with r_x, emphasizing that centrality's value is conditional on regain height.

These effects align closely with modern counterpressing theory: spatial control and immediate forward intent after regaining possession are key determinants of transition threat.

6. Application to Player Evaluation

Each regain event's predicted probability was aggregated by player to compute their **CPEV per 90 minutes:**

$$\text{CPEV/90} = \frac{\sum_i \hat{P}_i}{\text{minutes}/90}.$$

Only midfielders with at least 900 minutes and 50 regains were considered.

Table 3: Top 10 midfielders across Europe's top five leagues by Counter-Pressing Expected Value (CPEV) per 90 minutes.

Player	League	Minutes	CPEV	CPEV/90
Suat Serdar	Bundesliga	1668	3.10	0.167
Will Hughes	Premier League	974	1.46	0.135
Gonzalo Escalante	La Liga	1531	2.16	0.127
Lewis Holtby	Bundesliga	1176	1.49	0.114
Felipe Anderson	Serie A	1171	1.42	0.109
Robin Quaison	Bundesliga	1370	1.66	0.109
Morgan Sanson	Ligue 1	2120	2.56	0.109
Gastón Ramírez	Serie A	2183	2.61	0.108
Julian Baumgartlinger	Bundesliga	1376	1.59	0.104
Allan Marques Loureiro	Serie A	2845	3.25	0.103

The model highlights players excelling at turning regains into dangerous attacks. Interestingly, several of these (e.g., Will Hughes, Dele Alli) display aggressive pressing tendencies combined with proactive ball use.

When applied to other European leagues, the model identified **Suat Serdar (Schalke 04)** as a standout candidate—ranking highest in CPEV/90 across the top five leagues.

7. Discussion and Limitations

The CPEV model quantifies how efficiently players convert pressing moments into attacking threat. Its strong AUC and consistent coefficients show that event data can meaningfully represent counter-pressing effectiveness.

Strengths

- Fully interpretable model with transparent, tactically grounded features.
- Consistent with gegenpressing logic—aligned with Liverpool's playing philosophy.
- Robust validation and baseline checks ensure stability and reliability.

Limitations and Future Work

Despite promising performance, several methodological caveats remain:

- **High accuracy and possible leakage:** The model's ROC-AUC of 0.88 is unusually high for a linear model, raising concerns about hidden correlations. Precision–Recall AUC and a shuffled-label test ruled out random leakage. Excluding the `progressive_next_action` feature reduced AUC by only 0.02, indicating that the model's predictive power stems primarily from structural context rather than feature leakage.
- **Loss of temporal and sequential context:** The binary target (*shot within 10 seconds*) collapses complex attacking sequences into a single outcome, ignoring when and how the shot develops. By not modeling the sequence of passes or actions after the regain, the metric captures only the *immediate potential* of a recovery rather than the player's full contribution to the ensuing attack. Nevertheless, it remains a useful approximation for evaluating how specific regains translate into offensive threat.
- **Event-data limitations:** Without off-ball tracking, pressing triggers, runs, and defensive compactness remain unobserved—restricting tactical depth.

Future extensions could involve integrating tracking-based metrics such as **pitch control** or **Expected Threat (xT)** models to capture spatial and temporal interactions more comprehensively. Additionally, exploring non-linear learning methods (e.g., tree-based ensembles or neural networks) could better model the higher-order dependencies between space, opponent structure, and transition success.

Conclusion

This study developed the **Counter-Pressing Expected Value** (CPEV) model to quantify how effectively pressing recoveries translate into attacking threat. Using interpretable logistic regression and robust validation, the framework achieved strong predictive accuracy (AUC ≈ 0.88) while maintaining clear tactical relevance. CPEV offers a practical baseline for identifying players suited to high-intensity pressing systems. However, as discussed, future improvements should incorporate temporal dynamics, off-ball actions, and richer contextual features to enhance realism and tactical depth. Overall, CPEV bridges the gap between footballing intuition and quantitative analysis, offering a meaningful step toward data-driven pressing evaluation.