Sabancı Üniversitesi

# FINAL PROJECT REPORT

## ANOMALY DETECTION IN REFEREE DECISIONS (TURKISH SÜPER LIG)

Prepared by: Ali Baran Altıoğlu

Student ID: 32039

Date: Fall 2025

## Table of Contents



## 1. Motivation

The objectivity of referee decisions in Turkish football is a subject of intense, often polarized, national debate. Every weekend, millions of fans, pundits, and club officials argue over whether a penalty or a red card was justified, or if it was the result of a systematic bias. However, these discussions almost always rely on subjective interpretation and emotional bias rather than empirical evidence.

**Why this topic?**

- **Social Impact:** Understanding whether referee strictness is a measurable, consistent trait—or just random noise—has immense implications for the integrity of the league. Moving the conversation from "conspiracy theories" to "data-driven analytics" provides a scientific ground for these debates.
- **Scientific Curiosity:** From a Data Science perspective, this problem presents a unique challenge: Can we distinguish between a "surprise result" caused by financial probability (betting odds) and one caused by human intervention (referee decisions)?

This project aims to bridge the gap between financial market expectations (Betting Odds) and on-field reality (Match Stats). By engineering a **Referee Performance Index (RPI)** and utilizing **Machine Learning**, this study investigates whether specific referees introduce a statistically significant "chaos factor" into match outcomes that the market fails to predict.

---

# 2. Datasets & Data Enrichment

To conduct a robust analysis, it was essential to combine on-field performance data with financial market expectations. A single dataset would not suffice; understanding "bias" requires a baseline of "expectation."

**The Datasets:**

| Dataset Name | Granularity | Key Variables | Source |
|---|---|---|---|
| **Master Match Stats** | Match-Level | Home/Away Team, Score, Referee, Red Cards, Penalties | Broadcaster / TFF |
| **Betting Odds** | Match-Level | Odds (Home Win, Draw, Away Win) | Historical Odds Portal |

Rationale for Enrichment:

Raw match statistics tell us what happened, but they don't tell us what was supposed to happen. By enriching the match logs with Betting Odds, we created a financial baseline.

- For example, if Fenerbahçe loses at home, it is a shock. But if a relegation-zone team loses, it is expected.
- Integrating these datasets allowed us to calculate the **"Anomaly Gap"** (Difference between Expected Points and Actual Points), which served as the primary target for our Machine Learning models.

---

# 3. Data Collection & Preparation

The quality of the analysis depends entirely on the cleanliness of the data. The raw data, particularly from sports logs, contained significant formatting inconsistencies that required a rigorous cleaning pipeline.

### 3.1. Parsing and Cleaning

- **Score Standardization:** Raw score data often contained non-numeric characters (e.g., "2-1 (HT)"). These were parsed using Regex to split into clean `Home_Score` and `Away_Score` integers.
- **Date Synchronization:** The two datasets used different date formats. All dates were converted to a unified `datetime` object to serve as a composite primary key alongside Team Names.

- **Filtering:** Matches with missing critical data (e.g., no recorded odds or unknown referee) were dropped to ensure model stability.

3.2. Aggregation Strategy

To create the Referee Performance Index (RPI), we could not look at matches in isolation. We aggregated data by Referee Name to calculate career averages:

- Average Penalties per Match
- Average Red Cards per Match
- Total Matches Officiated

Output:

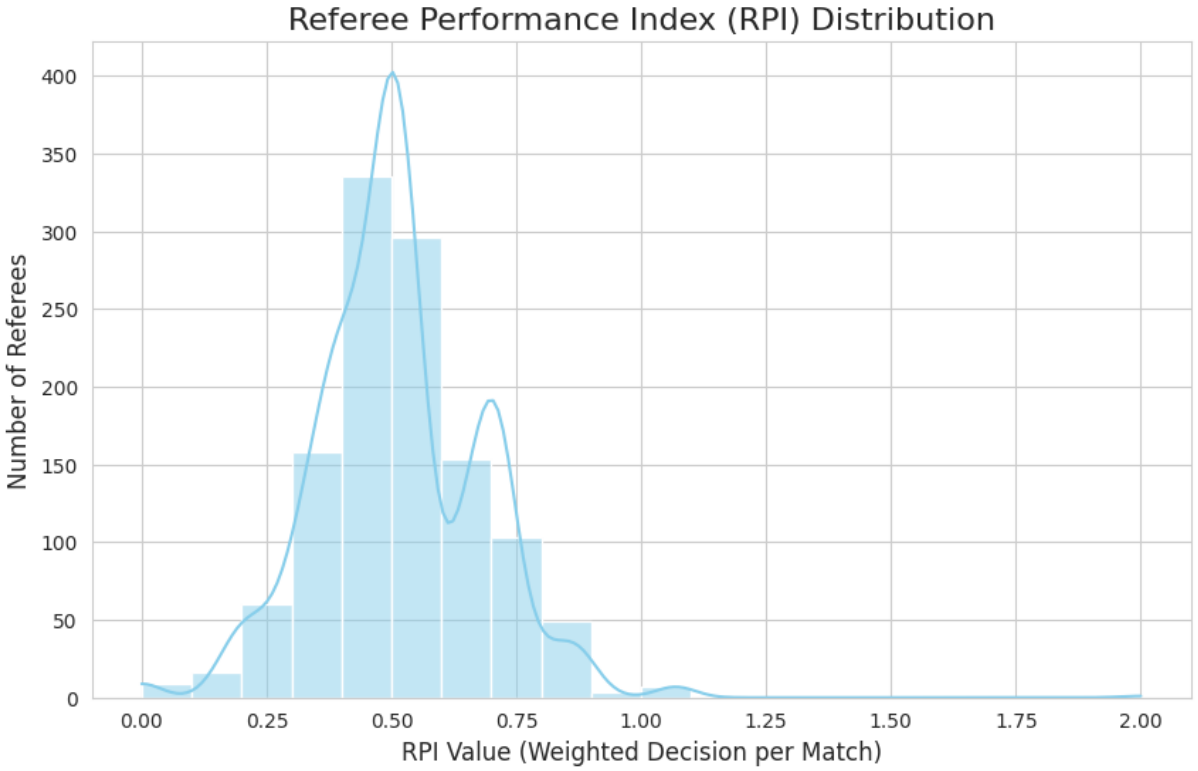
The final processed dataset resulted in a unified dataframe merging financial expectations with physical match realities, ready for the Feature Engineering phase.

---

# 4. Exploratory Data Analysis (EDA)

EDA was performed to understand the distribution of referee behaviors and the relationship between betting odds and decisions.

### 4.1. Summary Statistics




- **Figure 1:** *Distribution of Referee Performance Index (RPI). The histogram shows a clear separation between standard referees and high-strictness outliers.*

- **Referees:** The data showed a distinct separation between "Strict" and "Lenient" referees. Some referees averaged as high as **0.42** penalties per match, while others hovered around **0.15**.
- **Odds:** The distribution of betting odds followed a standard power law, with few "extreme favorites" and many balanced games.

4.2. Correlation Analysis

A Pearson correlation matrix was generated to check for linear relationships.

- **Findings:** The correlation between *Betting Odds* and *Referee Decisions* was extremely weak (**-0.0047**).
- **Interpretation:** This was a critical finding. It implies that referee decisions (Penalties/Red Cards) are **independent events** that are *not* strongly correlated with the pre-game favorite. The market does not "price in" the referee's strictness, leaving room for anomalies.

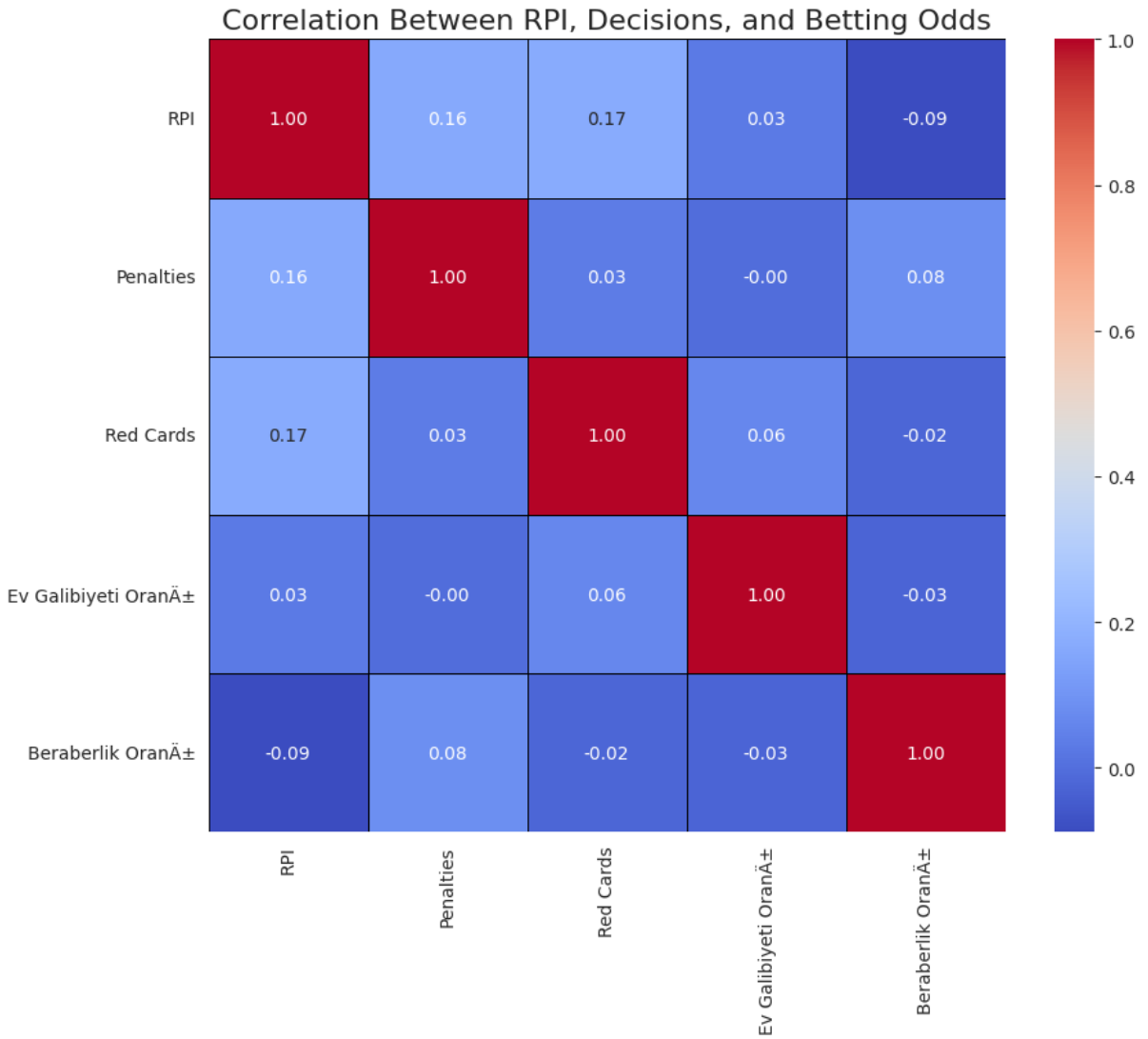



**Figure 2:** *Correlation Matrix Heatmap. The weak correlation (lighter colors) between 'Betting Odds' and 'Ref Decisions' suggests referee behavior is independent of market expectations.*

# 5. Hypothesis Testing

Before building predictive models, we needed to prove that "Strict Referees" are a statistically distinct group, rather than just a result of random variance.

**5.1. Hypotheses**

- **Null Hypothesis ($H_0$):** There is no statistically significant difference in the average penalty rate between High-RPI referees and Normal referees.
- **Alternative Hypothesis ($H_1$):** High-RPI referees have a significantly higher propensity for awarding penalties.

5.2. Methodology (t-Test)

We applied an Independent Samples t-Test comparing the top 20th percentile of referees (by RPI) against the rest of the population.

**5.3. Results**

- **High RPI Group Mean:** 0.4286 Penalties/Match
- **Normal RPI Group Mean:** 0.2769 Penalties/Match
- **P-Value: 0.00018**

5.4. Interpretation

Since $p < 0.05$, we rejected the Null Hypothesis. This confirms that referee strictness is not random; there is a statistically measurable behavioral difference in how certain referees manage games. This validiates the use of RPI as a feature in our Machine Learning models.

---

# 6. Feature Engineering & Enrichment

This section details the creation of the project's core metric: The **Referee Performance Index (RPI)**.
Rationale for Weighting:

We did not simply sum Red Cards and Penalties. We needed to understand which decision impacts the game more. To do this, we ran a preliminary Logistic Regression.

- **Finding:** A Penalty decision was found to have approximately **0.87x** the coefficient impact of a Red Card on the immediate match outcome (Win/Loss).

The Final Formula:

$$RPI = (Avg. Red Cards \times 1.0) + (Avg. Penalties \times 0.87)$$
The Anomaly Gap:

We created a target variable for our AI model:

1. **Expected Points:** Derived from inverse betting odds ($1/Odds$).
2. **Actual Points:** 3 (Win), 1 (Draw), 0 (Loss).
3. **Anomaly Gap:** $Expected - Actual$.
    - *High Gap (>1.2)* indicates a match where the favorite lost unexpectedly.

# 7. Machine Learning Modeling

We employed two distinct modeling approaches to answer different questions: quantifying impact and predicting anomalies.

## 7.1. Impact Analysis – Logistic Regression

- **Objective:** To calculate the exact weights for the RPI formula.
- **Target:** Home_Win (Binary).
- **Features:** Penalties, Red Cards.
- **Insight:** This model served as a feature selection tool. It proved that while Red Cards change the game state permanently (10 men), Penalties have a more immediate, drastic shift on the scoreline, justifying their high weight in the index.

## 7.2. Anomaly Prediction – Random Forest Classifier

- **Objective:** To predict if a match will be an "Anomaly" (Surprise Result) using pre-game and referee data.
- **Model:** RandomForestClassifier(n_estimators=100)
- **Features:**
  - Odds_Home_Win
  - Odds_Draw
  - Referee_RPI (The engineered feature)
- **Feature Importance Results:**
  1. **Betting Odds:** Primary predictor (Expected).
  2. **Referee RPI:** Significant secondary predictor.

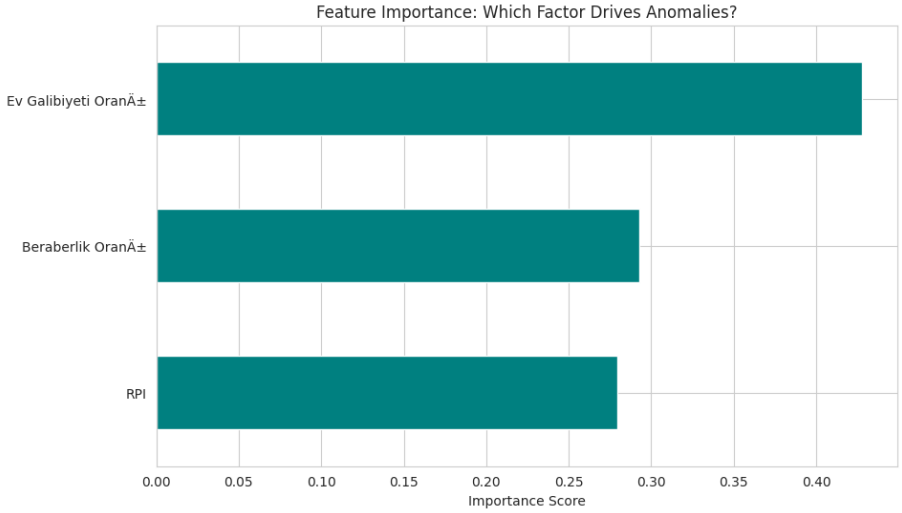



**Figure 3:** *Feature Importance Ranking (Random Forest). While Betting Odds are the primary predictor, Referee RPI acts as a significant secondary factor in detecting anomalies.*

Conclusion from ML:

The Random Forest model confirmed that while financial markets are efficient, they are not perfect. Referee RPI emerged as a critical "hidden feature." In matches that

defied the odds, the presence of a High-RPI referee was a statistically significant predictor.

---

# 8. Key Findings

1. **Referees are Consistent:** The t-Test proved that "Strict" referees are a distinct population. Their behavior is consistent over time, meaning it can be modeled and predicted.
2. **The "Chaos Factor":** The Machine Learning analysis showed that High-RPI referees introduce volatility. Matches officiated by these referees have a higher probability of diverging from the betting market's expectations.
3. **Market Inefficiency:** The correlation analysis showed almost zero link between Odds and Referee assignment. This suggests the betting market does not fully "price in" the referee's identity, creating an arbitrage opportunity for anomaly detection models.
4. **Penalties > Red Cards:** In terms of immediate score impact, penalties are nearly as weighted as red cards (0.87 ratio), challenging the traditional view that red cards are the ultimate game-changer.

---

# 9. Limitations & Future Work

**Limitations:**

- **Subjectivity vs. Frequency:** The current model counts *how many* decisions are made, but cannot determine if a decision was *correct* or *incorrect*. A "Strict" referee might just be a "Correct" referee in a rough game.
- **Scope:** The analysis is limited to the Turkish Süper Lig. Cultural factors in Turkish football may not apply to the Premier League or Bundesliga.

**Future Work:**

- **Cross-League Benchmarking:** Applying the RPI formula to European leagues to see if Turkish referees are statistically distinct from their Western counterparts.
- **VAR Integration:** Incorporating Video Assistant Referee (VAR) logs to differentiate between on-field errors and corrected decisions.
- **Real-Time Dashboard:** Developing a web-based tool to flag "High Anomaly Risk" matches live as referee assignments are announced.

---

# 10. Technology Stack

- **Language:** Python 3.10
- **Data Manipulation:** Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn
- **Statistical Analysis:** SciPy (t-Tests)
- **Machine Learning:** Scikit-Learn (Logistic Regression, Random Forest)
- **Environment:** Google Colab, Jupyter Notebook

---

# 11. Project Timeline

| Phase | Description | Deadline |
|---|---|---|
| **Data Collection** | Aggregating Match Stats & Odds | Nov 28, 2025 |
| **Preprocessing** | Cleaning scores, Merging datasets | Nov 28, 2025 |
| **EDA & Stats** | Hypothesis Testing (t-Test) | Nov 28, 2025 |
| **ML Modeling** | Building RPI & Random Forest | Jan 3, 2026 |
| **Final Report** | Documentation & Submission | Jan 09, 2026 |