# ML2 - Semestral Project Assignment

Pavel Zimmermann
zimmerp@vse.com

Karel Šafr
karel.safr@vse.com

Martin Bendík
benm35@vse.cz

October 8, 2025

## 1 Introduction

In the realm of sports, football is the most popular sport in the world. The game is played at a professional level all over the world, and millions of people regularly go to a football stadium to follow their favorite team, while billions more watch the game online. Many people bet on football matches, hoping to win some money. Their bets are often backed by data. **In this project**, we will simulate business decision making in the context of sports data analytics.

You are placed in the role of a data scientist in an imaginary betting company. Your role is to create predictions for football matches based on the data. Bookmakers use your predictive models for match outcomes to set betting odds.

The accuracy of these predictive models is the cornerstone of your company's revenue model. Higher model accuracy directly translates into reduced risk for bookmakers, enabling them to optimize their strategies. Therefore, the primary strategic objective is to continuously improve the accuracy of the model to maximize the value proposition of your business.

### 1.1 Business Model and Revenue Story

Bookmakers earn money by setting odds with a margin that covers their required profit, operating costs, and a risk premium. The risk premium compensates for uncertainty in predictions: the less accurate the model, the higher the risk of mispriced odds, which sharp bettors can exploit. When bookmakers have more accurate predictions, they can either:

- Lower the risk premium component, and hence the margin (making odds more attractive and increasing betting volume), or
- Maintain the margin and increase their required profit component as the risk premium component decreases, so they can enjoy higher profit per bet.

The increased model accuracy allows for optimization depending on the company strategy and corresponding target function (e.g. the total profit, or market share maximization).

The company overall profit is the volume of bets sold multiplied by the profit margin per bet. Better accuracy translates into higher overall profit, either by increasing the volume of bets or increasing the profit margin per bet, which creates a clear monetary value for accuracy improvement. This establishes a fundamental economic chain reaction:

1. **Higher Accuracy** → Better predictions reduce uncertainty in odds setting.
2. **Reduced Risk** → Lower risk of mispriced odds being exploited by sharp bettors.
3. **Lower Risk Margin** → Bookmakers can reduce the risk premium component $m_{\text{risk}}(A)$.
4. **Higher Profit** → Bookmakers can increase the overall profit either through increased betting volume or improved profit margins.

This chain creates a quantifiable relationship between prediction accuracy improvements and profit increases, allowing us to determine the maximum price justified for model creation expenses.

## 1.2 Business decision problem

Your company operates with a very limited set of historical data to make predictions, but was offered a richer set of data from a data provider recently. It contains more features than the current data set.

**The goal of this project** is to determine the maximum value of the improved data for the betting company, which is the maximum increase in profit caused by improved model accuracy. In turn, this translates into the maximum price that the betting company would pay their provider for the improved data.

# 2 Data and Resources

You are provided with a dataset containing 6 seasons of data from 21 top European football leagues from 11 countries. The data contains match statistics, results, betting odds, and other information. The data is provided in the form of `csv` files, one for each season and league. Each file contains information about all matches played in the given season in the given league.

```
england
├── 0
│   ├── 1920.csv
│   ├── ...
│   └── 2425.csv
├── 1
│   ├── 1920.csv
│   ├── ...
│   └── 2425.csv
└── ...
```

Data files are stored in the path `{country}/{league}/{season}.csv`, where `league` is a number (the lower the number, the higher the league) and `season` is a string representing the season in the format `{start_year}{end_year}`. For example, the file `england/0/1920.csv` contains data from the season 2019/20 from the highest English league - Premier League.

Together with the data, you can find the file `notes.txt`, which contains a description of the data including the meaning of each column.

The data is available at Shared Google Drive.

Note that the data available through the link is the *new* extended richer set the provider offers now. The current data your company works with contain only these columns: Div, Date, Time, HomeTeam, AwayTeam, FTHG, FTAG and FTR. This will serve as the foundation for your baseline model.

# 3 Mathematical Model

## 3.1 Margin Decomposition

As stated above, the betting company must cover its operations expenses, required profit and risk premium that compensates for uncertainty in predictions. So every bet is loaded with a margin $m$. We will work with relative margins that represent a margin per 1 USD of one bet. The margin $m$ is decomposed into three components:

$$m = m_{\text{profit}} + m_{\text{operations}} + m_{\text{risk}}(A)$$

where:

- $m_{\text{profit}}$: adjustable profit component
- $m_{\text{operations}}$: fixed operational cost margin.

| Description | Parameter | Value |
|---|---|---|
| Operational cost margin | $m_{\text{operations}}$ | 0.03 |
| Risk margin parameter | $k$ | 0.3 |
| Demand function parameter | $\alpha$ | 1000 |
| Demand function parameter | $\varepsilon$ | 3 |
| Average bet | $b$ | 12 USD |

Table 1: Values of the parameters assumed

- $m_{\text{risk}}(A) = k \cdot (1 - A)$: risk margin that decreases as accuracy $A$ increases

### 3.2 Demand function

Assume count of bets sold in one year, denoted as $V(m)$ and referred to as the volume. The competition and utility functions of the bettors cause a decrease in yearly volume $V(m)$ of bets sold for increasing margin $m$. Assume the following demand function

$$V(m) = \alpha \cdot m^{-\varepsilon},$$

where $\alpha$ and $-\varepsilon$ are parameters. This is a simple demand model common in theoretical models and dynamic pricing. It is based on the assumption of constant elasticity (power law).

### 3.3 Profit Function

The profit expected from one bet of 1 USD is represented by $m_{\text{profit}}$. ($m_{\text{operation}}$ are used to cover operating expenses, and $m_{\text{risk}}(A)$ is held to cover unexpected events connected with the uncertainty of the model.) The overall profit is

$$\Pi(m_{\text{profit}}) = V(m) m_{\text{profit}} b,$$

where $b$ is a parameter that represents the average bet amount in USD.

### 3.4 Parameter values

Use the following values of the parameters for the models stated above.

For simplicity, we provide fixed values of the demand function parameters. These values would practically be estimated on the basis of some historical data. Estimate of these parameters would typically be a result of 2-dimensional parametric optimization using some loss function.

The average bet and operational cost margin values can be thought of as the output of controlling department analysis.

# 4 Assignment Task

## 4.1 Problem Setup

Your task is to determine the maximum price your company can pay for additional data that improves model accuracy. The analysis involves:

- Assume that the bookmakers are currently operating with a margin $m_0$.
- Assume that the profit margin component of $m_0$ is optimal for given $m_{\text{operational}}$ and $A_0$, where $A_0$ is the precision of the current model (=model that does not use the improved data).
- If we can improve $A_0$ to some $A_1 > A_0$, the bookmakers' revenue may improve.
- The maximum price they would pay for the improvement equals their maximum profit improvement.
- Determine the maximum price your betting company can pay for the additional data that allows improving $A_0$ to $A_1$.

## 4.2 Model Requirements

For this analysis, focus on the popular bet type **Over/under 2.5 goals** scored in a match (binary classification). This is one of the most popular football betting type where bettors predict whether the total number of goals scored by both teams combined will be:

- **Over 2.5**: Three or more goals in the match (3, 4, 5, etc.)
- **Under 2.5**: Two or fewer goals in the match (0, 1, or 2)

The ".5" ensures there are no ties - a match cannot end with exactly 2.5 goals.

## 4.3 Tasks

1. **Machine Learning Model Optimization**
   - Train a reasonable classification or regression model. Use models taught in the course.
   - Use a reasonable loss function.
   - Train models on both baseline data (current features) and extended data (current + new features)
   - Evaluate and compare model accuracies $A_0$ vs $A_1$.

2. **Optimal Profit Increase**

   - As stated above, the profit margin corresponding to the current model is assumed to be optimized and the operation costs are fixed (=do not change if the accuracy changes). Assume that the accuracy of the model has increased thanks to the new features.
   - **Find new optimal profit margin** $m^*_{\text{profit}}$ that maximizes profit $\Pi(m_{\text{profit}}, A_1)$ for the new accuracy $A_1$.
   - Data is sold separately for each market (country) for a yearly subscription. You can choose to buy data for just selected markets. Determine the maximum subscription price for additional data based on profit improvement.
   - Provide both analytical solution (if possible) and numerical verification.

# 5 Submission Guidelines

You have to submit your well–documented code with your solution, preferably Jupyter Notebook(s) or Python scripts. Send your solutions to the following e–mail: `benm35@vse.cz`. The deadline for online submissions is November 16, 2025.

# 6 Evaluation and Grading

You can earn up to 30 points. The solution will be assessed based on the quality of the performed steps, selected methods, techniques, and reasoning applied throughout the project. The maximum rating will be awarded for demonstrating a thorough understanding of the problem, correct application of machine learning algorithms, appropriate data preprocessing, model selection, hyperparameter tuning, and evaluation methods. Clear and logical reasoning behind each decision, along with proper use of validation techniques and error analysis, will be key in earning full points in this section. In particular, you should focus on detailed error analysis because it is a cornerstone for answering the main question of the assignment.

Issues such as poorly justified decisions, incorrect use of methods or techniques, lack of clarity in reasoning, incomplete steps in the analysis, failure to properly validate models, neglecting essential steps in preprocessing or feature selection, or insufficient explanation of choices made during hyperparameter tuning and model evaluation could lead to a reduction in points.

Note: the semestral project contributes 30% to the final grade.