Chess Match Winning Chances Predictor and Skill Assessment

-Dhawal Arora (da812)
GITHUB: https://github.com/dhawal-arora/210final

Introduction

Chess is a universal game that has seen an increasingly significant online presence, especially post-COVID-19. Despite widespread interest, players often struggle to understand the factors influencing their chances of winning. This project aims to address that issue.

Project Definition

The goal of this project is to analyze my chess games on Chess.com and Lichess to derive meaningful insights about my performance and utilize machine learning techniques to predict the outcomes of future games. This involves:

- Evaluating win/loss/draw.
- Identifying patterns based on the time of play, openings used (via ECO codes), platforms, and opponent ratings.
- Developing a predictive model to estimate the likelihood of a win based on features such as time of day and opening (ECO code).

Strategically, this project aims to uncover how external factors, such as time of day and platform, influence performance. The data-driven insights gained can be used to refine strategies.

Novelty and Importance

This project is significant because it bridges a gap in personalized performance analysis for online chess games. While platforms like Chess.com and Lichess provide basic statistics, they lack:

- In-depth comparative analyses across platforms.
- Predictive modeling tailored to individual users.

I am particularly excited about this project because chess has been a long-standing passion of mine. It allows me to combine personal interest with technical expertise. The project's novelty lies in its integration of user-specific data with machine learning, enabling tailored insights and predictions that go beyond conventional analytics.

Challenges:

- Lack of Advanced Interpretable Analytics
- Limited Cross-Platform Comparisons

Existing Research Only Has:

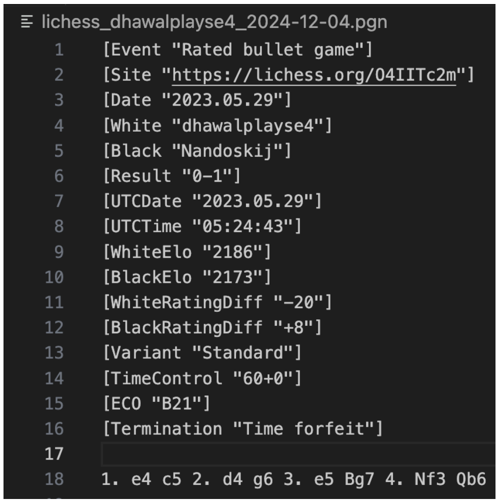
- Static data analysis.
- Win-prediction models specific to individual platforms.

Data Collection

The primary sources for online chess game data are **Chess.com** and **Lichess.org**, which both provide public REST APIs for retrieving game data.

1. **Lichess.org**:
    - API Endpoint: https://lichess.org/api/games/user/{username}
    - Purpose: Retrieves all games played by a specific user throughout their lifetime.
    - Format: The data is provided in **PGN (Portable Game Notation)** format, which is a standard for recording chess games.
2. **Chess.com**:
    - API Endpoint: https://api.chess.com/pub/player/{username}/games/{year}/{monthnumber}/pgn
    - Purpose: Retrieves games from a specific month for a given user.
    - Format: The data is also provided in **PGN format**.

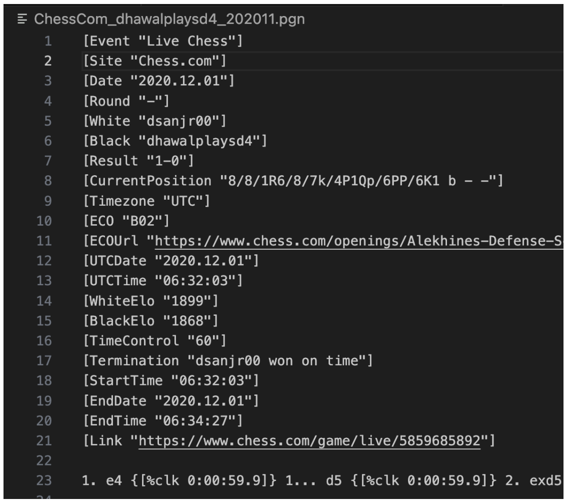
The use of PGN format ensures compatibility and easy parsing of game data for analysis.



lichess_dhawalplayse4_2024-12-04.pgn

```
[Event "Rated bullet game"]
[Site "https://lichess.org/04IITc2m"]
[Date "2023.05.29"]
[White "dhawalplayse4"]
[Black "Nandoskij"]
[Result "0-1"]
[UTCDate "2023.05.29"]
[UTCTime "05:24:43"]
[WhiteElo "2186"]
[BlackElo "2173"]
[WhiteRatingDiff "-20"]
[BlackRatingDiff "+8"]
[Variant "Standard"]
[TimeControl "60+0"]
[ECO "B21"]
[Termination "Time forfeit"]

1. e4 c5 2. d4 g6 3. e5 Bg7 4. Nf3 Qb6
```



ChessCom_dhawalplaysd4_202011.pgn

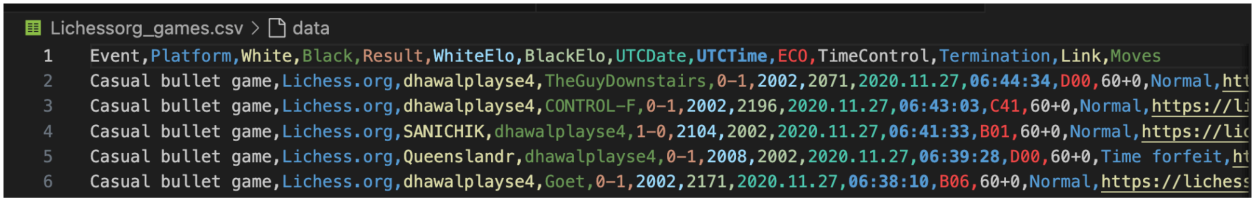
```
[Event "Live Chess"]
[Site "Chess.com"]
[Date "2020.12.01"]
[Round "-"]
[White "dsanjr00"]
[Black "dhawalplaysd4"]
[Result "1-0"]
[CurrentPosition "8/8/1R6/8/7k/4P1Qp/6PP/6K1 b - -"]
[Timezone "UTC"]
[ECO "B02"]
[ECOUrl "https://www.chess.com/openings/Alekhines-Defense-S
[UTCDate "2020.12.01"]
[UTCTime "06:32:03"]
[WhiteElo "1899"]
[BlackElo "1868"]
[TimeControl "60"]
[Termination "dsanjr00 won on time"]
[StartTime "06:32:03"]
[EndDate "2020.12.01"]
[EndTime "06:34:27"]
[Link "https://www.chess.com/game/live/5859685892"]

1. e4 {[%clk 0:00:59.9]} 1... d5 {[%clk 0:00:59.9]} 2. exd5
```

Data Cleaning

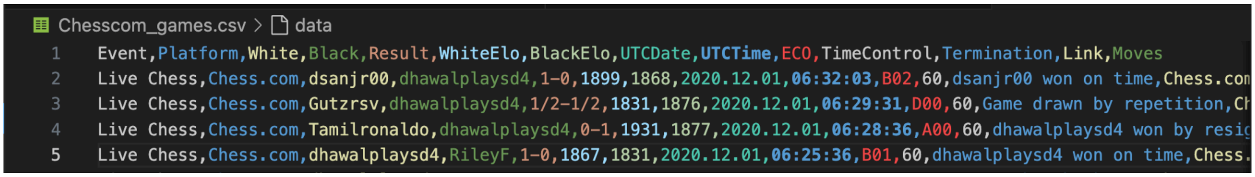
The data retrieved from the API calls of both websites were not consistent, and the **Chess.com** PGN appeared to be more detailed. For analysis, only common data was retrieved and stored in CSV files in the same order. Due to **Lichess.org** providing all game data and **Chess.com** only providing monthly data, with the number of games not being consistent, the analysis was done for the month of **November 2020**, when the chess boom was at its peak and I had played the most games during that month.

A new column was added to the **Lichess** CSV named **"Platform"**, and all data was given the value **"Lichess.org"**. For **Chess.com**, the **"Site"** column was renamed to **"Platform"**.



Lichessorg_games.csv > data

```
Event,Platform,White,Black,Result,WhiteElo,BlackElo,UTCDate,UTCTime,ECO,TimeControl,Termination,Link,Moves
Casual bullet game,Lichess.org,dhawalplayse4,TheGuyDownstairs,0-1,2002,2071,2020.11.27,06:44:34,D00,60+0,Normal,htt
Casual bullet game,Lichess.org,dhawalplayse4,CONTROL-F,0-1,2002,2196,2020.11.27,06:43:03,C41,60+0,Normal,https://li
Casual bullet game,Lichess.org,SANICHIK,dhawalplayse4,1-0,2104,2002,2020.11.27,06:41:33,B01,60+0,Normal,https://lic
Casual bullet game,Lichess.org,Queenslandr,dhawalplayse4,0-1,2008,2002,2020.11.27,06:39:28,D00,60+0,Time forfeit,ht
Casual bullet game,Lichess.org,dhawalplayse4,Goet,0-1,2002,2171,2020.11.27,06:38:10,B06,60+0,Normal,https://liches
```



Chesscom_games.csv > data

```
Event,Platform,White,Black,Result,WhiteElo,BlackElo,UTCDate,UTCTime,ECO,TimeControl,Termination,Link,Moves
Live Chess,Chess.com,dsanjr00,dhawalplaysd4,1-0,1899,1868,2020.12.01,06:32:03,B02,60,dsanjr00 won on time,Chess.com
Live Chess,Chess.com,Gutzrsv,dhawalplaysd4,1/2-1/2,1831,1876,2020.12.01,06:29:31,D00,60,Game drawn by repetition,Ch
Live Chess,Chess.com,Tamilronaldo,dhawalplaysd4,0-1,1931,1877,2020.12.01,06:28:36,A00,60,dhawalplaysd4 won by resig
Live Chess,Chess.com,dhawalplaysd4,RileyF,1-0,1867,1831,2020.12.01,06:25:36,B01,60,dhawalplaysd4 won on time,Chess.
```

Both **Lichess** and **Chess.com** used separate notations for chess moves, which had to be standardized for consistent analysis (code for this can be found on GitHub).

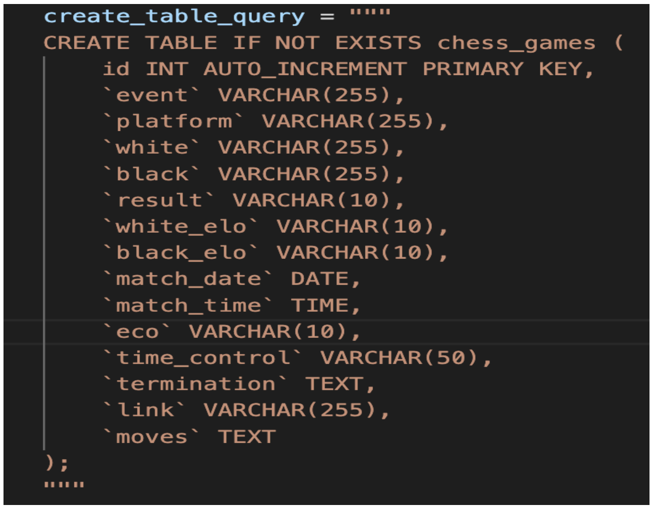
Columns in Both CSVs

The columns standardized and included in the final CSVs were:

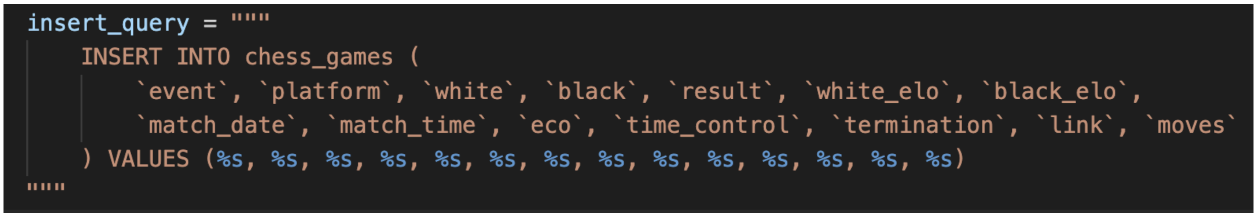
- **Event**: Indicates whether the game was part of a tournament or other mode.
- **Platform**: Specifies whether the game was played on **Lichess.org** or **Chess.com**.
- **White**: Username of the White player.
- **Black**: Username of the Black player.
- **Result**: Outcome of the game (1-0 for White win, 0-1 for Black win, or ½-½ for a draw).
- **WhiteElo**: Chess rating of the White player.
- **BlackElo**: Chess rating of the Black player.
- **UTCDate**: Date the game was played, in UTC.
- **UTCTime**: Time the game was played, in UTC.
- **ECO**: Chess opening played, represented by the corresponding code.
- **TimeControl**: Duration of the match.
- **Termination**: How the game ended (e.g., checkmate, stalemate).
- **Link**: URL to the game.
- **Moves**: All moves played during the game.

Data Storage (Database)

A database **chess_games** was created in MySQL to store all the info.



```
create_table_query = """
CREATE TABLE IF NOT EXISTS chess_games (
    id INT AUTO_INCREMENT PRIMARY KEY,
    `event` VARCHAR(255),
    `platform` VARCHAR(255),
    `white` VARCHAR(255),
    `black` VARCHAR(255),
    `result` VARCHAR(10),
    `white_elo` VARCHAR(10),
    `black_elo` VARCHAR(10),
    `match_date` DATE,
    `match_time` TIME,
    `eco` VARCHAR(10),
    `time_control` VARCHAR(50),
    `termination` TEXT,
    `link` VARCHAR(255),
    `moves` TEXT
);
"""
```



```
insert_query = """
    INSERT INTO chess_games (
        `event`, `platform`, `white`, `black`, `result`, `white_elo`, `black_elo`,
        `match_date`, `match_time`, `eco`, `time_control`, `termination`, `link`, `moves`
    ) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s)
"""
```

Both the CSV data were inserted in the database's table.

Due to the chess PGN format using special notations for the result column:

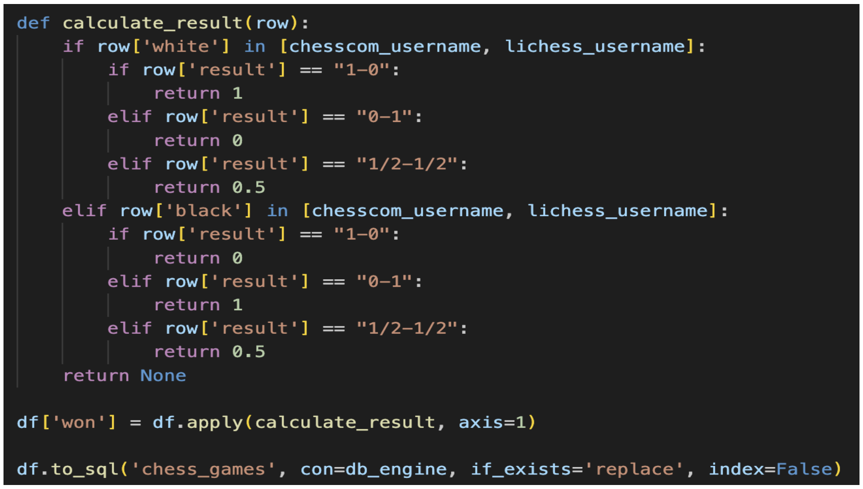
- **1-0** for White winning
- **0-1** for Black winning
- ½-½ for a draw

SQL queries were used to match this column with the actual result to determine if I was the winner. This task was challenging because the notations **1-0, 0-1,** or ½-½ only indicate which color won, not whether I won as White or Black.

To address this, a new column named **"won"** was added to each row in the database, with the following values:

- **1.0** if I won
- **0.0** if I lost
- **0.5** for a draw

This processed data was then used for analysis in data science and machine learning.



```
def calculate_result(row):
    if row['white'] in [chesscom_username, lichess_username]:
        if row['result'] == "1-0":
            return 1
        elif row['result'] == "0-1":
            return 0
        elif row['result'] == "1/2-1/2":
            return 0.5
    elif row['black'] in [chesscom_username, lichess_username]:
        if row['result'] == "1-0":
            return 0
        elif row['result'] == "0-1":
            return 1
        elif row['result'] == "1/2-1/2":
            return 0.5
    return None

df['won'] = df.apply(calculate_result, axis=1)

df.to_sql('chess_games', con=db_engine, if_exists='replace', index=False)
```

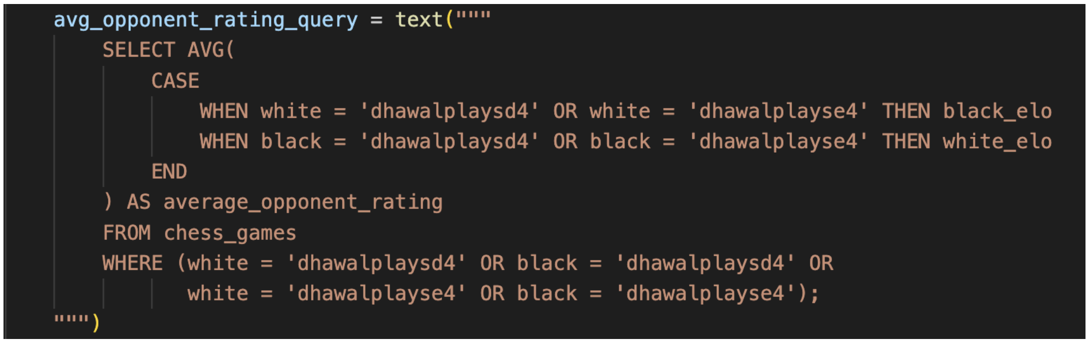
Data Science

The project covers comprehensive analyses that are essential for determining optimal chess strategies, such as the best openings to play, the most effective times to play, and overall performance improvement. These insights are particularly valuable for tournament preparation and enhancing ratings on leading This project analyzes chess gameplay to identify optimal strategies, best openings, effective play times, and performance trends, enhancing tournament preparation and ratings on platforms like Chess.com and Lichess.org.

Techniques and Insights

- **Exploratory Data Analysis (EDA):**
  - **Win/Loss/Draw:** Analyzed overall and platform-specific trends (Chess.com vs. Lichess.org).
  - **Win Percentages:** Evaluated across platforms, time controls (blitz, bullet, rapid), and times of day to identify peak performance periods.

  - **Opening Analysis:** Examined frequently played openings (ECO codes) and their win rates to prioritize effective openings.
  - **Average Opponent Rating Analysis:** Assessed performance against varying rating levels, revealing strengths and weaknesses by competition tier.



```
avg_opponent_rating_query = text("""
    SELECT AVG(
        CASE
            WHEN white = 'dhawalplaysd4' OR white = 'dhawalplayse4' THEN black_elo
            WHEN black = 'dhawalplaysd4' OR black = 'dhawalplayse4' THEN white_elo
        END
    ) AS average_opponent_rating
    FROM chess_games
    WHERE (white = 'dhawalplaysd4' OR black = 'dhawalplaysd4' OR
           white = 'dhawalplayse4' OR black = 'dhawalplayse4');
""")
```

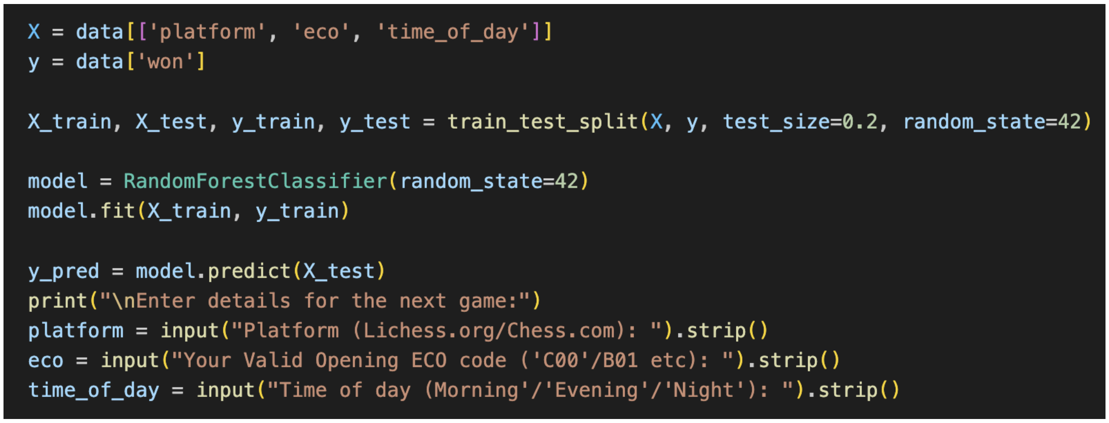
These analyses provided essential insights into my playing patterns, strengths, and areas for improvement. By understanding the impact of different factors such as the time of play, the platform used, and the choice of opening, I gained a strategic advantage that can be leveraged for improved performance and more effective tournament preparation.

Machine Learning

One main question that every chess player has is the likelihood of winning their next match. While there is no certainty in predicting game outcomes, machine learning techniques can be used to estimate the chances based on historical data. This project aimed to develop a model that predicts the probability of winning based on features such as the time of day the game was played, the platform used (e.g., **Chess.com** or **Lichess.org**), and the chess opening employed.

Approach and Features

- **Feature Selection**: The key features used to train the machine learning model included:
  - **Time of Day**: The specific time when the game was played, as this could influence player performance.
  - **Platform**: The platform on which the game was played, since performance could vary between **Chess.com**and **Lichess.org**.
  - **Opening (ECO code)**: The chess opening used, as certain openings may have higher win rates than others.



```
X = data[['platform', 'eco', 'time_of_day']]
y = data['won']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print("\nEnter details for the next game:")
platform = input("Platform (Lichess.org/Chess.com): ").strip()
eco = input("Your Valid Opening ECO code ('C00'/B01 etc): ").strip()
time_of_day = input("Time of day (Morning'/'Evening'/'Night'): ").strip()
```

- **Model Training**: A supervised learning approach was applied, using labeled data from past games to train the model. The target variable was set as the outcome of the game, represented as:
  - **1.0** if I won,
  - **0.0** if I lost,
  - **0.5** for a draw.
- **Model Evaluation**: Various machine learning algorithms were explored, such as logistic regression, decision trees, and more complex models like random forests or gradient boosting. The models were evaluated based on accuracy, precision, recall, and F1-score to determine their effectiveness in predicting the chances of winning.

These predictive models aimed to provide insights into how factors such as the time of day, platform, and opening choices could influence the outcome of the next game. By leveraging these models, chess players can make more informed decisions, potentially enhancing their strategic approach and improving their overall performance in future matches.

Major Findings

Win/Loss/Draw Analysis

Total Games Played Overall:

- Wins: 498
- Losses: 425
- Draws: 28

Win Percentages Overall:

- Chess.com: 55.19%
- Lichess.org: 52.20%

Best Time Control Analysis: I win the most 73% in 15 sec games on chess.com and 59% in 1 minute games on Lichess.org

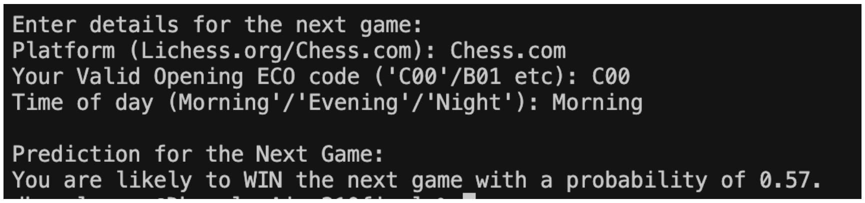
Time of Day Analysis: Most wins were recorded during "UTC mornings" on Chess.com, indicating that morning games may be more favorable for better performance.

Opening Analysis

- Black: The most frequently played opening was B01 (The Scandinavian Defense), which had the highest win rate.
- White: The C21 (Centre Game) opening had the most wins when playing as White.

Predictive Model

Based on the model, if playing on Chess.com, choosing the C00 opening and playing in the morning had an estimated win probability of 0.57 (57%).



```
Enter details for the next game:
Platform (Lichess.org/Chess.com): Chess.com
Your Valid Opening ECO code ('C00'/B01 etc): C00
Time of day (Morning'/'Evening'/'Night'): Morning

Prediction for the Next Game:
You are likely to WIN the next game with a probability of 0.57.
```

Advantages and Limitations

Advantages:

- Personalized Insights: Provides tailored data-driven insights across platforms to improve gameplay.
- High Model Interpretability: The Random Forest model allows easy interpretation of feature importance and decision-making.
- Integration of Temporal Factors: The model accounts for time of day, platform, and openings, offering actionable strategies.

Limitations:

- ECO Code Limitations: The use of ECO codes may lack the detail needed to fully capture complex positions and strategies.
- Time Zone Issues: The categorization of time_of_day may be affected by time zone inconsistencies.
- Static Model Assumptions: The model assumes consistent performance and does not account for potential changes due to fatigue or improvement over time.

Changes After Proposal:

The original proposal aimed to analyze extensive player data, including performance and ratings, and compare them to FIDE ratings. However, the limitations of the API prevented access to complete player information, altering the scope of the project.

Conclusion:

This project highlights the power of data-driven analysis in enhancing personal chess strategies. By employing EDA and predictive modeling, I was able to derive actionable insights and create a tool for estimating match outcomes. The results supported the hypothesis that the time of day, platform, and choice of opening have significant impacts on game outcomes. Despite certain limitations, the framework established can be further developed to incorporate more complex factors such as fatigue and psychological elements.

Future Work:

- Feature Expansion: Include more variables like time controls and psychological factors (e.g., stress levels or confidence).
- Web-Based Dashboard: Develop an interactive dashboard for real-time data visualization and analysis of game insights.