Predicting Scrabble Player Rating

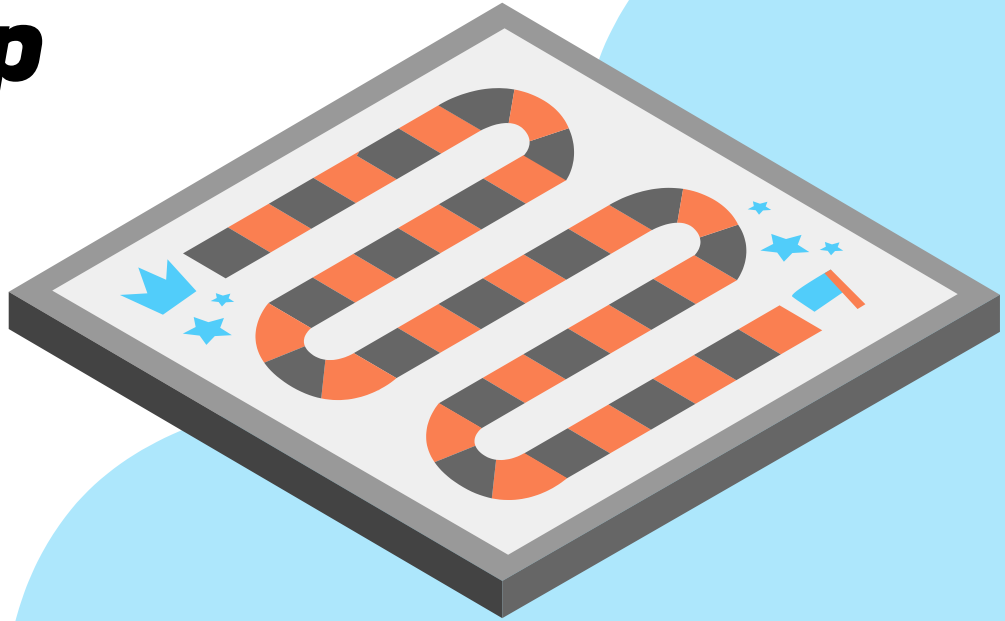
Alan Chen, Andy Byun, Claire Liu, Sally Lee, Eric Pan

AGENDA

1. Solution Map
2. Introduction to Datasets
3. Basic Features
4. Model Selection Process
5. Advanced Features and Improvement
6. Best Model and Final Score

01

Solution Map

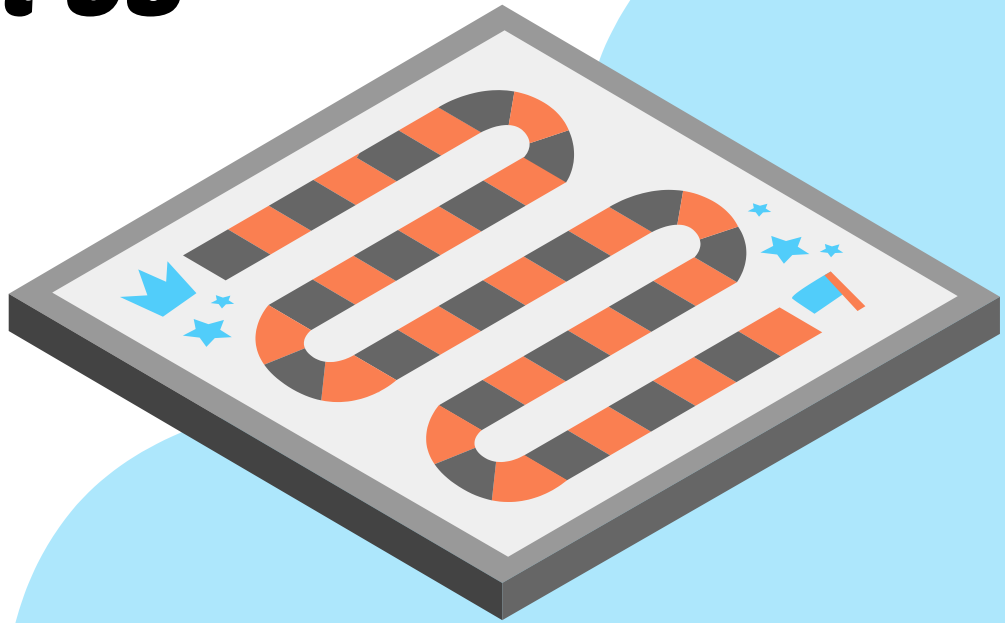


SOLUTION MAP

Dataset Introduction	What are the default features of datasets from Kaggle?	
Basic Features Added	What additional basic features might affect player ratings?	
Model Evaluation & Selection	What models perform the best?	Compare Root Mean Square Error (RMSE) of Linear Regression, k-Nearest Neighbor model, XGBoost, and Random Forest
Advanced Features Added	What advanced features affect player ratings?	Create features that represent a player's game play history
Final Deployment	Which model produces the lowest RMSE?	

02

Introduction to Datasets



INTRODUCE THE DATA SET

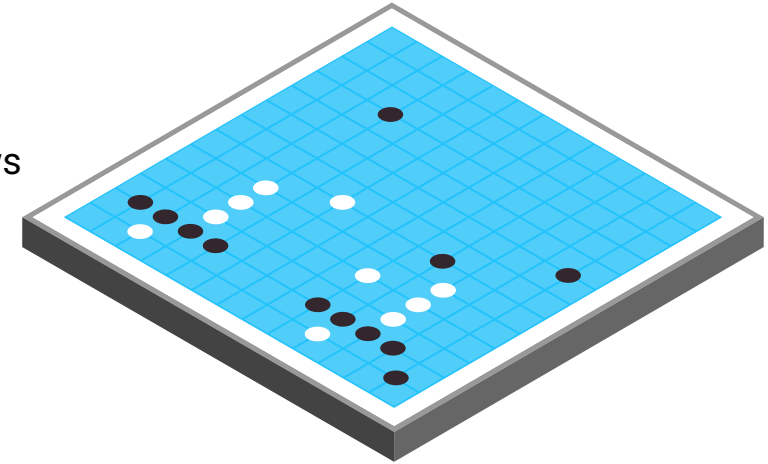
All data used for the analysis have been sourced from the **Kaggle** competition
The dataset comprises **four files** with one sample submission file

GAMES

12 columns, 72,772 rows

TURNS

9 columns, 2,005,497 rows



TRAIN

4 columns, 100,820 rows

TEST

4 columns, 44,726 rows

INTRODUCE THE DATASET

Detail information about the **files**

GAMES

- Metadata about each game
- game ID, game's duration etc

TURNS

- Detailed information about every turn in each game
- Points, current rack, moves etc

TRAIN

- Final scores and ratings for each player

TEST

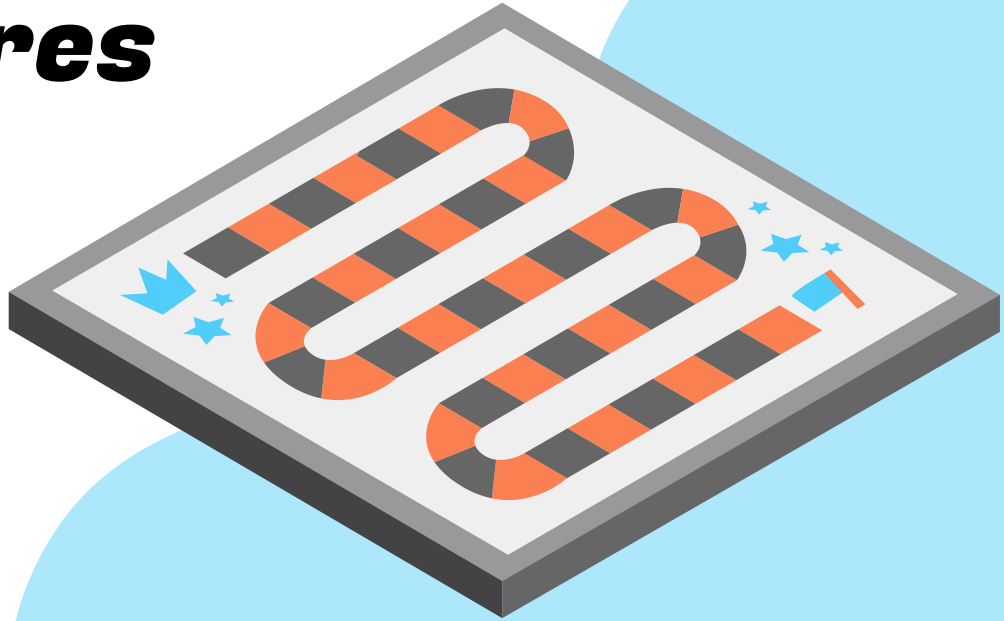
- Final scores and ratings for each player
- **Predicting the missing values in the test data**

SAMPLE SUBMISSION

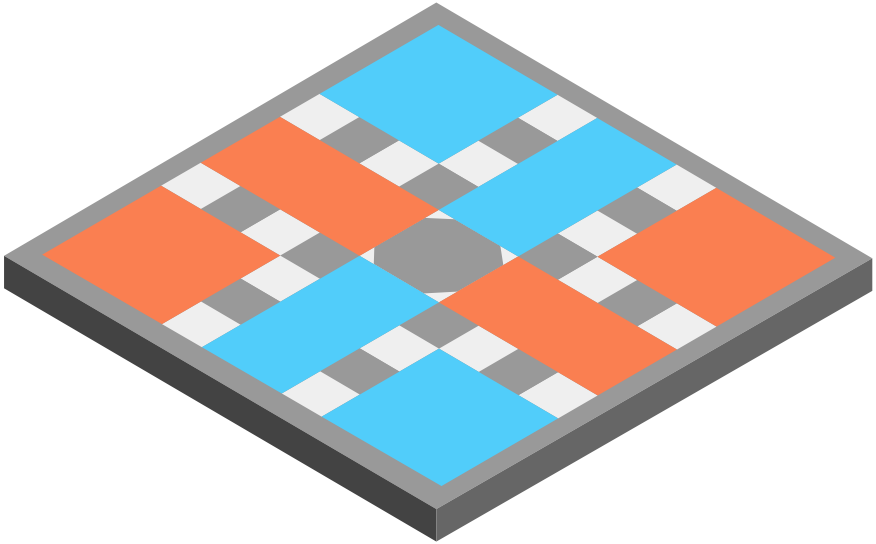
- Reference for the correct format when submitting predictions.

03

Basic Features



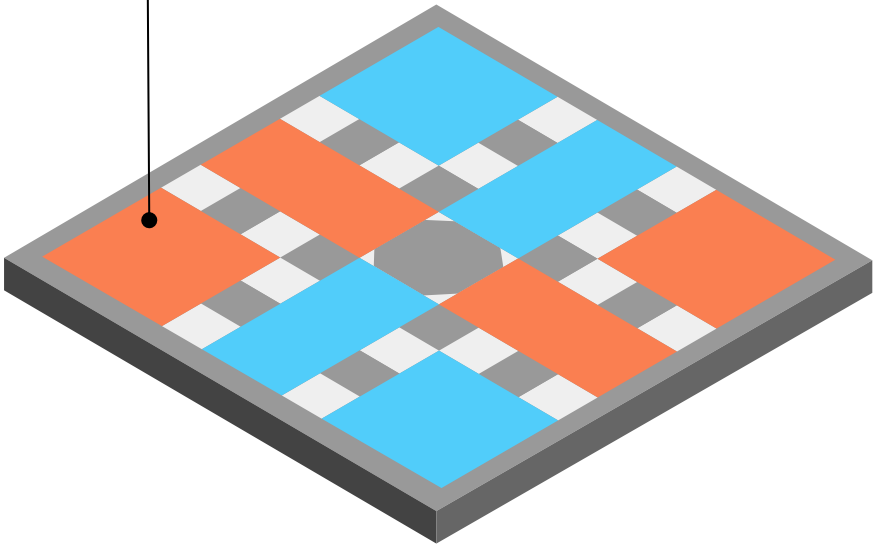
What features did we develop?



What features did we develop?

Letters Assemblement

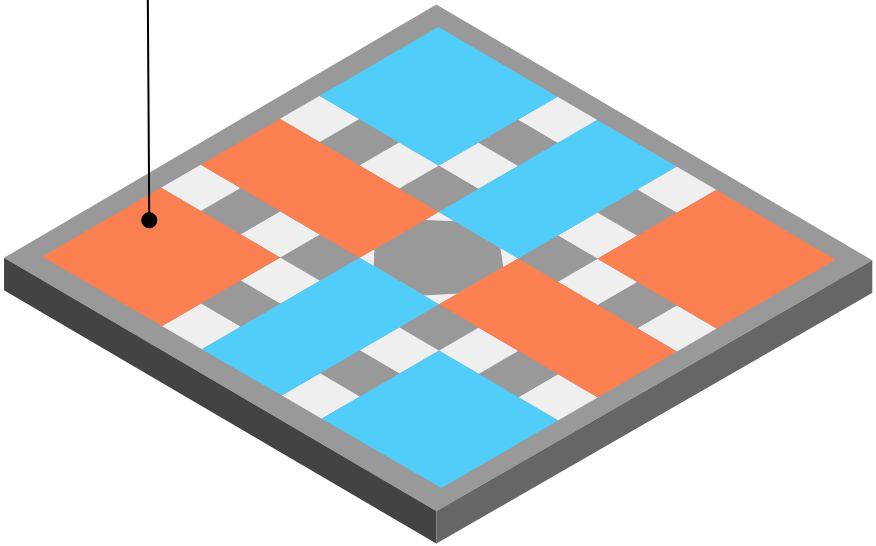
- Length of Moves



What features did we develop?

Letters Assemblement

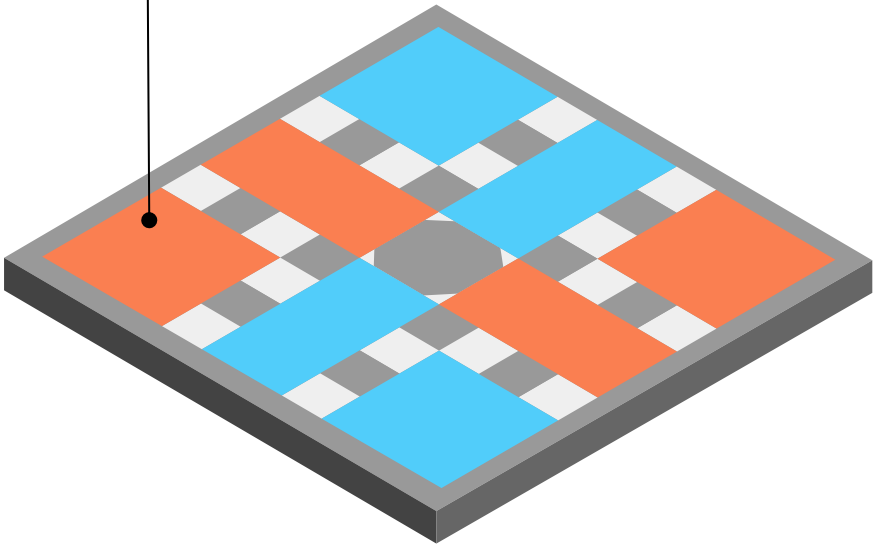
- Length of Moves
- Letters' Difficulty



What features did we develop?

Letters Assemblement

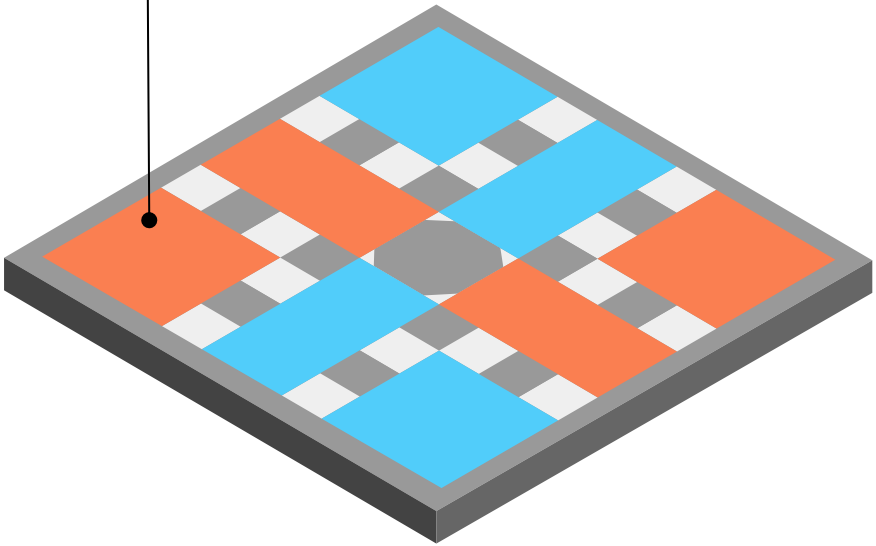
- Length of Moves
- Letters' Difficulty
- Blank Tiles Used



What features did we develop?

Letters Assemblement

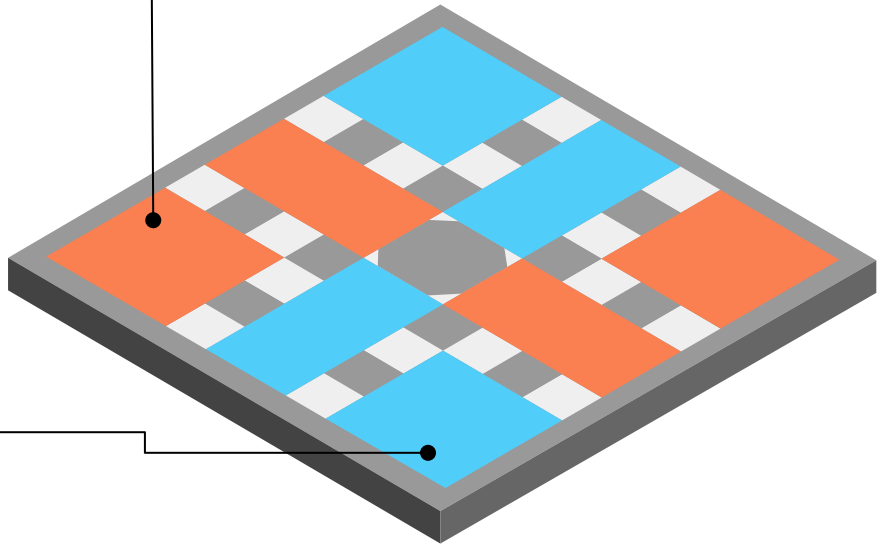
- Length of Moves
- Letters' Difficulty
- Blank Tiles Used
- Bingo Moves



What features did we develop?

Letters Assemblement

- Length of Moves
- Letters' Difficulty
- Blank Tiles Used
- Bingo Moves



Tiles Placement

- Location Bonus

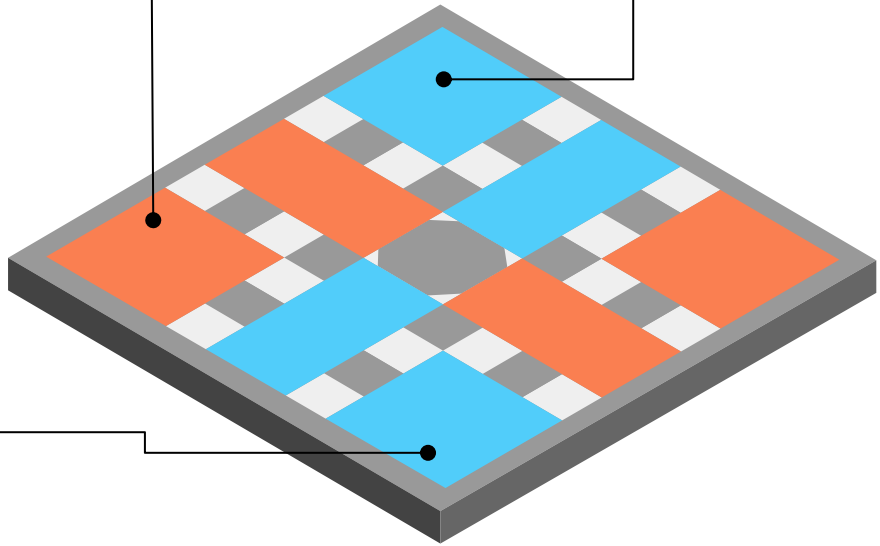
What features did we develop?

Letters Assemblement

- Length of Moves
- Letters' Difficulty
- Blank Tiles Used
- Bingo Moves

Historical Performance

- Win-Loss Rate



Tiles Placement

- Location Bonus

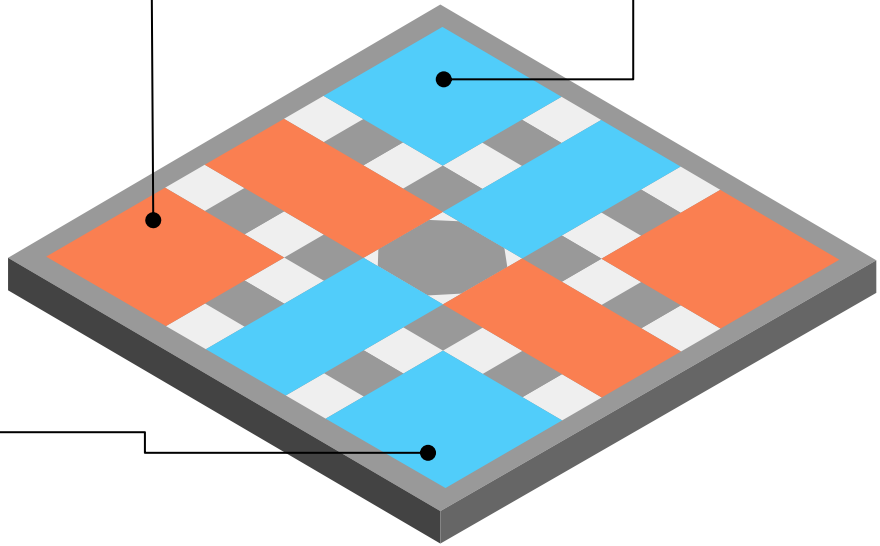
What features did we develop?

Letters Assemblement

- Length of Moves
- Letters' Difficulty
- Blank Tiles Used
- Bingo Moves

Tiles Placement

- Location Bonus



Historical Performance

- Win-Loss Rate
- Mean Score of Latest 10 Games

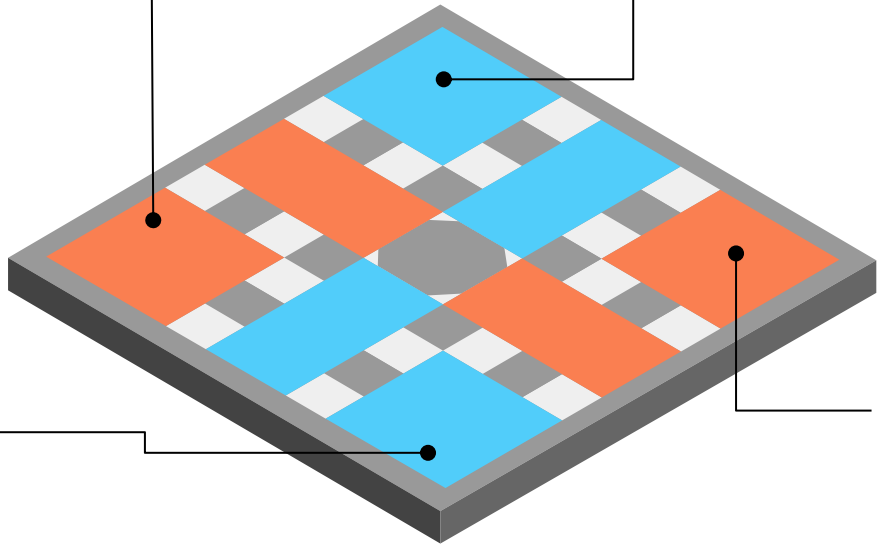
What features did we develop?

Letters Assemblement

- Length of Moves
- Letters' Difficulty
- Blank Tiles Used
- Bingo Moves

Historical Performance

- Win-Loss Rate
- Mean Score of Latest 10 Games



Tiles Placement

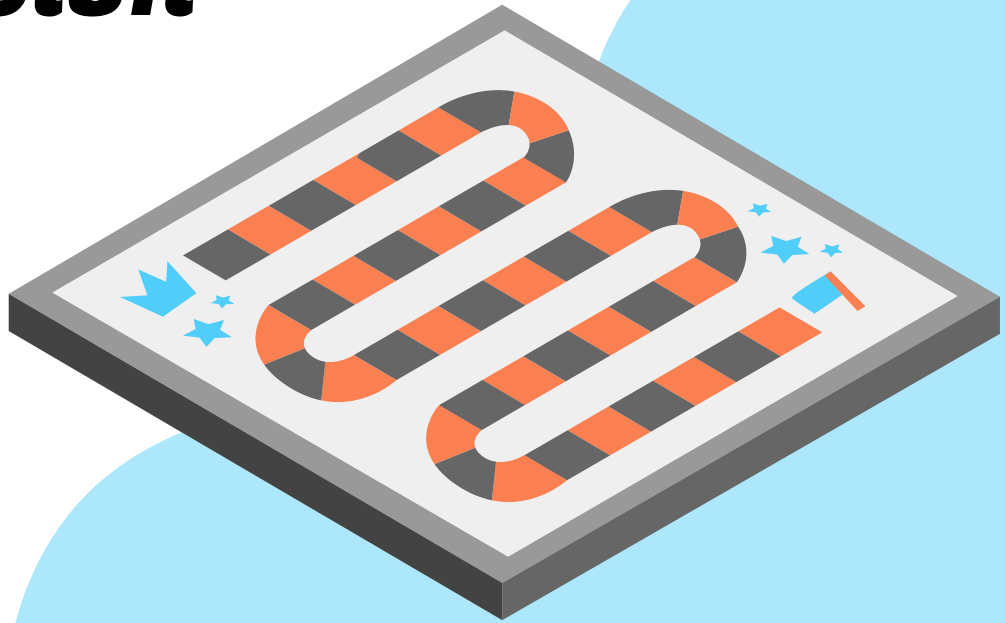
- Location Bonus

Competitor

- Game Level

04

Model Selection Process



How to evaluate model performances?

Root Mean Squared Error (RMSE)

Measure the **average difference** between predicted and actual values

Lower RMSE indicates **better** model performances

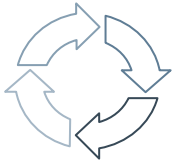
Sensitive to **outliers**

What is overfitting?

not generalizable

Data

=



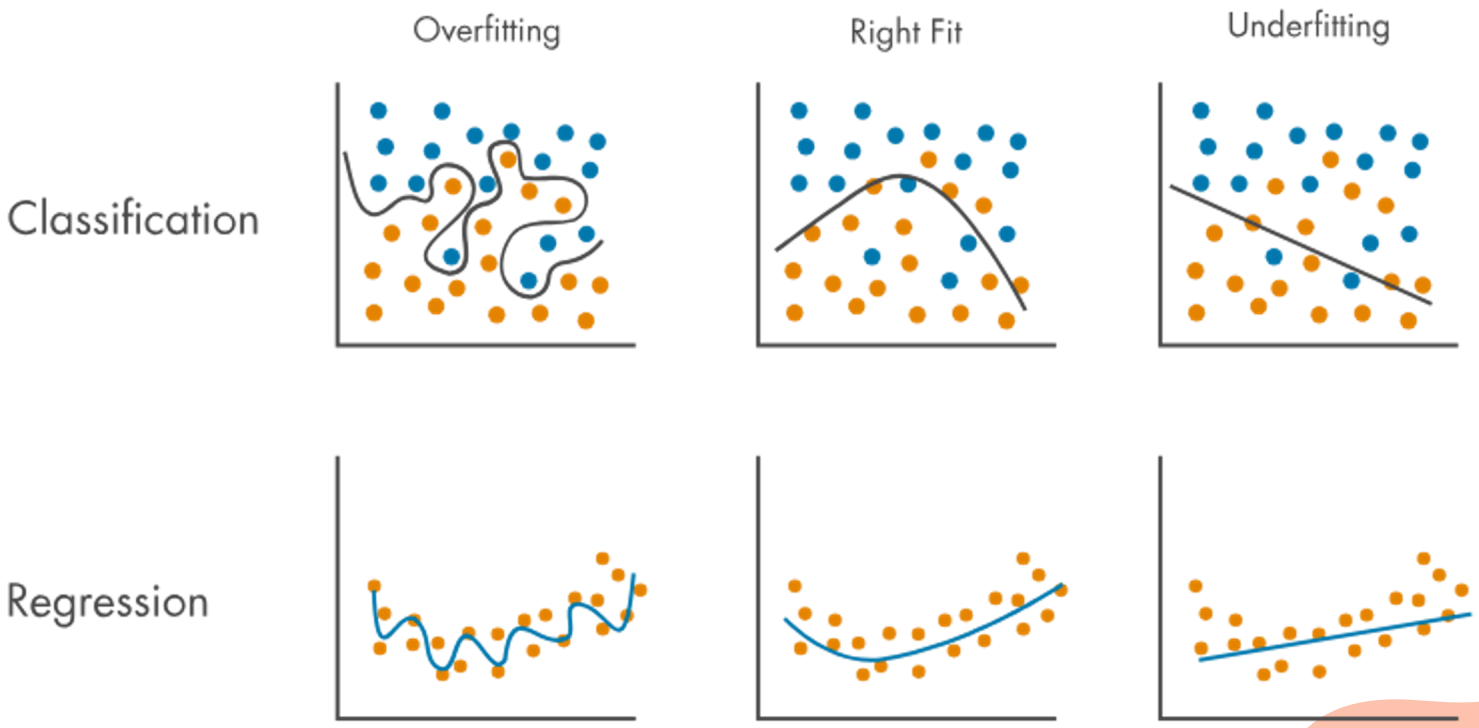
pattern

+

noise

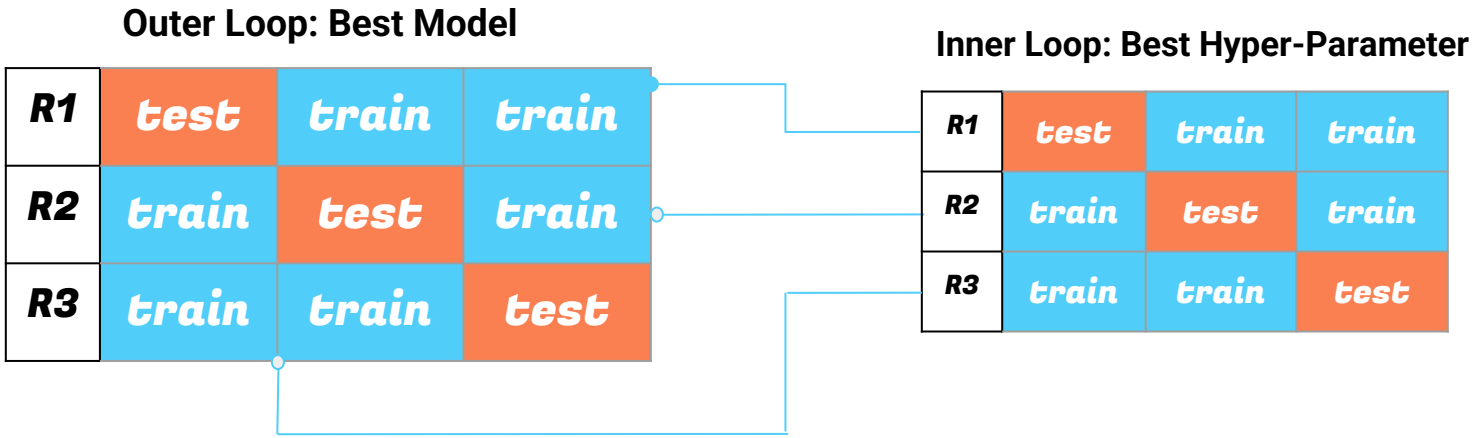
generalizable

What is overfitting?

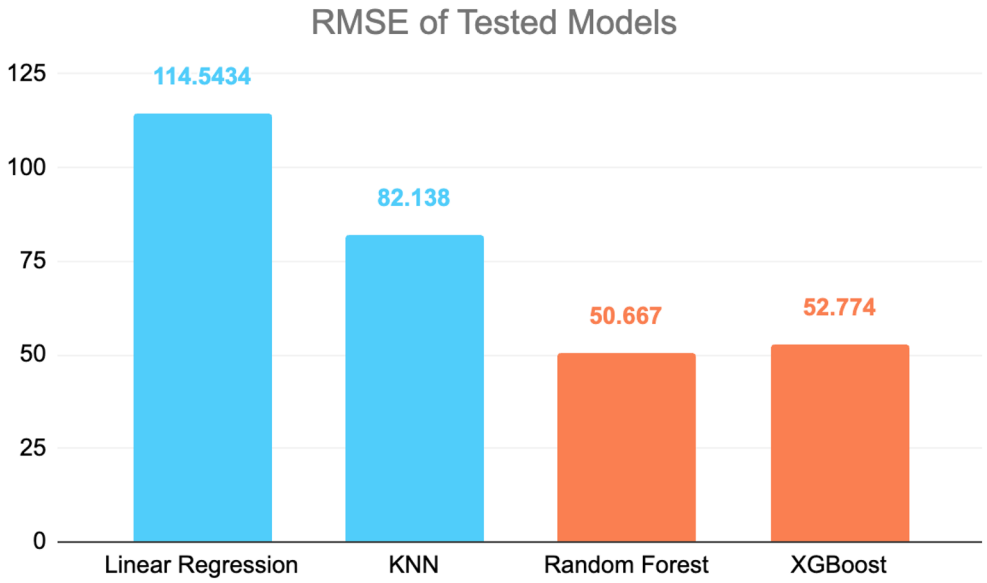


How to perform model selection?

Nested CV

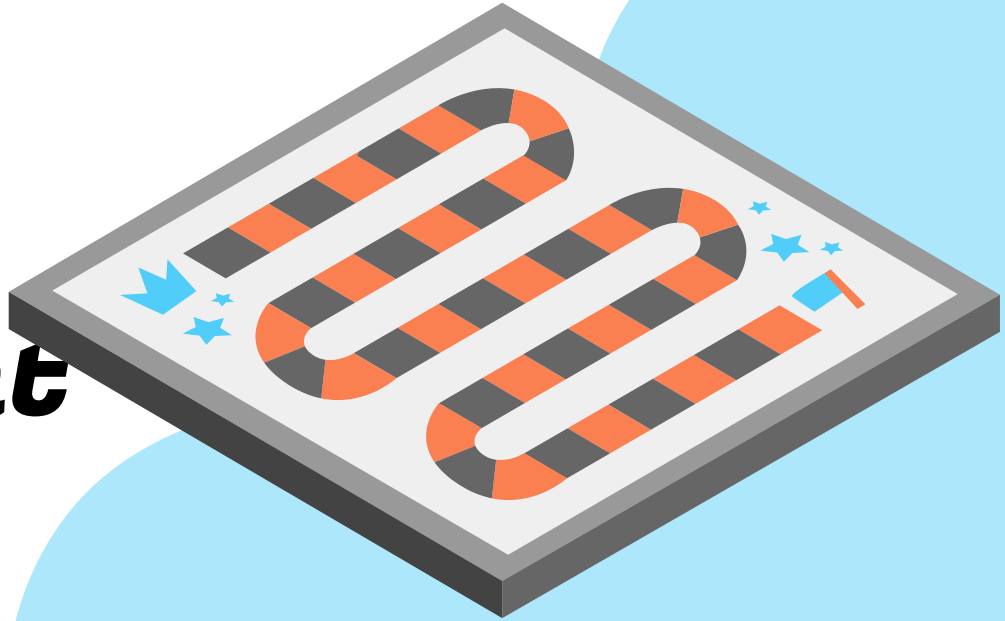


What models did we try?



05

Advanced Features and Improvement



Features Related Game History Added

Rolling Average Score

Rolling Win Rate

Rolling Average of Turn-based feature

Other Improvement

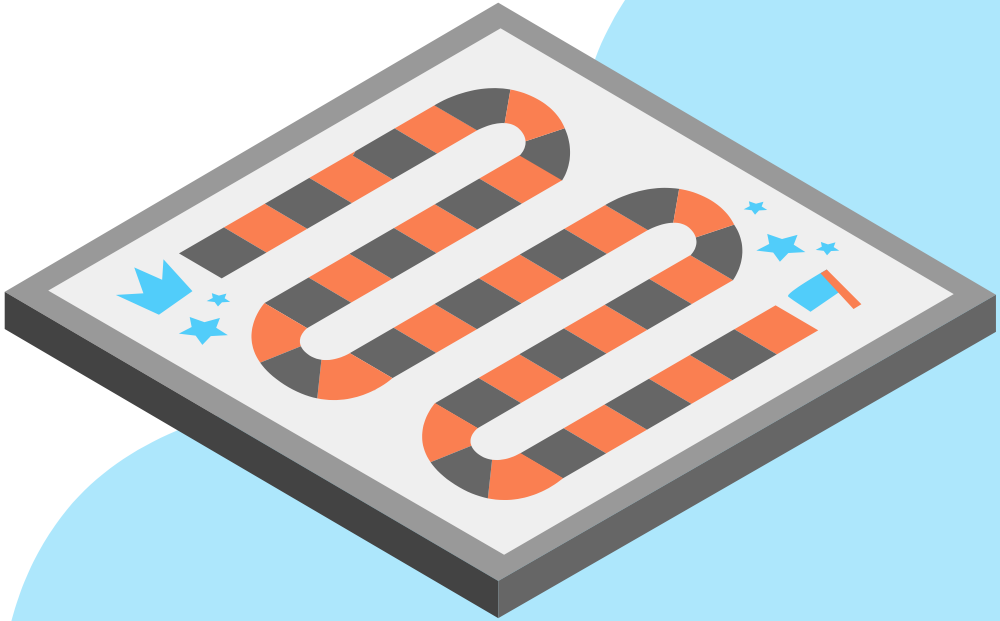
Build separate models for different game mode

Remove Anomalies

Combined models

06

Best Model and Final Score



Best Model: XGBoost

- Better performance
  - Stronger algorithm among all models
- Robust
  - Handles large amounts of information well
- Convenience
  - No need to normalize features before utilizing

Final Score

xgboost_submission_8_window30.csv Complete (after deadline) · 2d ago	109.50917	111.07828	

Thank you for your time