TECHNICAL MODEL EVALUATION

W210 Capstone Workshop

UC Berkeley School of Information

10/19/22

Mickey Hua
Jeffrey Budiman
Mrinal Chawla
Adi Khurana

Background

Technical Model Evaluation

- To access the performance of a model.
- The metric of choice is dependent on the task itself.

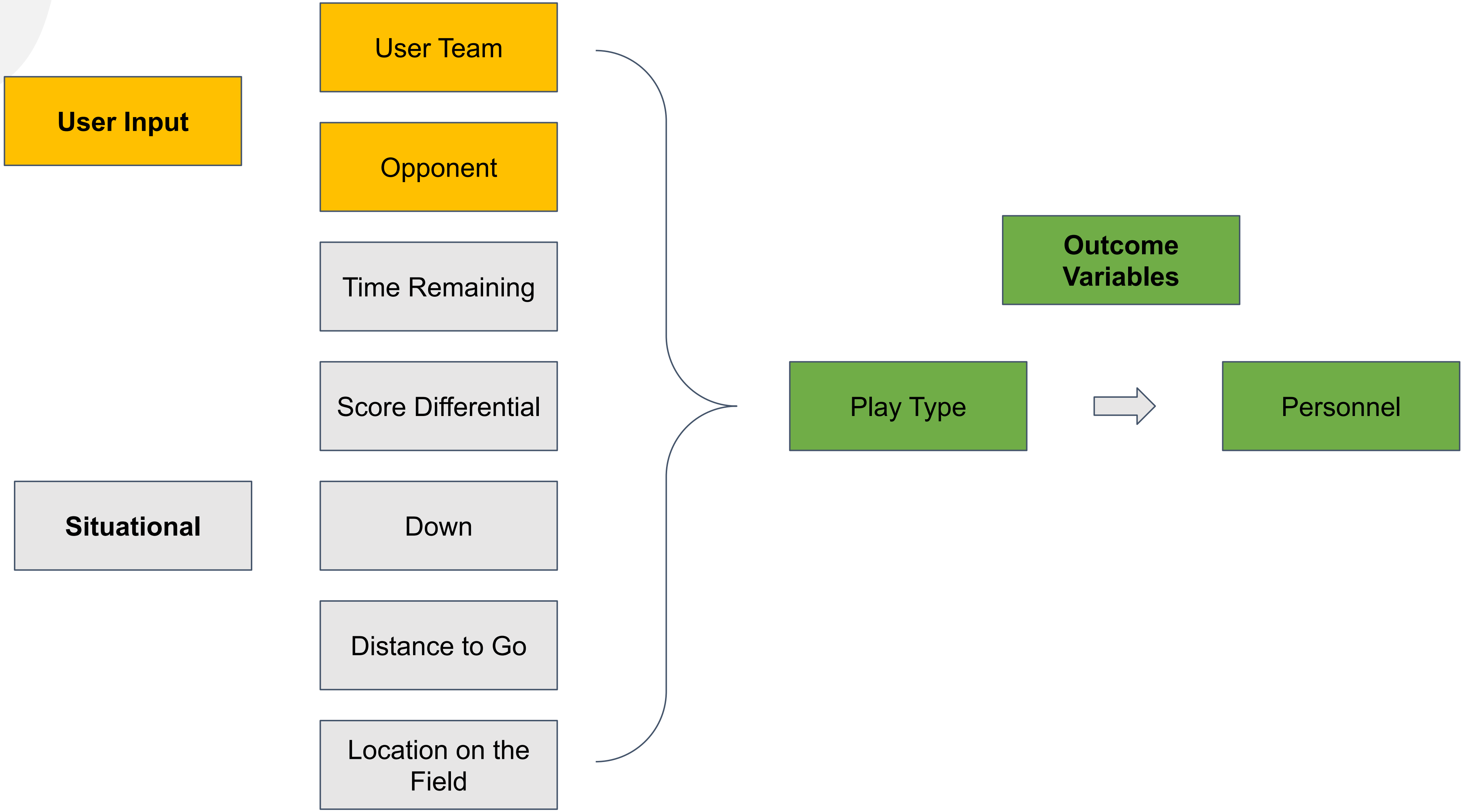
Techniques

- Train / Test Splitting
- Cross Validation
- Time Series Blocking
- Replay

Metrics

- Accuracy, precision, recall, F-score
- R-squared, MSE
- Speed, Memory
- Regret

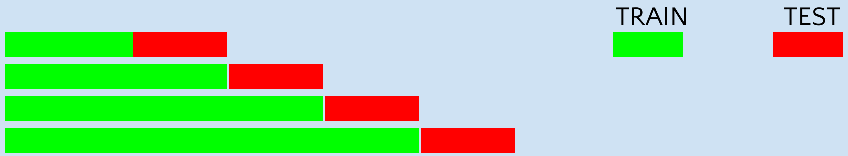
Our Project's Context



Challenges

- Reinforcement Learning - Contextual Bandit
  - Action variable: play type (pass / rush)
  - Rewards variables: yards gained, touchdown, 1st down
  - Iterations within offline process
- Traditional Classification vs Reinforcement Learning
- Each situation needs to prioritize different goals
  - Rewards are different

Things to be aware of

Potential Issue:	Treatment
Imbalanced Data	● Under-sampling ● Over-sampling ● SMOTE (Synthetic Minority Oversampling Technique) ● Adjust Metric
Data Leakage	Rolled out Cross Validation 
Overfitting & Underfitting	● L1 & L2 Regularization ● Hyperparameter tuning with Cross Validation ● Features set

Techniques

GOAL: How does the model perform in the real world?

- Train / Test Splitting
  - Test generalizability
- Cross Validation
  - Vary unseen data
  - Hyperparameter Tuning
- Time Series Blocking
  - Prevent leakage

- Replay
  - Select instances where prediction matches reality
  - Define success criteria
  - Divide successful by total matches

Replay

Actual Play:	PASS	PASS	RUSH	RUSH
Predicted Play:	PASS	RUSH	RUSH	RUSH
Used in Evaluation:	✓	✗	✓	✓
Successful Play:	✓		✓	✗

Replay: 2/3

Metrics

- What type of problem am I solving?
- Is my dataset balanced?
- What is the real world impact of a wrong prediction?
- Do I have an imposed threshold?
- What are my model training and deployment constraints?

Metrics

Classification

- Accuracy, Precision, Recall, F-score
- PR Curve, ROC Curve, AUC

Regression

- (Adjusted) R Squared
- (Root) Mean Squared Error
- Mean Absolute Error

Bandits

- Regret
- Total Reward
- Replay

Auxiliary

- Speed
- Memory

Improving Model

- Confusion Matrix
  - Analyze the kind of errors
- Data - Outliers/Missing/Incorrect Tagging
  - Dataset collated from multiple independent sources such as NFLFastR, NFL Data, Dynasty Process & Draft Scout. We have some outliers data points and may be incorrect play tagging.
- Standardizing Ranges
  - Standardizing the fields such as 'Yards to Goal' helped us standardize the features.
  - We will continue to standardize other features /data fields. '
- Player/Coach effects - Preferences
  - Players or coaches that prefer or excel at certain plays will likely execute them due to their preferences. For Example, Odell Beckham Jr. is exceptional as WR and will require to create play catering to his strengths.

Summary

- (Many) Metrics to consider *'More than one metric can be considered ideal'*
  - Metric determination should be based on the end goal of the product.
  - Many metrics; We concluded that in our situation, which metrics are less important than making the sure evaluation is done against a baseline?
- Analyzing Errors -*'What plays or situations the model incorrect predicted?'*
  - Without a detailed understanding of the errors, it's hard to know what to improve.
  - Also, the errors could be stemming from other unknown areas as the sport continues to evolve over the years.
- Situational plays - *'A solution needs to account for possible situations?'*
  - Given the nature of the game, the situation may dictate a certain play, and given the risk tolerance or coach's confidence in successful execution or fatigue or other factors may dictate a different play call. (Based on the SME interviews)

THANK YOU.

Questions & Answers

Have further questions?

Team Members::
Jeffrey Budiman
Mrinal Chawla
Mickey Hua
Adi Khurana

Discussion Questions

1. How do you decide the right metric to evaluate your model? What made you chose it?

2. What constraints prevent chasing a "perfect" model?

3. What makes your project unique to evaluate?