

# Introduction

## March Data Crunch Madness 2025 – An Overview

Our training dataset spans March Madness data from 2002 to 2024, capturing a comprehensive history of tournament outcomes and key statistical insights



### Game & Team Performance Metrics

Shooting percentages, time of possession, scoring & defensive efficiency, and more.



### Team & Coaching History

Seed rankings, NCAA & Final Four appearances, and coaching experience.



### Strategic Feature Selection

We excluded statistically insignificant variables, utilizing Principal Component Analysis



### Blending with NLP Insights

Integrated sentiment from Reddit comments, leveraging NLP scores to refine final predictions

# Key Sports Variables

01

Adjusted Efficiency  
Margin

02

Seed Differential

03

Point Differential

04

Free Throw Rate

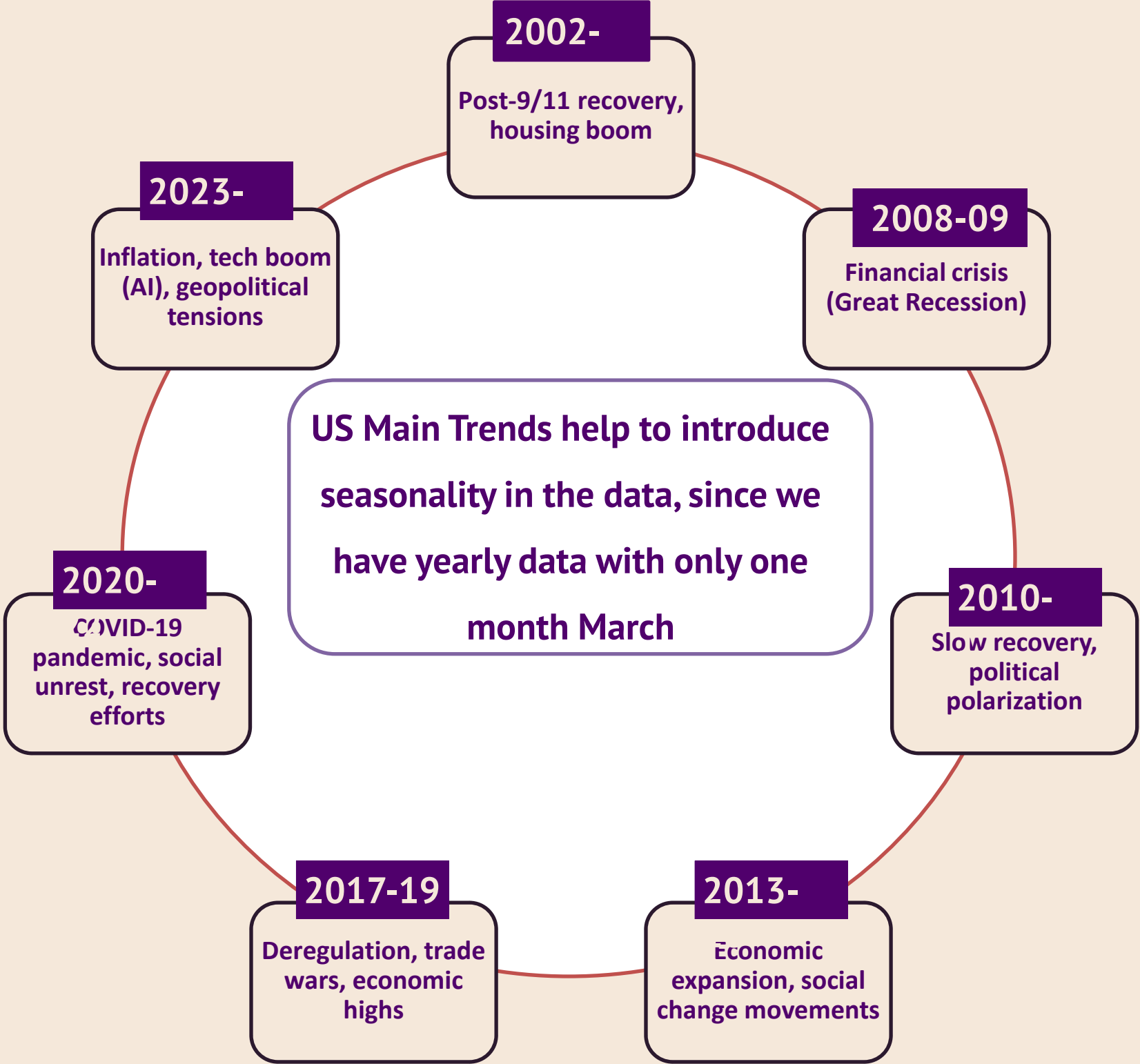
05

Turnover Margin

06

Coach Experience  
Score

# Macroeconomic Data



Year	March Airline Inflation %
2002	-2.3
2003	-1.8
2004	0.5
.	.
.	.
2023	17.7
2024	-7.1

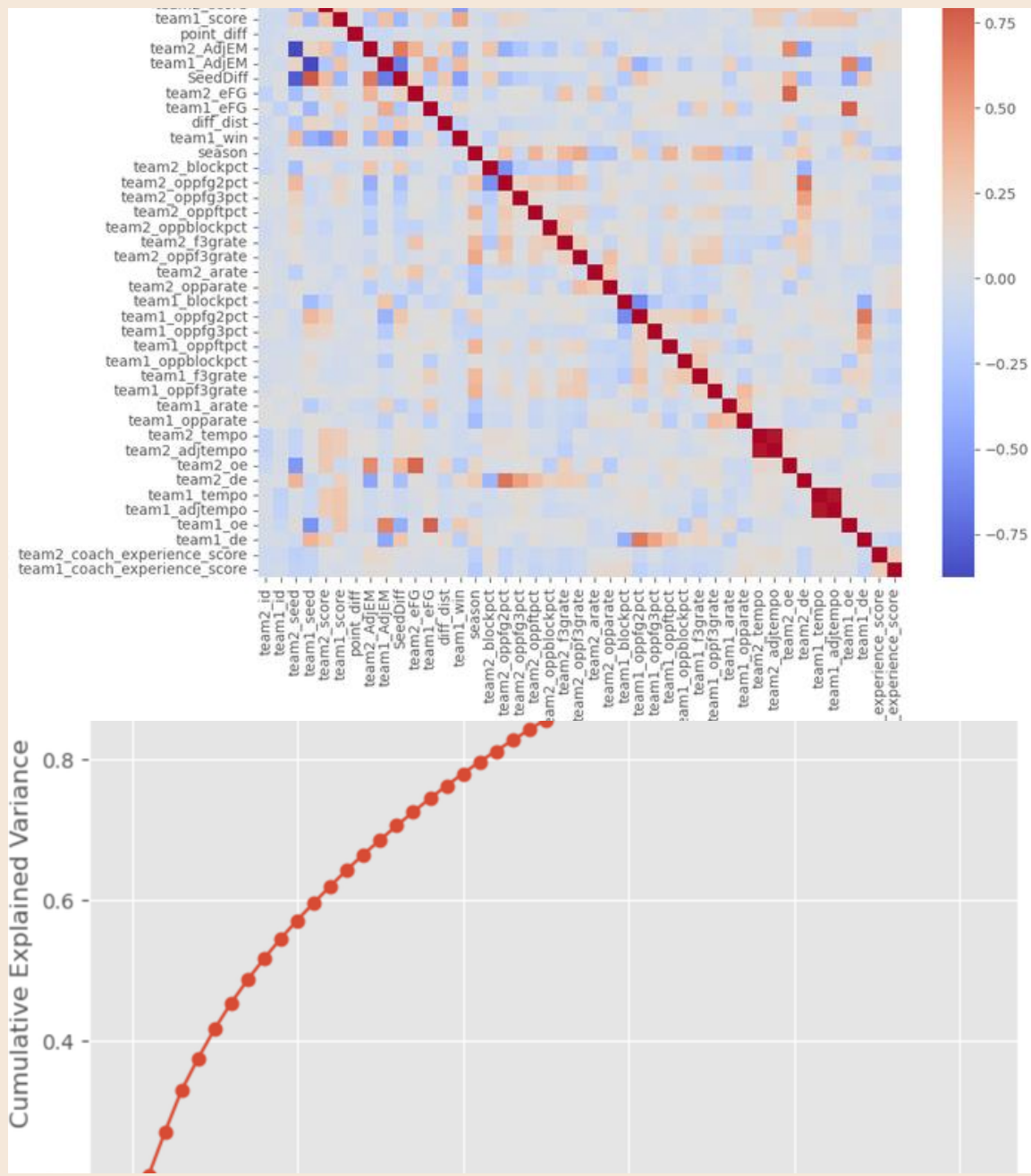
March Airline Inflation can be a proxy to capture overall external factors influencing game outcomes



Economic uncertainty can affect **players' mental health and contract negotiations**, leading to performance fluctuations

# Variable Selection


## Principal Component Analysis



44

38


## Optimized Variable Combinations


 **Innovative Approach:** Leveraging ChatGPT to dynamically generate **unique variable combinations** for model optimization

Started with 10 fixed variables

Randomly selected 3 to 7 additional variables from the remaining pool

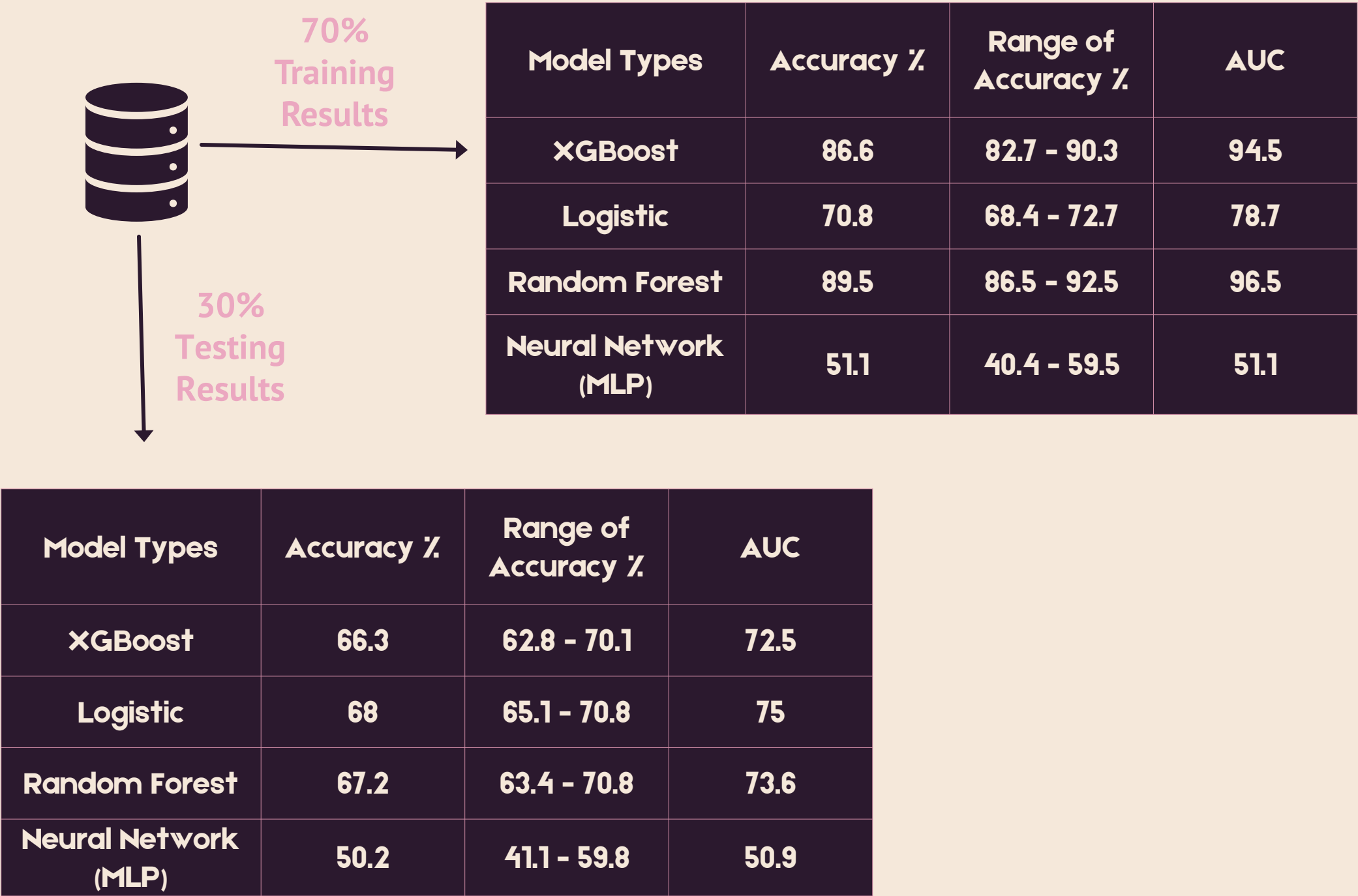
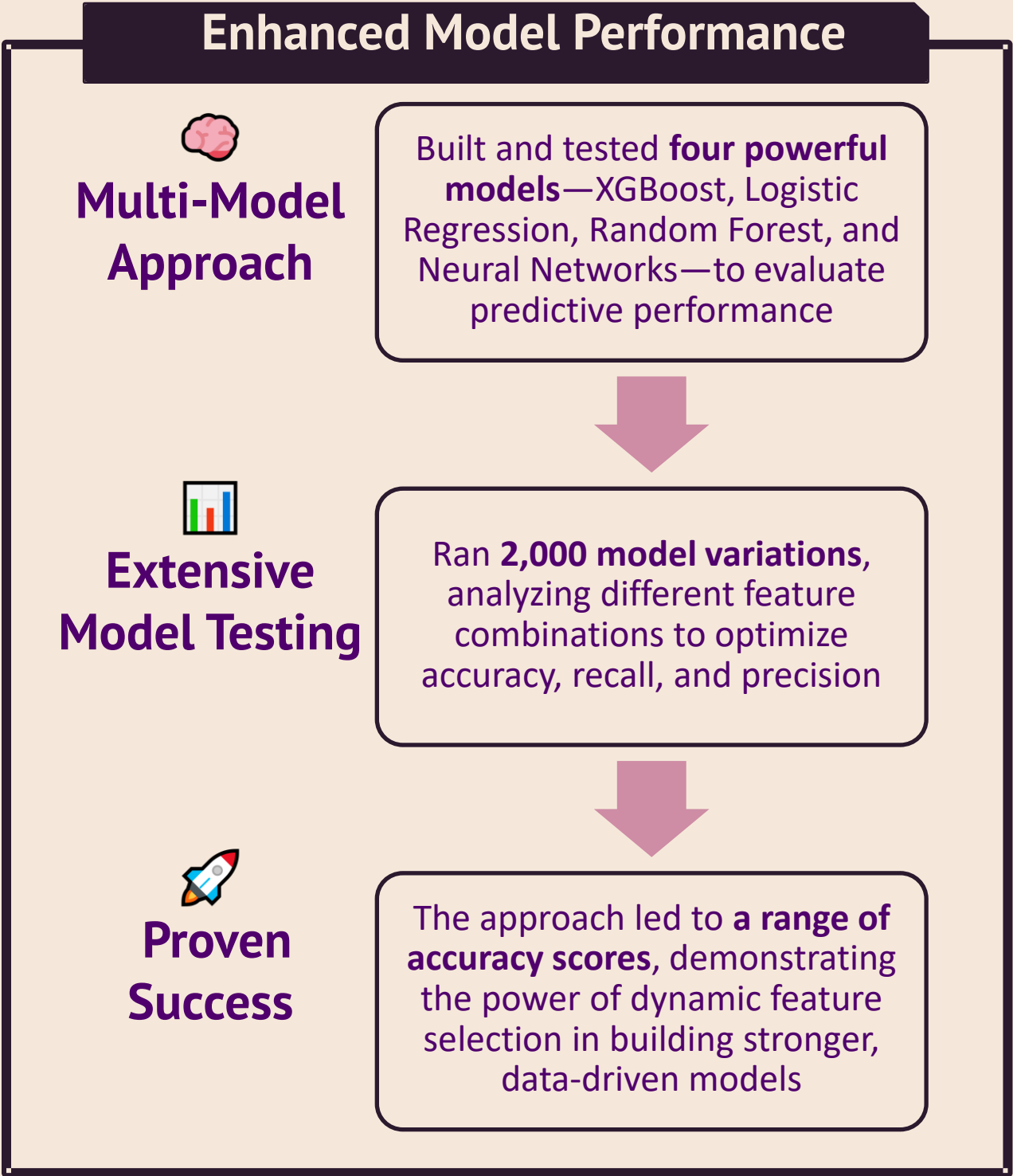
Generated 500 unique combinations to enhance model diversity

 **Optimizing feature selection for the best-performing models!**

  
Count of  
Modeling  
Variable before  
each step

17  
Max

# Model Building



# Model Selection

  
**Year-Based  
Model Validation**

This leads to better  
generalization of  
model for **2025**  
predictions

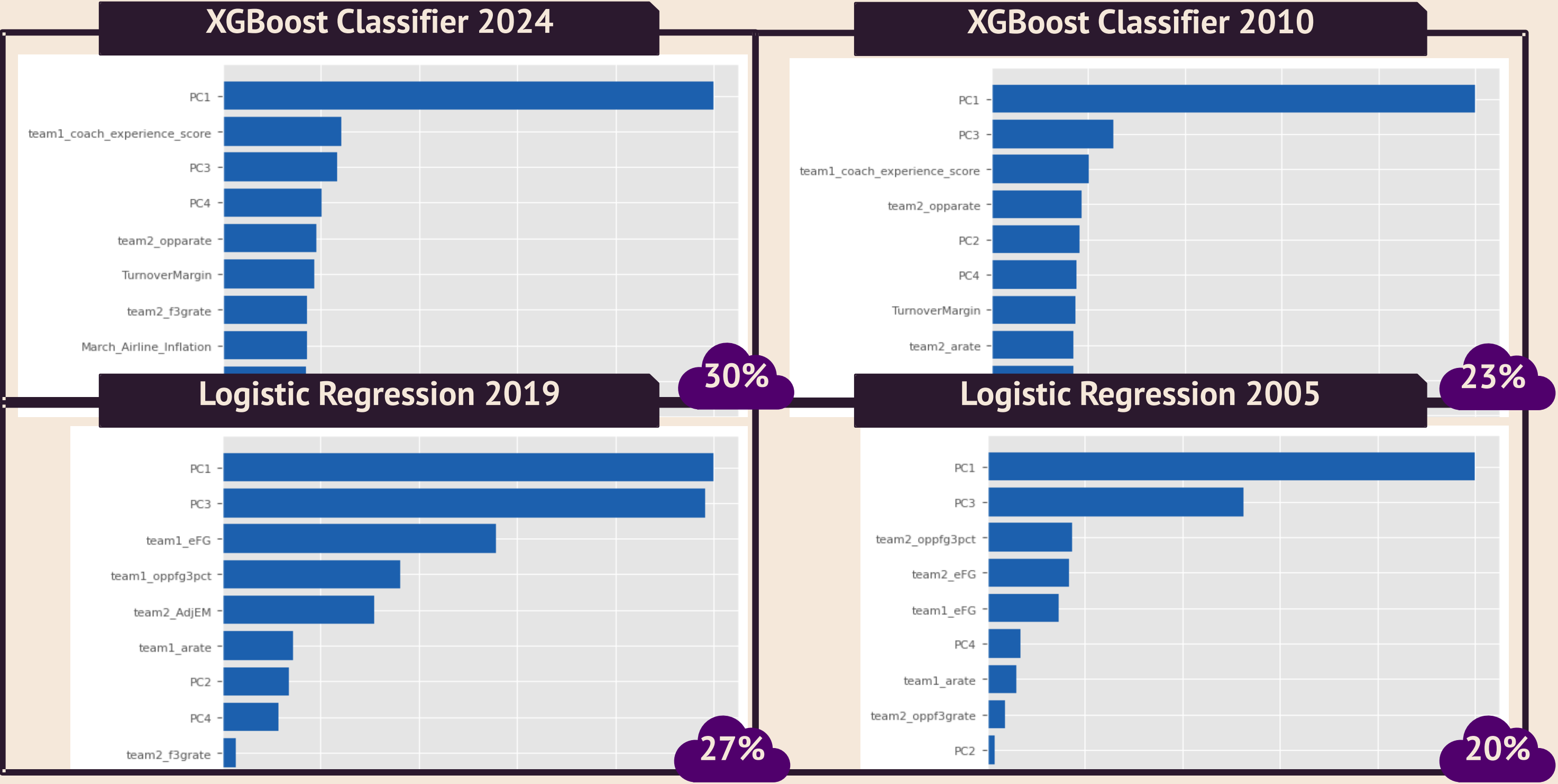
  
**Balanced Team  
Seeding Approach**


Yearly test data leads  
to a **consistent**  
distribution of team  
seeds

Test Year	Model Type	Random Variables	Accuracy	Precision	Recall	F Measure	AUC	Average
2005	Logistic Regression	'team2_oppf3grate', 'team2_eFG', 'team1_f3grate', 'team1_arate', 'team2_oppfg3pct']	75.0%	65.8%	89.3%	75.8%	82.2%	77.6%
2010	XGBoost Classifier	'team2_eFG', 'team2_oppblockpct', 'team2_opparate', 'team1_de', 'US_Main_Trend_2022_2025', 'team2_arate', 'team1_oppblockpct']	75.0%	76.5%	76.5%	76.5%	75.4%	76.0%
2019	Logistic Regression	'team1_arate', 'team1_oppftpct', 'team2_coach_experience_score', 'team2_f3grate', 'team2_AdjEM', 'team1_oppfg3pct']	74.6%	73.7%	80.0%	76.7%	83.8%	77.8%
2024	XGBoost Classifier	'team2_coach_experience_score', 'team2_f3grate', 'team1_oppf3grate', 'team2_opparate', 'team1_blockpct', 'team1_oppblockpct']	74.6%	73.0%	79.4%	76.1%	74.8%	75.6%

+  
**Fixed Variables:**  
['PC1', 'PC2', 'PC3', 'PC4', 'March\_Airline\_Inflation',  
'team1\_coach\_experience\_score', 'team1\_eFG', 'TurnoverMargin',  
'team1\_FTR', 'diff\_dist', 'team2\_coach\_experience\_score', 'team2\_f3grate',  
'team1\_oppf3grate', 'team2\_opparate', 'team1\_blockpct', 'team1\_oppblockpct']

# Feature Importance



  
Model Importance  
for  
Final Prediction

# NLP Deep Dive



Merged buzz scores into 2025 dataset

Weighted at 15% in the final predictive model