

Stable Signals from Noisy Resumes: Regularized NCAA Selection Models Under Feature Collinearity



Dinkins, A.; Kitson, C.; Mason, C.; Nickels, B.; Thompson, R.; Lanham, M.A.
Butler University Lacy School of Business
ajdinkins@butler.edu; banickels@butler.edu; cmason@butler.edu; rmthompson@butler.edu; ckiston@butler.edu



ABSTRACT

NCAA tournament resumes are built from performance metrics like NET rankings, strength of schedule, highly correlated quadrant records, inflating coefficients, and destabilizing standard linear regression. This research investigates whether penalized classifiers produce more reliable predictions under these conditions. Elastic-net linear regression models are compared against unpenalized linear benchmarks using leave-one-season-out cross-validation. The goal is to identify which variables remain persistently informative, not merely opportunistically significant, yielding a more defensible model for predicting bubble-team selection outcomes.

BUSINESS PROBLEM FRAMING

- NCAA tournament selection models often rely on highly correlated performance variables, which can inflate coefficients and make predictions unstable. This project asks which feature remains reliable under multicollinearity and whether regularized models like Elastic Net produce more stable and generalizable predictions than standard linear regression.
- This work is important because better predictions help analysts, media, and other stakeholders understand team selection more clearly. **Reducing unstable or misleading signals leads to more trustworthy insights** and less overfitting to noisy data.
- The main stakeholders include sports analysts, data scientists, media members, the NCAA selection committee, and fans or betting markets that rely on these predictions. Each group benefits from more accurate and consistent results.
- The project faces challenges such as limited historical data, many correlated features, and subjective human decisions. To address this, we use leave-one-season-out validation to test model stability.
- The **goal is to build more stable models**, identify truly important variables, and improve prediction accuracy. This will create more reliable insights and **better decision-making tools** for NCAA tournament predictions.

DATA

Notable Data Relationship

- Conference Finish + Winning percentage: Teams that finish higher in their conference typically have better records
- This creates another layer of **redundancy in measuring success**

Business Problem Framing Reflection

- By reducing the influence of unstable or misleading signals, the analysis produces more trustworthy insights and minimizes overfitting to noisy or redundant data.

Variable	Description	Type	Category
RecordID	Unique identifier for each team-season observation	Categorical (ID)	Contextual
Season	Year of the NCAA season	Categorical	Contextual
Team	Team name	Categorical	Contextual
Conference	Conference affiliation	Categorical	Contextual
Overall_Seed	NCAA tournament seed assigned to team	Numeric	Outcome
Bid_Type	NET Rank Evaluation Tool ranking	Numeric	Outcome
NET_Rank	NCAA Evaluation Tool ranking from presos	Numeric	Numeric
AvgOppNETRank	AvgOppNET Rank Average NET rank of opp.	Numeric (derived)	Numeric (derived)

Table 1. Data Dictionary

METHODOLOGY

- Defined business objectives followed by preprocessing of 10+ performance metrics. We employed **Elastic Net modeling** to isolate key selection drivers and validated the results against historical test data. This pipeline ensures a transparent, fair selection process.
- Developed in **Google Colab** using pandas, matplotlib, and seaborn for data manipulation and visualization.

PROCESS FOR DETERMINING NCAA AT-LARGE FAIRNESS.

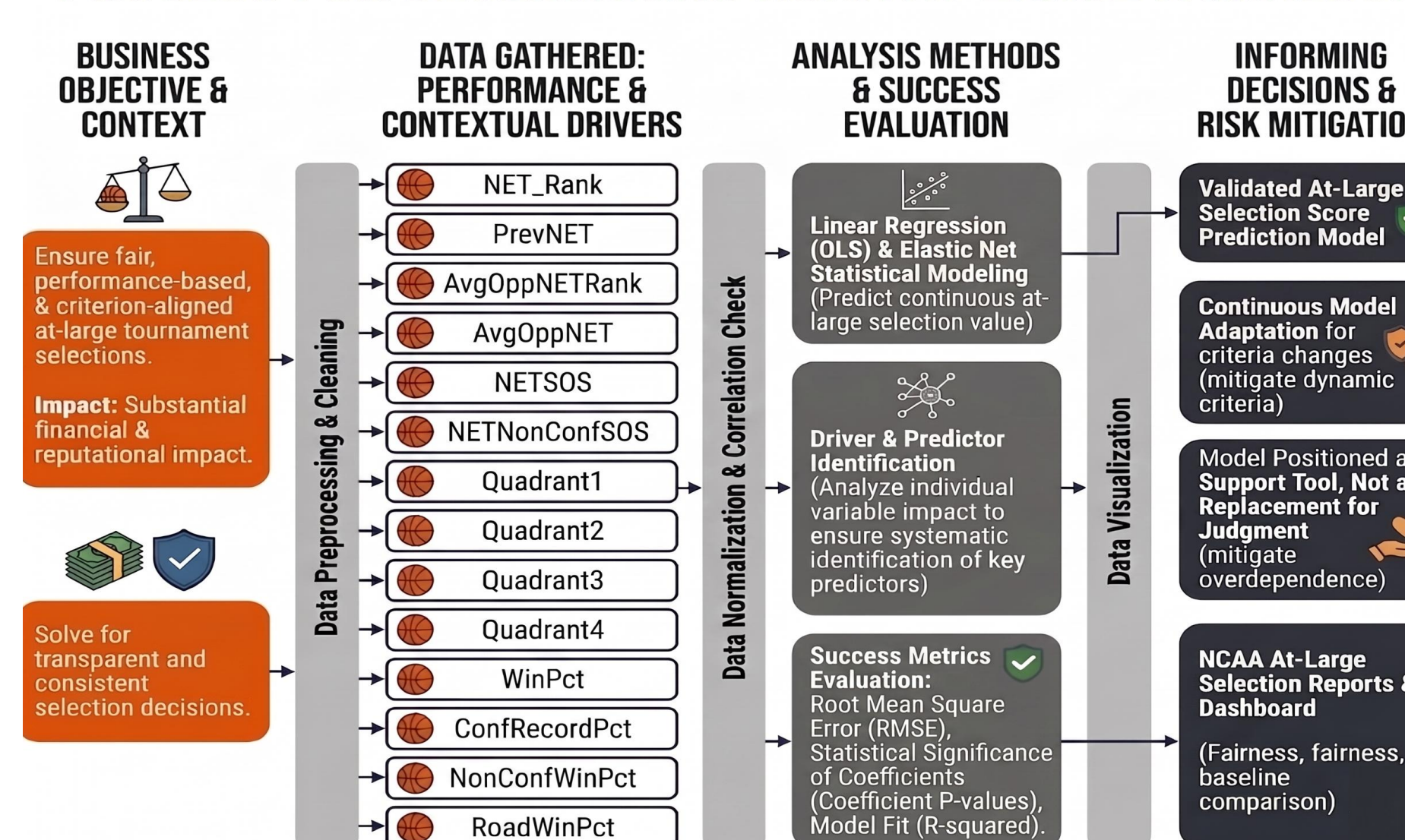


Fig 2. Methodology

MODEL BUILDING & EXPERIMENTAL RESULTS

The Linear Regression model is dominated by a single feature (NET_Rank) with an absolute coefficient value exceeding 70, indicating that the model places the vast majority of its predictive weight on this one variable. The remaining top features — AvgOppNET, WinPct, ConfRecordPct, and PrevNET — contribute comparatively little, suggesting the model may be highly sensitive to NET_Rank and potentially unstable if that feature is noisy or missing.

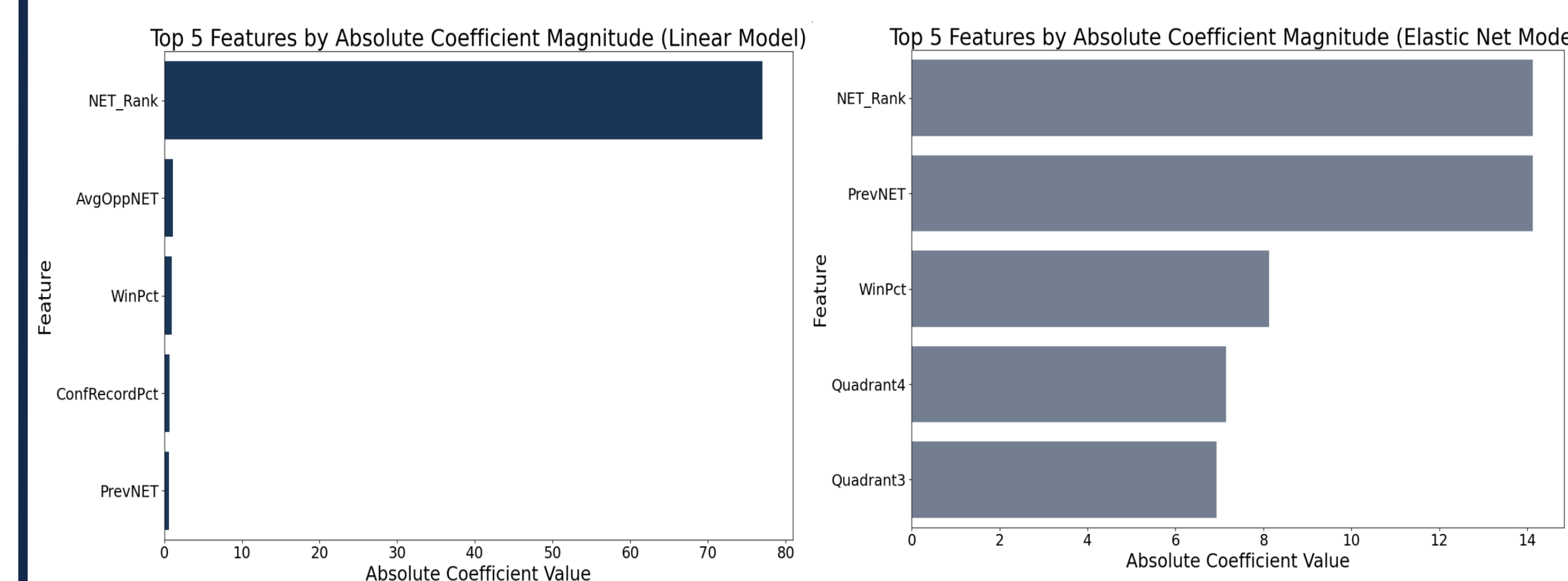


Fig 3. Estimate Parameter Coefficient Effects

- The Elastic Net Model was then tested against the data using leave-one out. It was extremely accurate in its predictions with an R-squared of 0.973. Some of the key factors driving the predictions of the Elastic Net Model include Net Rank (and Prev_Net), Win Percentage, and Quadrant wins.

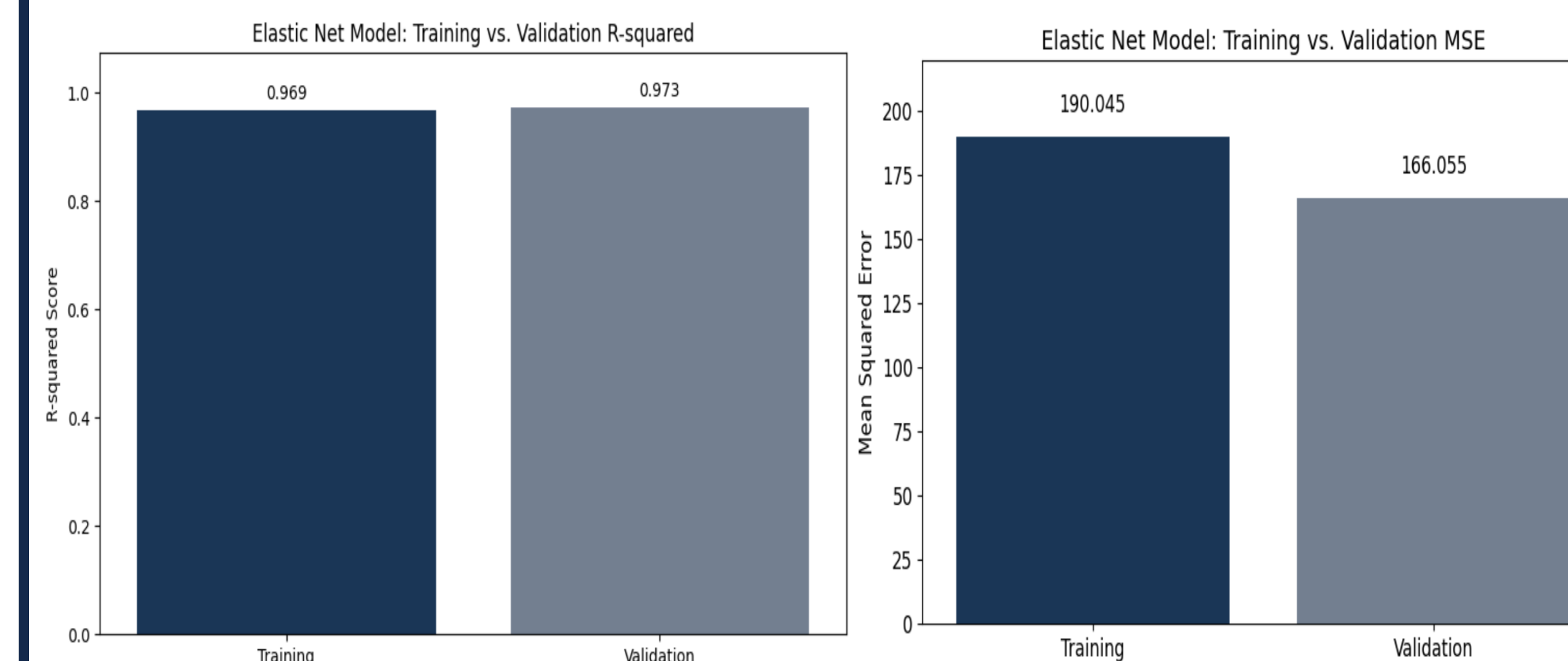
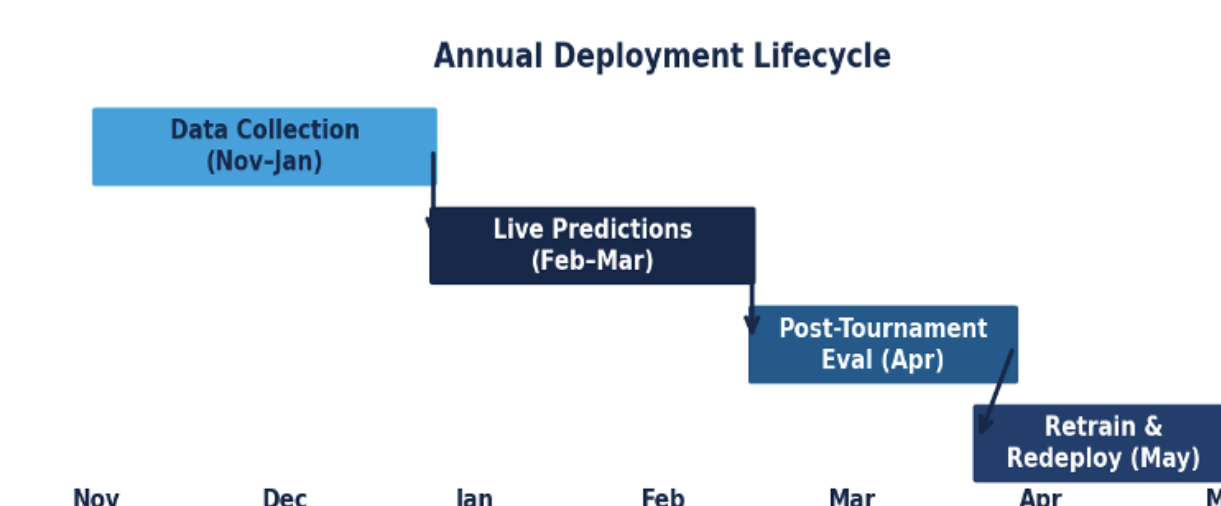


Fig 4. Experimental Results

DEPLOYMENT & LIFECYCLE MANAGEMENT

- Business validation was performed by testing predictions against actual Selection Sunday outcomes (2019–2024). The elastic-net model identified at-large bids with **greater consistency and stability** than unpenalized linear models, demonstrating **improved real-world decision support** for analysts and media.
- In practice, the model would run weekly from February through Selection Sunday, ingesting updated NET rankings and quadrant records. Outputs would power **bubble-team dashboards and bracket projections**, enabling real-time evaluation of at-large candidates.
- The model would be retrained **annually after each season** to incorporate new data and evolving selection trends. Additional recalibration would be required if **selection criteria or key metrics change**, ensuring continued accuracy and stability over time.



KEY TAKE-AWAYS

- Elastic Net models produced more stable and consistent predictions** across seasons, effectively handling multicollinearity among team performance metrics.
- Unpenalized linear models were less reliable**, showing greater variability in coefficient estimates and reduced stability in predictions.
- The modeling framework demonstrates how **penalized approaches improve decision-support in complex, correlated data environments**, such as NCAA tournament selection
- We competed in the Final Four Analytics Challenge, competing against a large field of undergraduate teams.

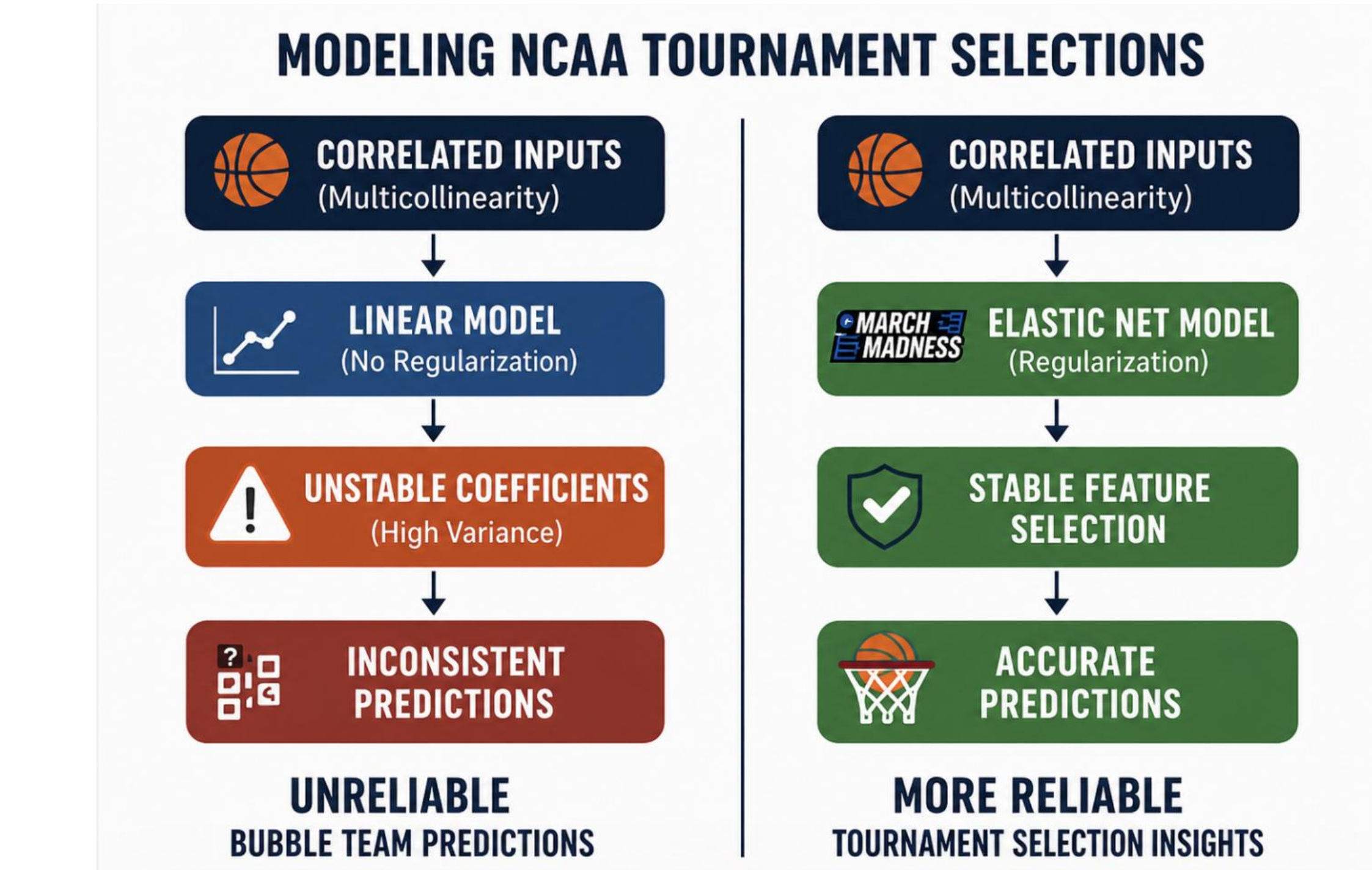


Fig 1. Business to Analytics Problem Challenge

ANALYTICS PROBLEM FRAMING

- Research Question**
 - Do penalized classifiers provide more stable and accurate predictions than unpenalized linear models when predictors are highly correlated?
- Clear Assumptions**
 - Historical selection decisions reflect underlying patterns that can be modeled
 - Input features contain multicollinearity
 - Regularization can improve stability and generalization
 - Past seasons are representative of future selection behavior
- Success Metrics**
 - R-Squared and MSE
 - Comparison of Absolute Coefficients
 - Model stability metrics
- Justification**
 - Elastic-net is chosen because it handles multicollinearity by shrinking and selecting correlated variables. Comparing it to unpenalized linear models allows evaluation of whether regularization improves both predictive performance and stability across seasons.

Personal Development & Outcomes

- Earned 6 new Datacamp certifications this year, significantly strengthening applied skills in Python, AI, and machine learning workflows.
- Enhanced ability to translate AI concepts into practical, real-world applications, improving data analysis and modeling effectiveness.
- Learned and applied the INFORMS Certified Analytics Professional (CAP) Framework, reinforcing structured, end-to-end analytics problem solving.
- Prepared for SAS Certified Specialist: Machine Learning Using SAS Viya, demonstrating continued commitment to technical and analytical development.

