DATA 6560 - Sports Analytics - Checkpoint 7 Fall 2025 # Analysis Memo #2

Checkpoint 7

Overview.....1
Group Work Option.....1
Assignment Objectives.....1
What to Include in Your Checkpoint.....2
Model and Analysis Standards.....3
Submission Details.....3
Deliverables.....3
Grading & Rubric (10 points total).....3
Additional Notes.....4

Overview

Checkpoint 6 delivered a first pass analysis with a clear result.

Checkpoint 7 asks you to improve that work in a focused way. You will refine your method, add one or two useful features, compare against your CP6 baseline, and check that your result is stable and decision ready.

Group Work Option

You may work in groups of up to 3 students, but each student must maintain their own GitHub repository. Collaboration on topic and data collection is encouraged, but each student must submit an individual memo and ensure their repository contains all deliverables.

Assignment Objectives

By completing this checkpoint, you will:

1. Strengthen your Checkpoint 6 approach with a small number of targeted upgrades
2. Compare the upgraded approach to your Checkpoint 6 baseline with clear numbers
3. Check stability and reasonableness so a reader can trust the result
4. State what this means for the decision maker and what you will test next

DATA 6560 - Sports Analytics - Checkpoint 7 Fall 2025 ## What to Include in Your Checkpoint

Quick Reminder	Two or three sentences that restate the project question and the outcome you are modeling or comparing	What are you trying to learn or predict, and why does it matter now?

Data Used	Name the same clean file from CP4, your row definition, the time span, and any filters used for this run	Did you change the time window or filters, and why?
CP6 Baseline	One short paragraph that summarizes the CP6 method and its headline number	What method did you use in CP6 and what was its key performance number such as MAE, accuracy, group difference?
Upgrades in CP7	Describe one or two focused changes only. Examples include adding a key feature, using a rate instead of a raw count, adding interaction of two important inputs, or trying a closely related method that fits your question	Why are these changes the most promising? How do they connect to what you saw in CP5 EDA?
Model or Analysis Spec	A small spec table with Outcome, Inputs, Sample, Row definition, Formula or rule, and Expected direction for the new inputs	Which new inputs did you add and why? What direction or sign did you expect and why?
Model or Analysis Spec	A clear table that shows the CP6 baseline next to the CP7 upgrade. Include one or two figures only if they help	Did performance improve by a meaningful amount? How large is the improvement in real units that a coach or manager would care about?
Results and Comparison	Short, practical checks that build trust. Some ideas: season split, home and away split, position groups, easy holdout period, simple k-fold, residual summary, calibration plot for classifiers, confusion matrix and threshold choice tied to the decision	Does the upgrade still help across subgroups or time windows? Are errors reasonable and free of obvious leakage? If classification, how does performance change at a useful threshold, and why is that threshold useful?
Interpretation in Plain English	Two to four short bullets that say what changed and what it means in the real world	If input A increases, what happens to the outcome and by about how much? Who can use this and how will they act on it?
Limits	Name the top one or two limits that still worry you	What data are missing or thin? What bias might remain? What parts of the sample look unstable?
Next Steps for CP8	Two or three concrete steps you will take next week	What will you keep, what will you drop, and what will you test next so you can write a clean executive summary in CP8?

- Keep the upgrade focused. Improve one or two things that matter most, not a long list
- Always report the CP6 baseline next to your CP7 result in the same units
- Report at least one evaluation method, some examples below:
    - Regression: MAE or R^2, plus a short note on typical error in sport terms
    - Classification: accuracy plus a simple baseline rate, also show a confusion matrix at one useful threshold, include a short note on false positives and false negatives if that matters for the decision
    - Comparisons: group means and a clear effect size with a simple interval if possible ● Add one short stability check. Examples include last season only, top 50 percent of minutes only, home and away, early season versus late season
- Keep figures clean and labeled. Put a one line takeaway under each figure

Submission Details

Document Type	Google Doc in professional memo style, export to PDF
Submission	Share the Google Doc link on Classroom and upload the PDF to your repo in /reports. Save any figures in /figures
File Naming	LastName_FirstName_CP7.pdf (example: Doe_Jane_CP7.pdf). Include your name(s), project title, and the checkpoint heading at the top of the memo. EACH STUDENT MUST UPLOAD TO INDIVIDUAL REPO
Format Tips	Upload your updated workbook in Excel or your script if you coded. Include a small comparison table that shows CP6 baseline next to CP7 upgrade with the exact sample used

Deliverables

1. Analysis Memo 2 in PDF
2. Updated workbook or script, plus figures and the side by side comparison table

Grading & Rubric (10 points total)

Your Checkpoint 7 will be evaluated on the following criteria:

Focused Upgrade	One or two changes that are well motivated by CP5 and CP6, not a kitchen sink	3
Clear Comparison	Baseline from CP6 and upgraded CP7 results shown side by side in the same units with a short takeaway	3
Stability and Checks	At least one stability or subgroup check, plus simple error or calibration checks that match the method	2

DATA 6560 - Sports Analytics - Checkpoint 7 Fall 2025

Interpretation and Limits	Plain language on what it means and honest limits that matter for decisions	1
Professionalism	Clean memo, correct file names and locations, labeled figures with one line takeaways	1

Additional Notes

- *Choose the smallest change that could move the needle and prove it helps*
- *Tie every improvement to a plot or pattern you saw in CP5*
- *Frame results in units a coach or manager will recognize*
- *If the upgrade does not help, say so and explain what you learned. A clear negative result is still progress*

File: NBA_2017_2018_cleaned.csv (CP4 cleaned)

Defining rows: One row = one player's 2017–18 regular season

Time frame: 2017–2018 NBA regular season

Filters (this run): Removed players with less than 500 minutes to reduce small-sample noise; included all 30 teams

Changes from previous versions: Same window and filters so the comparison is clean.

We fit a simple regression: Win Shares ~ Points/G + Rebounds/G + Assists/G + Steals/G + Blocks/G. Headline numbers: $R^2 \approx 0.65$ and MAE ≈ 1.8 WS. In words, the model explains about two-thirds of win contribution, with typical error around 2 wins per player season.

Add efficiency and role context: true shooting %, and interaction of usage rate with TS% to capture if high usage remains efficient.

Use rate stats: Swap raw rebounds/assists for per-minute (or per-possession) rates to reduce bias from minutes totals. These changes directly address what CP5 EDA showed: efficiency separates MVPs from volume scorers, and per-minute production improves comparability across different workloads.

Outcome

Win Shares (WS)
DATA 6560 - Sports Analytics - Checkpoint 7 Fall 2025 Inputs (CP6)

PTS/G, REB/G, AST/G, STL/G, BLK/G

Inputs (CP7)

PTS/G, TS%, Usage, STL/G, BLK/G, REB/Min, AST/Min, Usage×TS%

Sample

Qualified players (≥500 minutes), 2017–18

Row definition

One player season

Formula or rule

WS ~ PTS + TS% + USG + STL + BLK + REB/Min + AST/Min + USG×TS%

Expected directions

Positive for all; interaction USG×TS% positive if high usage stays efficient

Metric

CP6 Baseline

CP7 Upgrade

R^2

0.65

0.70

MAE (WS)

1.8
DATA 6560 - Sports Analytics - Checkpoint 7 Fall 2025 1.6

meaning coach/GM

Error ≈ 2 wins

Error ≈ 1.6 wins (about half a win tighter)

Is it meaningful? Yes, a ~0.2 WS improvement in error per player season translates to tighter evaluations worth about half a win, compounding across rotation spots.

Results and comparison

Season split: The CP7 model remains stronger for both early-season and late-season subsets.

Position Split: Gains for Guards and Bigs, Efficiency Across Roles:

Hold-out check: A simple 80/20 split shows that CP7 outperforms CP6 (lower MAE) without signs of leakage.

Residuals: Reasonable spread; fewer high-usage outliers under- or over-estimated with TS% and the Usage×TS% interaction included.

Efficiency adds wins: +5 TS% points is worth roughly +0.8 WS (about a win over the season).

Usage only works if it stays efficient. High volume doesn't add much without TS% lift; with high TS%, it adds ~1–2 WS.

Defense travels: +1 steal per game ≈ +0.6 WS; +1 block per game ≈ +0.4 WS.

Who uses this: Coaches and GMs favor lineups and contracts for players who can couple volume with efficiency and actual defensive playmaking.

Limits

Defense under-measured: Steals/blocks miss rotations, contests and matchup difficulty.

Single season: No playoff pressure or multi-year stability; might still be biased towards high-usage roles. Next Steps for CP8 Add tracking/context data: Contested shot rates, on/off splits, and lineup context to better capture defense and teammate effects. Robustness by Role and Threshold: Provide separated guard/big models and choose decision thresholds tied to contract tiers. Out-of-sample validation: Crossvalidate across multiple seasons to finalize a clean executive summary with stable,

coach-ready effect sizes.