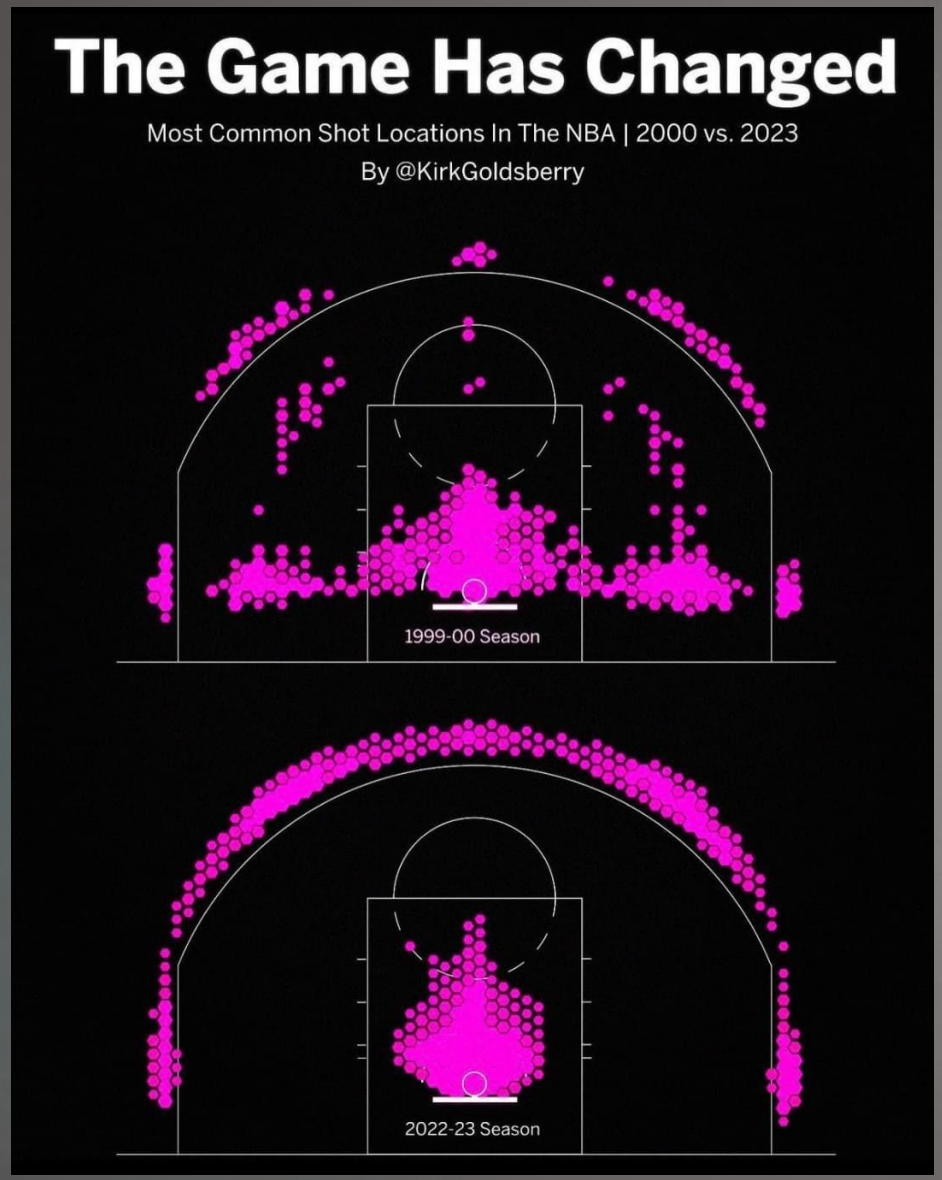


NBA Shot Efficiency

CONNOR, OM, TITO, & PURVA

Goals & Motivation

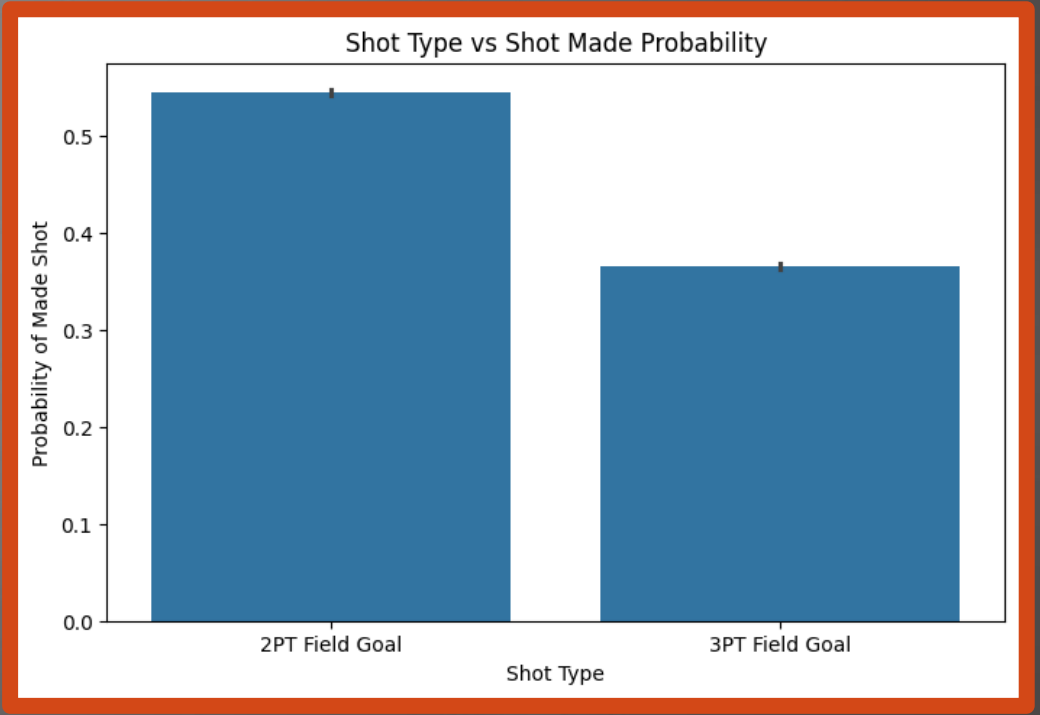
- Goals:
  - Research Goal:
    - Create a predictive model that predicts an NBA shot is made (SHOT_MADE) using five key shot features for new datasets.
  - Secondary Goals:
    - See the joint effects on how each variable is affected. (E.g. how distance & shot type changes
    - Ideally our model provides recommendations for NBA analysts to design or train players to increase point potential.
- Why it matters:
  - Helps coaches & analysts improve shot selection and game strategy
- Useful for:
  - NBA coaches designing plays, players training for higher-percentage shots

Dataset Dissection / Cleaning:

- **Variables**
  - **Response:** SHOT_MADE (0 = missed, 1 = made)
  - **Predictors:** SHOT_TYPE, ZONE_ABB, SHOT_DISTANCE, QUARTER, POSITION
- **Initial Dataset:** Over 218,000 rows
- **Dropped:** 1,264 rows missing player position info.
- **Data Cleaning**
  - No rare categories in Zone_ABB
  - Kept valid outliers in SHOT_DISTANCE (e.g. half-court shots)
- **Outliers Dropped**
  - **Shots made after the 4th Quarter**
  - **Hybrid positioning roles (few instances) – e.g. SG-PG**
- **Final Dataset:** 216261 rows

Preliminary Analysis – Strong Predictors

- Image 1: SHOT_DISTANCE
  - Made shots – Q2 = ~ 5 ft., while **missed shots** – Q2 = ~ 18–20 ft. (Reference: NBA 3PT ~23 ft)
  - More shots are made within 5ft from the basket
- Image 2: SHOT_TYPE
  - 2pt – more accurate (~55%) than 3pt (~36%).

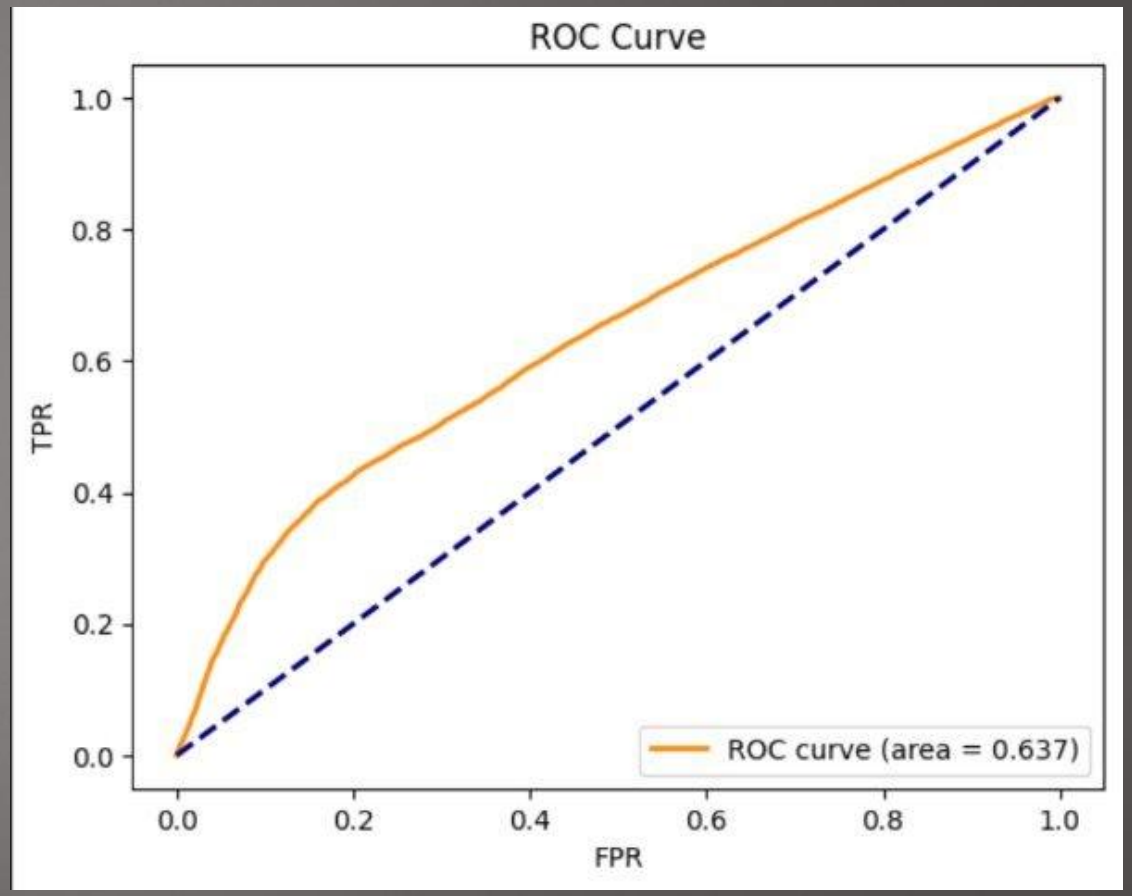


Model Data Preprocessing

- Pulled our relevant response variable and selected explanatory variables from the cleaned dataset
- We extracted our response variable, creating a target array and a feature matrix
- Non-numeric variables were converted into dummy variables
- Numeric variables in the features matrix were standardized to ensure all features were evenly weighted.
- The final shape of the matrix was 210,584 rows and 12 columns

k-Fold Cross-Validation Technique:

- Model Selected: Elastic Net
  - combines both L1 and L2 penalties, allowing for coefficient shrinkage and sparsity
  - LASSO has potential to drop significant yet correlated dummy variables
  - Ridge models may retain uninformative features
  - Our best model regularization strength of lambda = 0.001 and showed a mean AUC of 0.638 during 5-fold cross-validation



Our Best Model:

- Elastic Net Logistic Model was the best model since it includes both penalty terms
- Test AUC: 0.638
- Top Predictor: Shot Distance
- Other Helpful Predictors: Shot Type and Zone
- TPR: 80.1 %, of made shots were correctly predicted
- FPR: 68.7%, of missed shots were incorrectly predicted as made

Best Model Equation:

- **Made Shot** = **-0.1078** - **0.7768**(SHOT_DISTANCE) - **0.0107**(QUARTER) + **0.3081**(SHOT_TYPE_3PT Field Goal) + **0.5576**(ZONE_ABB_C) + **0.3196**(ZONE_ABB_L) + **0.4161**(ZONE_ABB_LC) + **0.3234**(ZONE_ABB_R) + **0.4025**(ZONE_ABB_RC) - **0.0308** (POSITION_PF) - **0.0383**(POSITION_PG) - **0.0516**(POSITION_SF) - **0.0434**(POSITION_SG)

Conclusion

- Our best model had an AUC score of approximately 0.638
  - Moderate predictive power
- People who can gain from results
  - Basketball coach
  - Data analyst
  - Basketball scouts
- Shortcomings
  - Possible confounders (ie: player fatigue or defender distance)
  - Future improvement: player or position-specific models