**Final Report: NBA Jersey Sales Analysis and Prediction**

**1. Introduction**

This project explores the factors that influence NBA jersey sales rankings. Our primary objective was to analyze what drives a player's popularity in terms of jersey sales using both statistical testing and predictive machine learning models. The dataset spanned five seasons and integrated various metrics, including performance statistics, player popularity, and physical attributes.

**2. Data and Preprocessing**

**Data Sources**

* **Hoopshype**: Top 8 jersey-ranked players per season (used as the target variable)
* **Basketball Reference**: Player performance data such as points per game (PTS), minutes played (MPG), games played (GP), and award counts
* **Google Trends**: Weekly search interest per player, averaged over each season
* **Kaggle Dataset**: Physical attributes like height and weight, used to compute BMI

After merging all sources, the final dataset included:

* **Target variable**: Jersey Rank (1 to 8)
* **Features**: Points/Game, Games Played, Minutes/Game, Award Count, Google Trends Score, Height (length), BMI

All variables were cleaned and joined on the basis of player\_name and season. The dataset was enriched with both continuous and categorical columns for downstream analysis.

**3. Hypothesis Testing**

Using Pearson and Spearman correlation tests, we examined the relationships between each metric and jersey rank.

**Key Findings**

| **Metric** | **Pearson r** | **p-value** | **Interpretation** |
| --- | --- | --- | --- |
| Google Trends Score | –0.374 | 0.017 | Moderate inverse correlation, statistically significant |
| BMI | –0.440 | 0.009 | Moderate inverse correlation, statistically significant |
| PTS/Game, GP, MPG | < ±0.15 | > 0.35 | Weak or no correlation |

Only Google Trends and BMI showed significant inverse correlation with jersey rank. Performance metrics such as points per game or games played were surprisingly weak predictors.

**4. Unsupervised Learning**

We employed the following unsupervised methods:

* **Principal Component Analysis (PCA)**: Helped reduce feature dimensionality and revealed clustering tendencies across metrics.
* **K-Means and Hierarchical Clustering**: Did not show strong grouping according to jersey rank, but identified clusters based on physical and popularity attributes.

These insights guided our feature selection process for the machine learning stage.

**5. Supervised Learning**

**5.1 Classification: Jersey Sales Tier Prediction**

We converted jersey rank into three tiers:

* **High (1–3)**
* **Mid (4–6)**
* **Low (7–8)**

**Models Tested**

* Logistic Regression
* Random Forest
* k-Nearest Neighbors (k-NN)
* Support Vector Machine (SVM)
* Gaussian Naïve Bayes
* AdaBoost

**Evaluation**

* Metrics: Accuracy, Macro-F1, Weighted-F1
* Best model: **Random Forest**
  + Test Accuracy: ~56%
  + Cross-Validation Accuracy: ~50% ± 0.32
* Feature importances indicated that BMI, Google Trends, and Points/Game were the strongest predictors of jersey tier.

**5.2 Regression: Continuous Jersey Rank Prediction**

We then treated jersey rank as a continuous variable and implemented regression models.

**Models Used**

* **Linear Regression**
* **Gradient Boosting Regressor**

**Experiment: Varying Feature Sets**

To improve model performance, we tested three different combinations of features:

| **Feature Set** | **MAE** | **RMSE** | **R²** |
| --- | --- | --- | --- |
| All 7 features | 2.42 | 2.76 | –0.33 |
| Google Trends, BMI, Points/Game | 1.69 | 2.03 | 0.13 |
| Google Trends and BMI (**Best Result**) | 1.48 | 1.85 | 0.18 |

**Motivation for Multiple Tests**

Our full-feature model yielded a **negative R²**, indicating that it performed worse than simply predicting the mean jersey rank. This highlighted the importance of reducing feature noise. By limiting the model to high-correlation features, we achieved a **positive R² of 0.18**, improving model validity and interpretability.

**Example Prediction**

A sample player profile with:

* 24.5 PTS/Game, 72 Games Played, 34.1 MPG, 1 award
* Google Trends Score: 18.2, Height: 200 cm, BMI: 24.5  
  was predicted to have a **jersey rank of ~2.94** using the most optimized model.

**6. Project Coverage**

This project includes the following components:

* Data Enrichment and Preprocessing
* Hypothesis Testing (Pearson and Spearman)
* Exploratory Data Analysis (EDA)
* PCA and Clustering for Unsupervised Learning
* Classification and Regression for Supervised Learning
* Model Evaluation with Visualization and Interpretation

**7. Conclusion**

This project demonstrates that:

* Performance metrics alone are **not sufficient** to predict jersey popularity.
* **Public visibility** (via search interest) and **physical branding** (BMI) have **greater predictive power**.
* Gradient Boosting, combined with selective feature use, provides the most reliable predictions for jersey ranks.
* A multi-phase regression approach—starting with full features and progressively trimming them—shows how **feature selection enhances model quality**.