Business Problem, Insights, and Methodology Report for Mobile Games Sales Prediction

1. Business Problem Description

1.1 Context

The mobile gaming industry is a highly competitive and lucrative market, with thousands of games launched annually across various platforms. Companies, including game developers and publishers, face the challenge of accurately predicting a game's total sales (sales_total) to optimize resource allocation, marketing strategies, and development efforts. The dataset, comprising 16,598 mobile game records, includes features such as title_name, device_type, launch_year, game_genre, publisher_name, and regional sales (sales_usa, sales_europe, sales_asia, sales_misc). The primary business problem is to develop a predictive model that estimates total sales based on these features, enabling stakeholders to make data-driven decisions.

1.2 Problem Statement

The objective is to build a robust machine learning model to predict sales_total for mobile games with high accuracy, targeting an R^2 score of at least 0.80 (explaining 80% of the variance in sales). The current RandomForestRegressor model achieves an R^2 of 0.70, indicating a gap in predictive power. Accurate predictions will help:

- **Publishers**: Prioritize marketing budgets for high-potential games.
- **Developers**: Identify successful game genres or platforms.
- **Investors**: Assess the viability of funding specific game projects. Key challenges include handling skewed sales data, missing values (271 in launch_year, 58 in publisher_name), high-cardinality categorical features (e.g., 11,493 unique title_name values), and capturing non-linear relationships.

1.3 Business Objectives

- Achieve an R^2 score of 0.80 or higher on the test set to ensure reliable predictions.
- Identify key features driving sales to provide actionable insights for stakeholders.
- Develop a scalable model pipeline that can handle new data for future predictions.
- Minimize prediction errors (e.g., RMSE) to support confident decision-making.

2. Insights from Data Exploration

2.1 Dataset Overview

- **Size**: 16,598 entries, 10 columns.
- **Numerical Features**: launch_year, sales_usa, sales_europe, sales_asia, sales_misc, sales_total.
    - Skewed distributions: sales_total mean: 0.5374, max: 82.74; regional sales show similar skewness (e.g., sales_usa max: 41.49, 75th percentile: 0.24).
    - Missing values: 271 in launch_year.
- **Categorical Features**: title_name (11,493 unique), device_type (31 unique), game_genre (12 unique), publisher_name (578 unique).
    - Dominant categories: game_genre (Action, 3,316 entries), device_type (DS, 2,163 entries), publisher_name (Electronic Arts, 1,351 entries).
    - Missing values: 58 in publisher_name.
- **Target Variable**: sales_total, highly skewed, suggesting the need for transformation or robust modeling techniques.

2.2 Key Insights

- **Skewness and Outliers**: Sales columns exhibit extreme outliers (e.g., top games like Wii Sports with sales_total of 82.74), which may distort model predictions unless addressed through transformations (e.g., log) or robust scaling.
- **Feature Correlations**: Regional sales (sales_usa, sales_europe) likely have strong correlations with sales_total, as they are components of the target. A correlation heatmap would confirm this.
- **Categorical Impact**: High-cardinality features like publisher_name and title_name suggest potential overfitting if encoded directly. game_genre and device_type may capture market trends (e.g., Action games or DS platform dominance).
- **Temporal Trends**: launch_year (mean