Mobile Games Sales Prediction Model Report

1 Overview

This report details the development and evaluation of machine learning models to predict video game sales based on historical data. The sections include:

1. Introduction
2. Model Selection
3. Why These Models Were Selected
4. Model Parameters
5. Performance of Models
6. Selection of Best Model

2 Introduction

The primary goal of this project is to develop a predictive model for future video game sales using historical data. This model aims to assist game developers in creating better games, understanding current market trends, and devising effective promotional strategies.

3 Model Selection

Model selection is a critical step in the machine learning pipeline, as it dictates theapproachtohandlingthedataset'scharacteristics. AfterperformingExploratory Data Analysis (EDA), we identified key challenges: the dataset contains both numerical(e.g., `launch_year`, `sales_usa`, `sales_europe`, `sales_asia`, `sales_misc`) and categorical features (e.g., `device_type`, `game_genre`, `publisher_name`), and the target variable `sales_total` exhibits significant right skewness (mean: 0.537, max: 82.74). Additionally, the dataset includes potential non-linear relationships and outliers.

To address these challenges, the following models were selected:

- RandomForestRegressor
- GradientBoostingRegressor
- XGBRegressor
- LGBMRegressor
- VotingRegressor (Ensemble of LightGBM, GradientBoosting, and XGBoost)

4 Why These Models Were Selected

The models were chosen for their ability to handle the dataset's complexities. Below are the reasons for their selection:

1. **Robustness to Mixed Data Types**: The dataset includes both numerical andcategoricalfeatures. Tree-basedmodels(RandomForest, GradientBoosting, XGBoost, LightGBM) effectively handle mixed data types after preprocessing with `OneHotEncoder` for categorical features and `RobustScaler` for numerical features. These models do not assume specific data distributions, unlike linear regression.
2. **Ability to Capture Non-Linear Relationships**: Video game sales likely depend on complex, non-linear interactions (e.g., between `game_genre` and `sales_usa`). Tree-based models excel at capturing such patterns without extensive feature engineering.
3. **RobustnesstoSkewedDataandOutliers**: Thetargetvariable`sales_total` ishighlyskewed, andregionalsalesfeaturescontainoutliers(e.g., `sales_usa` max: 41.49). Tree-based models are less sensitive to skewness and outliers, as they use feature thresholds for splits.
4. **FeatureImportanceCapabilities**: Allmodelssupport`feature_importances_`, enabling identification of key predictors (e.g., `sales_usa`, `game_genre`). Thisenhancesinterpretabilityforstakeholders, supportedbyvisualizations using seaborn bar plots.
5. **EnsembleApproachforImprovedPerformance**: RandomForestusesbagging, while GradientBoosting, XGBoost, and LightGBM use boosting. The VotingRegressor combines the top three models (LightGBM, GradientBoosting, XGBoost) to balance stability and accuracy, reducing variance and bias.
6. **ScalabilityandEfficiency**: XGBoostandLightGBMareoptimizedforspeed, handling the dataset's 16,598 entries efficiently. LightGBM's native categorical feature processing and XGBoost's optimized boosting enhance computational performance.
7. **Hyperparameter Tuning Support**: The use of `RandomizedSearchCV` allowedoptimizationofmodelparameters(e.g., `n_estimators`, `max_depth`), improving accuracy. The ensemble model leverages the strengths of tuned individual models.
8. **Industry Relevance**: Tree-based and ensemble models are widely used in industry for regression tasks like sales prediction due to their accuracy, interpretability, and versatility.

5 Model Parameters

Hyperparameters were tuned using `RandomizedSearchCV` to optimize performance while balancing accuracy and generalization. Below are the key parameters for each model:

5.1 RandomForestRegressor

- `n_estimators`: 200 (number of trees for robust predictions)
- `max_depth`: None (allows full tree growth to capture complex patterns)

- `min_samples_split`: 5 (minimum samples to split a node, balancing flexibility)
- `max_features`: 0.8 (fraction of features per split for diversity)

5.2 GradientBoostingRegressor

- `n_estimators`: 300 (number of boosting stages for accuracy)
- `learning_rate`: 0.1 (step size for updates)
- `max_depth`: 5 (controls tree complexity)
- `subsample`: 1.0 (fraction of samples per tree, reducing variance)

5.3 XGBRegressor

- `n_estimators`: 200 (number of trees for boosting)
- `learning`$_r$*ate* : 0.1(*stepsizeforupdates*)`max_depth` : 5(*depthtocapturepatterns*)
- `subsample`: 1.0 (sample fraction to prevent overfitting)
- `colsample_bytree`: 0.8 (feature fraction for efficiency)

5.4 LGBMRegressor

- `n_estimators`: 300 (number of trees for robust boosting)
- `learning`$_r$*ate* : 0.05(*stepsizeforconvergence*)`max_depth` : 7(*limitstreedepth*)
- `num_leaves`: 31 (leaves per tree, capturing detailed patterns)
- `subsample`: 0.8 (sample fraction to reduce overfitting)
- `colsample_bytree`: 0.8 (feature fraction for speed)

5.5 VotingRegressor (Ensemble)

- Combines`LGBMRegressor`, `GradientBoostingRegressor`, and`XGBRegressor` with tuned parameters as above, using equal weights to leverage their collective strengths.

6 Performance of Models

Models were evaluated using the following metrics:

- **RMSE (Root Mean Squared Error)**: Measures prediction error, emphasizing larger deviations; lower is better.
- **MAE (Mean Absolute Error)**: Average absolute prediction error, robust to outliers.
- **R^2 Score**: Proportion of variance explained; closer to 1 is better.

Cross-validation ensured robust performance estimates.

6.1 RandomForestRegressor

- **Description**: An ensemble of decision trees using bagging, robust to noise and outliers in skewed sales data.
- **Performance**:
  - RMSE: 0.0599
  - MAE: 0.0230
  - R^2: 0.9671

6.2 GradientBoostingRegressor

- **Description**: Sequentially builds trees to correct errors, effective for complex patterns.
- **Performance**:
  - RMSE: 0.0555
  - MAE: 0.0228
  - R^2: 0.9717

6.3 XGBRegressor

- **Description**: Optimized gradient boosting, fast and regularized, suitable for large datasets.
- **Performance**:
  - RMSE: 0.0558
  - MAE: 0.0248
  - R^2: 0.9714

6.4 LGBMRegressor

- **Description**: Fast, histogram-based boosting model, efficient for complex relationships.
- **Performance**:
- **Performance**:
  - RMSE: 0.0539
  - MAE: 0.0208

– R^2: 0.9734

6.5 VotingRegressor (Ensemble)

- **Description**: Combines predictionsfrom LightGBM,GradientBoosting, and XGBoost to leverage their strengths, reducing variance and improving accuracy.
- **Performance**:
  – RMSE: 0.0538
  – MAE: 0.0211
  – R^2: 0.9734

6.6 Sample Predictions

The ensemble model's predictions on a test set sample (first 10 instances) are shown below, with actual values and percentage errors:

• Sample 1: Predicted: 0.116024 | Actual: 0.122218 | Error: 5.1% • Sample 2: Predicted: 0.914119 | Actual: 0.924259 | Error: 1.1% • Sample 3: Predicted: 0.133093 | Actual: 0.122218 | Error: 8.9% • Sample 4: Predicted: 0.858542 | Actual: 0.854415 | Error: 0.5%

• Sample 5: Predicted: 0.160110 | Actual: 0.157004 | Error: 2.0%

• Sample 6: Predicted: 0.037774 | Actual: 0.019803 | Error: 90.8%

• Sample 7: Predicted: 0.799278 | Actual: 0.756122 | Error: 5.7% • Sample 8: Predicted: 0.505743 | Actual: 0.500775 | Error: 1.0%

• Sample 9: Predicted: 0.162748 | Actual: 0.157004 | Error: 3.7%

• Sample 10: Predicted: 0.286073 | Actual: 0.285179 | Error: 0.3% **Prediction Statistics:**

• Min: 0.037774

• Max: 0.914119

• Mean: 0.397350

• Std: 0.324810 **Actual Values Statistics:**

• Min: 0.019803

• Max: 0.924259

• Mean: 0.389900

• Std: 0.341151

7 Selection of Best Model

The **VotingRegressor** (ensemble of LightGBM, GradientBoosting, and XGBoost) was selected as the best model, achieving an RMSE of 0.0538, slightly outperforming the best individual model (LightGBM, RMSE: 0.0539). The ensemble's superior performance is attributed to its ability to combine the strengths of boostingbased models, reducing both bias and variance. The high R^2 score (0.9734) indicates excellent explanatory power, and the low MAE (0.0211) reflects robust average prediction accuracy. Hyperparameter tuning via `RandomizedSearchCV` and the ensemble approach ensured optimal performance, making this model suitable for real-world video game sales prediction.