Playtime Matters: Analyzing Steam Games

DSA210 Project Presentation

Esra Esen
32640

Project Overview

- **Objective:** Investigate how game genres, prices, playtime lengths, and popularity are related on Steam.
- **Datasets:**
  - Steam dataset (Kaggle): Game details (genres, prices, estimated owners).
  - HowLongToBeat (HLTB) dataset: Completion times for games.
- **Goal:** Analyze the influence of playtime on popularity and identify trends in genres and pricing.
- **Hypothesis:**
  - H0: No significant relationship between playtime and popularity.
  - H1: Longer playtime increases popularity.

Data Preparation

- **Merging Datasets:** Combined Steam and HLTB datasets using game names, resulting in `merged_data.csv`.
- **Cleaning:**
  - Removed unnecessary columns (e.g., appid, developer).
  - Renamed `steamspy_tags` to `genres`.
- **Time Formatting:**
  - Converted Steam playtime (minutes to hours).
  - Rounded HLTB time columns for consistency.
- **Handling Missing Values:**
  - Dropped columns with 50% missing data (e.g., `main_story`, `completionist`).
  - Removed rows with missing time values (12,089 to 6,271 rows).
- **Final Dataset:** Saved as `clean_merged_data.csv` with 12 columns (e.g., `name`, `genres`, `positive_ratings`, `average_completion_time`).

Hypothesis Testing

- **Test:** Pearson correlation between `average_completion_time` and `positive_ratings`.
- **Results:**
  - Correlation: 0.165 (weak positive relationship).
  - P-value: 9.86e-40 (0.05, statistically significant).
- **Conclusion:** Rejected H0, accepted H1. Longer playtime slightly increases popularity, but the effect is weak, suggesting other factors may be more influential.

Visualizations: Playtime and Popularity

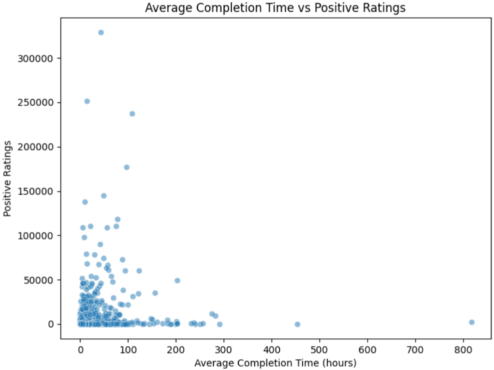


Figure: Scatter Plot: Average Completion Time vs Positive Ratings

- **Scatter Plot (Completion Time vs Positive Ratings):**
  - Shows a weak positive trend (correlation coefficient: 0.165).
  - Most games cluster below 200 hours with ratings under 50,000.

Visualizations: Playtime and Popularity

- **Scatter Plot (Completion Time vs Positive Ratings) - Detailed Analysis:**
  - The weak positive trend (correlation coefficient: 0.165) indicates that longer playtime is associated with slightly higher positive ratings, though the relationship is not strong.
  - Most games cluster below 200 hours with ratings under 50,000, suggesting that the majority of Steam games have moderate playtimes and popularity.
  - Outliers with higher ratings (up to 300,000) for longer playtimes (up to 800 hours) may represent content-rich games (e.g., MMORPGs or strategy titles), appealing to niche audiences.
  - The sparse data points beyond 200 hours imply that very long games are rare, and their popularity might depend on factors like genre or community engagement beyond playtime alone.

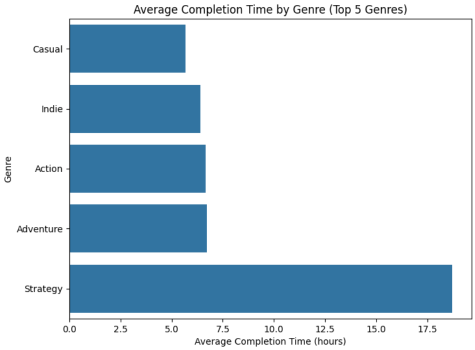


Figure: Bar Chart: Average Completion Time by Genre (Top 5 Genres)

- **Bar Chart (Average Completion Time by Genre):**
  - Strategy: 18.5 hours (longest).
  - Casual: 5 hours (shortest).
  - Action, Indie, Adventure: 5.5–7 hours.

- **Bar Chart (Average Completion Time by Genre) - Detailed Analysis:**
  - Strategy games take 18.5 hours on average, reflecting their complex mechanics and deep strategic elements.
  - Casual games, averaging 5 hours, are designed for quick and accessible play sessions.
  - Action, Indie, and Adventure genres (5.5–7 hours) balance challenge and accessibility, appealing to a broader audience.
  - The significant gap between Strategy and Casual highlights how genre design (depth vs simplicity) influences playtime.
  - This suggests players prefer shorter sessions for casual gaming, while strategy titles cater to longer, more engaged play.

Visualizations: Average Ownership by Genre

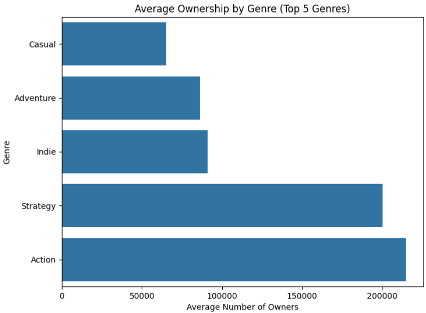


Figure: Bar Chart: Average Ownership by Genre (Top 5 Genres)

- **Bar Chart (Average Ownership by Genre):**
  - (Approximate Numbers)
  - Action: 200,000 owners (most popular).
  - Casual: 50,000 owners (least popular).
  - Strategy, Indie, Adventure: 175,000, 100,000, 75,000 owners.

- **Bar Chart (Average Ownership by Genre) - Detailed Analysis:**
  - Action games lead with 200,000 owners, likely due to their fast-paced gameplay and wide appeal.
  - Casual games, with only 50,000 owners, target a niche market seeking quick entertainment.
  - Strategy, Indie, and Adventure genres (175,000, 100,000, 75,000 owners) show varying market penetration, possibly due to marketing differences.
  - Action games' dominance suggests dynamic gameplay drives ownership, while Indie games' lower numbers may reflect limited visibility.
  - This distribution underscores the importance of genre-specific marketing to boost ownership numbers.

- **Data Preparation:**
  - Loaded `clean_merged_data.csv`.
  - Converted genres to numerical (one-hot encoding).
  - Normalized `owners` to `owners_numeric`.
- **Decision Tree Model:**
  - Train-test split (80%-20%, `random_state=42`).
  - Trained a Decision Tree Regressor (`max_depth=10`).
  - Results: $R^2 = 0.574$, RMSE $= 15{,}416$.

Machine Learning: Decision Tree Results

Feature Importance:

- `owners_numeric`: 0.642.
- `average_completion_time`: 0.195.

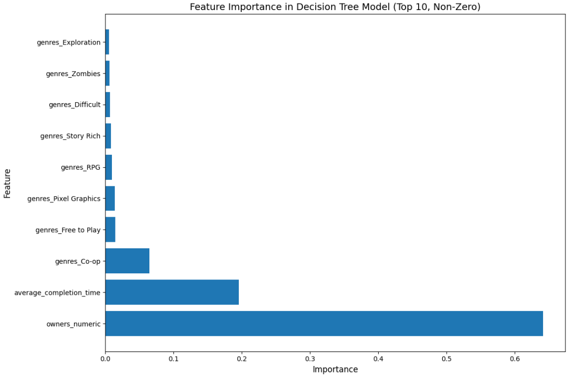
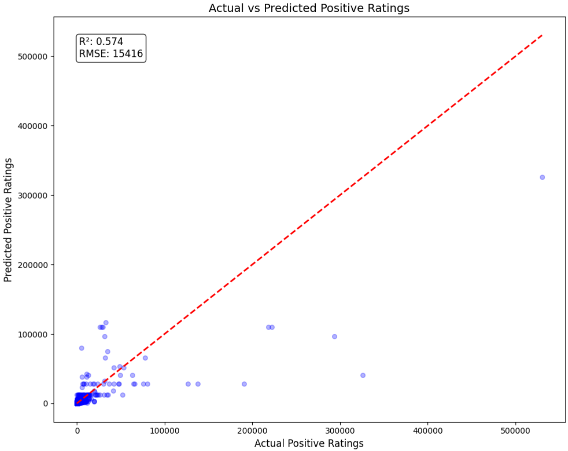


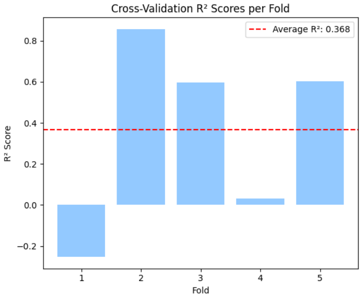
Figure: Feature Importance

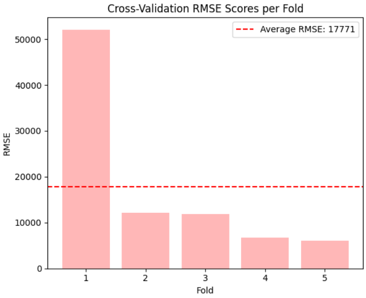
Machine Learning: Decision Tree Results

- **Actual vs Predicted Plot:**
  - Good performance for low ratings, struggles with high ratings ($>$400,000).



- **Cross-Validation (5-Fold):**
  - Average R^2: 0.368, RMSE: 17,770.
  - High variability (std dev R^2: 0.412, RMSE: 17,355).
  - Indicates potential overfitting.





- **Random Forest Model:**
  - Same train-test split as Decision Tree.
  - Trained a Random Forest Regressor (`n_estimators=100`, `max_depth=10`).
  - Results: $R^2 = 0.564$, RMSE = 15,595.
- **Actual vs Predicted Plot:**
  - Similar to Decision Tree, struggles with high ratings.

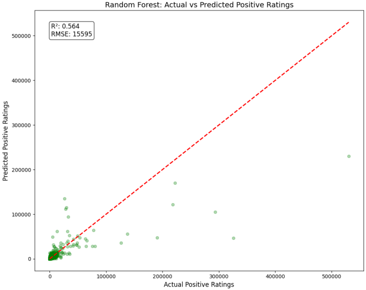


Figure: Random Forest: Actual vs Predicted Positive Ratings

- **Cross-Validation (5-Fold):**
  - Average R^2: 0.479, RMSE: 16,119.
  - Lower variability (std dev R^2: 0.132, RMSE: 11,286).
  - Better generalization than Decision Tree.

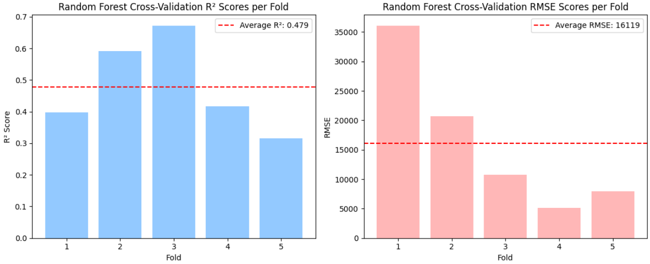


Figure: Random Forest Cross-Validation Scores

Conclusion

- **Summary:**
  - Playtime has a weak positive effect on popularity (correlation: 0.165).
  - Strategy games have the longest playtime (17.5 hours), Action games the highest ownership (200,000).
  - Decision Tree: $R^2 = 0.574$ (train-test), 0.368 (cross-validation).
  - Random Forest: $R^2 = 0.564$ (train-test), 0.479 (cross-validation).
- **Model Comparison:**
  - Random Forest outperforms Decision Tree in cross-validation (R^2: 0.479 vs 0.368, RMSE: 16,119 vs 17,770).
  - Random Forest shows better generalization (lower variability).
- **Future Improvements:**
  - Hyperparameter tuning (e.g., `max_depth`, `n_estimators`).
  - Collecting more data or engineering new features.

Future Improvements and Limitations

- **Limitations:**
  - High missing data: Reduced dataset from 12,089 to 6,271 rows after dropping rows/columns with >50% missing values (e.g., `main_story`, `completionist`).
  - Imbalanced data: Most games under 200 hours and 50,000 ratings, affecting model performance on outliers.
  - Missing external factors: Marketing, updates, and community effects not captured in the dataset.
  - Limited data scope: Only Steam and HLTB datasets used, potentially missing broader trends.
  - Model performance: High variance in Decision Tree (R^2: 0.368, std dev: 0.412) indicates overfitting.

- **Future Improvements:**
  - Collect more data: Include broader datasets (e.g., other platforms) and reduce missing values.
  - Feature engineering: Add external factors like marketing spend, update frequency, or community metrics.
  - Advanced modeling: Explore deep learning or hybrid models for better prediction.
  - Address imbalance: Use techniques like SMOTE or weighted loss functions for imbalanced data.

Thank You and Note

Thank you for reviewing this presentation!

Note: For detailed information, including datasets, code, and additional analysis, please refer to the .md files in my GitHub repository: https://github.com/esraesen/DSA210-Spring2025-Project.