Video Game Sales Prediction - Project Documentation

Introduction

Video Game Sales Prediction is a machine-learning project that forecasts whether a new video game will achieve GOOD or BAD global sales based on its platform, genre and publisher.

Using historical sales data, engineered statistical features and an ensemble of models, the system classifies potential releases so publishers can better prioritise marketing and distribution efforts. This document outlines the purpose, data preparation, modelling approach, evaluation and deployment details of the project.

Purpose

The aim of this project is to demonstrate how advanced AutoML techniques combined with careful feature engineering can deliver accurate sales predictions for the video game industry.

Accurate sales forecasts help stakeholders make informed decisions about inventory, marketing and release timing. Studies show that video game sales forecasting benefits the industry by identifying patterns and predictors that enable companies to adjust strategies, manage inventory and optimise release schedules.

Machine learning algorithms can analyse large amounts of data, identify patterns and make highly accurate predictions by considering historical sales, market trends and consumer behaviour.

Project Overview

This project takes a supervised classification approach. A tabular dataset of historical video game sales is used to train models that predict whether a future game's global sales will be above or below a threshold (GOOD vs BAD).

The workflow includes exploratory data analysis (EDA), data preprocessing, feature engineering, automated model benchmarking via AutoGluon, manual retraining of the best model (LightGBM) and deployment in a Streamlit web app. Users input the platform, genre and publisher of a proposed game, and the model outputs a prediction along with the probability and decision threshold.

Data

The dataset used is a widely known video game sales dataset scraped from VGChartz via Kaggle. It contains over 16,000 observations and 11 features including Rank, Name, Platform, Publisher, Genre, Year of Release, sales in North America, Europe, Japan and other regions, and total global sales. Missing values in the year and publisher columns are handled through imputation or left as NA, depending on the context. The dataset provides a comprehensive view of sales performance across multiple platforms and genres.

In a separate research study, a Kaggle video game sales dataset containing 16,719 games was analysed with machine learning algorithms such as linear regression, support vector machines, k-nearest neighbours, random forests and gradient boosting. After data preprocessing and feature selection, the gradient-boosted model surpassed other algorithms in delivering precise sales predictions. This finding supports our choice of LightGBM, a gradient boosting framework, as the final classifier in our project.

Modelling Methodology

The project uses AutoGluon-Tabular, an open-source AutoML framework that trains highly accurate models on raw tabular datasets with just a single line of Python code. AutoGluon ensembles multiple models and stacks them in layers, making better use of training time than searching for a single best model.

Experiments across public benchmarks show that AutoGluon is faster, more robust and more accurate than other AutoML frameworks. We feed the cleaned data and engineered features into AutoGluon to automatically evaluate a variety of algorithms and identify the strongest performer.

Feature Engineering

Feature engineering is crucial for boosting predictive performance. The following types of features were created:

- Statistical features: summary statistics of regional sales (mean, median, sum) and ratios between regions.
- Interaction features: combinations of platform, genre and publisher to capture interaction effects.
- Ranking features: ranks of games within their platform or genre based on historical global sales.
- Aggregated features: average sales per publisher or genre, enabling the model to learn typical performance patterns.

Model Selection and Retraining

After benchmarking, the LightGBM classifier achieved the best macro F1-test score (0.74). LightGBM is an open-source, high-performance framework that uses gradient boosting decision trees; it constructs strong learners by sequentially adding weak learners and is designed for efficiency, scalability and high accuracy.

A manual retraining step was performed outside AutoGluon to reduce memory usage and fine-tune hyperparameters. The final model yields an accuracy of 73% and a macro F1-score of 0.73, with a decision threshold of 0.28 for the GOOD classification.

Workflow

1. Exploratory Data Analysis (EDA) – Visualise distributions, detect outliers and understand relationships between sales, platform, genre and publisher.
2. Data Preprocessing – Clean the dataset, handle missing values, remove invalid entries and split data into training and test sets.
3. Feature Engineering – Generate statistical, interaction, ranking and aggregated features from the raw data.
4. AutoML Benchmarking – Use AutoGluon-Tabular to automatically train and evaluate multiple models, selecting the best based on F1-score.
5. Manual Retraining – Retrain the selected LightGBM model separately to optimise memory usage and performance.
6. Deployment – Integrate the final model into a Streamlit app that accepts platform, genre and publisher as inputs and returns the prediction, probability and decision threshold.

Technologies Used

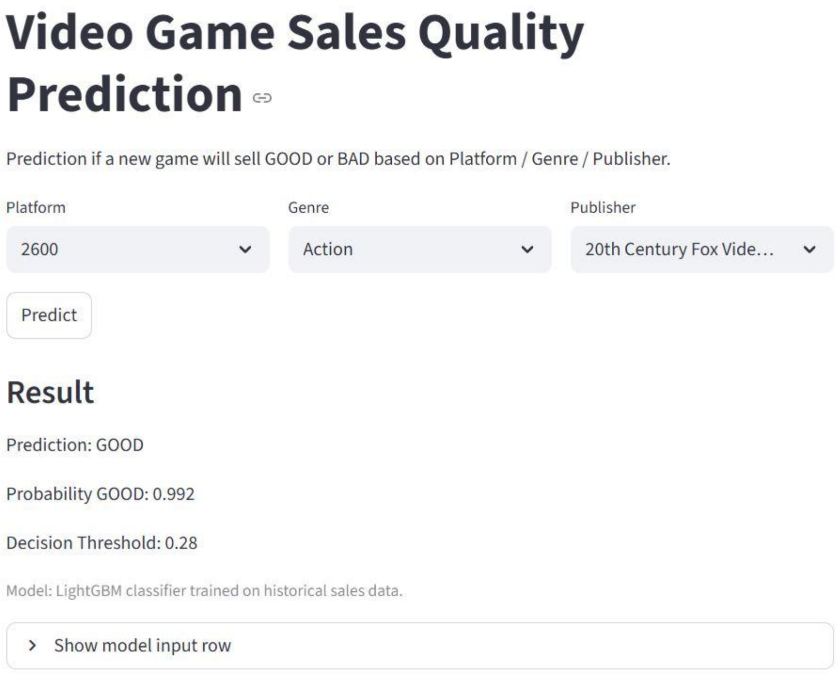
- Python, Pandas, NumPy – core programming and data manipulation tools.
- Matplotlib, Seaborn – data visualisation for EDA.
- AutoGluon, LightGBM, scikit-learn – AutoML benchmarking and final model training.
- Streamlit – deployment of the model as a web app for real-time predictions.

Results

The LightGBM classifier achieved a macro F1-test score of 0.74 during AutoGluon benchmarking and an overall accuracy of 73% on the hold-out test set after retraining.

These results highlight the effectiveness of gradient boosting algorithms for sales prediction tasks. The Streamlit app exposes the model via a simple interface where users can select the platform, genre and publisher to receive a GOOD/BAD prediction, the probability and the decision threshold.

App Preview



Video Game Sales Quality Prediction

Prediction if a new game will sell GOOD or BAD based on Platform / Genre / Publisher.

Platform: 2600

Genre: Action

Publisher: 20th Century Fox Vide…

Result

Prediction: GOOD

Probability GOOD: 0.992

Decision Threshold: 0.28

Model: LightGBM classifier trained on historical sales data.

Video Game Sales Predictor App Preview

Conclusion

The Video Game Sales Prediction project demonstrates how an end-to-end machine-learning pipeline can be built to forecast video game success.

By combining thorough data cleaning, sophisticated feature engineering and automated model selection via AutoGluon, the project achieves robust predictive performance. The final LightGBM model provides accurate classification of prospective video games as GOOD or BAD in terms of sales, enabling publishers and developers to make informed strategic decisions.

This approach can be extended to include additional features (such as critic scores or social media buzz) and to predict actual sales figures rather than binary outcomes.