MACHINE LEARNING PROJECT PROPOSAL

John McGinnes

03/09/2025

A. Project Overview

This proposal describes a machine learning project that aims to predict the most popular video game genre in future years based on historical sales data. The project will use the Video Game Sales dataset from Kaggle, which includes game sales, genre, platform, and release year. This data will be used to build a predictive model.

A.1. Organizational Need

Video Game companies often face difficulties in predicting which genres will be successful. By forecasting trends in game popularity, organizations can make informed decisions about game development and marketing strategies. The current process of predicting genre success is based only on historical trends without using a data-driven, predictive model.

A.2. Context and Background

The video game industry has grown rapidly over the past several decades, with different genres emerging as trends change. As video game developers focus on building games that align with expected market demands, predicting these trends can help maximize profits and enhance a company’s reputation. This machine learning project will utilize historical sales data from the Video Game Sales dataset available on Kaggle, which includes information about sales, game genres, platforms, and regions to develop a model capable of predicting which genres will be most successful in the future.

A.3. Outside Works Review

1. Using the power of machine learning in sales research: process and potential:

In this paper, Glackin, C. E. W., & Adivar, M. (2023) explore the potential of machine learning algorithms in predicting sales and sales performance. The study focuses on the process of applying machine learning to sales forecasting, specifically using regression and classification techniques. While it is not specific to the video game industry, the principles outlined in this paper can be directly applied to predicting the success of video game genres. By leveraging these machine learning methods, it becomes possible to create a predictive model using the Gradient Boosting Regressor for sales success in various video game genres.

2. Video Game Software Publishing in the US:

The report written by Burns, J. (2025) provides an industry-focused perspective on the state of video game publishing, examining emerging trends and business model changes. While this report does not directly incorporate machine learning, it is crucial for understanding the broader industry dynamics that could influence the predictive model in this project. By combining insights from this analysis with machine learning methods, companies could create better-informed predictions regarding which genre could be popular in the future. Factors identified in the report could serve as features in a machine-learning model to predict genre success.

3. Trends in Machine Learning Applied to Demand & Sales Forecasting: A Review

Usuga Cadavid, J.P., Lamouri, S., & Grabot, B. (2025) provide a comprehensive review of the latest trends in machine learning applied to demand and sales forecasting. The paper highlights various machine-learning techniques, such as support vector machines, Gradient Boosting Regressor, and deep learning, used to improve forecasting accuracy. These methods can be directly adapted to predict sales trends and genre popularity in the video game industry. The study emphasizes the importance of choosing the right forecasting techniques, which can guide the selection of a method for our predictive model.

4. Machine-Learning Models for Sales Time Series Forecasting

Pavlyshenko, B.M. (2019) examines the application of machine learning models in sales time series forecasting, with a focus on the use of stacking methods to improve model performance. This paper demonstrates how ensemble models can enhance predictive accuracy, an approach that can be applied to video game sales forecasting.

A.4. Solution Summary

The proposed machine-learning solution aims to predict the most popular video game genres in future years using historical data. We will train a supervised learning model using the Gradient Boosting Regressor algorithm, which is known for handling large datasets and providing insights into feature importance. The model will use factors such as genres, sales, platform, and release year to make predictions. By evaluating the model with metrics like accuracy and F1-Score, we aim to provide actionable insights for video game publishers to guide development and marketing strategies in alignment with predicted genre trends.

The solution will be implemented using Python, a versatile and widely used programming language in data science. Jupyter Notebooks will serve as the primary IDE, providing an interactive environment that is ideal for experimentation and analysis. The following libraries and tools will be used throughout the project:

- Pandas: For data manipulation and preprocessing, such as handling missing values and encoding categorical values.
- Scikit-learn: For implementing the Gradient Boosting Regressor, training the model, and evaluating performance metrics like accuracy, F1-Score, and cross-validation.
- Matplotlib and Seaborn: For visualizing data, model performance, and feature importance to gain insights into which factors drive the genre predictions.

A.5. Machine Learning Benefits

Machine learning will help to automate and enhance the prediction process, allowing for data-driven decisions on game development and marketing strategies. The model’s predictions will enable businesses to focus on creating games in the genres that are forecasted to succeed, improving potential sales and reducing the risk of investing in a less successful genre.

B. Machine Learning Project Design

B.1. Scope

In Scope:

- Data preprocessing, including cleaning, handling missing data, and feature engineering.

- Model selection and training using the Video Game Sales dataset.
- Evaluation of model performance using relevant metrics.
- Deployment of the trained model to make predictions about future popular genres.

Out of Scope:

- Real-time data collection or integration (there will be no live data streaming from external sources).
- Predictions for individual game success based on anything besides genre (e.g., price points, ESRB ratings).

B.2. Goals, Objectives, and Deliverables

Goals

- Improve game development strategies by predicting future popular genres.
- Provide data-driven recommendations to game publishers about the best genres to focus on.

Objectives

- Achieve at least 80% accuracy in genre classification.
- Achieve an F1-Score of 0.75 or greater for the model’s classification performance, balancing precision and recall.

- Develop a model that can predict the top 3 most likely successful genres for the next year. This will be measured using cross-validation.

Deliverables

- A trained machine learning model that can predict popular genres based on historical data.
- A report detailing the methodology, findings, and recommendations for future game development.

B.3. Standard Methodology

The CRISP-DM methodology will guide the implementation of this project to predict the most popular video game genres based on historical data.

- Business Understanding: The objective of this project is to build a predictive model that forecasts the most popular video game genres, aiding game developers in making data-driven decisions for game development and marketing strategies.
- Data Understanding: We will begin by exploring the Video Game Sales dataset from Kaggle, which includes information such as genre, sales, platform, and release year. We'll identify any data quality issues, such as missing values or inconsistencies, and analyze the data to predict trends in genre popularity over time.

- Data Preparation: In this phase, the dataset will be cleaned and preprocessed. Missing values will be handled, variables like genre and platform will be encoded, and the data will be prepared for modeling.
- Modeling: We will apply the Gradient Boosting Regressor algorithm to train the model with historical sales data to predict the most popular genre for future years.
- Evaluation: The model's performance will be evaluated using metrics like accuracy, precision, and recall. We will assess whether the model meets the business goal of accurately forecasting the top genres for the next year.
- Deployment: We will present the trained model's predictions in a report, which will provide insights to video game companies about which genres are likely to succeed in the future.

B.4. Projected Timeline

3/20/2025 – Project Kickoff Activities.

3/25/2025 – Data cleaning, exploration, and preprocessing. Prepare data for modeling and handle any missing values.

3/30/2025 – Initial model training using Gradient Boosting Regressor and evaluation with F1-Score and Accuracy metrics.

4/5/2025 – Model refinement and hyperparameter tuning.

4/10/2025 – Final model testing and evaluation.

4/15/2025 – Report and project submission to stakeholders. Include performance metrics and model results in the final report.

Sprint Schedule

Sprint	Start	End	Tasks
1	3/20/2025	3/25/2025	-Data cleaning and preprocessing. -Explore the dataset to understand its structure. -Prepare the data for machine learning model training.
2	3/26/2025	3/30/2025	-Initial model training using the Gradient Boosting Regressor algorithm. -Evaluate the model performance with F1-Score and accuracy. -Perform cross-

			validation to check for overfitting.
3	4/1/2025	4/5/2025	-Refine the model based on evaluation results. -Tune hyperparameters and experiment with different feature engineering methods. -Improve model accuracy and train with updated features.
4	4/6/2025	4/10/2025	-Final model testing and evaluation. -Prepare the final report with model performance metrics and recommendations.
5	4/11/2025	4/15/2025	-Finalize the report based on model

			results and insights. -Submit the project and present findings to stakeholders.

B.5. Resources and Costs

Resource	Description	Cost
Kaggle Dataset	Access to Video Game Sales dataset	$0.00
Development Tools	Python libraries (Pandas, Scikit-learn, etc.)	$0.00
Computational Power	Local machine computations will require an Intel i7 CPU and 16GB RAM Lenovo Legion 7i laptop or comparable recommended.	$1500.00
Data Scientist	Responsible for model development, training, and optimization. Labor cost for Data Scientist (40 hours at	$2000.00

	$50/hour)	
Data Analyst	Responsible for data cleaning, preprocessing, and initial data exploration. Labor cost for Data Analyst (20 hours at $30/hour)	$600.00
Project Manager	Responsible for coordinating the project and communicating progress to stakeholders. Labor cost for Project Manager (10 hours at $40/hour)	$400.00
	Total	$4500.00

B.6. Evaluation Criteria

Objective	**Success Criteria**
Achieve at least 80% accuracy in genre classification.	The model achieves 80% accuracy or higher on the validation set.
Develop a model	The model consistently predicts the top 3 genres

that can predict the top 3 most likely successful genres.	that align with historical data (accuracy measured by 5-fold cross-validation).
The model should achieve an F1-Score of 0.75 or greater.	The model achieves an F1-Score of 0.75 or greater for the top 3 predicted genres, balancing precision and recall.

C. Machine Learning Solution Design

C.1. Hypothesis

I hypothesize that the historical data on video game sales, including factors like genre, platform, and release year, can be used to predict the most popular video game genres in the future. I will be testing this hypothesis by training a model with historical data and evaluating its accuracy in predicting popular genres.

Success will be measured by the model's ability to predict the top genres with an accuracy of 80% or higher and its ability to generalize to unseen data using cross-validation. The F1-score will also be a metric used to verify success as a balance of precision and recall, aiming for 0.75.

C.2. Selected Algorithm

I will be using the Gradient Boosting Regressor algorithm.

C.2.a Algorithm Justification

Gradient Boosting Regressor is a robust and efficient algorithm that handles large datasets with complex relationships between features. It works well with both numerical and categorical data and is less prone to overfitting than other algorithms.

C.2.a.i. Algorithm Advantage

I am very confident that the algorithm will succeed. Gradient Boosting Regressor can handle large data sets efficiently, requires minimal hyperparameter tuning, and provides feature-importance insights.

C.2.a.ii. Algorithm Limitation

While Gradient Boosting Regressor is efficient, it may not always provide the highest accuracy on small datasets or when the number of features is extremely large.

C.3. Tools and Environment

- Operating System: Windows 11
- Programming Language: Python
- Libraries (Python): Pandas, Scikit-learn, Matplotlib, Seaborn (for visualization).
- Hosting Platform: Kaggle (for the dataset), Jupyter Notebook IDE, Local machine (Lenovo Legion 7i Laptop).

C.4. Performance Measurement

Model Performance will be evaluated using multiple metrics:

- Accuracy: The percentage of correct predictions for genre classification. The model should achieve an accuracy of 80% or greater, meaning the model correctly predicts the genre for at least 80% of the test dataset.

- Precision: The ability of the model to correctly identify relevant genres from the predicted top genres. The model will aim for at least 75% precision in measuring how many predicted successful genres are successful.
- Recall: The model's ability to correctly identify all the top genres in the dataset. The model will also aim for 75% recall for identifying the top genres.
- F1-Score: A metric that considers both precision and recall. A score of 0.75 will indicate a balance between precision and recall and will be ideal for this application

We will also use 5-fold cross-validation to ensure the model generalizes well to unseen data and isn't overfitting the training set.

A confusion matrix will also be used to analyze the classification performance further. This will allow a better assessment of the model's ability to correctly classify each genre.

D. Description of Data Sets

D.1. Data Source

The data will be extracted from Kaggle's Video Game Sales dataset, which contains historical sales data, including genre, platform, and release year. The dataset can be accessed directly on Kaggle.com using this link: https://www.kaggle.com/datasets/ulrikthygepedersen/video-games-sales

D.2. Data Collection Method

Data will be extracted from Kaggle's public API.

D.2.a.i. Data Collection Method Advantage

The dataset is readily available, cleaned, and formatted for machine learning purposes.

D.2.a.ii. Data Collection Method Limitation

The dataset may be limited in terms of platform variety and specific regions.

D.3. Quality and Completeness of Data

The dataset will be cleaned for missing values, outliers will be identified and removed, and categorical data will be encoded for use in machine learning models.

D.4. Precautions for Sensitive Data

There are no sensitive data concerns in this project, as the dataset focuses on publicly available game sales and metadata.

References

1. Glackin, C. E. W., & Adivar, M. (2023). Using the power of machine learning in sales research: Process and potential. *Journal of Personal Selling & Sales Management, 43(*3*),* 178-194. https://doi.org/10.1080/08853134.2022.2128812

2. Burns, J. (2025). Video Game Software Publishing in the US. *IBISWorld*. https://my.ibisworld.com/us/en/industry/51121e

3. Usaga Cadavid, J.P., Lamouri, S., & Grabot, B. (2025). Trends in machine learning applied to demand & sales forecasting: A review. *International Conference on Information Systems, Logistics, and Supply Chain.* https://hal.archives-ouvertes.fr/hal-01881362

4. Pavlyshenko, B.M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4(*1*)*, 15. https://doi.org/10.3390/data4010015

5. Pederson, U.T. (2021). *Video Game Sales*. Kaggle

https://www.kaggle.com/datasets/ulrikthygepedersen/video-games-sales