Report

Business understanding

Identifying the business goals

Background

The video game industry is a massive global market, which means that to succeed, it is crucial to know the factors which contribute to the success of a product. This project aims to evaluate the influence of several factors, including genre, release timing, critic and user scores and ESRB rating, on the sales in Europe, North America, Japan as well as globally. This goal of this analysis is to allow stakeholders to make more informed decisions to achieve more successful game releases.

Business goals

The project has no explicit business goals. The purpose of this project is learning and obtaining experience working on data mining projects. The main goals are developing a predictive model for video game sales with an accuracy of at least 80% and presenting the results and process in a visually understandable way.

Business success criteria

The accuracy of the model can be verified with RMSE. The quality of the presentation will be assessed by the course graders.

Assessing the situation

Inventory of resources

People: 3 data miners

Data: 1.62 MB dataset in .csv format, which contains information about the sales and relevant features of video games released between the years 1980 and 2016

Hardware: 3 personal computers

Software: The modeling and data visualisation will all be done in Python.

Requirements, assumptions, and constraints

The project needs to be completed by 11th December. At the point of the deadline all data mining activities need to be done and the results as well as the methodologies need to be presented as a poster in a visually appealing way. The results must contain visual material, which explains the work. All figures must be understandable without any oral explanations.

Risks and contingencies

Risk: Internet outage

Contingency: continue work somewhere with a working Internet connection, i.e using university Wi-Fi

Risk: Poor data quality

Contingency: use scraping to fill the missing values or find another dataset, which contains the missing values

Terminology

Critic score - aggregate score compiled by Metacritic staff, on a scale of 0-100

User score - average score by Metacritic subscribers

Rating - the ESRB maturity rating

Costs and benefits

Costs: The project will not cost anything and will not generate revenue. The expected time expenditure is 30 hours of work per person, which accumulates to 90 hours of work in total.

Benefits: Improved ability to predict successful video game attributes, leading to better investments and development decisions.

Defining your data-mining goals

Data-mining goals

Identify patterns and trends that consistently correlate with high-performing games.

To develop a predictive model that can forecast video games success based on various metrics such as review scores and sales data.

Data-mining success criteria

Ability to identify key metrics that are most indicative of a game's potential success.

High accuracy in predicting the success of new video games.

Data understanding

Gathering data

Data requirements

The project requires data in .csv format. Every game in the dataset needs to have the following features: name, platform, release date, genre, publisher, developer, North American sales, European sales, Japanese sales, other sales, global sales, user score, user count, critic score, critic count and rating. The release date values should fall between the years 1980 and 2016.

Data availability

The data is available in a Kaggle dataset. The features user score, user count, critic score, critic count, developer and rating contain missing values. The rows with missing values will either have the values substituted by scraping data from the Metacritic website or be discarded. Also, the dataset contains only a year of release, the release date will be obtained by scraping from the Metacritic website.

Selection criteria

The primary dataset to be used can be found here: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings. The missing values will be scraped from the Metacritic website.

Describing data

The dataset contains 16 719 cases and 16 features.

1. Name:
   Description: Name of the game.
   Data Type: Text/String
2. Platform:
   Description: Console on which the game is running.
   Data Type: Categorical
3. Year_of_Release:
   Description: Year of the game's release.
   Data Type: Numeric (Year)
4. Genre:
   Description: Game's category or genre.
   Data Type: Categorical

5. Publisher:
   Description: Publisher of the game.
   Data Type: Text/String
6. NA_Sales:
   Description: Game sales in North America (in millions of units).
   Data Type: Numeric (Millions)
7. EU_Sales:
   Description: Game sales in the European Union (in millions of units).
   Data Type: Numeric (Millions)
8. JP_Sales:
   Description: Game sales in Japan (in millions of units).
   Data Type: Numeric (Millions)
9. Other_Sales:
   Description: Game sales in the rest of the world, excluding NA, EU, and JP (in millions of units).
   Data Type: Numeric (Millions)
10. Global_Sales:
    Description: Total sales worldwide (in millions of units).
    Data Type: Numeric (Millions)
11. Critic_Score:
    Description: Aggregate score compiled by Metacritic staff on a scale of 0-100.
    Data Type: Numeric
12. Critic_Count:
    Description: The number of critics used in coming up with the Critic_Score.
    Data Type: Numeric
13. User_Score:
    Description: Score by Metacritic's subscribers on a scale of 0-10,0.
    Data Type: Numeric
14. User_Count:
    Description: Number of users who gave the User_Score.
    Data Type: Numeric
15. Developer:
    Description: Party responsible for creating the game.
    Data Type: Text/String

16. Rating:

    Description: The ESRB rating (e.g., Everyone, Teen, Adults Only, etc.).

    Data Type: Categorical

Additionally, the following feature will be scraped.

17. Release_date:

    Description: The date of the game's release

    Data Type: Date

Exploring data

Attributes

- Platform consists of 31 categorical values
- Year_Of_Release consists of 40 unique discrete values in the range [1980 - 2020]
- Genre consists of 13 unique categorical values
- Publisher consists of 582 unique categorical values
- NA_Sales consists of 402 unique discrete values in the range [0 - 41.36]
- EU_Sales consists of 307 unique discrete values in the range [0 - 28.96]
- JP_Sales consists of 244 unique discrete values in the range [0 - 10.22]
- Other_Sales consists of 155 unique discrete values in the range [0 - 10. 57]
- Global_Sales consists of 629 unique discrete values in the range [0.01 - 82.53]
- Critic_Score consists of 83 unique discrete values in the range [13 - 98]
- Critic_Count consists of 107 unique discrete values in the range [3 - 113]
- User_Score consists of 96 unique discrete values in the range [0 - 9.7]
- User_Count consists of 889 unique discrete values in the range [4 - 10665]
- Developer consists of 1696 unique categorical values
- Rating consists of 8 unique categorical values

Trend analysis over time: Investigate how metrics have changed over time.

- How have sales trends changed across different regions?
- Have certain genres become more popular?

Hypotheses:

- Games rated for mature audiences have become more popular
- Some publishers/developers will have high sales regardless of user scores because of their past reputation. Meaning that a publisher can make multiple bad games in a row but it will still have high sales because of making a lot of great(very highly rated) games in the past.

- Majority of games have moved to PC
- Video game sales are dominated by a handful of publishers. Meaning that highly rated games may not have a lot of global sales since they are not well known.

Verifying data quality

The following attributes have 100% coverage, meaning that there are no NaN's in the column of that attribute.

- Platform
- Year_of_Release
- Publisher
- NA_Sales
- EU_Sales
- JP_Sales
- Other_Sales

The following don't have 100% coverage. In parentheses the coverage has been brought out in percentages or in amount (if the percentage is insignificantly small)

- Genre (2 missing out of ~16700 entries)
- Critic_Score (51% missing)
- Critic_Count(51% missing)
- User_Score(40% missing)
- User_Count(55% missing)
- Developer(40% missing)
- Rating(40% missing)

None of the aforementioned attributes that are missing are really important to our task. But in case we think of a way to use them, data about them can be scraped from the website www.metacritic.com.

All of the attributes that would contribute to our primary goal are 100% covered. And future goals that require us to analyze the other attributes that have missing entries can be repaired by web scraping which is relatively easy to carry out.

Project plan

Task 1: Data collection and preparation

- Collect sales data, genre information and release dates
- Clean and preprocess the data
- Estimated time: 18 hours (Robert Ivask: 6 hours, Marko Peedosk: 6 hours, Henrik Innos: 6 hours)

Task 2: Statistical analysis of genre trends

- Analyze trends in game popularity by genre over time
- Use statistical methods to identify significant changes and patterns.
- Estimated time: 12 hours (Robert Ivask: 2 hours, Marko Peedosk: 2 hours, Henrik Innos: 8 hours)

Task 3: Visualization of trends

- Create visual representations of genre trends over time.
- Estimated time: 12 hours (Robert Ivask: 2 hours, Marko Peedosk: 2 hours, Henrik Innos: 8 hours)

Task 4: Analysis of user scores and engagement metrics

- Analyze correlation between user scores and metrics like na_sales, eu_sales, jp_sales, other_sales, global_sales, achievements_count, suggestions_count
- Identify patterns indicating user satisfaction.
- Estimated time: 12 hours (Robert Ivask: 2 hours, Marko Peedosk: 8 hours, Henrik Innos: 2 hours)

Task 5: Reporting on user engagement insights

- Report relationships between user engagement and game satisfaction
- Include visualizations and key findings
- Estimated time: 10 hours (Robert Ivask: 2 hours, Marko Peedosk: 6 hours, Henrik Innos: 2 hours)

Task 6: Developing predictive model

- Use machine learning techniques to create models predicting future sales based on historical data.
- Test different algorithms for best accuracy.

- Estimated time: 16 hours (Robert Ivask: 10 hours, Marko Peedosk: 3 hours, Henrik Innos: 3 hours)

Task 7: Model evaluation

- Evaluate model performance using RMSE.
- Refine and tune models for better accuracy
- Estimated time: 10 hours (Robert Ivask: 6 hours, Marko Peedosk: 2 hours, Henrik Innos: 2 hours)

Task 8: Final report

- Compile all findings, insights, and model details into a comprehensive final report.
- Estimated time: 12 hours (Robert Ivask: 4 hours, Marko Peedosk: 4 hours, Henrik Innos: 4 hours)