1 December 2019

HW10 Team M2: Video Game Sales Predictions

Sabina Mammadova, Heba Elshatoury, Maria Medina

Task 1. Setting up (0.5 points)

Project repository: https://github.com/DataScienceProject2019/M2_VIDEOGAME_PROJECT

Task 2. Business understanding (1 point)

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-analysis, data-mining or machine learning goals and producing your project plan. For this exercise, please, develop a business understanding of your project. According to CRISP-DM, you should report the following:

- **Identifying your business goals**
  - **Background:** Our database show the sales of video games in multiple locations around the world along with the publisher, genre and platform among other information. The video game industry is very huge with so many sales happening around the world. Some games are very popular among people in all countries and make so much profit. The popularity of a certain game varies among generations and change over the years.
  - **Business goals:** Our goal in this project is to predict the global sales in the coming years for some video games based on some features like genre, publisher and platform. The idea is to analyse what kind of features make the sales highest.
  - **Business success criteria:** This model can help game developers design games that would sell most. Our analysis works as a guide to increase the sales of video games in the coming

years while taking certain features into mind. Multiple visualisations of the data will also give an insight about correlation between features in the data.

- **Assessing your situation**
  - **Inventory of resources:** We will use the Video Game Sales with Ratings dataset from Kaggle. The data is a .csv file format. To train and test our model we will work with Python programming language. For data visualisation we are going to use Tableau Software. For technical support, we will ask the supervisors for advice.
  - **Requirements, assumptions, and constraints:** The project due date is the 19th of December, we will present a poster that showcase our idea, problem statement and results.
  - **Risks and contingencies:** One risk that we might encounter is that some information in the data are missing and can create an unbalanced dataset. We will solve this issue by assessing if the feature is not completed properly in all the instances we will exclude from the data.
  - **Terminology:** The topic we chose is not specific and there are no particular terminology that need to be explained.
  - **Costs and benefits:** Since the data is public and available for use we don't have cost associated to the project we also have the software and hardware needed to complete this project. The benefits of our project is that companies can use the model to predict the sales and increase their profit for the next products.

- **Defining your data-mining goals**
  - **Data-mining goals:** The resources and work done will be available for supervisors on our repository in Github. Final presentation of the work would be the poster session which will be displayed not only for supervisors but also for all course attendees.
  - **Data-mining success criteria:** The assessment of the project can be based on the accuracy of our model. Additionally, it would be interesting to verify if our results match with other sources that have documented the highest video game sales per year, such as: https://www.absentdata.com/global-video-games-meta-critic-scores/

Task 3. Data understanding (2 points)

1. Gathering data

Outline data requirements: To address our data mining goals, we need information about past video game sales, including at least:

1. Amount of sales: It could be either amount of units sold or amount of income gained.

2. Time range of a year: Video game sales are highly seasonal and peak in December [1], just like in many other domains, so in order to have an unbiased approach, sales per year is the best option.

According to the "Six rules for effective forecasting", [2] researchers should "always look back at least twice as far as you are looking forward"; since we would like to predict video game sales for next year, we would need data from at least 2018 and 2019, however, in order to produce a more clear overview of how video game sales have evolved, we would like to include data from the last 10 years.

Verify data availability: Since our instructor recommended us Kaggle as a reliant data source, we searched for "video game sales" and found a really good database with many features available: https://www.kaggle.com/gregorut/videogamesales

We downloaded the zip file that contained a .csv with the data. The file contains many features about video game sales, such as the year and the number of sales in Europe and North America. So in our case, we have found available and sufficient data with all the features we needed and more.

Features:

1.Rank - Ranking of overall sales

2.Name - The games name

3.Platform - Platform of the games release (i.e. PC,PS4, etc.)

4.Year - Year of the game's release

5.Genre - Genre of the game

6.Publisher - Publisher of the game

7.NA_Sales - Sales in North America (in millions)

8.EU_Sales - Sales in Europe (in millions)

9.JP_Sales - Sales in Japan (in millions)

10.Other_Sales - Sales in the rest of the world (in millions)

11.Global_Sales - Total worldwide sales.

Define selection criteria: The fields that are relevant to this project are the **names of the games, the genres, the NA_Sales, the EU_Sales, JP_Sales, Other_Sales and Glogal_Sales, as well as the genres**, which are the same through the years. By using this information, we could conduct a regression analysis based on the genre. Using regression analysis, for example, with KNN, is a task that has been posed by another user but not yet completed: https://www.kaggle.com/gregorut/videogamesales/discussion/116998?fbclid=IwAR2kaLXDKnFV3ifw02PFeoAdHoiqJ6wZeadmGt1dArLjab8GUyYe2xZpSxg

2. Describing data: Bartos Trzaskowski [3] created various visualizations of the database, where it is possible to see the global sales per year in millions, sales per region, number of releases per year, the publishers that released the most, the releases per year, etc. There are many other sources of Kaggle available where many plots allow us to understand the data

better: https://www.kaggle.com/neilslab/seaborn-visualization

3. Exploring data: As we were exploring the data, we realized that the field "Year" does not refer to the sales performed on that entire year, instead, it refers to the year of release, so that instead of having information of the sales of each game over the years, we have the information of many games and how much they sold on their year of release only. This is a main constraint that would only allow us to perform regression using the genre, but we could still apply several methods to it, such as KNN and the DecisionTree. This risk was addressed in our **Risks and contingencies** section in point 2.

4. Verifying data quality: Albeit we have found certain constraints, the dataset has proven to be useful and to contain adequate data, with minor exceptions:
-One of the games has the release year of 2020, so only the *presales* are available, which are out of our scope.
-268 instances do not include the year, so we will either erase them or find the year manually. Rather, the first, since the quantity of the data is already so immense.

References:

[1] Video Game Sales Are Extremely Seasonal:
https://www.statista.com/chart/16211/monthly-video-game-industry-sales/
[2] Six rules for effective forecasting: https://hbr.org/2007/07/six-rules-for-effective-forecasting
[3] Bartos' Analysis: http://www.bmtfx.com/video-games-sales-1980-2017-2/?lang=en

Task 4. Planning your project (0.5 points)

- **Task1:** Selecting Data, Data cleaning, Integrating Data, Formatting Data *(Repeatable task)*
  Hours contributed: ~5 hours each (total 15 hours)
  Tools used: Python - Jupyter Notebook
- **Task2:** Predictions of sales by genre using KNN, DecisionTree and RandomForest *(Repeatable task)*
  Hours contributed: ~5 hours each (total 15 hours)
  Tools used: Python - Jupyter Notebook
- **Task3:** Visualization of correlation in data *(Repeatable task)*
  Hours contributed: ~5 hours each (total 15 hours)
  Tools used: Tableau Software
- **Task4:** Evaluation and analysis of results
  Hours contributed: ~4 hours each (total 12 hours)
  Tools used: Online sources of video game sales, such as: https://www.absentdata.com/global-video-games-meta-critic-scores/
- **Task5:** Poster design
  Hours contributed: ~6 hours each (total 18 hours)
  Tools used: Google Slides , PDF

  The hours are approximate, also there are some repeatable tasks that could take longer hours. We also already worked on discussing and choosing the topic and finding the data. Adding to that we prepared for presenting our idea and preparing this report. (~ 8 hours) Some of these tasks we will do together and some would be divided or done separately. We do regular meetings to update each other about the progress that we have made and put things together.