Project Proposal

Video Game Predictive Analytics on Genre Sales

June 13th, 2020

Gabriel Valenzuela

**Introduction**

Background

Video game companies are currently in the process of introducing new video game platforms such as the PlayStation 5 and the Xbox Series X. With the future release of these products, they are providing consumers with various new and immersive experiences throughout several genres from action to sports. Therefore, it would be valuable for future video game publishers to understand which genres to start or continue developing for its current gamers and future ones. This analysis will reach millions of customers not only in the United States but around the world.

The goal of this project is to provide video game publishers such as Electronic Arts and Activision with a new found means to expand their markets by focusing their growth efforts on specific genres in the video game industry. Data visualization and statistical techniques will be used to determine patterns among: video game sales across several markets and different video game platforms.

Problem Statement

Utilizing video game sales data, to investigate correlations between video game publishers and genre, genres and sales across various markets including United States and European. I believe that video game publishers and developers would be interested to determine which genres will produce higher sales towards the future for the upcoming releases of new platforms.

Scope

The project will focus on the video game data set provided by *Kaggle* describing the platforms, genres, and sales across the globe.

Within Scope:

* Constructing a model capable of inferring sales for genres across different platforms
* Model should only consider video game sales data, platforms, and genres

Outside of Scope:

* Data not pertaining to other platforms or developers outside of ones mentioned within the data set

Document Overview

The remainder of this document details the requirements, technical approach, expected results, and project management plan. These sections will describe how I plan on providing a solution to the problem statement, and achieve the scope designated for this project.

**Preliminary Requirements**

Requirement 1: Utilize the data provided by *Kaggle*

The type of data from *Kaggle* includes data about the rank of the video game including its name, platform of the game, the year it was released, the publisher of the game, and the genre it belongs too. At the same time, it shows the sales of the video in North America, Japan, Europe, other areas, and a global total.

Requirement 2: Develop methods to identify patterns in sales by genres

The methods used will be developed based on statistical and data-mining principles and techniques. These methods shall consistently identify patterns in sales by genres.

Requirement 3: Develop methods to identify patterns in sales by publishers (developers)

The methods used will be developed based on statistical and data-mining principles and techniques. These methods shall consistently identify patterns in sales by publishers.

Requirement 4: Develop a model for determining the future sales of a genre

I will attempt to develop a LSTM model for determining the future sales of a genre.

Requirement 5: The model shall predict the sales of a genre for the next five years

The model will be developed based on patterns identified techniques to predict the sales of a genre for the next five years.

Requirement 6: The model shall provide accuracy of the prediction of sales

The model will produce accuracy of its predictions for each genre

**Technical Approach**

Data Source

Video Game Sales

The data set is pulled from the website *VGChartz* using Python’s BeautifulSoup library which is located on the website *Kaggle* which contains the data set that was made to analyze sales data for more than 16,500 games.

Data Set Located: <https://www.kaggle.com/gregorut/videogamesales?select=vgsales.csv>

Data Source Originally pulled from: <https://www.vgchartz.com/gamedb/>

Analysis

For me to gain a stronger foundation of knowledge on how to utilize prediction analytics for this video game project, I believe that other aspects need to be studied and evaluated. Also, I will need to understand and research typical video game sales for certain genres and developers since there may certain timelines and typical sale trends within the market. I will need to perform more research into the most appropriate data mining techniques to be able to develop and deploy the planned models. Before I am able to determine the forecast sales for genres of video games, I will need to analyze all of these different aspects while incorporating all of the data given at the time.

Requirements Development

In the requirements development, it will be require the use of a Python and R environment in order to collect and prepare the data as well as developing the model to make the predictions for the sales. The current plan is to format the data in R and deploy and develop the model within Python while possibly using both for visualizations.

Model Development

The first step in this process will be the selection of features as inputs to the model. In order to determine which features to use, the data will be investigated to determine which data will be used to make the future predictions. Trends in sales for different platforms, developers, and genres will be plotted to show past or possible new patterns. Throughout the process of the development, I will apply agile methodology to make adjustments on the model to ensure that it can be at its more successful point of predicting. The adjustments will be expected based on future changes and updates to the model to ensure that the algorithm is correct and is inputting the best information.

Testing and Evaluation

Once the model is able to perform, an evaluation on the model’s prediction of future video game sales for genres will be employed. The testing of accuracy will be conducted on a hold of sample data for video game sales from 2010-2020, while the model will be constructed on data from 1980-2009. I will input the hold data set into the model without looking at the sales across all of the genres and verify that the model consistently outputs the accurate results. After being able to run the model with the test data, I think it would be important to switch out different features in order to see how of an affect it would have on the results. Therefore, it will increase the understanding of relationships between all of the variables as well as being more certain about the results.

Delivery

The final delivery of the results will be presented in a presentation and technical paper to describe the results of the video game sales prediction analysis and use of the methods learned throughout the project to those of a technical and non-technical background.

**Expected Results**

The project will produce two major deliverables:

* Methods developed to identify patterns in video game sales
  + In the presentation and paper, I will submit a narrative on what methods were used, why they were chosen, and how they were used to locate possible patterns
* Model to predict video games sales by genre
  + A model will be presented that will predict the next five years of video games sales by genre. A summary will be shown on how the model was developed with an analytical approach including the used input and the output created from the model

**Management Approach**

Project Plan

|  |  |  |
| --- | --- | --- |
| Task | Start Date | Finish Date |
| **Decide on business case** | 6/7/2020 | 6/8/2020 |
| **Locate data set for business case** | 6/8/2020 | 6/9/2020 |
| **Draft project proposal** | 6/9/2020 | 6/14/2020 |
| **Data preparation - R** | 6/14/2020 | 6/21/2020 |
| **Perform exploratory data analysis - R** | 6/21/2020 | 6/24/2020 |
| **Preliminary analysis report** | 6/25/2020 | 6/28/2020 |
| **Model selection and evaluation - Python** | 6/29/2020 | 7/5/2020 |
| **Test and Visualize Results – Python & R** | 7/6/2020 | 7/12/2020 |
| **Presentation of intermediate results** | 7/13/2020 | 7/19/2020 |
| **Summarize results and revise final model** | 7/20/2020 | 7/26/2020 |
| **Outline final presentation and paper. Creating and organizing necessary materials** | 7/27/2020 | 8/2/2020 |
| **Final Paper** | 8/3/2020 | 8/7/2020 |
| **Final Presentation** | 8/3/2020 | 8/7/2020 |

Project Risks

Risk 1: Insufficient amount of data

Due to the data being pulled off a dataset website source, it is not a recognized central source of information on all video game sales across different marketplaces and the world resulting in possible inaccurate information on certain sales.

Risk 2: Expertise in Predictive Analytics

Since I am student in the process of learning the field, I do not have expertise knowledge on forecasting sales and only a fair amount of exposure to data mining techniques.