Anomaly Detection Techniques to Find Influential Users

Shawn L. Watts

Western Governors University

Table of Contents

[A. Proposal Overview 4](#_Toc2137083979)

[A.1 Research Question or Organizational Need 4](#_Toc1815560889)

[A.2 Context and Background 4](#_Toc1384268585)

[A.3 and A3A Summary of Published Works and Their Relation to the Project 4](#_Toc2079458267)

[Review of Work 1 5](#_Toc694593955)

[Review of Work 2 5](#_Toc475195288)

[Review of Work 3 5](#_Toc1156669356)

[A.4 Summary of Data Analytics Solution 5](#_Toc374341712)

[A.5 Benefits and Support of Decision-Making Process 6](#_Toc1116633891)

[B. Data Analytics Project Plan 6](#_Toc1561262715)

[B.1 Goals, Objectives, and Deliverables 6](#_Toc1217623361)

[B.2 Scope of Project 7](#_Toc427690838)

[B.2.A Included in Project Scope 7](#_Toc1243851001)

[B.2.B Not included in Project Scope 7](#_Toc2036717847)

[B.3 Standard Methodology 7](#_Toc204762054)

[B.4 Timeline and Milestones 8](#_Toc1049840195)

[B.5 Resources and Costs 8](#_Toc313275103)

[B.6 Criteria for Success 9](#_Toc1337049307)

[C. Design of Data Analytics Solution 9](#_Toc764513894)

[C.1 Hypothesis 9](#_Toc1169566000)

[C.2 and C.2.A Analytical Method 9](#_Toc1622461191)

[C.3 Tools and Environments 10](#_Toc53928856)

[C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance 10](#_Toc1386867513)

[C.5 Practical Significance 11](#_Toc157862960)

[C.6 Visual Communication 11](#_Toc1263386613)

[D. Description of Dataset 12](#_Toc34323025)

[D.1 Source of Data 12](#_Toc1867631323)

[D.2 Appropriateness of Dataset 12](#_Toc418252330)

[D.3 Data Collection Methods 12](#_Toc1665164903)

[D.4 Observations on Quality and Completeness of Data 12](#_Toc57902398)

[D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances 13](#_Toc1649406101)

[References 14](#_Toc1322159537)

[Appendix A 14](#_Toc1046046101)

[Title of Appendix 15](#_Toc804972029)

[Appendix B 15](#_Toc128761729)

[Title of Appendix 16](#_Toc1200922991)

[Appendix C 16](#_Toc1897962072)

[Title of Appendix 17](#_Toc1862588384)

[Appendix D 17](#_Toc1292381938)

[Title of Appendix 18](#_Toc390418831)

# A. Proposal Overview

## A.1 Research Question or Organizational Need

Which users are influencing others the most with their reviews and recommendations of video games on the Steam platform?

## A.2 Context and Background

In this fictitious scenario based on real a real company, Valve Corporation and Steam gaming platform need help discovering user patterns in their recommendation and review data to better understand who their most influential users are. They have asked for a short list of the most influential users for study by their selves and other game development companies that use the Steam platform to sell their games. The hope is that this will help game companies develop better games and marketing strategies to increase sales. Since this data is not known at this time, they have asked to do some discovery and provide them a list of users with proper rational and methods. The value of this study could improve success for all parties.

## A.3 and A3A Summary of Published Works and Their Relation to the Project

### Review of Work 1

Game Recommendations on Steam (Kozyriev, 2023) has done a lot of work screen scraping the Steam platform. This data will be useful for this project and be the basis for our analysis and deliverables. Using an already curated data set will save us and our client valuable time. There are also many recommendation models that people have created for this data. We will be repurposing the data to look for influencers.

### Review of Work 2

This article discusses several use cases for sales and marketing using machine learning and data mining (Glackin & Adviar, 2023). While most of this work discusses supervised machine learning techniques, the processes it lays out will be a good foundation for our project.

### Review of Work 3

Data Profiling and Machine Learning to Identify Influencers from Social Media Platforms from the journal of ICT Standardization (Elbaghazaoui, Amnai, & Fakhri, 2022), details methodologies like we plan to do with this project. We will use domain knowledge, statistics, and machine learning insights to achieve our objective.

## A.4 Summary of Data Analytics Solution

The dataset has over thirteen million users. Steam is interested in the most impactful. I will use basic Exploratory Data Analysis (EDA) to understand the dataset. I will use Principal Component Analysis (PCA) to reduce the number of dimensions to one and take everything on the right tail at ninety five percent or more. I will use a second PCA model with two components to visualize the results. To be complete, I will also include a t-Distributed Stochastic Neighbor Embedding (t-SNE) model with two components to see if it agrees with the PCA visualization. After this selection and inspection, I will use the Isolation Forest algorithm to discover the outliers using the relevant features discovered during EDA. I will choose an appropriate contamination parameter to curate a dataset of under one thousand users. This will be saved to excel for further analysis by the client.

## A.5 Benefits and Support of Decision-Making Process

The benefits of this decision-making process will bring knowledge about which users to spend Steam’s research and marketing efforts first. Using statistics and machine learning techniques together will yield a high probability that the user list generated will be of have the greatest value to focus on first. The consequences of not doing it in a methodical manner could mean wasted time, effort, and money and opportunities could be missed if guessing starts to occur. The process laid out can be repeated simply by removing the current list from the sample and repeating the steps to again get the next list to focus on.

# B. Data Analytics Project Plan

## B.1 Goals, Objectives, and Deliverables

* Goal 1: The goal is to select a set of users with the most influence among the others.
  + Objective 1.0: Client discussions centering around plan, design, and expectations.
    - Deliverable 1.0.1: Meeting notes and transcript with client.
    - Deliverable 1.0.2: Project plan and timelines.
  + Objective 1.1: Acquire the data
    - Deliverable 1.1.1: The data will be downloaded in its original CSV format.
    - Deliverable 1.1.2: The data will be saved in parquet format for faster processing.
  + Objective 1.2: Analyze available features and clean the data.
    - Deliverable 1.2.1: Features will be selected with descriptive statistics, correlation, and basic understanding of the data structure used to get a sense of the data.
  + Objective 1.3: Sample enough data to run machine learning algorithms on.
    - Deliverable 1.3.1: A sample of data with the highest impactful users will be selected and stored in memory.
    - Deliverable 1.3.2: The sample will be much smaller than the original for faster processing during machine learning.
  + Objective 1.4: Reduce the data to two dimensions and visualize for an understanding of the outliers.
    - Deliverable 1.4.1: PCA will be used to reduce dimensions to two dimensions and a scatter plot displayed to show the pattern of our sample.
    - Deliverable 1.4.2: t-SNE will be used to reduce to two dimensions and a scatter be displayed of the results to see how it agrees with the PCA version.
  + Objective 1.5: Create an Isolation Forest model.
    - Deliverable 1.5.1: An isolation forest model shall be created using the relevant features and appropriate contamination parameter.
    - Deliverable 1.5.2: A pickle file shall be saved of the model to disk.
  + Objective 1.6: Select the outliers and ensure count is below one thousand.
    - Deliverable 1.6.1: A dataset shall be created of the outliers predicted by the model.
    - Deliverable 1.6.2: The dataset shall be saved as an Excel file per the client.
  + Objective 1.7: Meet with clients and deliver the result.
    - Deliverable 1.7.1: Obtain sign-off on work or feedback about revisions.
    - Deliverable 1.7.2: Wrap up project or deliver a new plan based on feedback.

## B.2 Scope of Project

### B.2.A Included in Project Scope

The scope of the project will be limited to de-identified user data with minimal features. The only hard deliverable here will be the final list of users in Excel format. No more than one thousand. If the client is not satisfied, we will review the issues and iterate again.

### B.2.B Not included in Project Scope

Bringing in identifiable data and doing further analysis will be the responsibility of the client. This data is not provided and thus would be impossible to analyze.

### B.3 Standard Methodology

This team will use the Rapid Application Development process. We have well-defined objectives, and this is time sensitive. Also, the user will likely have feedback and will let us know when they are satisfied. During the analysis and curation part of this project, we will define the requirements, go through a user design session, then begin constructing the solution to deliver the product. If the user is satisfied, we will end the engagement. If not, we will go back to a user design session and construction phase until the client is satisfied.

## B.4 Timeline and Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Deliverable 1.0.1 | 16 hours | *June 24th 2024* | *June 25th 2024* |
| Deliverable 1.0.2 | 24 hours | *June 26th 2024* | *June 28th 2024* |
| Deliverable 1.1.1 | 2 hours | *July 1st 2024* | *July 2nd 2024* |
| Deliverable 1.1.2 | 2 hours | *July 1st 2024* | *July 2nd 2024* |
| Deliverable 1.2.1 | 8 hours | *July 1st 2024* | *July 2nd 2024* |
| Deliverable 1.3.1 | 7.5 hours | *July 3rd 2024* | *July 3rd 2024* |
| Deliverable 1.3.2 | 0.5 hours | *July 3rd 2024* | *July 3rd 2024* |
| Holiday | 0 hours | *July 4th 2024* | *July 7th 2024* |
| Deliverable 1.4.1 | 4 hours | *July 8th 2024* | *July 8th 2024* |
| Deliverable 1.4.2 | 4 hours | *July 8th 2024* | *July 8th 2024* |
| Deliverable 1.5.1 | 15 hours | *July 9th* | *July 10th 2024* |
| Deliverable 1.5.2 | 1 hours | *July 10th 2024* | *July 10th 2024* |
| Deliverable 1.6.1 | 6 hours | *July 11th 2024* | *July 11th 2024* |
| Deliverable 1.6.2 | 2 hours | *July 11th 2024* | *July 11th 2024* |
| Deliverable 1.7.1 | 6 hours | *July 12th 2024* | *July 12th 2024* |
| Deliverable 1.7.2 | 2 hours | *July 12th 2024* | *July 12th 2024* |

## B.5 Resources and Costs

1. Laptop (already owned): No cost
2. GitHub: $4.00
3. Windows OS: No cost
4. WSL/Unix: No Cost
5. Miniconda: No Cost
6. Python: No Cost
7. Scikit-learn: No Cost
8. Pandas: No Cost
9. Polars: No Cost
10. Seaborn: No Cost
11. Matplotlib: No Cost
12. Jupyter Notebook: No Cost
13. 100 work hours: $2,000 ($200/hour)

## B.6 Criteria for Success

For this project to be a success, the number of influential users delivered must be one thousand or less. The users should be the most significant and most impactful. The final dataset can be evaluated using the Mann-Whitney U test to determine if the outliers are greater than the non-outliers with a p-value of less than the alpha 0.05. This will be the right tail version of the test.

# C. Design of Data Analytics Solution

## C.1 Hypothesis

Analysis will reveal a short list of anomalous users with the most influence on the Steam platform.

## C.2 and C.2.A Analytical Method

I will explore the raw data and try to understand what is going on and what I need to do. I will summarize the user data set, explore, and clean. I will create a single PCA component from the relevant features. There are over 14 million users in the summarized data set. There are multiple columns I will need to rank on, and it would be too much trial and error without a method. I will want the right tail of this data.

I will take everything to the right of 95th percentile using that single PCA component. We know we want users in the right tail. It would be appropriate to choose users we know we want to target, so that my algorithms downstream can run faster. From here I will analyze the data for the most useful features.

I will then train an Isolation Forest algorithm with the appropriate contamination rate set and features to obtain enough users for our final sample and label them as outliers or not. Isolation Forests are a good algorithm for finding outliers in data. We are most interested in the extreme outliers. The contamination rate we choose will be appropriate to find enough outliers to satisfy our requirements.

I will finally use the Mann-Whitney U Test to compare the outliers with the non-outliers. This is an appropriate tool to compare two skewed distributions. A p-value of .05 or less will give strong evidence in support of the alternative hypothesis. I will perform this on the single PCA component as the final compare. However, for completeness, I will run the test for each feature independently.

## C.3 Tools and Environments

* **Windows 11**: This is the main operating system of my laptop.
* **Windows Subsystem for Linux (WSL) with Ubuntu 20.04.6 LTS**: I want to isolate my process from my main OS. I will run the project in a Linux virtual machine.
* **Miniconda**: This is a free installer for a minimal version of conda. This is a package and environment management system. We will want to create our conda environment separate from any others. We will also need it to install Jupyter Notebooks and other python packages.
* **Python 3.10.14**: This is the python version we will use for this project.
* **GitHub**: GitHub is where we will keep all our work. We need some place to stored our code and changes and be able to roll back to a previous version if needed.
* **Jupyter Notebook**: This tool helps present your data process, visualizations, and run most any python script needed for this project. Our EDA and final file creation will be all here.
* **Matplotlib**: For visualization of data.
* **NumPy**: It has many mathematical functions that will be useful for data analytics.
* **Pandas**: It helps read and store data in a structure that is ready to be consumed by python.
* **Scikit-learn**: This is a popular machine learning library for python, and we will use it for PCA, t-SNE, and Isolation Forests.
* **Seaborn**: This extends Matplotlib with a richer feature set and a much better presentation.
* **SciPy**: This has many uses. Among them is the test we plan to use to compare to skewed distributions. Another is the bootstrapping we will use to simulate normal distributions for our skewed data.
* **Polars**: A faster implementation and alternative for large data sets. It will be used to import the raw data.
* **OpenPyXL**: Needed to write our file to Excel. It extends Pandas and adds an Excel method to export the data frame.

## C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

**Standardization**

* Unsupervised statistical method for scaling data. Fitted before use of PCA.

**PCA**

* Dimension reduction unsupervised.
* Principal Component Analysis algorithm
* Features included in the fitting of the PCA component will rely on correlation metrics of less than 0.60 and removal of features that don’t make sense. Feature importance can be viewed to see how each feature is weighed. We expect social features to be heavily weighted and will remove features that are not adding value. We will repeat this process until we get a valid feature selection.
* No benchmark for this. In a different scenario, the number of components could be determined by the amount of variance explained by the components. This could be viewed using a scree plot, but this is not necessary for this application as we only want one PCA component to decide the tail of the data.

**Percentile Method**

* n = (P/100) x N where P=percentile and N = number of values in the dataset sorted from greatest to least using our single PCA component.
* We will use the percentile method to sample everything right-of and including the 95th percentile.
* We will measure the count of the sample size to ensure it is large enough.
* Greater than 10,000 should be sufficient.

**Isolation Forest**

* Unsupervised
* Isolation Forest algorithm
* We will measure the count of the outliers
* between 500 and 1,000 users should be sufficient.

**Mann-Whitney U Test**

* I will use the Mann-Whitney U test to validate that my sample of outliers is greater than the non-outliers.
* The metrics computed from the Mann-Whitney U test are the U-Statistic and the p-value.
* I will use an alpha of 0.05 to measure this.
  + H0: There is no difference in the means between the outlier users from the non-outlier users.
  + H1: The outlier user means are greater than the non-outlier user means.

Standardization is recommended for PCA. PCA operates under the assumption that the data is normally distributed. Our data we are sure is not. PCA is necessary here to gain a single value that we can judge variance by and sort. We are looking for extreme cases of user activity, but our original sample is so large that it will not be easy to visualize or run through machine learning algorithms. We also don’t need anyone with low activity.

Once we have our user sample, we can use Isolation Forest to separate out the strong cases from the others. To ensure we get enough users, we will have to set the contamination parameter correctly. Isolation Forests is a well-known method for discovering outliers in a dataset. Once we visualize the data with PCA and t-SNE with two components, we will hopefully see a clear separation of data. More importantly the outliers should be more on the right side of the distribution.

## C.5 Practical Significance

Our solution will provide the answers the client is expecting. It will reduce the sample users to a size that the client can handle and can begin marketing to. Follow-up metrics by the client can provide insight into the success of marketing efforts. PCA and Isolation Forests are well documented and are used in many applications and thus been proven. Comparing our outliers with non-outlier’s distributions is common practice and can be evaluated using the Mann-Whitney U Test.

## C.6 Visual Communication

**Rubric C6:** The submission describes key details about *each* tool and graphical representation that will visually communicate the findings of the data analytics solution, and the described tools and graphical representations will effectively communicate the expected findings.

Task 3, the Project Report, must include graphic visualizations (at least two) for visually communicating elements of your project (see Task 3: G2). Describe a plan to include at least two visualizations of the data, statistical test(s), or model(s). Specifically, name the types of graphs, what they will visualize, and the tools you’ll use to generate the images.

# D. Description of Dataset

## D.1 Source of Data

**Rubric D1:** *Each* source of the data is correctly identified.

Identify each data source. The minimum number of data sources is one.

## D.2 Appropriateness of Dataset

**Rubric D2:** The discussion provides reasons why the dataset is appropriate for the stated goals of the project.

Describe why each data source provided in section D.1 is appropriate for supporting the research question or organizational need from section A.1.

## D.3 Data Collection Methods

**Rubric D3:** The described data collection methods are thorough.

Describe how each data source listed in section D.1 was collected, e.g., “the data was collected by downloading the .csv file from www.kaggle.com/data\_source\_link.html.”

## D.4 Observations on Quality and Completeness of Data

**Rubric D4:** The summary includes logical and accurate observations on *both* the quality and completeness of the data.

Describe both the quality and completeness of the data and any accommodation needed. Often, data is already clean and complete, but it is still necessary to comment on both.

## D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

**Rubric D5:** The discussion accurately addresses the data governance, privacy and security, ethical, legal, and regulatory compliance considerations, and *all* of these relate to the dataset and the proposed project.

Specifically, address how *each* of the following relates to your data and project:

* Data governance
* Privacy
* Security
* Ethical, legal, and regulatory compliance considerations

**Rubric D5A:** *Each* described precaution includes specific details about working with and communicating about the data, and there is a precaution described for *each* of the considerations discussed in part D5. *Each* precaution reasonably manages the risk associated with the considerations discussed in part D5.

Describe any necessary precautions. In cases where an item is not relevant, you must explain why. You only need to discuss measures for handling human data if you collected that data.

References

Anton Kozyriev. (2023). Game Recommendations on Steam [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DS/2871694>Scribbr. (2022, December 21).

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.

Elbaghazaoui, B. E., Amnai, M., & Fakhri, Y. (2022). Data Profiling and Machine Learning to Identify Influencers from Social Media Platforms. Journal of ICT Standardization, 10(2), 201–218.

# Appendix A

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix B

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix C

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix D

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.