**Data Analytics Capstone Topic Approval Form**

**Capstone Topic Approval Form**

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information*.*

**Student Name:** Shawn Watts

**Student ID:** 00126321

**Capstone Project Name:** Anomaly Detection to Find Influential Users

**Project Topic:** Utilizing anomaly detection to find influential users based on product reviews and recommendations.

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

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

**Context:** 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 hopes is that this will help game campanies 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.

**Data:**

All data sets are located here on Kaggle under the CCO Public Domain license agreement:  
Anton Kozyriev. (2023). Game Recommendations on Steam [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DS/2871694>

* Games.csv is a tabular file that contains the title of the games, date released, which platforms they support, rating category, number of user reviews and pricing information.
* Games\_meta.json is a JSON file containing the app\_id, game description, and genres within a tags list.
* Users.csv is a tabular file that has a deidentified user id along with how many products they have purchased, and how many reviews they published.
* Recommendations.csv is a tabular file that has the most records which identify whether or not the user recommends the game, how many people found the recommendation helpful or funny, the review date, and number of hours user played the game.

**Data Gathering:**

Data required will be downloaded from Kaggle. Datasets pertinent to this study are provided by Valve Corporation for their Steam game distribution platform. Four datasets will be downloaded and combined with appropriate features relevant to the study.

**Data Analytics Tools and Techniques:** The dataset has over 13 million users. We will first perform outlier detection to discover the users with the most reviews using the IQR method and taking the upper bound. Those will be mined to understand how others vote with them. Finally, the data will be transformed and run through several unsupervised machine learning anomaly detection models to pick the best performing and uncover the users with the highest influence.

**Justification of Tools/Techniques:** *Explain why the data-analysis technique you chose is an appropriate technique to analyze the data collected*. With over 13 million users, we really only care about the users with the most activity. The IQR method is used as a pretty good rule to find these users. Once we have them narrowed down, engineering features will be much easier. Once the data has been appropriately wrangled, it can go through the machine learning algorithms without issue. The correct algorithm will need to be chosen. There are several unsupervised methods to choose from such as, we will use three popular algorithms: Isolation Forests, KNN in unsupervised mode, and Local Outlier Factor. Once discovered which performs the best, use that model to predict the most influential users. These will be the upper echelon of the users with the most influence, thus we are looking for the anomalies here.

**Application Type, if applicable (select one):**

mobile

web

stand-alone



**Programming/Development Language(s), if applicable:** SQL, Python

**Operating System(s)/Platform(s), if applicable:** Windows, Linux (WSL)

**Database Management System, if applicable:** Microsoft SQL Express

**Project Outcomes:** Likely outcomes will be a lis of influential anomalous users that can be targeted for study by Steam and game developer marketing organizations.

**Projected Project End Date:** 7/31/2024

**Sources:**

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

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.

**Human Subjects or Proprietary Information**

Does your project involve the potential use of human subjects? (Y/N): **N**

Does your project involve the potential use of proprietary company information? (Y/N): **N**

**STUDENT SIGNATURE**

**­­­­­­­­­­­­­­­­­­­­­­­­­­­­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**



**By signing and submitting this form, you acknowledge** that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

**TO BE COMPLETED BY AN INSTRUCTOR**

**The capstone topic is approved by an instructor.**

**INSTRUCTOR’S NAME AND SIGNATURE:**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**INSTRUCTOR APPROVAL DATE:**

**Project Compliance with IRB (Y/N):**