**Machine Learning Final Assignment**

Movie theaters are slowly dying in attendance. 61% of Americans have not attended a movie in a theater in the past year.

Monetary State University has over 60,000 students. The college has a conducted a survey and realized that most students are commuting over an hour in travel time to be able to view movies at the nearest cinema, additionally most of the students have indicated that they do not go to cinema because they cannot afford it and they have no mode of transportation.

To provide this valuable source of entertainment for their students MSU has opted to build an in-house cinema utilizing Netflix as their temporary streaming source for the media entertainment. The plan for the college is to collect the data from the Netflix views to gauge what movies/shows attract students to begin their own in-house program and make a profit from that instead of relying on the Netflix catalogue.

To solve this problem. MSU’s film department has collected a data sample of Netflix’s movies and TV shows and plans to run a yearlong program for the students. The data the college plans to collect and study is what shows have the highest attendance, which genres bring in the most students? This information will be based on the audience trailer views conducted on a weekly basis as well as the IMDB rating of the movies. The university must determine if there is a correlation between these two factors to be able to establish what shows they will use to draw in a mass wave of students, additionally the college must factor in runtimes of the shows to be able to show the maximum amount of quality products in a day and charge an affordable admission price.

To be able to achieve their goals MSU will have to understand the market segmentation of film attendees.

When it comes to movies, analyzing text taken from a script is limiting because it only provides a skeleton of the story, without any of the additional dynamism that can entice an audience to see a movie. A trailer release for a new movie is a highly anticipated event that can help predict future success, so it will behoove MSU to ensure the trailer is hitting the right notes with moviegoers and this data metric can be measured by the number of views the trailer generates in a specific time period.

The first step MSU will have to take is to determine which machine learning tool they will utilize for the greatest success. Cloud ML Engine and TensorFlow are two of the best but expensive options.

Cloud ML engine will allow the university to focus on further building the deep learning model for the program, because it automates the monitoring and provisioning. This will give the university a much more in depth understanding of the entire process. A data science team will be sufficient in monitoring the day-to-day maintenance of the ML engine, as it both easy and simple to understand.

To start its analysis of the data, Monetary State University can utilize YouTube 8M (a free/public available dataset of YouTube videos). The University can release a trailer for one of their own films at a specific time and day and begin its analysis for whichever time period they prefer to use. The YouTube 8M dataset comes with a pre-trained model from Google. The pre-trained model comes with the ability to analyze specific video features such as illumination, color, faces, landscapes, and thousands of objects.

The first step in the design architecture is to filter out the predefined characters to determine which elements of the released trailers are most predictive of the potential customers’ preferences.

Example, the movie blockbuster action movie “Logan” stars Hugh Jackman as Wolverine. Now it’s the university’s job to determine if a student who watches mostly male action lead movies are they still likely to continue that trend and see another male action lead flick?

Continuing with the Logan example. Fox studios compared the analysis of that movie trailer with other trailers of similar genre and tone and identified two challenges. The first is that it matters when the label of movie makes its first occurrence in the trailer. The 2nd is the high dimensionality of the data, for each movie there are hundreds of elements that can gauge the interest of an audience, and the ML engine must analyze all of these simultaneously. The great advantage of the Cloud ML Engine is its elasticity which allows the data science team to iterate and run the tests quickly and efficiently without compromising the integrity of the deep learning model. This enabled the system to become a production-ready tool within a matter of days, something that the university will need because a semester is only three to four months.

A brief overview of the entire ML engine process:

The data science team develops a neural network which the analysis pipeline feeds all the components of a trailer, in this case the label of the film. The model begins to learn the temporal sequencing (Example short vs long shot) of labels in the trailer. The sequencing is then able to convey data about the movie type, character roles and plot as well as the director’s cinematography choices. Then the data scientists will combine this new data with the consumers historical data to predict the customer’s behavior. Additionally, the ML engine includes a logistic regression layer and distance-based collaborative filtering model, which can combine all the model outputs to predict the probability of an audience attendance for each movie the university plans to release. The data team then refreshes the data pipeline weekly to add in the new trailers released. The final step is the university using BigQueryML to merge millions of the customer predictions with other data sources which enables them to draw up useful reports to efficiently prototype media avenues for marketing campaigns.

Using the Logan movie example below is the data prediction for customers who were also inclined to see a male action lead film. The top 5 movies were predicted to be true and did occur to be true as well. This is the information the university will need to collect and study to ensure a successful theatrical campaign.

Graphical user interface, application, table

Description automatically generated

**The Approach**

I believe the K means clustering will be the best approach for Monetary State University. This theorem is utilized in many real-life applications and comes with multiple advantages such as It can be used for binary as well as Multi-class Classifications. As the movie theaters starts out small and grows, it will generate more and more movie categories and will need to be able to make quick predictions on which shows to release.

K means clustering is an unsupervised learning algorithm that has the capability to group the dataset into multiple clusters. For my data set I will be utilizing two clusters to divide the dataset.

Additionally, I will also use the silhouette score to measure the performances of the clusters.

By clustering several consumers, the university will gain an even more in-depth knowledge of which movies a general audience will be more predicted to come watch, instead of trying to cater to each consumer’s selection which will be impossible for a theatrical release.

To make predictions using cluster will require averaging. The university will collect data of the same category from a certain number of students and use the average of those selections to be able to make their screenings.

Below is an example of a study conducted on if a film audience preferred “The Matrix” or the “The Matrix 3” Clustering is a key component of widely used recommendation engines embedded into the products of Amazon, Netflix, and many others that millions of people use every single day

Chart, line chart

Description automatically generated

The reason Monetary State University will be utilizing Netflix is that the algorithm has such an impressive accuracy that at least 80% of all audience streams on its platform are solely due to the recommendations from the data gathered and studied.

The netflix algorithim is integrated with its user interface and splits recommendations into categories such as “similar to what you watched.” The algorithm behind the user interface dives deeper and captures different data such as when the consumer abandoned the show/movie and the time passed since the consumer viewed the show/movie. Studying this data will give the university team a broader of understanding of what college aged adults are looking and willing to pay for.

Netflix also utilizes a two tier ranking system which the university may look into adopting as well. Within each row the strongest recommendations are listed on the left and across the rows the strongest recommendations are listed on top. There are two advantages to this system. For a consumer it is easier to understand when the rows of similar shows are presented together, enabling them to decide if they are interested in viewing that specific category. For the university it is simpler to collect data/feedback as scrolling down and right scrolling would indicate the viewer is not interested in that category.

Below is a table for two of Netflix’s biggest movies the Greatest Showman and Love Simon. Again you can see the prediction model was utilized using the collected data to determine an audience’s interest in both movies.

A screenshot of a computer

Description automatically generated with low confidence

**Admission and showtimes**

Admission tickets are based on several factors:

* Format: Which formats will university display 2D, 3D and even IMAX?
* Sound: Does the university plan to install the highest quality sound speakers such as Dolby Atmos?
* Seat capacity and seat style: How many seats do they plan to install? Will these also be reclining seats?
* Discount days: Does the university plan to run a half price day like most theatre companies.
* Advertisements: Companies looking to lure in college students to buy their products and services can offer financial compensation to the university to run ads.

So, while the university will have to recoup its finances of building the theatre, they must still be able to attract college students who don’t have the biggest financial resources.

Showtimes: As the general audience will be college students the university will have to factor in class scheduling. Most students have classes in the mornings and afternoons, while graduate students are mostly in the evenings. The best approach to this is for the university to send out a survey to every student who is interested in the film program and have them fill out their class schedule on that survey. The university will then base their movie showtimes off that schedule.

**Additional algorithm models which MSU can consider:**

Monetary State University does not have to adopt the Netflix two row ranking based algorithm model, there are much more options such as:

* Personalized Video Ranking: An algorithm which filters down the catalog by a certain criterion (US TV shows, comedy, and gore), and certain side features including user popularity.
* Top-N Video Ranker: Like PVR except that it only looks at the head of the rankings and looks at the entire catalog.
* Trending Now Ranker — This algorithm captures temporal trends which Netflix deduces to be strong predictors. These short-term trends can range from a few minutes to a few days. These events/trends are typically seasonally (Christmas/Halloween), short term (disasters) and episodic content (Drama 22 episodic shows), non-episodic (Shows in which episodes are independent from each other such as Hawaii five 0)

A relatively unknown algorithm based on a presentation conducted by Justin Basilico involves the use of RNNs in a limited time sequence prediction. It was devised that a consumer’s previously watched conducted alongside the contextual information can be used to predict what the consumer might play next.

The machine learning approach is simple but effective. The company will aim to create a ranking function by training the system model to use historical information from the data gathered from their consusmers, which includes what they watched and how they watched it, and using that to find new customers and collect their data and historical patterns over a time in order to gain new consumers in one cycle as seen below.

Graphical user interface, text, application, chat or text message

Description automatically generated

In today’s world data analytics is not a mandatory expensive study. The university will have access to multiple resources of freely available big data online which they can utilize to unlock doors to a wider range of the student population and give them the best potential at limitless success as there will always be college students so long as they are colleges to attend. In addition, the university can also investigate bringing in current data science student majors to alleviate costs and gain more in-depth knowledge as to what students are looking for, as well as provide the students real practical training they can apply to their resume and later job experiences. So, while there are multiple approaches to the prediction theory, the k means clustering will be the best approach for this case as it will give the most accurate results.

**Results & Conclusion**

* **Graph of the optimal k valueChart, line chart

  Description automatically generated**
* **Graph that represents the genres of cluster 1 vs cluster 2**

**Chart

Description automatically generated**

Both clusters represent a strong interest in the students love for action and thriller genre movies. A key area for the university to focus on. This graph also shows that cluster 1 contains the most interest in each genre with Action being the highest. Analyzing the data, it shows that even though Cluster 2 had more audience views for the movie trailers. The trailers in cluster 1 were of higher quality and that's what drew in the students to those films.

Interpretation of the data:

Based on the findings of the data, The IMDB ratings are highest in the cluster one group which is indicative of why most genre categories have the highest view in cluster one, because Cluster one contains the best selection of movie choices. The university can base their approach on which movies to stream not only on the genre of well-liked movies but also the IMDB ranking of those movies.

References:

* <https://www.thesmartcube.com/resources/blog/datawatch-lights-camera-analytics-the-secret-role-data-plays-in-movie-making/>
* <https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48>
* <https://raindance.org/big-data-analytics-filmmaker-future-proofing/>
* <https://cloud.google.com/blog/products/ai-machine-learning/how-20th-century-fox-uses-ml-to-predict-a-movie-audience>
* <http://cs229.stanford.edu/proj2008/KammHuangSathi-TheNetflixChallenge.pdf>
* <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>
* <https://www.marketplace.org/2019/03/14/peek-behind-curtain-movie-ticket-prices/>
* <https://news.gallup.com/poll/388538/movie-theater-attendance-far-below-historical-norms.aspx>
* <https://www.scienceabc.com/innovation/netflix-get-right-content-time.html>
* https://recoai.net/netflix-recommendation-system-how-it-works/