OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS

EXAM ASSIGNMENT COVER PAGE

Module Name and Code: **Decision Theory and Analysis (MS5104)**

Student Name and ID: **Dilip Venkatesan Sankar 22225743**

"In submitting this work I confirm that it is entirely my own. I acknowledge that I may be invited to undertake an online interview if there is any concern in relation to the integrity of my submission."

Overview:

In Europe, **Vidflix** is a popular over-the-top (OTT) video service for offering streaming services. The company classifies its user based on the genres they watch, and we have information on ten categories of users, and this is achieved through analytics.

Although Vidflix dominates the market in Europe, it has a relatively modest market share in the US and is working to grow its use and popularity. The managing directors of Vidflix are intensively looking for external solutions and currently are in touch with an outside vendor calling FilMine to offer them a solution to draw customers from the US and Europe by suggesting movies and television shows based on their demographics and launching new services.

The start-up company **FilMine's** suggestion mechanism to address the issue is used by Vidflix's main rival in Europe. The business analyst for FilMine stated the following:

1. "**We already know which are the right variables to select and build a new and accurate AI-enabled film recommendation system**"
2. "**Even after Vidflix selects to purchase our solution, we will not reveal how our algorithm works as this is a company secret**"

Question 1: The CEO of the company asked your opinion on FilMine's BA head statements.

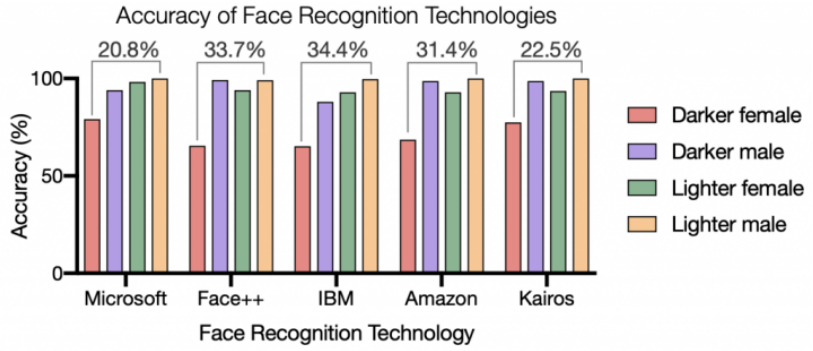
Answer: Vidflix's primary goals are to expand its US services and offer new features and services to both the US and Europe. The required algorithm and created system recommendation are available from an outside source by the name of FilMine, enabling Vidflix to accomplish its goal. However, the BA at FilMine has released a few statements and the CEO of Vidflix is looking for opinions and recommendations which will help her to decide.

Statement 1: "We already know which are the right variables to select and build a new and accurate AI-enabled film recommendation system"

Based on the above statement from FilMine BA, we can see that **biasing** is present. As humans, we have several biases that might alter our perspective and lead to poor decisions. **"The major cause of making the wrong decision is called Bias".** Biased data indicates that the complete dataset was not taken into account and examined. For instance, Racial Discrimination in Face

Recognition Technology demonstrated a bias in classifying gender. It demonstrated prejudice and inconsistencies in categorizing the precision of facial recognition based on various skin tones and sexes.

The graph below demonstrates how the algorithm created by IBM, Microsoft, etc. performed worse at recognizing darker-skinned people than lighter-skinned people.



There are 5 types of biasing involved in data analysis,

Selection Bias: When using samples that are not typical of the population, selection biases develop. This indicates that the data were chosen arbitrarily and that the sample was not a representative sample of the population. In most cases, it occurs due to selecting and working with data that is easy to access

Stability Bias: Inertia prevents effective learning in an unpredictable setting. When there is hesitation to include fresh data for analysis, some of the prospective data is excluded.

Interpretation Bias: Tendency to analyze ambiguous events and stimuli in a positive or negative fashion

Emergent Bias: To construct algorithms based on historical data. This can sometimes be a negative as the behavior can change over time

Confirmation Bias: Usually, it concentrates on presumptions. Some of the variables are left out of the study in order to support the assumption.

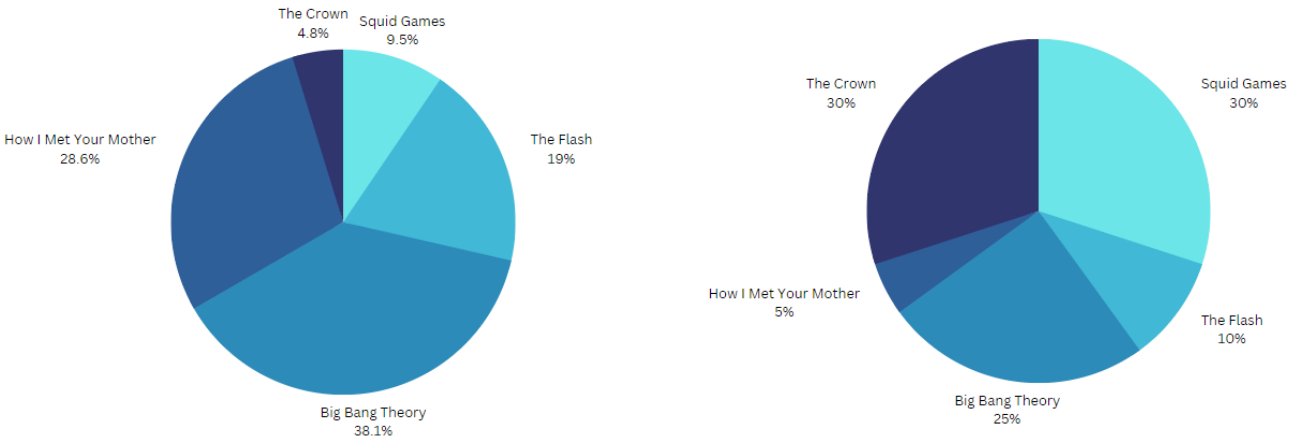
Based on the statements from FilMine's BA we are looking into Selection Bias and Confirmation Bias to determine how they came into the picture.

Selection Bias: The third-party provider FilMine claims to have created a recommendation system using an algorithm that is appropriate for Vidflix. The product and usage are only restricted to Europe, despite the fact that FilMine is aware of the appropriate factors to be chosen for the AI-enabled recommendation system. For Vidflix, this is a huge setback because we are not sure how accurate the same algorithm prediction might work in the US. Because preference may differ depending on region, the company cannot employ the same algorithm in the US.

For example, A tv show like 'The Crown' will be preferred by people in Europe as it involved the life of Queen Elizabeth II which is of no interest to the people in the US whereas a series like 'Mr Robot' will interest the people in the US as it involves crimes and conspiracies related to the US.

Even though FilMine claims to have the right variables for a recommendation system, it has only been deployed in Europe and used by Vidflix's competitor which is based in the Europe region, and it has not taken a global perspective which leads to **selection bias**. The FilMine algorithm offers the possibility of more growth and success in the Europe region, but there won't be much of an accuracy of prediction and impact in the US region because they can receive irrelevant information. In order to overcome selection bias, FilMine should consider the entire population (US+Europe) instead of focusing on the subset (Europe).

<u>Preferences of Tv shows in the US vs Europe</u>

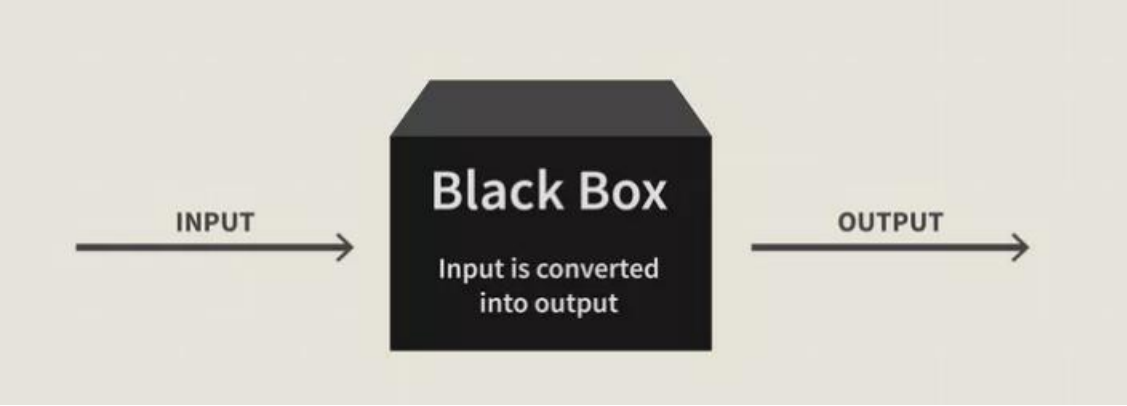


Confirmation Bias: The goal of Vidflix is to expand its market beyond Europe and into the US. The main rival of Vidflix has adopted the FilMine approach, but its deployment focuses only on the European market focusing on specific consumers by eliminating factors. FilMine is confident in its ability to offer the proper recommendations based on the presumptions, but since the algorithm is once again only employed for the European region, it won't be the best choice in the US because of their very different preferences and point of view.

For example, from Fig 1.1 we can see 'How I met your mother' is preferred less in the European region than in the US as it deals with the lifestyle of the people in the US. Thus, it is impossible to apply the same procedure to both sections. Instead of concentrating on past successes, all the aspects and variables should be taken into account in order to overcome confirmation bias.

Statement 2: "Even after Vidflix selects to purchase our solution, we will not reveal how our algorithm works as this is a company secret"

We can deduct from the aforementioned statement that FilMine is prepared to offer the solution without disclosing how its internal operations even after the purchase. It's known as **Blackbox**. The Blackbox receives information, processes it, and outputs something, but the method is kept secret,



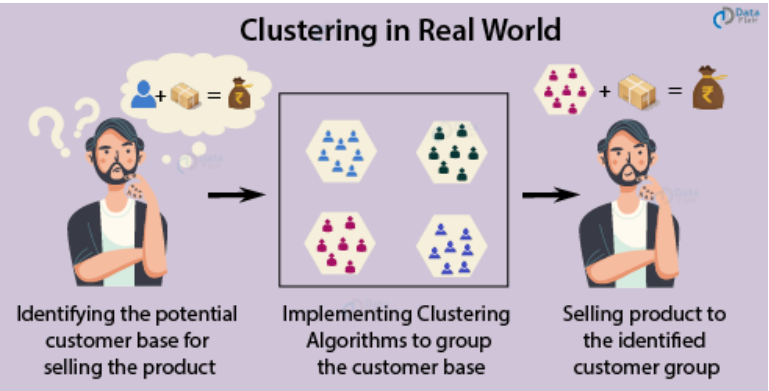
Even though FilMine treats its algorithm as a trade secret, Vidflix can use this to further increase its market share in Europe. However, the algorithm needs to be altered and tested to advertise it on a worldwide scale, which will demand more time, resources, and investment. With limited flexibility and access, Vidflix can only act as an observer and inform the changes required which

cannot be achieved without extensive usage of time and resources, so this needs to be discussed before buying the product. Another crucial element we must take into account is how rapidly users' behaviors change. As a result, Vidflix needs flexible control over the algorithm to be able to make frequent modifications depending on ongoing observation.

Conclusion: According to FilMine's claims and Vidflix BA's analysis, the working model may be a success in the European region. If Vidflix wants to further increase its market share in Europe, it can buy this product. However, if it wants to expand to the US, it will need to make changes to the algorithm based on regional preferences, which will require more time, money, and resources. Additionally, Vidflix will only be given very limited access to make any changes since it wants to maintain the concept as a "**Blackbox**". And this will be a significant downside because user preferences, regardless of region, vary frequently, and the model might need frequent revisions to consistently make more money. Before buying the product, the CEO of Vidflix should consider all of these issues and speak with FilMine about the additional requirements to ensure that all parties are on the same page.

Question 2: How can Vidflix further exploit the datasets they already have to provide value to their customers and support decision-making?

Answer: By examining the current information, we can deduce that the data is created based on the Type of User, Series preferences, and Personas, which suggests that **clustering** is occurring. Vidflix is attempting to increase the accuracy of its forecast. This kind of learning, also known as unsupervised learning, often uses unlabeled data.



There are two types of clustering,

1. K- means — Partitional clustering
2. Agglomerative — Hierarchical Clustering

For the understating of clustering, we are assuming that each user ID, movie ID, demographics, and each movie in the dataset has a rating given by each user.

K-means – Partitional clustering:

This method classifies the groups based on their similarities and characteristics of data where k represents the number of groups. In this method, each cluster is represented by its **centroid** which corresponds to its mean which helps group people (according to different criteria such as willingness, purchasing, interest, etc.)

For example, which series will the user1 watch next given that he has seen and rated 'Squid Games' previously? Which is also seen and rated by another user. In this scenario, the algorithm will group users with similar choices where the similarity is measured based on the distance between two points. Below are the steps based on assumptions,

1. In order to know how much is user1 likely to watch 'Stranger Things' calculate the query instance of q(user1, Stranger Things) with mean and standard deviation. Let's suppose user 1 has watched 'Squid Games' and has rated the series as 5 stars then user 1 will belong to the '5 scale' group in 'Squid Games'.
2. The next step is to calculate the distance (Euclidean distance) between user 1 and each center of the group in the 'Squid Games' domain which determines the closest neighbor, so it will decide which cluster the user belongs to.
3. However, in some cases, some individuals may be conservative, and a few may be very generous, about the movie rating, thus it's vital to normalize the data to prevent inaccurate predictions due to personal bias.
4. The final step would be to calculate the prediction output using the below formula,

   (distance * Standard deviation) + Mean ratings.

So based on cluster assignment and value, the algorithm will recommend the best-rated movies to user 1.

Agglomerative — Hierarchical Clustering

In this approach, each user is given a cluster, the similarity (or distance) between each cluster is then calculated, and the users are then combined into a single cluster.

Let us assume that each user watches different series are assigned to an individual cluster based on demographic and genre. So, we can construct large clusters from small clusters based on either genre, demographic, or both. For instance, user 1 rates a series in 'Geeks' genre called 'Squid games' then we convert the relationship from **user→series** to **user→genre** and then find out the nearest cluster (k=1) and combine them into one cluster. This can be achieved by either a single linkage or a complete linkage. So, the same series will be present in 'Trendies' genre as well and the algorithm will recommend other series like 'The Sandman' and 'Stranger things' to user 1.

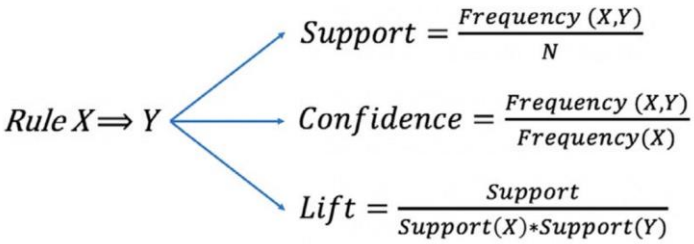
Real World Cases: Due to the primary factor known as a recommendation system, Disney+ Hotstar, Netflix, Amazon, and SonyLiv have had enormous development in the OOT platform for many years. Based on the consumers' demographics and preferred genre, this algorithm recommends TV shows and movies to them. This means that users can watch series or movies based on their viewing history, or if they travel to a new location, their preferences will alter based on the region while taking into account their viewing history.

Question 3: Vidflix's CEO is interested in exploring the following association rule: The Sandman→ Squid Game & Stranger Things

Answer: Based on the provided data, the CEO of Vidflix is looking to figure out on what basis users watching 'The Sandman' also watch 'Squid games' and 'Stranger things'. So the main objective is to figure out patterns, correlations, and associations. Hence Association rules come into the picture.

The Association rule is a process of finding relationships amount huge datasets. It is also called market-based analysis because it helps in figuring out the relationship between the product and how many times the user has purchased it. For example, in a grocery store, a customer who buys bread also buys jam, so the association rule shows 80% of checkout includes bread and jam.

Association rules involve three types,



1. **Support**

   Support gives an idea of how frequent an itemset is in all the transactions. And it is represented using the below formula,

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

   Where X and Y = Items in the dataset

   In our dataset, X = The Sandman, Y = Squid Game & Stranger Things, and the Total number of transactions = 10 because our dataset contains 10 users

   The Support is calculated below,

   Support(The Sandman)=P(The Sandman) = 4/10 = 0.4

   Support(Squid Game & Stranger Things) = P(Squid Game & Stranger Things) = 3/10 = 0.3

   Support(The Sandman→ Squid Game & Stranger Things) = 3/10 = 0.3

2. **Confidence**

   Confidence refers to the likeliness of the same item to consequently occur when the item has existed before. And it is represented using the below formula,

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

   Where Numerator represents 'The Sandman' association with 'Squid game' and 'Stranger things' and the denominator represents 'The Sandman'

   Confidence = Support(The Sandman→ Squid Game & Stranger Things)/ Support(The Sandman) = 0.3/0.4 = 0.75

3. **Lift**

   Lift has a literal meaning which denotes the increase in confidence of Y that is provided by X. And it is represented by the below formula,

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

   The Numerator represents the confidence values, and the denominator represents the support value

   Lift = Confidence / Support(Squid Game & Stranger Things) = 0.75/0.3 = 2.5

Conclusion: From the analysis above and based on the values obtained from Support, Confidence, and Life we can conclude that since the value =2.5 which is greater than 1 it indicates cross-selling opportunity which signifies users watching The Sandman also tends to watch stranger things and squid games.

Based on this we can for example analyze how 'The Big Bang Theory' is associated with 'How I met your Mother' and if the association rules satisfy we can add a recommendation section for the users based on genres which will open doors for new opportunities

References:

https://nuigalway.blackboard.com/webapps/blackboard/execute/content/file?cmd=view&content_id=_3039937_1&course_id=_147476_1

[1]Bias

https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/
https://www.codecademy.com/article/bias-in-data-analysis
https://cmotions.nl/en/5-typen-bias-data-analytics/

[2]Clustering

https://medium.com/@chitu_rk/clustering-algorithm-89f79c456336
https://medium.com/@rohanjoseph_91119/learn-with-an-example-hierarchical-clustering-873b5b50890c
https://www.scitepress.org/Papers/2010/27376/27376.pdf

[3]Association

https://www.geeksforgeeks.org/association-rule/