Magid

Emotion, Engagement and Evaluation Modeling

B412 Capstone Project
George Brown College

Contents of the Report

1: About the Project

- Magid
- Objectives

Magid

- About the Company:
  - A leading consulting and strategy company.
  - Presence in the United States, where it is located, and globally across over 40 countries.
  - services include Market & Consumer Intelligence, Product Development & Optimization

- Business Problem:
  - Find new and innovative ways to recommend, market and substitute content.
  - Keep users within the same platform.

Objectives

Create new genres based on emotional data

Develop a model to predict scores and evaluation

Develop a recommendation engine

2: METHODOLOGY

- Analytical Approach

Methodology

Data Preparation

- Data Audit

Creation of Analytical File

- Streamlining of current variables through correlation
- Creation of New Variables through Correlation and Factor Analysis
- Cluster sentiments
- Customer Segmentation
- Correlation between Sentiments and Ratings

Model Building

- Data Pre-processing
- XDG Boost

Performance Evaluation

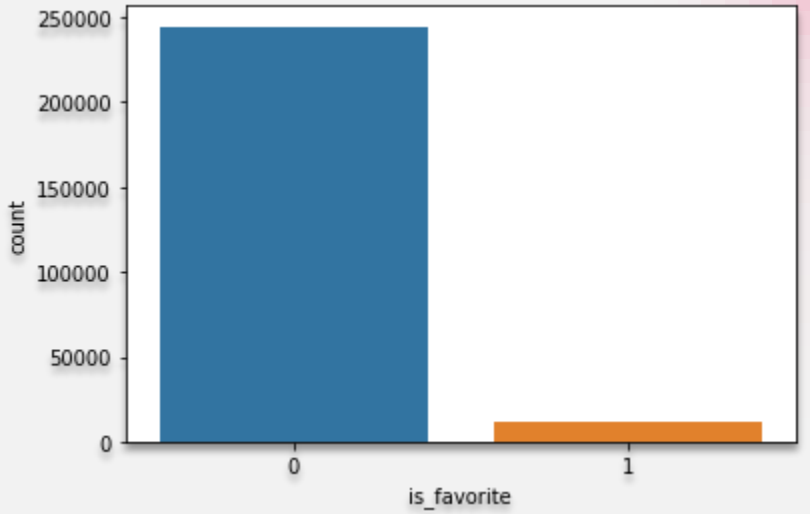
- Cumulative Gain Chart
- Lift Chart

3: Key Findings on Data Audit Results

- Data Audit
- Summary of Key Findings

Data Audit

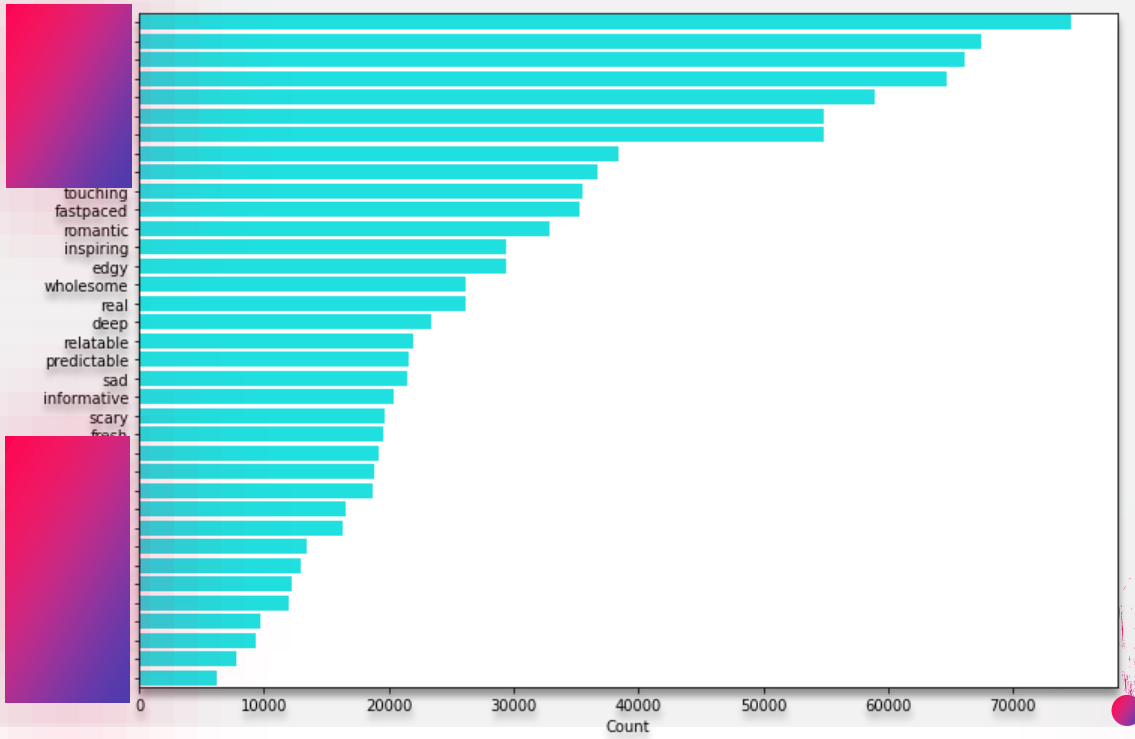
- 3 files were provided – Movie Attributes, Movie Intent and Movie Ratings
- Check missing values, formats
- Understand the data that was provided
  - Client Brief
  - Check Descriptive Statistics
  - Data types
  - Count




```
movie_df.dtypes
```

```
survey_date           object
etl_run_guid          object
content_type          object
respondent_id          int64
movie_id               int64
movie_name            object
wave_id                int64
age                    int64
zip_code               int64
household_size        object
live_with_children      bool
gender_name           object
age_group_bracket     object
```

	age_group_bracket	counts
2	18 to 24	5841
1	25 to 34	6814
0	35 to 44	6895
5	45 to 49	2723
4	50 to 54	3161
3	55 to 64	5234
7	65 and Over	701
6	Under 18	2679



Key Findings on Data Audit

Missing Values

- Education_level 1.2%
- Ethnicity_name 0.6%
- Income_bracket 7.6%

Categorical Variables

- Income
- Age

Understanding the Variables

- User Profile – User demographics (e.g. *gender, age, ethnicity, income*)
- User Intent – User behavior and attitude towards the movies (*home_surfing_channels, home_background, all_soon_as_available, all_no_rewatch*)
- User Response – *is_favorite*
- Movie Features – Movie Sentiments (e.g. *real, sad, informative, suspense*)

4: Creation of Analytical File

- Update Data Frame
- Creation of New Variables

Update the Data Frame

From Key Findings:

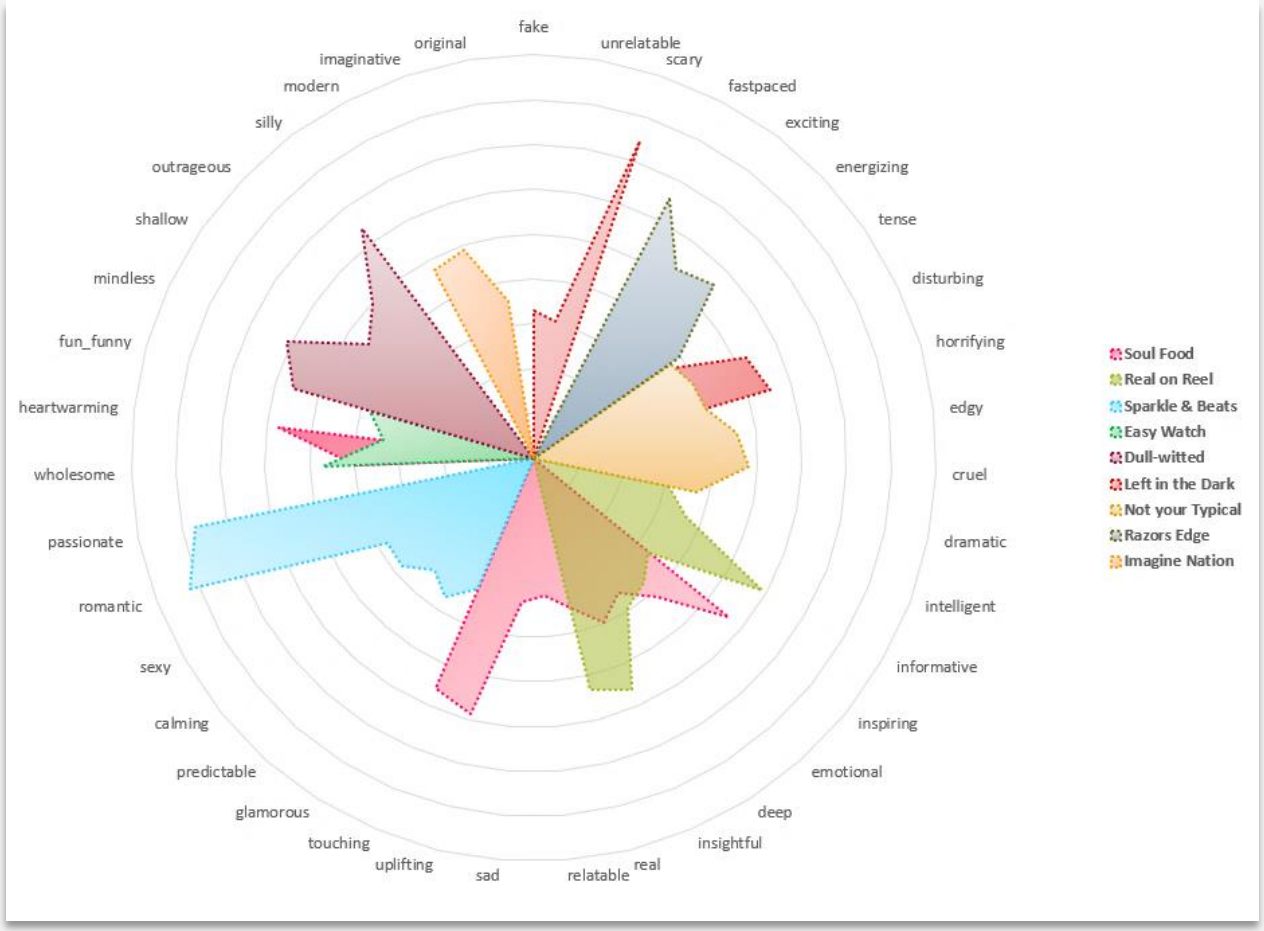
Missing Values	Use MODE to fill-in missing values
Categorical Variables *(ordered variables)*	Transformed into Numerical Variables (Income, Age, Education)
Categorical Variables that can't be transformed into Numerical Variables	Create Dummies (Ethnicity & Clusters)

Creation of New Variables

New Variable	Averaged variables
info_real_int_insp	informative, real, intelligent, inspiring
fun_outrag_mindless	fun_funny, outrageous, mindless
fastpaced_thrilling_exciting	fastpaced, thirlling_susp, exciting

- Merged correlated variables to avoid multicollinearity
- Creation of 8 new variables from **Factor Analysis** using all sentiments using correlation between factors' loading and sentiments

Creating New Variables: Clustering Sentiments



- Created new movie genres through the most related sentiments

- Used K-mean clustering

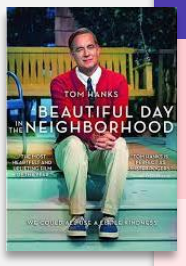
  Problem encountered and solutions:

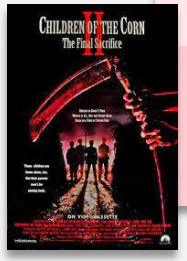
  - *2 romantic clusters were present*
    - ✓ *merge 2 clusters*
  - *Dumpster cluster*
    - ✓ *random forest classifier to redistribute the movies*

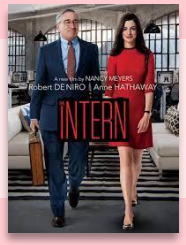
Result: 90% on both training and validation

Creating New Variables: New Genres

Movie Genre	Most Related Sentimets
SOUL FOOD	uplifting, heartwarming, inspiring, touching and emotional
Easy watch	wholesome, picker-upper, heartwarming, silly and uplifting
Sparkles & Beats	romantic, passionate, calming, sexy, glamorous
IMAGINE nation	imaginative, modern, original, exciting and intelligent
LEFT IN THE DARK	scary, disturbing, tense and fake



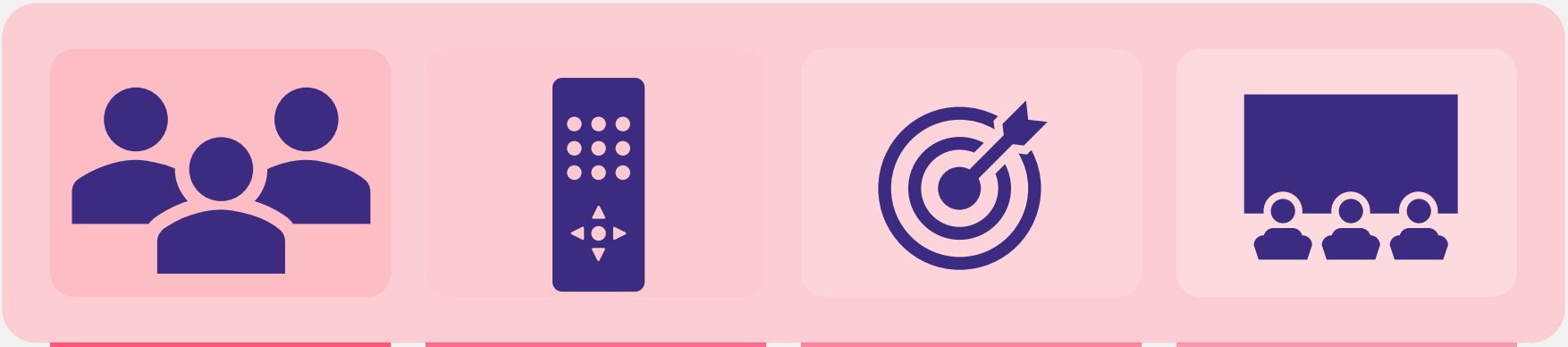




Movie Genre	Most Related Sentimets
NOT YOUR TYPICAL KIND OF MOVIE	cruel, edgy, horrifying, disturbing
ON RAZORS EDGE	fast paced, energizing, tense
DULL-WITTED	silly, mindless, picker-upper, outrageous, shallow
REAL or REEL	informative, insightful, real, deep and emotional

Group 4

Creating New Variables: Customer Segmentation



Social Viewers

Users who talk with other people about the shows they watch

Casual Viewers

Users who end up watching the movie mostly by chance or zapping

Focused Viewers

Users who don't want to be disturbed while watching the movie

Followers

Users who follow the movie relaeses and the movies in general

Summary of Variables Classified

Name	Type	Name	Type
freq	derived		
movie_id	source		
movie_name	source		
gender_name	source		
age_group_bracket	source		
fresh	source		
mindless	source	Final_Cluster	Derived/Target
unrelatable	source	social_people_2	Derived
imaginative	source	casual_viewer_2	Derived
wholesome	source	serious_2	Derived
inspiring	source	follower_2	Derived
shallow	source	edgy	source

Sourced Variables = 41

Derived Variables = 17

Group 4

Analytical Results

- Identify Key Variables
- EDA of Top 5 Correlated Variables
- Model Building
- Key Insights - Analytical Results

Identify Key Variables

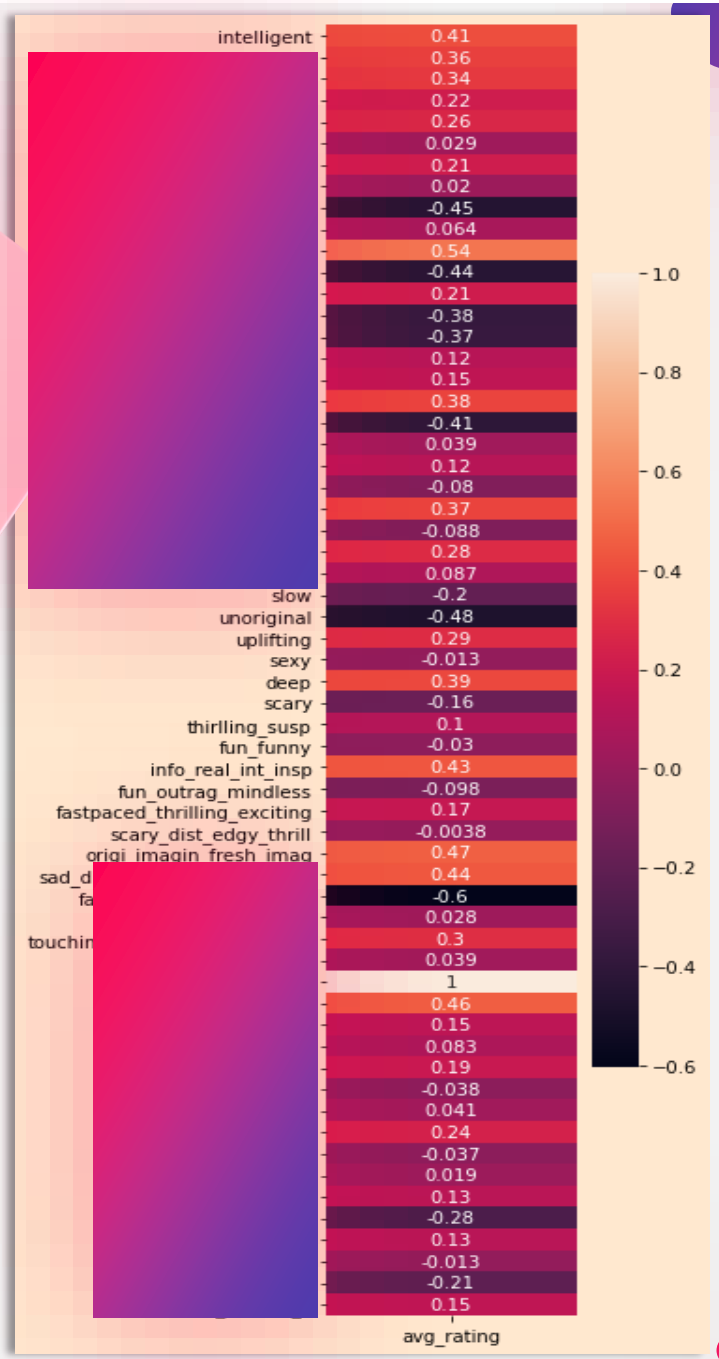
1) Retain and eliminate variables

2) Analyze ***how they felt*** and ***how they watched*** the movie

- Social people
- Casual viewers
- Serious viewers
- Followers
- Frequent users

3) Source and derived metrics were grouped into movies

4) Extract dataset with target variable and all other metrics



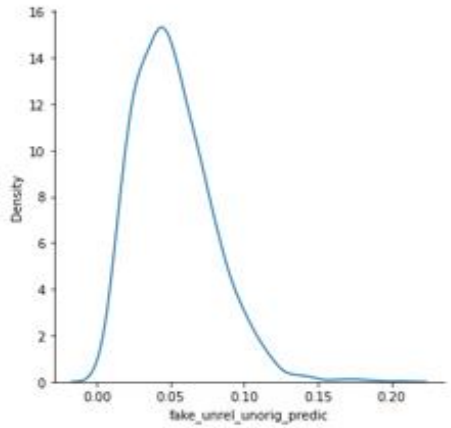
Exploratory Data Analysis (EDA)
Top 10 Correlated Variables

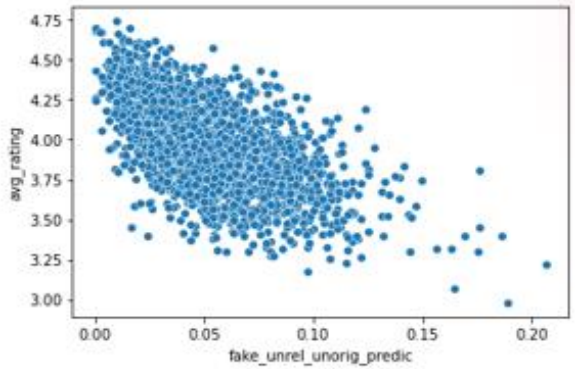
	is_favorite_avg								
count	2618.000000	2618.000000	2618.000000	2618.000000	2618.000000	2618.000000	2618.000000	2618.000000	2618.000000	2618.000000
mean	0.147332	0.111036	0.149760	0.045087	0.162029	0.267257	0.051938	0.036119	0.048129	0.086355
std	0.101229	0.074968	0.085626	0.043564	0.062125	0.099019	0.026871	0.029384	0.039579	0.053627
min	0.000000	0.000000	0.015152	0.000000	0.030864	0.030000	0.000000	0.000000	0.000000	0.000000
25%	0.071429	0.056250	0.089506	0.012821	0.115854	0.195122	0.032051	0.012658	0.020202	0.049383
50%	0.123711	0.091804	0.129630	0.035294	0.153125	0.256098	0.048387	0.030612	0.039604	0.078431
75%	0.202101	0.145408	0.189394	0.060241	0.199356	0.329412	0.067708	0.051546	0.066667	0.113924
max	0.610000	0.527500	0.563291	0.411765	0.458861	0.683544	0.206633	0.181818	0.367347	0.376812

Top 10 Correlated Variables			
original	0.5395		
Is_favorite_avg	0.4575	fake	-0.4476
		unoriginal	-0.4751
		fake_unrel_unorig_predic	-0.6041

Group 4

Exploratory Data Analysis (EDA)
In-depth: Top 5 Correlated Variables

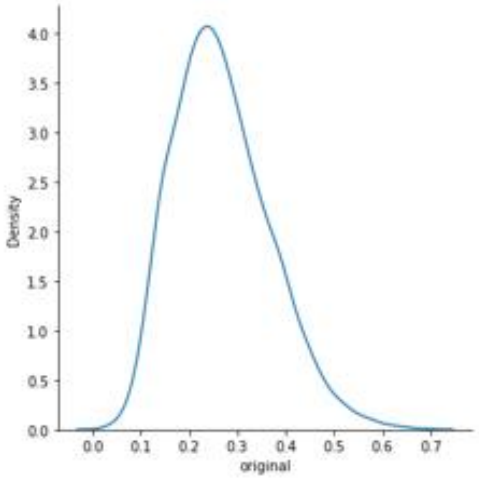


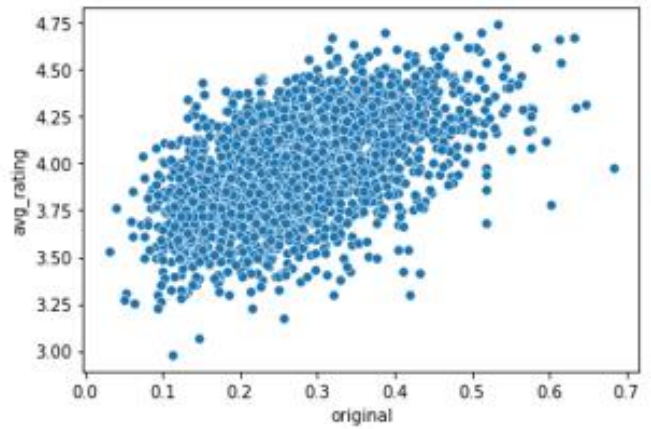


- Skewed to the right frequency distribution; strong negative relationship
- Consider all averaged variables: *fake, unrelatable, unoriginal, predictable*

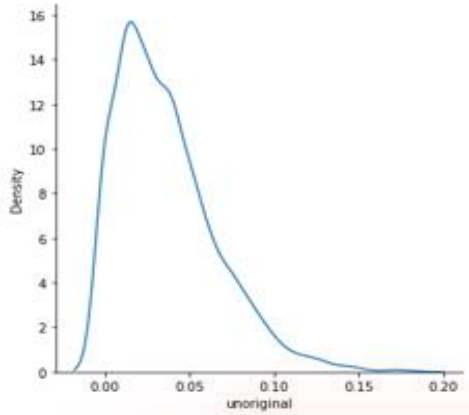
<u>original (0.54)</u>

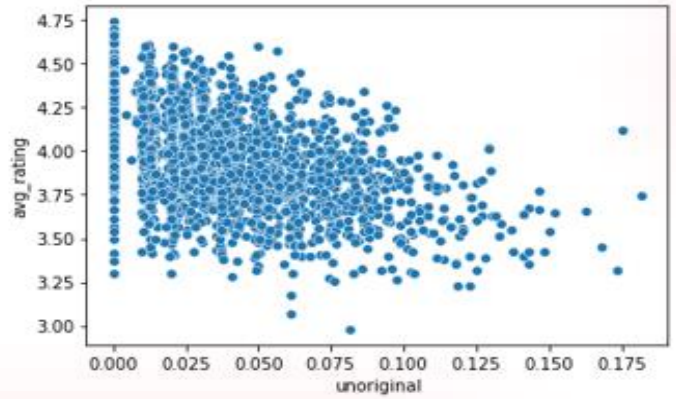
- Skewed to the right frequency distribution; strong positive relationship
- EDA suggests that it is more strategic to produce movies with new plots and characters instead of big-budget movie sequels





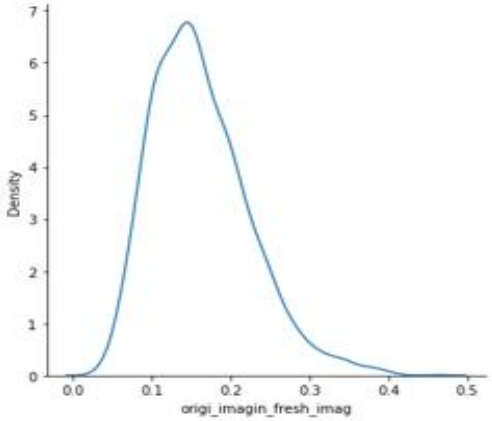
Group 4

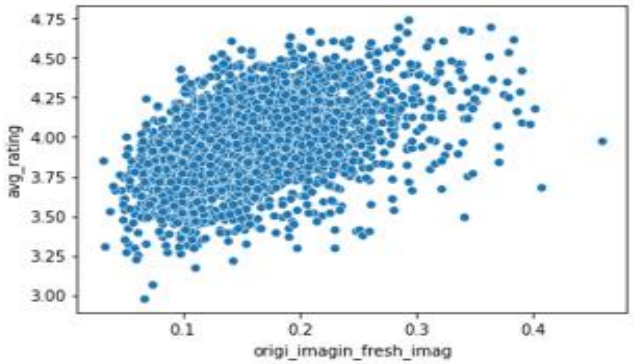


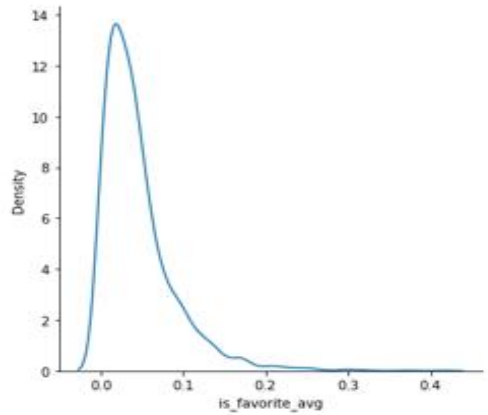


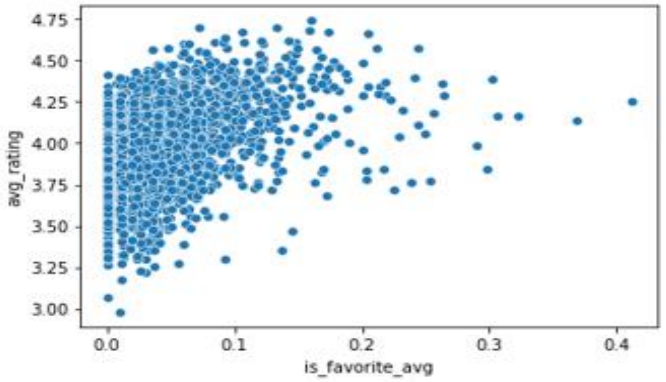
- Skewed to the right frequency distribution; strong negative relationship
- Magnifies viewer's dissatisfaction with reuse of old plots and rehashed storylines

- Skewed to the right frequency distribution; strong negative relationship
- Aside from originality, make use of imaginative thematic plots, fresh narratives, and intelligent scripts





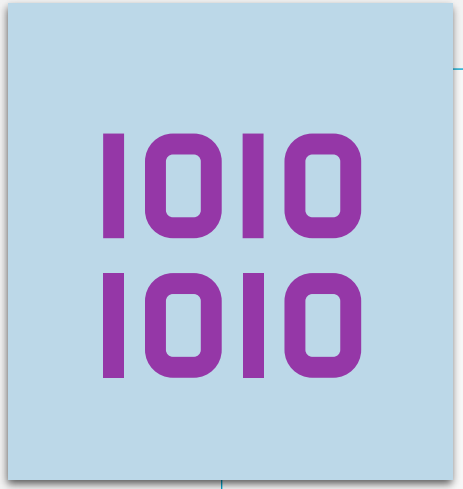




is_favorite_avg (0.46)

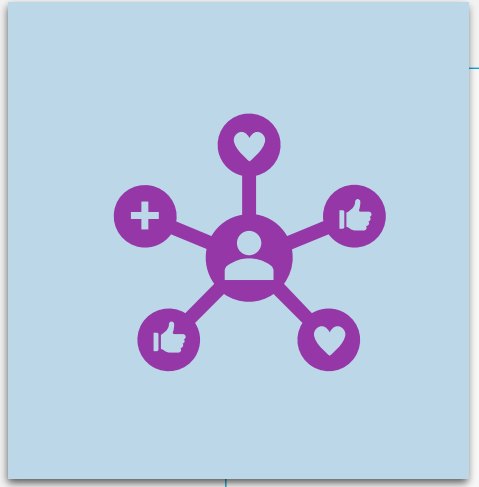
- Skewed to the right frequency distribution; strong positive relationship
- Results are expected since viewers will most likely give a high rating for their favorite movies

Building the Model



Data Pre-processing

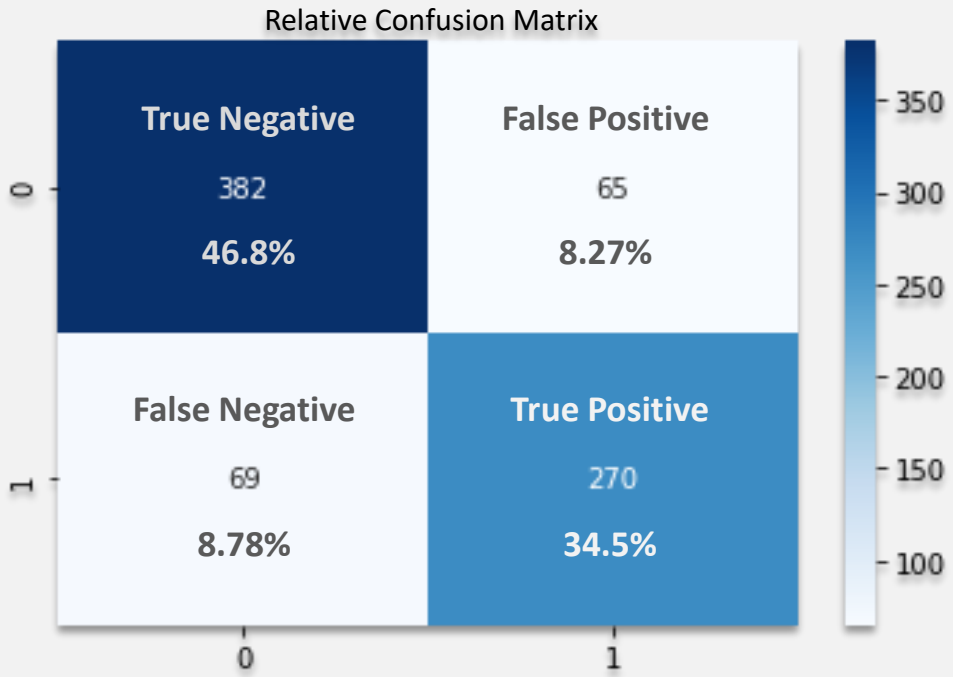
- Standardize numerical variables
- Convert data frame to Sparse Matrix
  - Easily compressed
  - Requires less storage
  - a format XGBoost can read (it doesn't work with pandas dataframe)

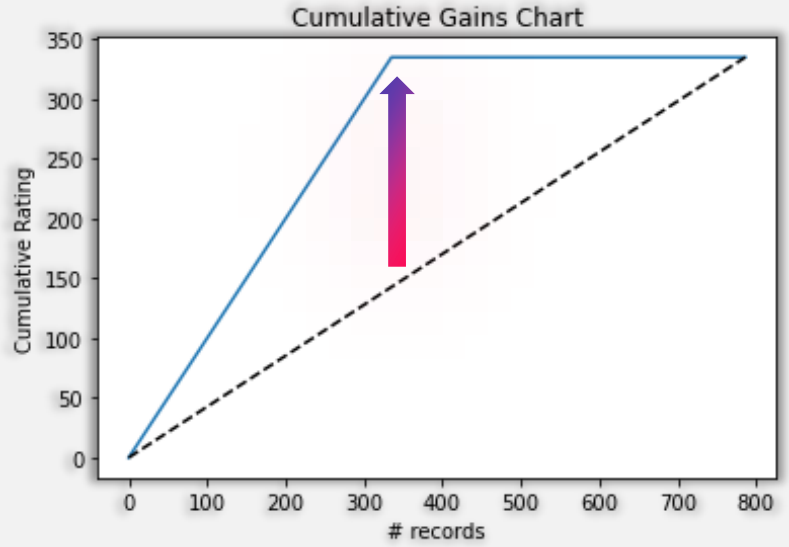


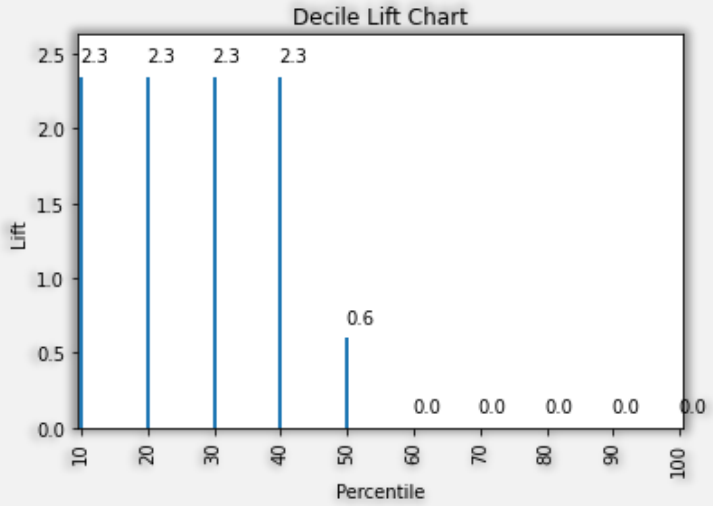
Model Building

- **Extreme Gradient Boosting** Algorithm
  - Based on Boosting with decisional trees as weak learners
  - Uses a Gradient descent optimisation to minimize the loss function

Performance Evaluation of the Model







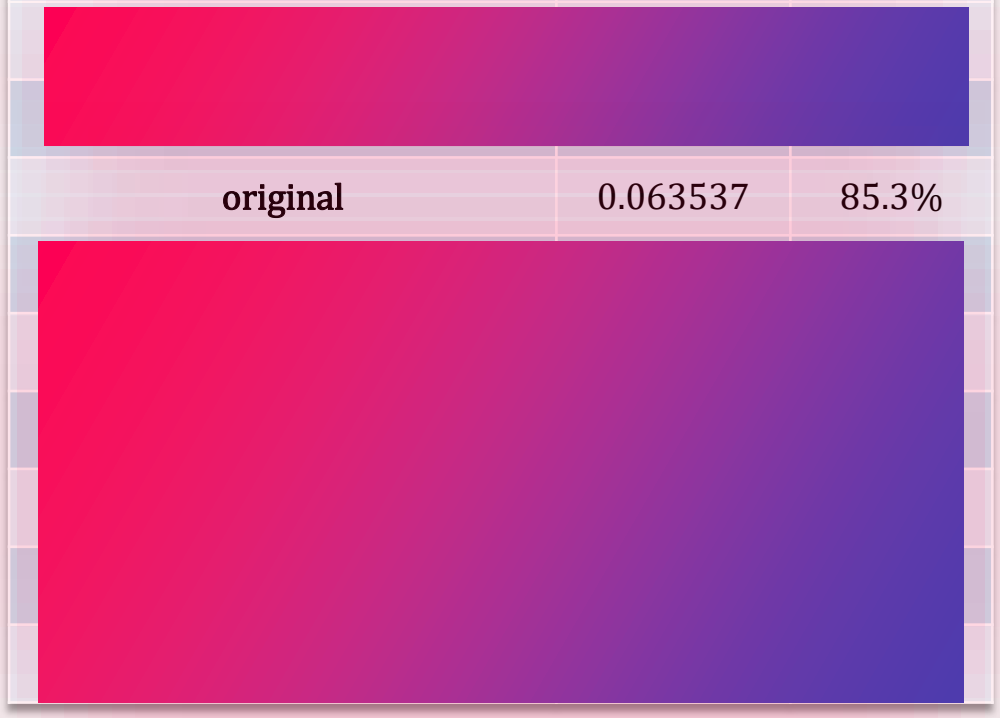
- Accuracy of Model: 82.95%.
- Almost equal percentages for False Negative and False Positive, 8.78% and 8.27% respectively

A lift of 2.3 at 10% for category “1” (yes) means that by using 10% of the dataset the model performs 2.3 times better than random guessing

Key Findings from XGBoost

Most Important Variables According to the Model

Variables	Importance	imp_score
is_favorite_avg	0.074497	100.0%
original	0.063537	85.3%



- Most Important Features are the most (positively or negatively) related to the average rating

- We created **Importance Score (imp_score)** which refers to the relative importance of the variable based on the top-most important feature which is *is_favorite_avg*.

Key Insights from Correlation and XGBoost

What variables affect the average movie rating?

Variables	Importance	imp_score
is_favorite_avg	0.074497	100.0%
original	0.063537	85.3%

Originality of the Movie

Original (+0.54), and *o* being on top of both correlation and feature importance implies that rating is a very important feature to average rating. Originality of the plot brings positive impact on users that is reflective on average rating

Percentage of users who classify the movie among their "FAVORITES"

Definitely one of the key variable for average rating prediction is *is_favorite_avg* (+0.46). This can only be given by a user to (1) movie which gives the impression that the movies chosen for this category are the best movies according to the users. More users that mark the movie as their favorite, translates to a higher the average rating of the movie.

Predictability of the Movie

have a very high *negative* correlation against average movie rating. The latter also comes in second for feature importance. This implies that users are not impressed with stories that has obvious story outcomes.

Key Insights from Correlation and XGBoost

What variables affect the average movie rating?

Reflection of Reality

There is a strong correlation between [redacted] and average rating. This suggests that real-life stories which brings out real-life emotions are appealing to users. At the same time we can see a negative correlation on *unrelatable* (-.37) which speaks of the same point. Both of these features are part of the most important features that affect average rating.

Dragging Movies

The feature [redacted] is one of the features that came out in the feature importance. This shows us that a slow-paced story telling technique have negative implications on how the user views the movie and pulls down the average rating of the movie accordingly.

Dramatic Component of a Movie

The dramatic component of a movie has a positive influence on the average rating. [redacted] are part of the most important features and is positively correlated to the ratings. Users are giving importance to the serious tone and rollercoaster of emotions that a movie is bringing.

Recommendation and Summary

- Recommendation Engine

Recommendation: Create a Recommendation Engine

- Create a recommendation engine using the selected key variables
- Create Similarity Matrix

Jaccard Index

- Produced inconsistent recommendations
- Some movies recommended are not aligned with original movie
- All results from test sets had at least one erroneous recommendation

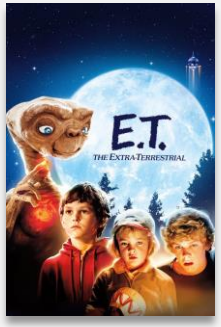
Cosine Similarity

- Produced Consistent and Acceptable results
- All results from test sets had acceptable results
- Yielded best results among the 3 methods.

Euclidean Distance

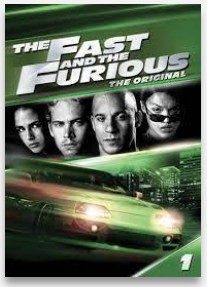
- Produced consistently erroneous results.
- Movie recommendations are not acceptable.
- All test sets did not produce good recommendations

Sample Test Sets



Jaccard	Cosine	Euclidean
Fantastic Mr. Fox	The Iron Giant	Saw III
BlacKkKlansman	Kubo and the Two Strings	Annabelle Comes Home
Where the Wild Things Are	Big Hero 6	Saw IV
Beauty and the Beast [2017]	Brave	The Texas Chainsaw Massacre: The Beginning
Alexander and the Terrible, Horrible, No Good,...	Coco	The Nun
Anacondas: The Hunt for the Blood Orchid	The Lion King [1994]	The Possession
Alpha	The BFG	Fifty Shades Freed
Dr. Dolittle 3	How to Train Your Dragon: The Hidden World	Halloween II
Clockstoppers	Big Fish	Friday the 13th
Bad Boys I	WALLÂ·E	Hoste

Jaccard	Cosine	Euclidean
The Philadelphia Story	The Upside	Saw III
The Wizard of Oz	Secondhand Lions	Predator
Galaxy Quest	A Dog's Journey	The Autopsy of Jane Doe
Forever Young	Yesterday	Saw IV
Stronger	Selena	The Texas Chainsaw Massacre: The Beginning
A League of Their Own	Spirited Away	Hostel
Black Swan	Last Christmas	The Possession
Enemy of the State	Good Will Hunting	A Quiet Place
The Lord of the Rings: The Two Towers	Where the Wild Things Are	Mad Max: Fury Road
Big Fish	A League of Their Own	Insidious



Jaccard	Cosine	Euclidean
Edge of Tomorrow	Fast Five	A Walk to Remember
Flight	Furious 7	War Room
Captain America: The First Avenger	Lethal Weapon	Steel Magnolias
Panic Room	2 Fast 2 Furious	Forrest Gump
Surf's Up	Die Hard 2	The Passion of the Christ
Bruce Almighty	xXx	Overcomer
Cocktail	Fast & Furious Presents: Hobbs & Shaw	Loving
Doctor Strange	The Last Boy Scout	A Dog's Purpose
Divergent	Gone in Sixty Seconds	The Wedding Singer
Dirty Grandpa	The Wolverine	Sense and Sensibility

Summary

1) Find new movie genres based on sentiments	Clustering and result modification	
2) Identify sentiments and metrics most affecting ratings	XGBoost Model for ratings	1. Originality 2. Favorites 3. Dramatic Component 4. Reflection of Reality 5. Dragging 6. Dramatic Component
3) Build a recommendation engine	Recommender system based on cosine similarity	

Thank You!

Leonardo Patricelli
Team Leader

Olusegun Ajao

Ira Martina Balmes

Anna Bianca Ongtengsiem

Cenith Ramirez

B412 Capstone Project
Analytics for Business Decision Making
George Brown College