University Of Liverpool Management School

BY

NIKITA GHULE

CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

The report provides a succinct explanation of the significance of analytics approaches, such as regression and visualisation, for the executives of over-the-top (OTT) platforms. Colossal Entertainment, one of the many players in the fiercely competitive OTT sector, needs a cutting-edge method of revenue and data analysis to boost sales. The research provides a case study of an established, sizable OTT platform, like Netflix, that makes use of artificial intelligence and machine learning algorithms to promote customer interaction. It also addresses how data visualization tools offer a simple and intuitive approach to comprehend and view the raw data as graphs, charts, and geospatial data. Focusing on supervised learning methods, regression, and classification also aids in understanding the key factors that influence user engagement and business models. The stakeholders might then utilise this information to enhance the streams and get profound understanding about user behaviors and preferences. The last section of the paper explains how Colossal Entertainment can put these concepts into action, give their customers the finest service possible, and grow their network.

MOTIVATION

The Over-the-Top (OTT) streaming platforms are primed to post a compound annal growth rate of ~29% during 2021-2028, which would be the top highest in among the media and entertainment segments. By 2025, it is anticipated that the worldwide market would have grown further and generated US $271,837m in sales. (Chakraborty et al., 2023) The market for streaming services and online content is expanding due to the advent of digital technology. The Internet of Things, Augmented Reality, and Virtual Reality, which are easily accessible, have elevated the OTT platform and allowed consumers to obtain material that fits their personal tastes. Visualization will help the organization to improve the customer engagement using the analysing metrics. The metrics helps to analyse the views, interactions, ratings and help to stop trends and preferences to propose similar pattern. It also can help in the areas of the improvements for the organization. Churn prediction would benefit from supervised algorithms like classification and regression, as well as content performance analysis to examine trends and norms and user watching patterns. These analytics can also be leveraged to target advertisements, which will enhance the income model and aid to make ad campaigns more effective. Colossal Entertainment should also inculcate these visualization and regression models to understand the users and boost customer engagement.

LITERATURE REVIEW

Data visualisation is the process of displaying data as visuals. The data is extracted from the raw data and presented as a scatterplot, statistical summary, or a histogram. Data cleansing, trend detection, local pattern detection, development of modelling outputs, and presentation of results are accomplished through visualisation. (Antony Unwin,2020) The key factors that should be included in the report of the OTT platforms are Customer Lifeline Value (CLV) for understanding how much a customer has spent. Churn rate to get the unsatisfactory users. Identifying new users and getting a general understanding of the business requires using monthly active users (MAU). Monitoring user engagement, how much time they spend on the app or website, and what they view, is important for business. By sending emails or alerts to promote the app, inactive users can be reactivated. Comparable metrics such as Subscriber Return on Investment (ROI), Monthly Recurring Revenue (MRR), Network experience, Customer experience, and View Duration would contribute to the analysis for the data associated with users. (Naresh, 2022) With the help of this data, businesses may concentrate on audience segmentation, content analytics, and market optimization. They can learn more about their rivals' strategies, that will allow them to increase the return on investment (ROI) of their marketing investments.

Regression analysis is a method for calculating the information in the data that must be sorted out to make inferences and forecast the future. Finding a straight line that can effectively describe the relation between two or more variables is the objective of linear regression. Linear and logistic modelling approaches are widely used regression techniques. (Gallo, 2015) These techniques can help in identifying the key drivers by using predictive analytics when the users are more active and how much time, demand analysis to understand the market when the new users are likely to take the subscription. While it is simple to demonstrate the relationship between two variables, regression can be used in sales forecasting to assist the business understand sales more precisely. Based on the data gathered, correlations provide a clear picture of cause and effect in the business. With new insights, hidden patterns are revealed, and the audience is better understood. It provides connections between various elements and suggestions for engaging the audience through predictive analysis. This

research, for instance, reveals which genre is popular and in demand on particular days of the week. (Gogtay, Deshpande and Thatte, 2017)

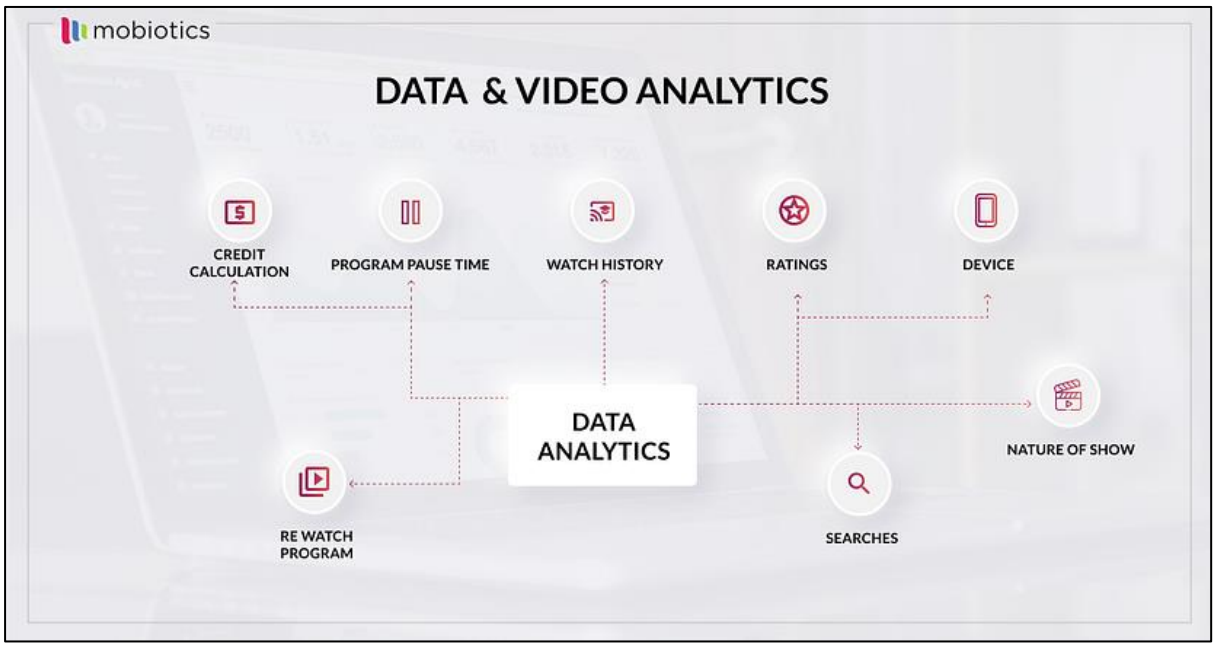


Fig 1: Data and Video Analytics of OTT platforms

CASE EXAMPLE

CASE 1: Application of visualization for improving Netflix's operations

The capacity of Netflix to use its data to make wise decisions is a crucial element of its success. A new UI with video-forward features, navigation choices, and digital experience for consumers was developed after analysing the users' points of view and the unfriendly experience back in 2010.

Netflix invested in creating its own data visualisation platform, Lumen, to provide relevant data to stakeholders in real-time. Data sources, Visualizations, Mappers, and Variables make up the bulk of Lumen dashboards. (Netflix Technology Blog, 2020)

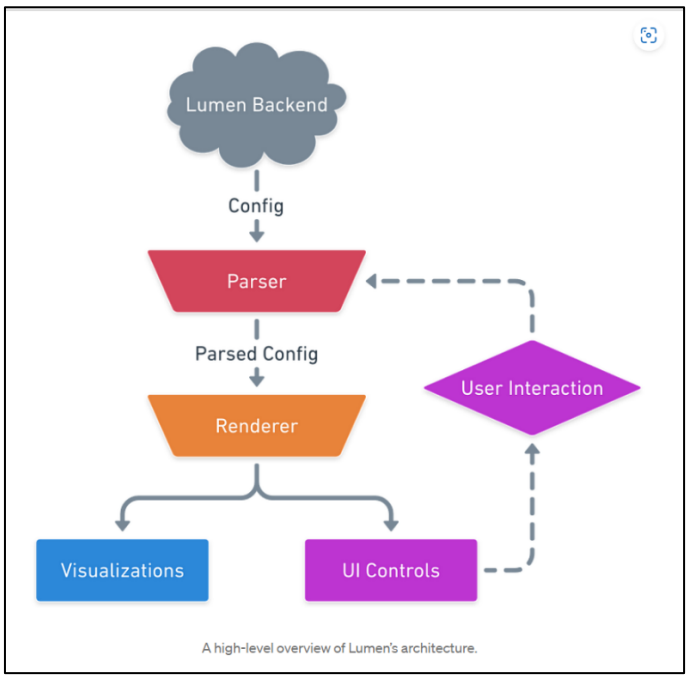


Fig 2: Architecture of Lumen

Netflix uses A/B tests to make decisions that improved the product. It divided the users into subsets and provide current product experience to one and improved to other. After analysing the response from the user, they deliver the feature with highest statistics. (Netflix Technology Blog, 2018)

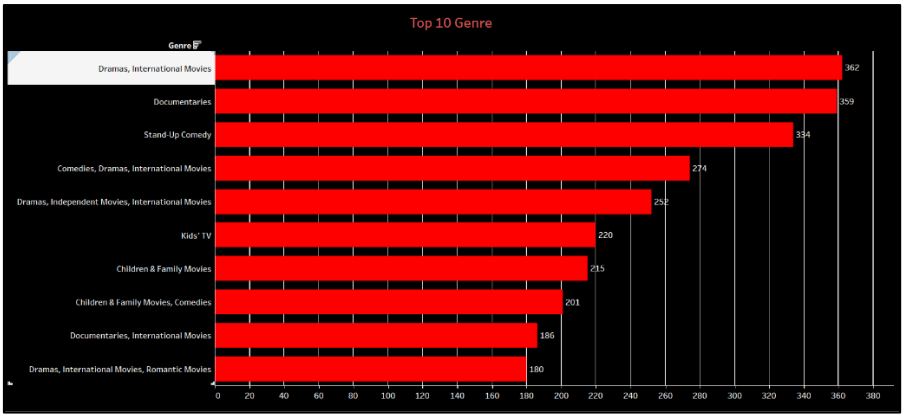


Fig 3: Top 10 Genre on the Netflix

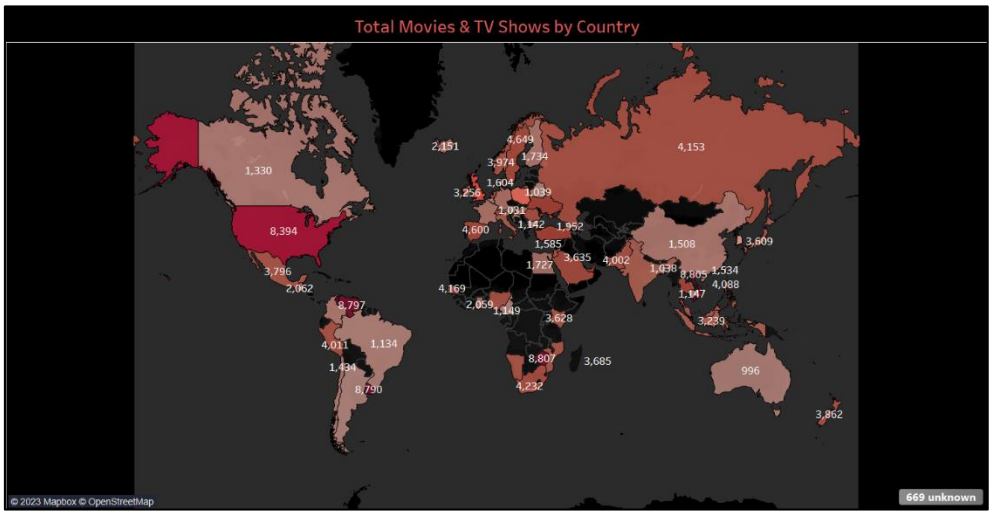


Fig 4: Country wise Contribution in content of Netflix

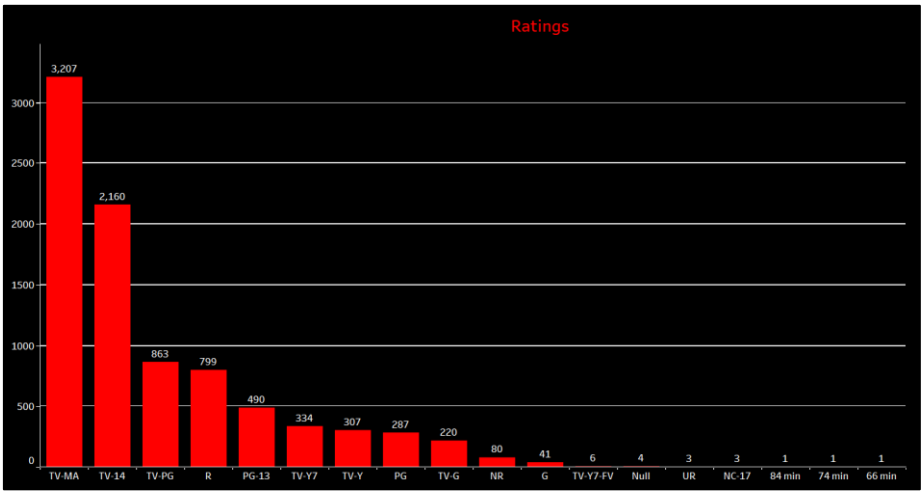


Fig 5: Rating given by the viewers

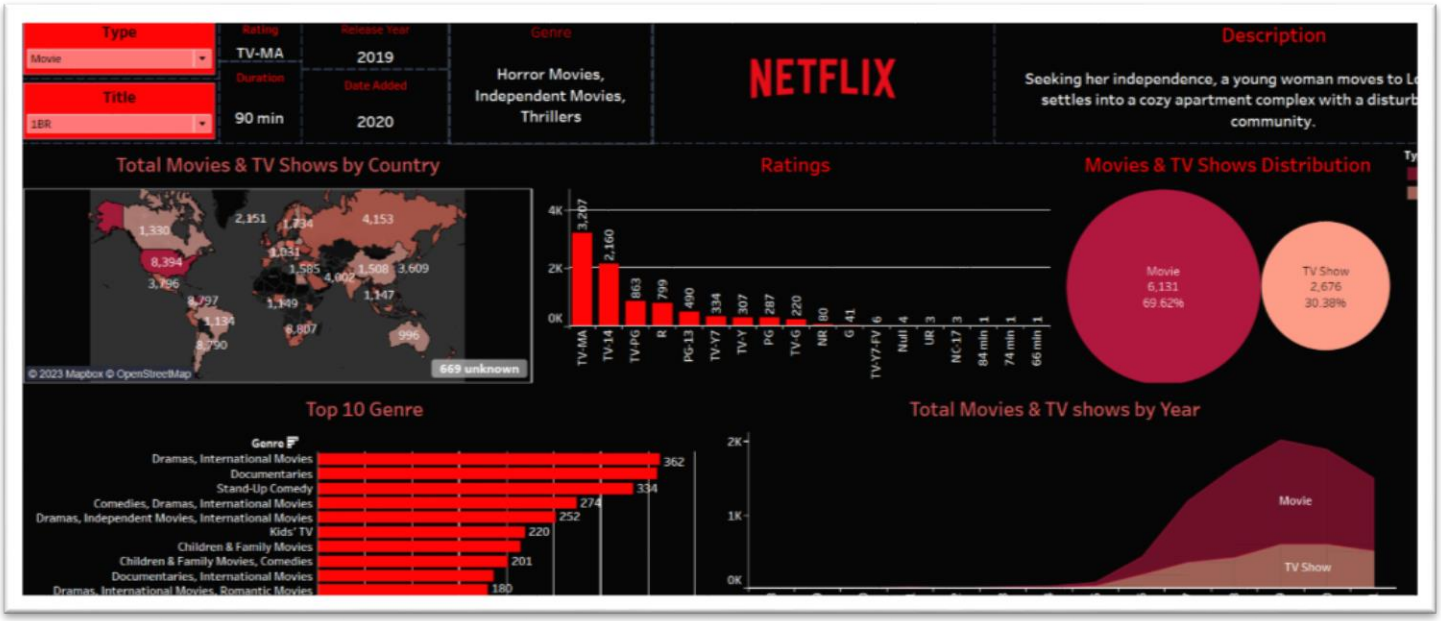


Fig 6: Dashboard of the Netflix dataset

Managers can find hidden insights using dynamic comparison and cross-referencing techniques, which can be obtained by data visualizations. From the above statistics stakeholders can understand the market and customers in better way. For example, in figure 3 shows that which genre are most watched and then the more content related to this is released. Figure 4 explains the country-wise contribution of the content to the Netflix. And the figure 5 provides the stats about the rating to fetch insights over the maturity rating level for a TV show/movies, which the user watches and that helps the Netflix to monitor the rating type of content which is sold on average. Figure 6, Dashboard created using Tableau to analyse the dataset uncover insights about the most popular shows, genre, years which helps the Managers making concrete decisions.

CASE 2: Application of Regression Technique in Netflix Recommendation System

Netflix's whole business model is the suggestion of content to customers. To get insights about viewer behaviour, they employ supervised techniques classification and regression, and unsupervised approaches clustering, anomaly detection, dimension reduction, and topic modelling.

To track time, location, and device, Netflix employs a customised recommendation engine. It tracks platform searches, whether the content was paused, rewound, rewatched or fast-forwarded using regression analysis. Personalized Video Ranking, Trending Now Rankers, Continue Viewing Rankers, and Video-Video Similarity Rankers are examples of algorithms used in recommendation engines. Content based filtering based genre, language, region, age, sex, and maturity content is used. This classification makes sure that right content is suggested to the audience. Both user-based and item-based collaborative filtering algorithms are used by Netflix. In user-based, the system recommends material to the user based on their watching habits and the viewing preferences of other users. In contrast, item-based recommendations are based on the user's prior behaviour.

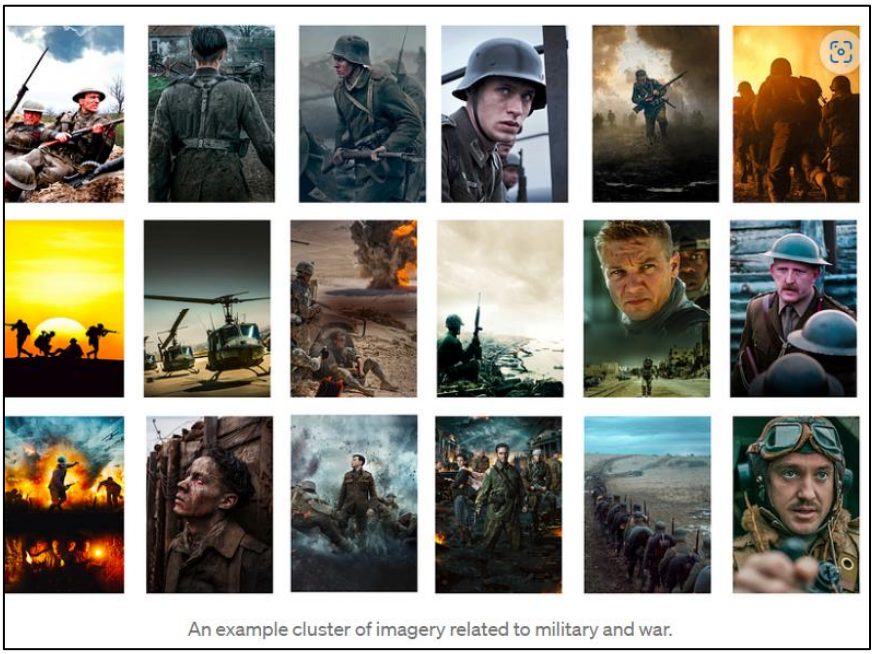


Fig 7: Clustering of images in Netflix for recommendation

Netflix makes extensive use of the clustering method. They gather information or comparable photos to produce expert-free patterns and ideas for groupings and suggestions for the audience. (Netflix Technology Blog, 2018)

```
In [2]: # libraries Imported
        # data analysis and wrangling
        import pandas as pd
        from datetime import datetime
        # visualization
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        # machine learning
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import classification_report
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.naive_bayes import GaussianNB

In [3]: netflix = pd.read_csv(r'C:\Users\Nikita\Downloads\netflix_titles.csv')
        netflix.head()
```

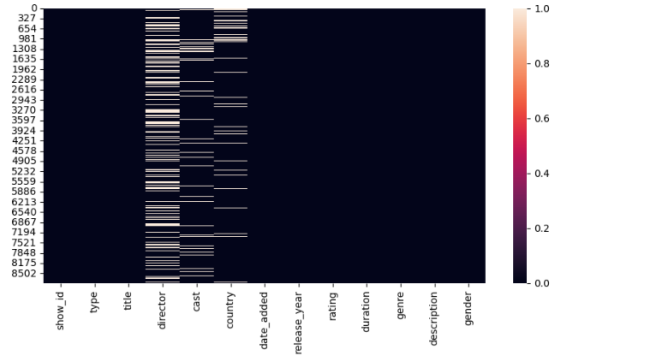
Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	genre	description	gender
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	male
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	male
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	male
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...	male
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...	male

Fig 8: Importing the dataset

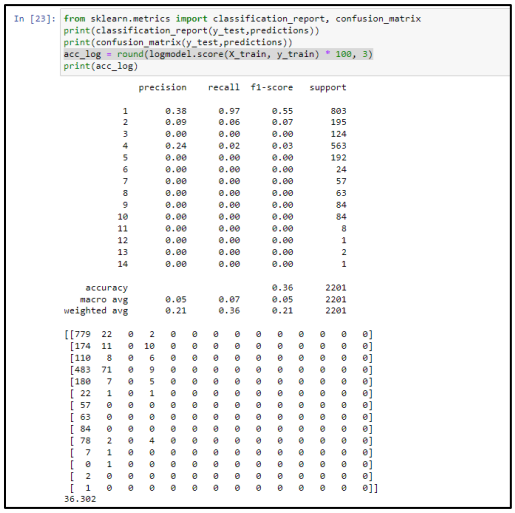
```
In [6]: plt.figure(figsize=(10, 5))
        sns.heatmap(netflix.isnull())
        plt.show()

        for i in netflix.columns:
            null_rate = netflix[i].isna().sum() / len(netflix) * 100
            if null_rate > 0:
                print("{}'s null rate :{}%".format(i, round(null_rate, 2)))
```




```
director's null rate :29.91%
cast's null rate :9.37%
country's null rate :9.44%
date_added's null rate :0.11%
rating's null rate :0.05%
duration's null rate :0.03%
```

Fig 9: Removing the Null values



```
In [23]: from sklearn.metrics import classification_report, confusion_matrix
         print(classification_report(y_test,predictions))
         print(confusion_matrix(y_test,predictions))
         acc_log = round(logmodel.score(X_train, y_train) * 100, 3)
         print(acc_log)

                       precision    recall  f1-score   support

                    1       0.38      0.97      0.55       803
                    2       0.09      0.06      0.07       195
                    3       0.00      0.00      0.00       124
                    4       0.24      0.02      0.03       563
                    5       0.00      0.00      0.00       192
                    6       0.00      0.00      0.00        24
                    7       0.00      0.00      0.00        57
                    8       0.00      0.00      0.00        63
                    9       0.00      0.00      0.00        84
                   10       0.00      0.00      0.00        84
                   11       0.00      0.00      0.00         8
                   12       0.00      0.00      0.00         1
                   13       0.00      0.00      0.00         2
                   14       0.00      0.00      0.00         1

            accuracy                           0.36      2201
           macro avg       0.05      0.07      0.05      2201
        weighted avg       0.21      0.36      0.21      2201

        [[779  22   0   2   0   0   0   0   0   0   0   0   0]
         [174  11   0  10   0   0   0   0   0   0   0   0   0]
         [110   8   0   6   0   0   0   0   0   0   0   0   0]
         [483  71   0   9   0   0   0   0   0   0   0   0   0]
         [180   7   0   5   0   0   0   0   0   0   0   0   0]
         [ 22   1   0   1   0   0   0   0   0   0   0   0   0]
         [ 57   0   0   0   0   0   0   0   0   0   0   0   0]
         [ 63   0   0   0   0   0   0   0   0   0   0   0   0]
         [ 84   0   0   0   0   0   0   0   0   0   0   0   0]
         [ 78   2   0   4   0   0   0   0   0   0   0   0   0]
         [  7   1   0   0   0   0   0   0   0   0   0   0   0]
         [  0   1   0   0   0   0   0   0   0   0   0   0   0]
         [  2   0   0   0   0   0   0   0   0   0   0   0   0]
         [  1   0   0   0   0   0   0   0   0   0   0   0   0]]
        36.302
```

Fig 10: Logistic Analysis

```
# Naive Bayes algorithm:
gaussian = GaussianNB()
gaussian.fit(X_train, y_train)
Y_pred = gaussian.predict(X_test)
acc_gaussian = round(gaussian.score(X_train, y_train) * 100, 3)
print(acc_gaussian)
```

```
21.469
```

```
from sklearn.linear_model import LinearRegression
```

```
lr= LinearRegression()
lr.fit(X_train,y_train)
predicted = lr.predict(X_test)
```

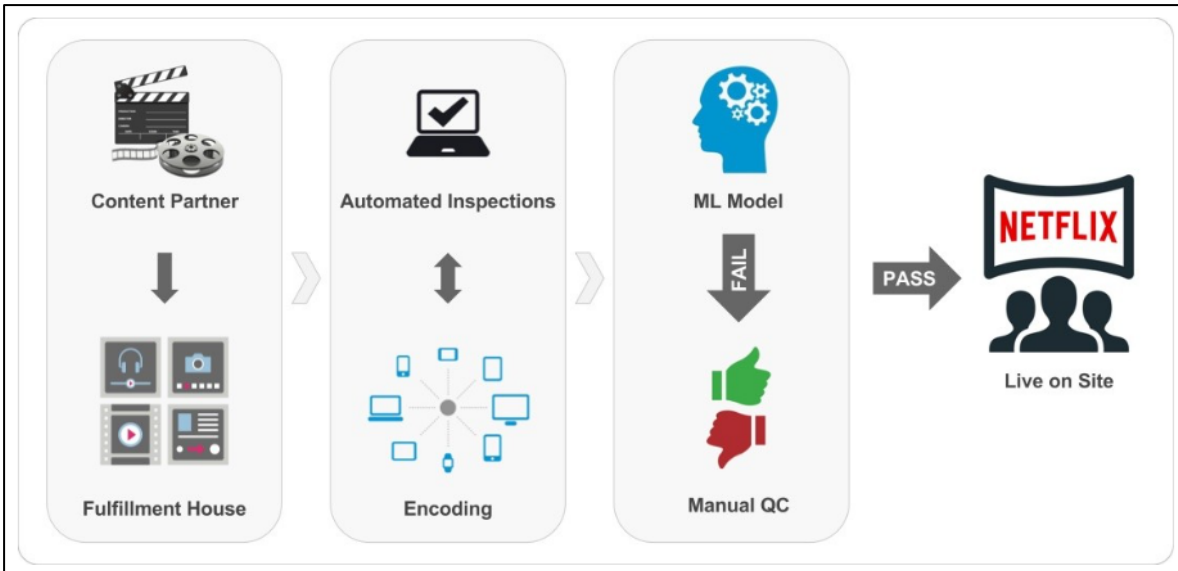
```
acc_lr = round(lr.score(X_train, y_train) * 100, 3)
print(acc_lr)
```

```
11.157
```

```
# model Evaluation
models = pd.DataFrame(
    {"Model": ["Logistic Regression","KNN", "Naive Bayes","LinearRegression"], "Score": [acc_log,acc_knn, acc_gaussian,acc_lr]}
)
models.sort_values(by="Score", ascending=False)
```

	Model	Score
1	KNN	60.981
0	Logistic Regression	36.302
2	Naive Bayes	21.469
3	LinearRegression	11.157

Fig 11: Linear Regression, Naïve bayes algorithm



Netflix Supply Chain with Predictive Quality Control

Fig 12: Predictive Quality Control Analysis

Every month, billions of hours are streamed by over 69 million individuals worldwide. The platform's material must be of a high standard at all times. To do this, Predictive Quality Control, a supervised ML approach, is employed to anticipate a failure, if failed returned to the manual team for analysis. Checking for faulty assets before distributing them to the consumers is one of this module's main objectives. (Nirmal Govind , 2015)

BARRIERS

These are some restrictions and constraints that might preclude the adoption of visualisation**.** (Hoffman, 2021)

- False correlation: Owing to incorrect interpretations and a lack of knowledge of linkages, visualisations may portray incorrect factors. This may lead to incorrect inferences and conclusions, as well as erroneous correlations between the data. (Gemini Data, 2021)
- It provides an approximation outcome instead of precise result.
- Design flaw: Although visualisation is a form of communication, it might be problematic to communicate if the design is done incorrectly.
- The data that is displayed may be skewed. For instance, the individual gathering the data could evaluate only a subset of the data and leave out the other significant data, which could provide biased conclusions.
- It is difficult to draw maeaningful visulaizations for multivariated data with many variables.

Similarly, Regression analysis has the following limitations:

(Anon, 2020) (Vidyashri M. H, 2021)

- Low-quality data cannot be used to create regression models. If the duplicated data and missing values are not properly removed during data preparation, this might result in an imbalanced data distribution and incorrect findings.
- The models work on independent variables,giving a strong linear correlation between them. If the variables selected are not correct might not give the correct predictive analysis.
- Regression analysis makes the assumption that the independent variables have a minimal correlation with each other. It may be challenging to isolate each independent variable's unique impact on the dependent variable if there is significant multicollinearity.

- The working of regression model is inversely proportion to the number of variables i.e as the number of variables increases the reliability of the regression model decreases.
- The process of regression analysis is quite time-consuming, intricate, and made up of several computations and analyses.

RECOMMENDATIONS & ROADMAP

With the growth of OTT streaming services like Netflix and Amazon individuals may now watch an infinite number of shows whenever and wherever they want. Similar to this, Colossal Entertainment may use data visualisation to gather information for reporting, communication, and sales augmentation. Using real-time data, Managers can evaluate user information to produce customized recommendations, including browsing history, ratings, and search queries. Colossal Entertainment can keep an eye on user patterns and preferences, which will make it easier to see new trends and move swiftly when the market changes. Spending on marketing and promotions, Netflix attracts users. Similarly, Colossal Entertainment can analyse the effectiveness of marketing initiatives using visualisation tools. The managers will be able to target the appropriate audience and optimise promotion campaigns as an outcome. Visuals also can be used to track their sever performance, network traffic and metrices to ensure a seamless user experience. (Sahu, Gaur and Singh, 2022) (Lock and Araujo, 2020) (Chakraborty et al., 2023)

Regression models can further integrated into Colossal Entertainment's recommendation engine to help the company understand behaviour patterns, demographics, preferred content, engagement, and revenue model. Similar to how Netflix creates customised thumbnails to delight clients, our business may use ML algorithms to extract high-quality data from the video files. Reviewing the data and identifying user preferences can be aided by text mining, text analysis, clustering, content-based filtering, collaborative filtering, linear regression, logistic regression, Poisson regression. These methods will assist in better understanding viewer and prospect profiles, customer history analysis, market trends and their implementation over time, feedback and review monitoring utilising sentiment analysis, and user understanding. (Martínez-Sánchez, Nicolas-Sans, and Bustos Díaz, 2021)

Colossal Entertainment may enhance its content strategy, boost user engagement and retention, optimise operations, enhance the content offerings, and discover new growth prospects by implementing these visualisation and regression tools.

CONCLUSION

Key Takeaway's:

- The identification of trends, and correlations in user behaviour, content performance, and profitability can be facilitated by data visualisations for OTT businesses. With this, data-driven decisions can be made to enhance performance.
- Regression modelling has a strong degree of predictability. It can be used to predict sales for short and long term. By properly analysing the outcomes of the decisions, it may aid in the correction of errors.
- Findings from regression models can be used for better understand consumers, content performance, and what marketing tactics to increase sales and attract customers.
- Studying the OTT platform, obtaining undiscovered information, and real-time visualisation of this data are crucial for staying ahead of the competition. This can assist viewing new patterns, addressing problems early, and making choices right away.

Report's Limitations:

- Other analyses, such as cohort analysis, can be employed in addition to regression analysis to comprehend how users behave. Age, gender, geography, subscription plan, and other variables can be used to group content and make it available according to user preferences.
- Natural language Processing technique and sentiment analysis used for monitoring brand reputation, provide recommendations, and evaluate reviews, and social media posts for areas that can be improved.
- ML approaches, are not included in report, can be used to develop models, and make automated decisions. These include collaborative filtering, reinforcement learning, neural networks, and matrix factorization. (Rojas, Gallón and Corrales, 2018)

REFERENCES

- Chakraborty, D., Siddiqui, M., Siddiqui, A., Paul, J., Dash, G. and Mas, F.D. (2023). Watching is valuable: Consumer views – Content consumption on OTT platforms. *Journal of Retailing and Consumer Services*, 70, p.103148.
- Unwin, A. (2020). Why is Data Visualization Important? What is Important in Data Visualization? *Harvard Data Science Review*, [online] 2(1). Available at: https://hdsr.mitpress.mit.edu/pub/zok97i7p/release/4.
- Naresh, S. (2022). *Data Reporting & Visualization for OTT*. [online] mobiotics. Available at: https://medium.com/mobiotics/data-reporting-visualization-for-ott-6345c50abad5 [Accessed 19 Mar. 2023].
- Gogtay, N., Deshpande, S. and Thatte, U. (2017). Principles of Regression Analysis. *Journal of The Association of Physicians of India* ■, [online] 65. Available at: https://www.kem.edu/wp-content/uploads/2012/06/10-Principles_of_regression_analysis-1.pdf.
- Gallo, A. (2015). *A Refresher on Regression Analysis*. [online] Available at: https://online210.psych.wisc.edu/wp-content/uploads/PSY-210_Unit_Materials/PSY-210_Unit12_Materials/Gallo_HBR_Regression_2015.pdf.
- Gemini Data. (2021). *The Pros and Cons of Data Visualization*. [online] Available at: https://www.geminidata.com/pros-and-cons-of-data-viz/.
- www.linkedin.com. (n.d.). *Optimizing Content Quality Control at Netflix with Predictive Modeling*. [online] Available at: https://www.linkedin.com/pulse/optimizing-content-quality-control-netflix-predictive-nirmal-govind/ [Accessed 22 Mar. 2023].
- Netflix.com. (2020). *Netflix Research*. [online] Available at: https://research.netflix.com/research-area/experimentation-and-causal-inference.

- Karad, V., Shah, H., Jadhav, V., Gharte, T., Wattamwar, S. and Naik, V. (n.d.). *EVALUATION OF DIFFERENT OTT PLATFORMS WITH DATA ANALYTICS TECHNIQUES FOR RECOMMENDING PERSONALIZED CONTENT TO THE USERS*. [online] *International Journal of Research and Analytical Reviews*. Available at: https://www.ijrar.org/papers/IJRAR1CNP012.pdf.
- Kim, R. (2021). *Data Analysis on OTT Platforms: Which Service Should I Choose?* [online] Medium. Available at: https://towardsdatascience.com/data-analysis-on-ott-platforms-which-service-should-i-choose-8eed953ff7d2.
- Hoffman, J. (2021). *Pros and Cons of Data Visualization Explained*. [online] WisdomPlexus. Available at: https://wisdomplexus.com/blogs/pros-cons-data-visualization/.
- www.ablison.com. (2023). *Pros and Cons of Regression Analysis 2023 - Ablison*. [online] Available at: https://www.ablison.com/pros-and-cons-of-regression-analysis/ [Accessed 20 Mar. 2023].
- Anon, (2020). *Regression Analysis: Types, Importance and Limitations*. [online] Available at: https://commercemates.com/regression-analysis/?utm_content=cmp-true [Accessed 20 Mar. 2023].
- H, V.M. (2021). *Advantages and Disadvantages of Regression Model*. [online] VTUPulse. Available at: https://www.vtupulse.com/machine-learning/advantages-and-disadvantages-of-regression-model/.
- Choi, J. and Kim, Y. (2020). Time-Aware Learning Framework for Over-The-Top Consumer Classification Based on Machine- and Deep-Learning Capabilities. *Applied Sciences*, 10(23), p.8476.
- Rojas, J.S., Gallón, Á.R. and Corrales, J.C. (2018). Personalized Service Degradation Policies on OTT Applications Based on the Consumption Behavior of Users. *Computational Science and Its Applications – ICCSA 2018*, pp.543–557.

- Lock, I. and Araujo, T. (2020). Visualizing the triple bottom line: A large-scale automated visual content analysis of European corporations' website and social media images. *Corporate Social Responsibility and Environmental Management*.
- Acuvate. (2018). *5 Amazing Business Benefits Of Data Visualization | Blog*. [online] Available at: https://acuvate.com/blog/5-amazing-business-benefits-of-data-visualization/.
- Martínez-Sánchez, M.E., Nicolas-Sans, R. and Bustos Díaz, J. (2021). Analysis of the social media strategy of audio-visual OTTs in Spain: The case study of Netflix, HBO and Amazon Prime during the implementation of Disney +. *Technological Forecasting and Social Change*, 173, p.121178.
- Chakraborty, D., Siddiqui, M., Siddiqui, A., Paul, J., Dash, G. and Mas, F.D. (2023). Watching is valuable: Consumer views – Content consumption on OTT platforms. *Journal of Retailing and Consumer Services*, 70, p.103148.