TEAM 8 PROJECT REPORT

TITLE: McDonalds Store Reviews Sentiment and Location Based Analysis.

TEAM MEMBERS: Evan Chan, Farhan Rasheed Chughtai

1. Problem Statement/Motivation:

There are many branches of McDonalds in the United States, with an unending quantity of reviews available on the internet. The ability to leverage consumer reviews to gain valuable insights into customer experiences is a useful tool for this fast-food chain. We hope to facilitate this process by using machine learning techniques to bolster the analysis and utilization of a sample of customer reviews by location.

2. Dataset/EDA:

The dataset we used for the project consists of 33,000 McDonald's reviews that were scraped from Google reviews from all over the USA. The dataset contains the following columns: reviewer ID, store name, category, store address, latitude, review time, review, and the rating the customer gave, which is a 1–5 star rating. The dataset contains reviews from 39 stores in 26 cities and 11 states in total.

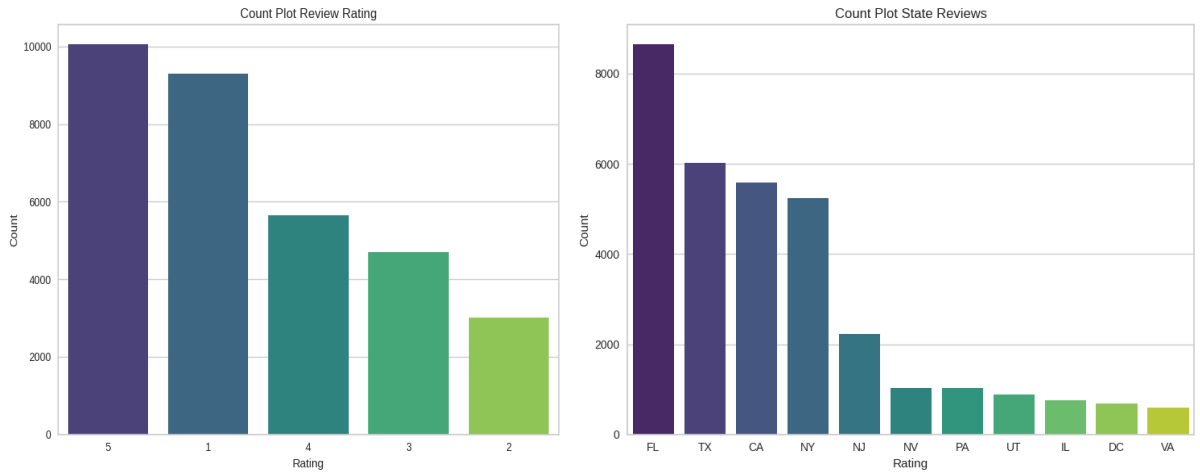

Figure 1 Shows Count Plot of Ratings on the Left and on the Count Plot of Reviews for each state in the dataset

Looking at the dataset above, you can see that the amount of 5-star and 1-star reviews is pretty equal, which means there are a good amount of positive and negative reviews. Furthermore, looking at the diagram on the left, you can see that in this dataset, Florida has the most reviews—around 8500 reviews out of the 33,000 reviews that are in the dataset. While DC and Vancouver have the least number of reviews in the dataset.

3. Methodology:

Moving on to the methodology part of our project, we will discuss what we did for each category of our risk goals. For our low-risk goals, we did a location- and NLP-based analysis.

We did that by looking at average ratings across different states and cities and forming word clouds of reviews for different categories, like worst states and best states, to better understand what the driving factors were for each sentiment. Moving on to our medium-risk goals. A detailed diagram is present in the index, which shows the methodology, which goes over each step in more detail, but we trained five models: Logistic Regression, Random Forest, MultiNomialNB, Bernouli, and SVC. We selected the best model based on our validation test results and then took that model to test its performance on the test set. After that, we took a pretrained LLM model, BERT, and used that to get embeddings for our reviews. After getting the embeddings, we retrained our best performance model on those embeddings and then checked its performance on the test set. For our High risk goals we did analysis based on the date column of the dataset and saw trends in ratings over time.

4. Low Risk Goals and Results:

For our Low Risk Goals we did a location and NLP based analysis on the reviews and got the below results.

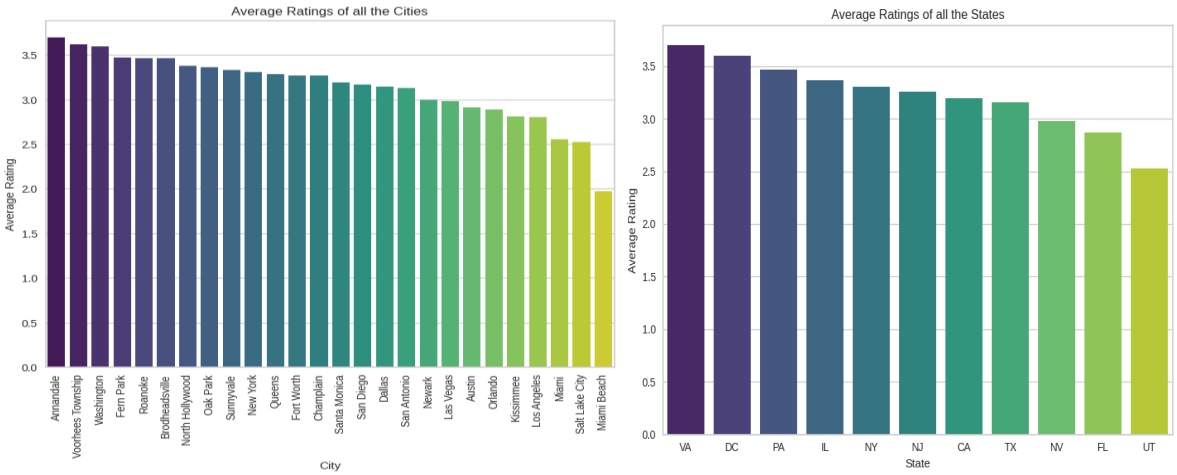

Figure 2 shows the Average Rating so all the different cities and states in our dataset

Looking at the above graphs, you can see that Annadale has the best rating across all the 26 cities we have in our database, with Miami Beach and Salk Lake City having the worst review ratings. Similarly, we can see that Florida and Utah are the worst-performing states as well, while Vancouver is the best state. So, our suggestion would be to focus more on the states like Florida and Utah and improve the McDonald's experience there, as they badly need it.

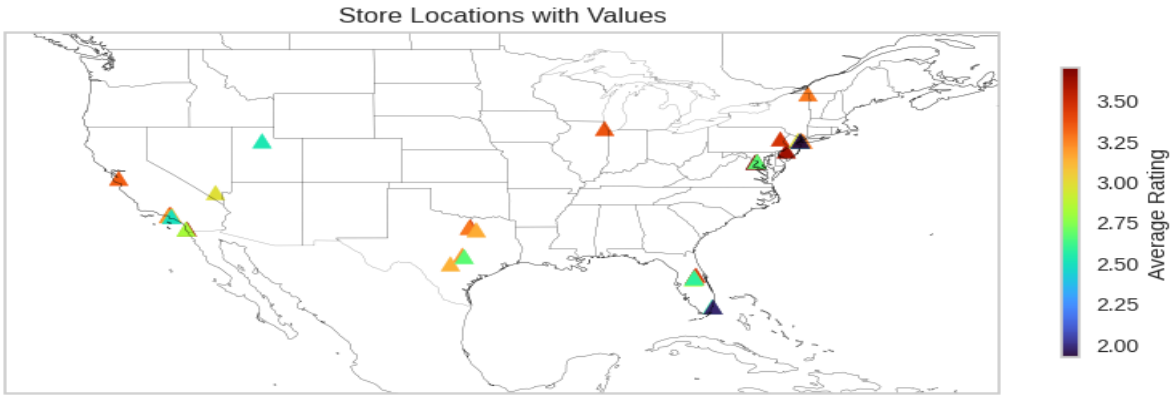

Figure 3 Shows the Heat Map of all the locations in our dataset Based on their Average Ratings.

Looking at the heat map of the USA for all our store locations, we can see that. The best and worst McDonald's reside on the east coast of the USA, while the places in central America have average ratings across the board. While the West Coast is also struggling in ratings, more emphasis should be given to branches on the West Coast than on the East Coast, but Florida and Utah should be given more attention as these states are struggling across the board.

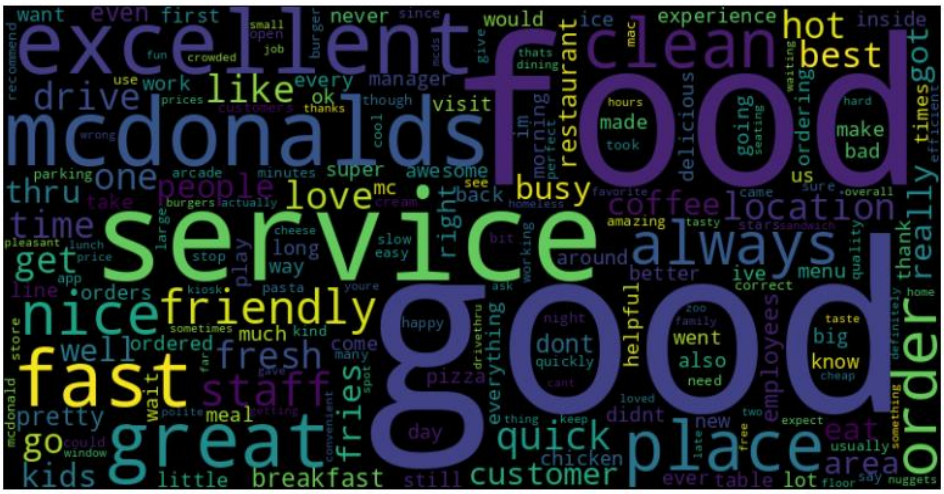

Figure 4 shows word cloud of Positive Reviews.

Looking at some word clouds of all the positive reviews we can see that service food and fast delivery are key words that lead to positive sentiment so Mc Donalds should focus on these aspects while trying to improve the ratings of places which are suffering.

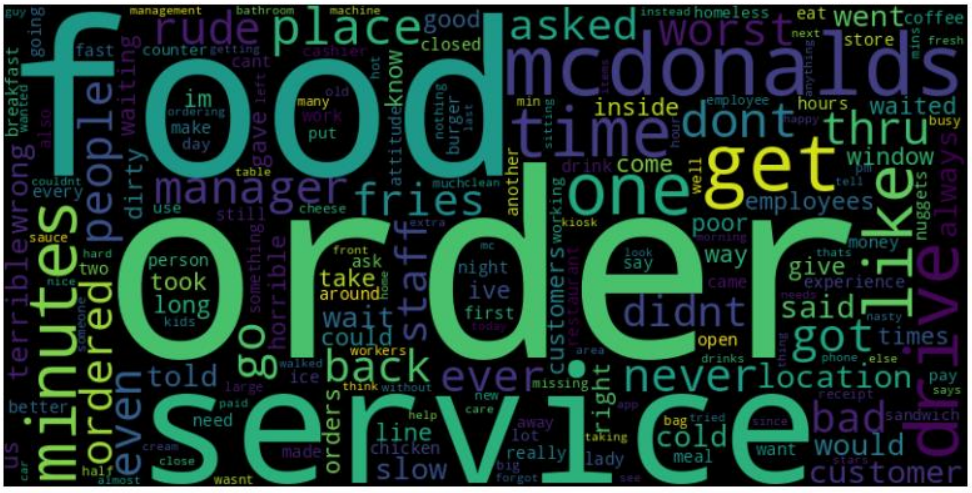

Figure 5 Shows word Cloud of Negative Reviews.

Looking at the word cloud of Negative Reviews you can see that rude bad and drive, service, order and food are the most prominent words which further signifies the importance of staff attitude and the service are critical things that drive the sentiment of the reviews.

5. Medium Risk Goals and Results:

For our medium-level goals, we are training five models. A detailed list of all the model performances is available in the index, along with the results we tried with adding different lexicons to our training data as well. For our evaluation of the model's performance, we are converting the 1–

5-star rating scale into three categories. 1, 2-star ratings are negative, 3 stars are neutral, and 4- and 5-star ratings are positive. After looking at the validation results, we found out that the Random Forest Model was our best-performing model, so we tested its performance on our test set and got the results below.

Classification Report on Final Test Set:

	precision	recall	f1-score	support
negative	0.82	0.91	0.86	2465
neutral	0.85	0.40	0.54	941
positive	0.85	0.91	0.88	3142
accuracy			0.84	6548
macro avg	0.84	0.74	0.76	6548
weighted avg	0.84	0.84	0.82	6548

Figure 6 Shows the Result of Random Forest on Final Test Set.

Over all, the performance is pretty good to identify the positive and negative sentiment, but we are struggling with identifying the neutral sentiment, which has a good precision score but a 0.40 recall score, which is very not that good. To further improve the performance, we leveraged a pretrained LLM model called BERT and used that to get embeddings for our reviews. We then tested the performance again on our test set and got the results below.

	precision	recall	f1-score	support
negative	0.93	0.93	0.93	2710
neutral	0.89	0.67	0.77	809
positive	0.90	0.95	0.92	3029
accuracy			0.91	6548
macro avg	0.90	0.85	0.87	6548
weighted avg	0.91	0.91	0.91	6548

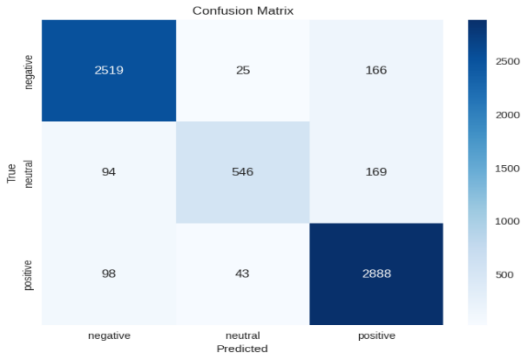

Figure 7 shows the classification report and Confusion matric of Random Forest with BERT embeddings on Test Set.

As you can see, we saw improvement across the board with the embeddings from BERT. The performance of the neutral sentiment was also improved a lot.

6. High Risk Goals and Results:

For our high-risk goals, we tried to do a time-series-based analysis. We came up with some graphs to see the trends of ratings over time and then forecast the ratings, but due to the dataset having review dates that are not spread evenly across the years, the results are a little skewed and should be looked at with that in mind. We tried to forecast using ARIMA model but that also failed due to this issue. All the results can be seen in the index of our report. But we can summarize them here in this section. So according to our time-based analysis that we were able to perform, the ratings of McDonald's stores were increasing over the years from 2017 to 2020. But as soon as COVID hit, the ratings saw a drop, and this was the case in 2021, 2022, and 2023. So, this implies that McDonald's online and drive-through capabilities should be improved and looked at, which could be the cause of this decrease in ratings after the year 2020.

7. References:

- https://www.kaggle.com/datasets/nelgiriyewithana/mcdonalds-store-reviews/data the dataset used.
- https://huggingface.co/docs/transformers/en/model_doc/bert

8. Appendix:

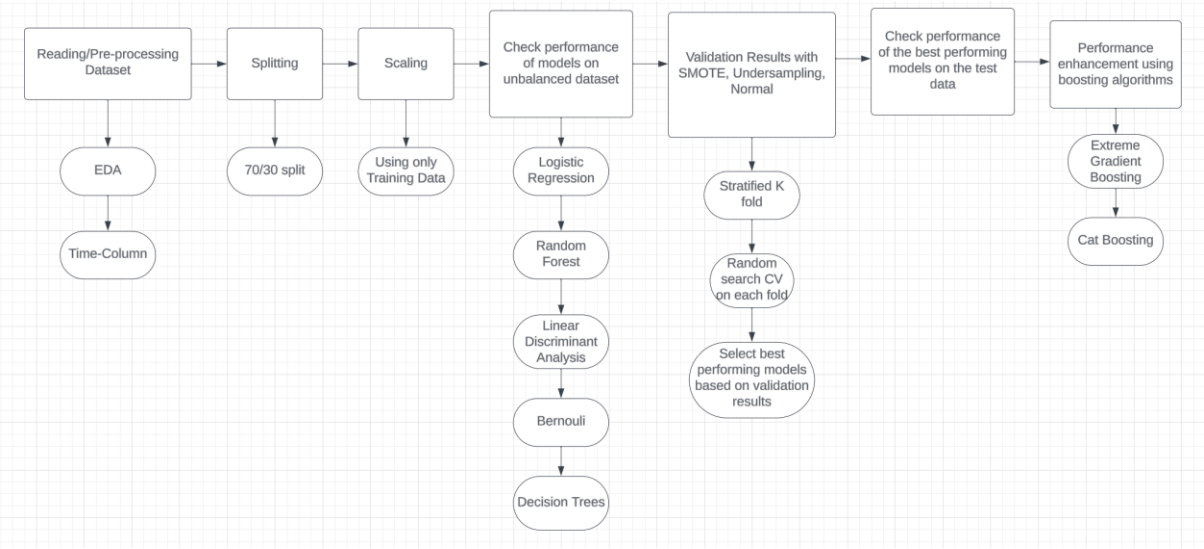

Figure 8 shows the Methodology we used for the project.

	Model	Class	Precision	Recall	F1-Score	Support
0	Logistic Regression	postive	0.837845	0.902889	0.869138	2512.6
1	Logistic Regression	negative	0.807305	0.894219	0.848508	1972.0
2	Logistic Regression	neutral	0.716342	0.328552	0.450413	753.0
3	Random Forest	postive	0.847864	0.898033	0.872215	2512.6
4	Random Forest	negative	0.80224	0.90284	0.849536	1972.0
5	Random Forest	neutral	0.838422	0.397078	0.538669	753.0
6	BernouliNB	postive	0.672119	0.907267	0.77217	2512.6
7	BernouliNB	negative	0.778757	0.612779	0.685737	1972.0
8	BernouliNB	neutral	0.709607	0.276228	0.397464	753.0
9	MultinomialNB	postive	0.820758	0.896362	0.856872	2512.6
10	MultinomialNB	negative	0.770702	0.911156	0.835013	1972.0
11	MultinomialNB	neutral	0.984436	0.211421	0.347931	753.0
12	SVC	postive	0.84869	0.913954	0.880099	2512.6
13	SVC	negative	0.818574	0.908824	0.861314	1972.0
14	SVC	neutral	0.811939	0.368924	0.507231	753.0

Figure 9 Shows the Results of all our models without using any lexicons in our Training Data

	Model	Class	Precision	Recall	F1-Score	Support
0	Logistic Regression	postive	0.84135	0.904242	0.871645	2512.6
1	Logistic Regression	negative	0.805692	0.893611	0.847354	1972.0
2	Logistic Regression	neutral	0.715002	0.332005	0.453298	753.0
3	Random Forest	postive	0.845446	0.90878	0.875957	2512.6
4	Random Forest	negative	0.811475	0.902333	0.854473	1972.0
5	Random Forest	neutral	0.850564	0.388313	0.533018	753.0
6	BernouliNB	postive	0.700134	0.904959	0.789458	2512.6
7	BernouliNB	negative	0.787985	0.657505	0.716756	1972.0
8	BernouliNB	neutral	0.644307	0.294024	0.403661	753.0
9	MultinomialNB	postive	0.831211	0.890154	0.859637	2512.6
10	MultinomialNB	negative	0.75918	0.919473	0.831618	1972.0
11	MultinomialNB	neutral	0.992418	0.208234	0.344159	753.0
12	SVC	postive	0.841283	0.909258	0.873929	2512.6
13	SVC	negative	0.805954	0.903144	0.851758	1972.0
14	SVC	neutral	0.793539	0.328552	0.464549	753.0

Figure 10 Shows the Results of all our Models with adding Features using the lexicons

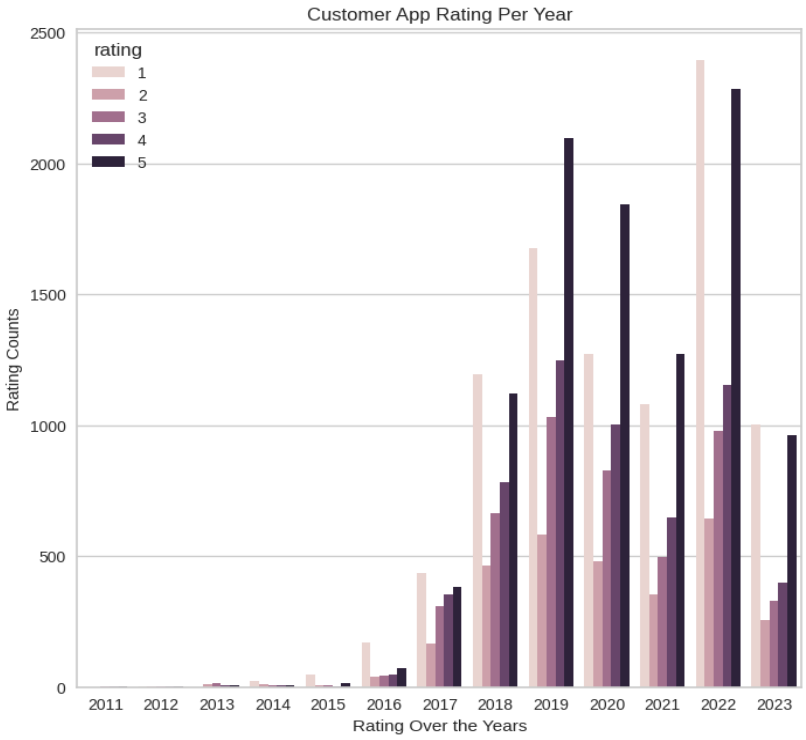

Figure 11 Shows the Average Ratings over Time broken down by Rating.

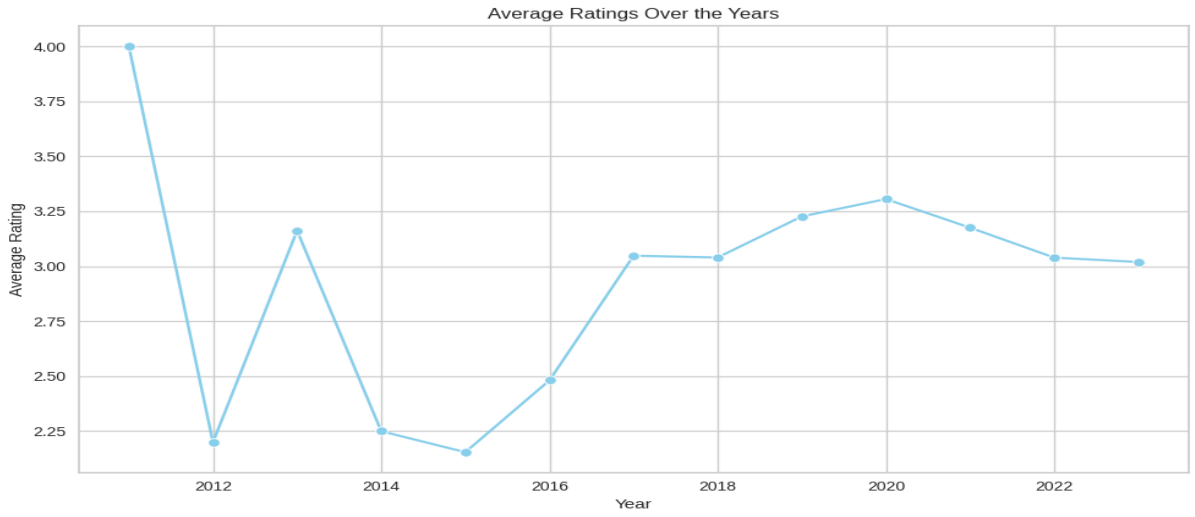

Figure 12 Shows the Average Ratings over Time.

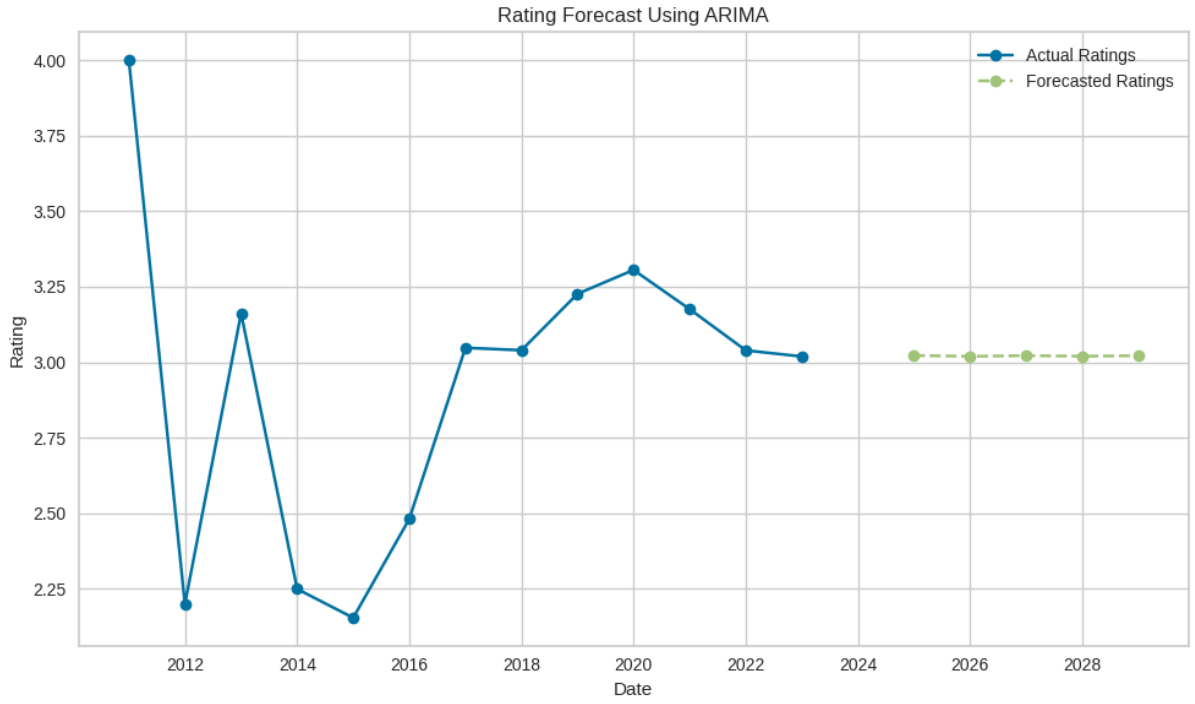

Figure 13 Failure of Forecasting