

**A**

**PREDICTIVE DOCUMENT**

**ON**

**MOVIELENS DATASET**

**Prof. Yegin Genc**

**Submitted By:**

* **Akshit Shah**
* **Aman Trivedi**
* **Sujan Reddy**
* **Nazeer Budhwani**

**INTRODUCTION**

After doing successful descriptive and explanatory analysis, now we will perform Predictive Analysis. In Predictive Analysis, we make predictions about our dataset. We make the equations that can explain the relationship between the two variables and we see that how these variables are related to each other and how they are being affected by the dataset.

First step we did is to merge movies.csv and ratings.csv. We did this because the predicting variable is in ratings.csv and details about the movie like name year released and genres are present in movies.csv so we merged on based on movie ID Since, they are common throughout the different file.

## **Model Building and Model Results**

We have taken following dependent variables: -

1) Ratings

After exploring the dataset, we came to know that following variables plays a key role in our dataset.

1) Year

So, our next step is to make predictions on these variables and to find out the relationship between them. This will help us to figure out how these variables play key role in our dataset.

As the data in our dataset is based on time series since we have a data from a wide range of year we will be using time series models in our project. We have eliminated using of Regression and Classification because we have less explanatory variables and the accuracy of model would not be high enough.

We have used two forecasting algorithms i.e. ARIMA and HoltWinter.

ARIMA stands for Autoregressive Integrated Moving Average models. Univariate (single vector) ARIMA is a forecasting technique that projects the future values of a series based entirely on its own inertia.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

* p is the number of autoregressive terms,
* d is the number of nonseasonal differences needed for stationarity, and
* q is the number of lagged forecast errors in the prediction equation.

The forecasting equation is constructed as follows. First, let y denote the dth difference of Y, which means:

If d=0: yt = Yt

If d=1: yt = Yt - Yt-1

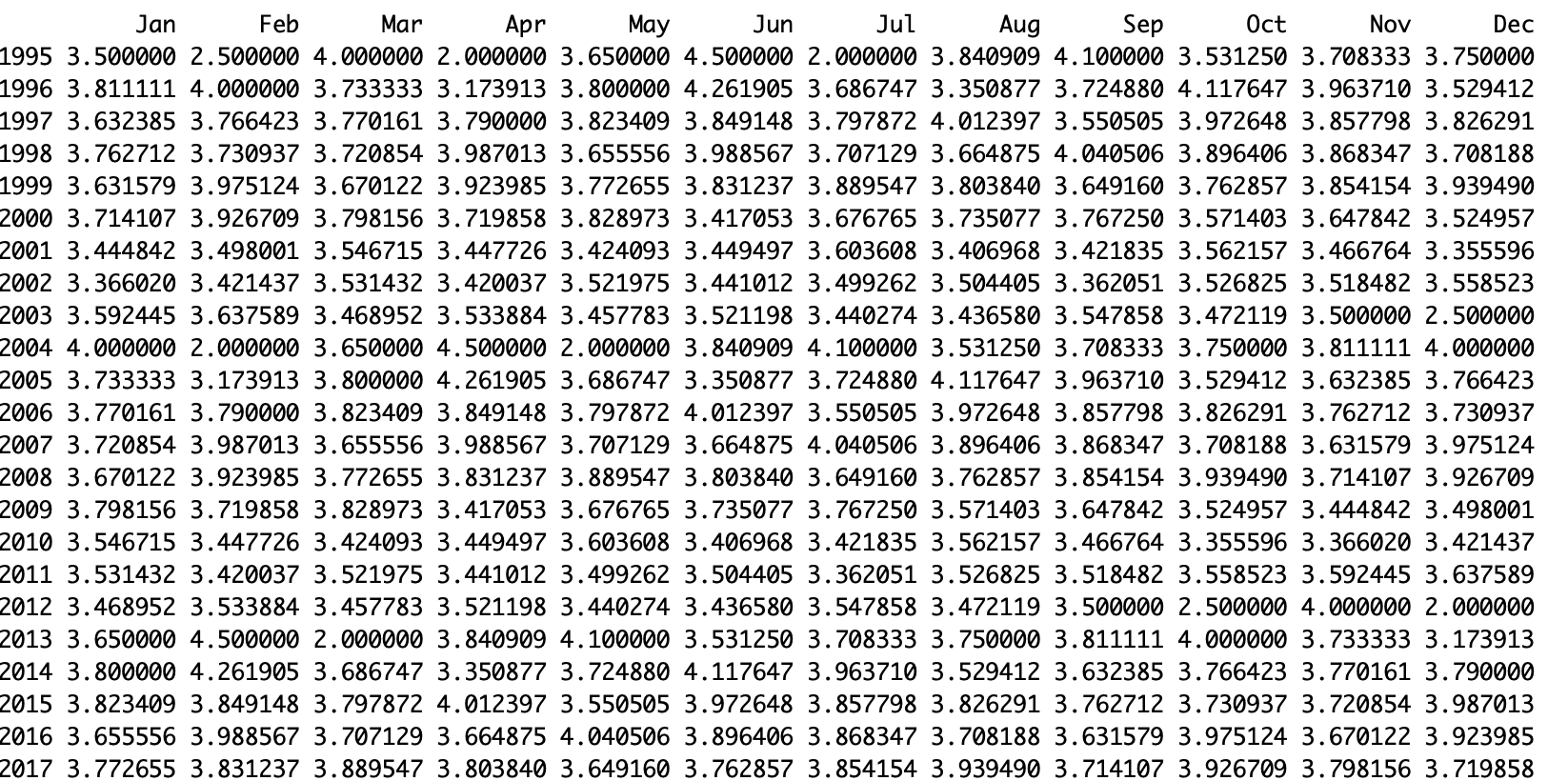
If d=2: yt = (Yt - Yt-1) - (Yt-1 - Yt-2) = Yt - 2Yt-1 + Yt-2

Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict “typical” values for the present and future

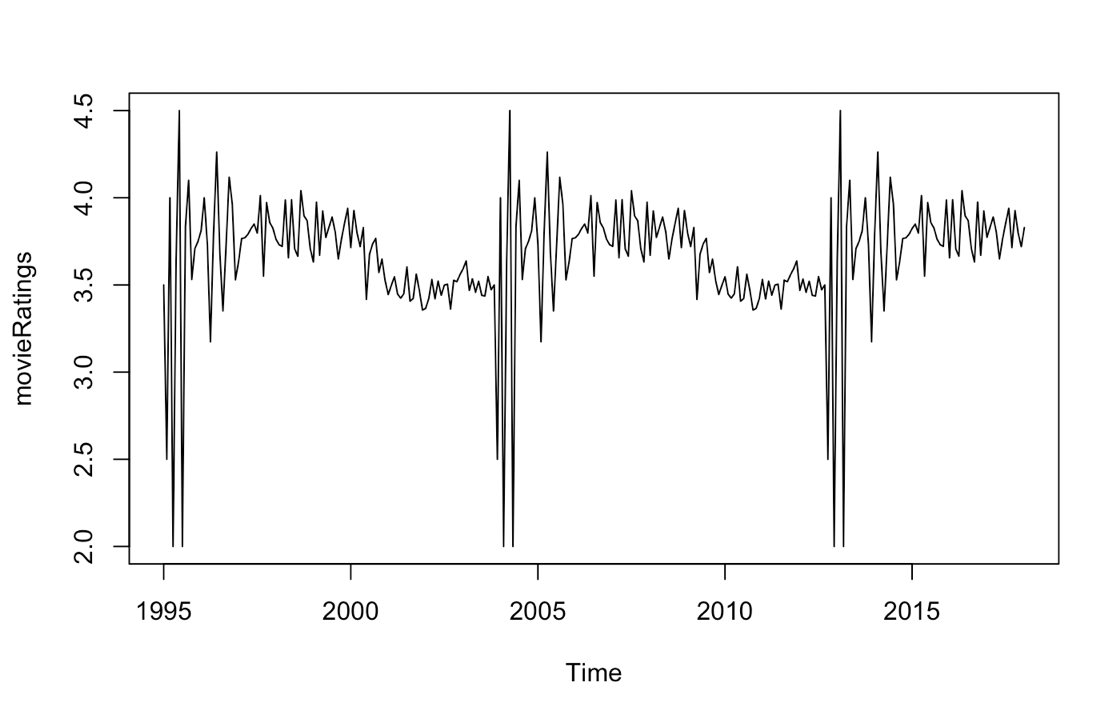
The three aspects of the time series behavior—value, trend, and seasonality—are expressed as three types of exponential smoothing, so Holt-Winters is called triple exponential smoothing. The model predicts a current or future value by computing the combined effects of these three influences. The model requires several parameters: one for each smoothing (ɑ, β, γ), the length of a season, and the number of periods in a season

**Process/Work Completed**

In Descriptive We merged the movies.csv and ratings.csv into one file and separated the year from movie\_title. Here then we are Preparing the time-series dataset by calculating the average ratings of each individual year we then exported into another dataset named as Year\_by\_ratings. We then developed a time series object us ts() function from library fpp which would help us in plotting time series graph

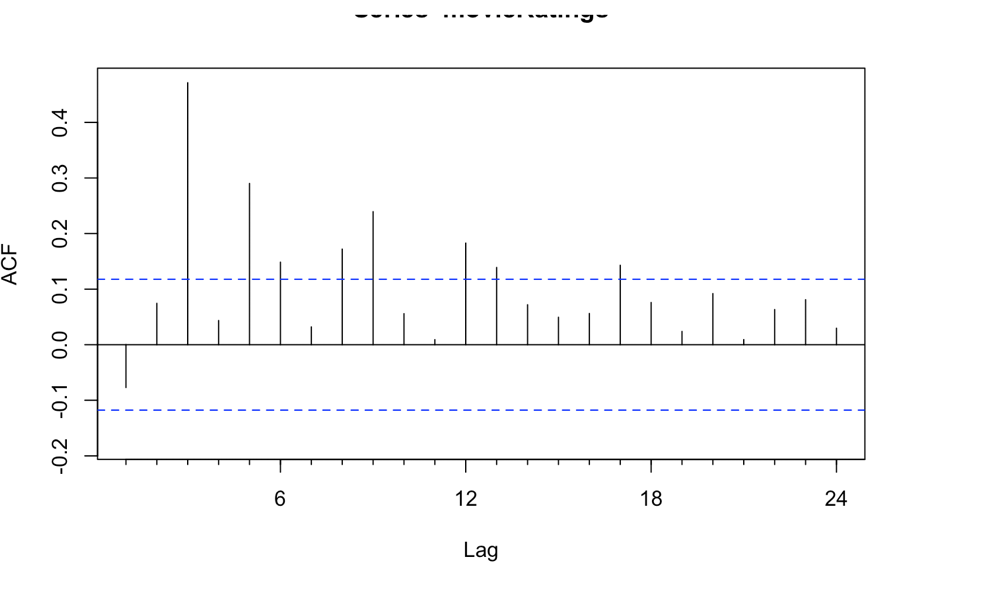


Time series object snapshot using ts function

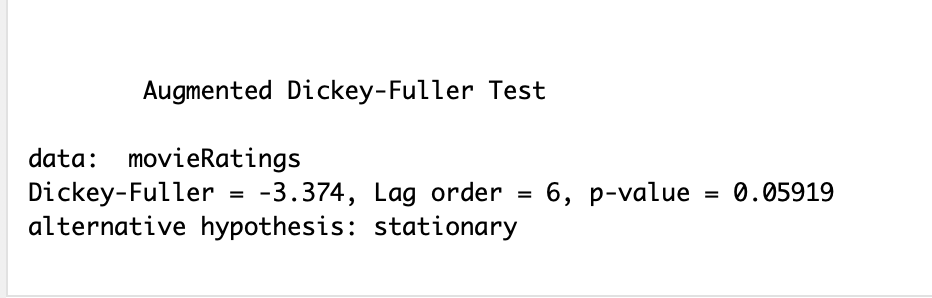


Visual representation of the time-series object

We have selected the year from 1995 to 2018 because the variations before 1995 were not too high and it did not added value into the model. So we selected from 1995 to 2018 and we added frequency =12 because it’s a yearly data.

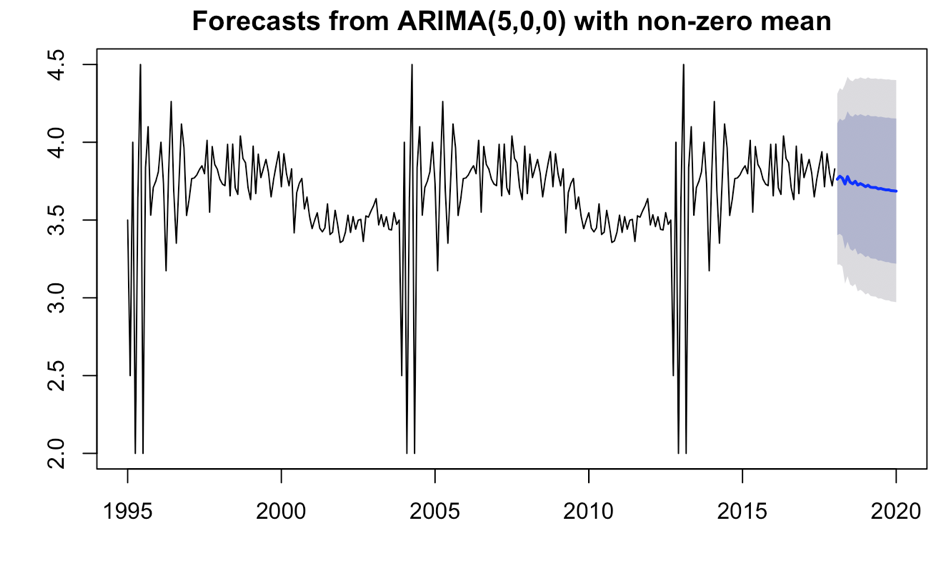


In the above graph we are checking the Trend and Seasonality of the Dataset because if the whole data is between the 0.1 and -0.1 then there are high chances that the whole data has random values and cannot be further to progress, But In our dataset the data has a Seasonality so it can be loaded into the time series.



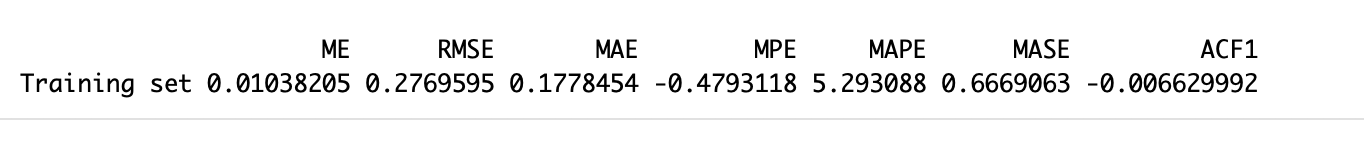
We conducted the Augmented Dickey-Fuller Test to check that if we have unit root present in the time-series model, If the p-value is very large compared to significant p-value we would have rejected the data but it’s very close and alternative hypothesis is also stationary. So we can say our data rejects null hypothesis.

We then developed ARIMA model using the library forecast to find the predictions of rating for future years

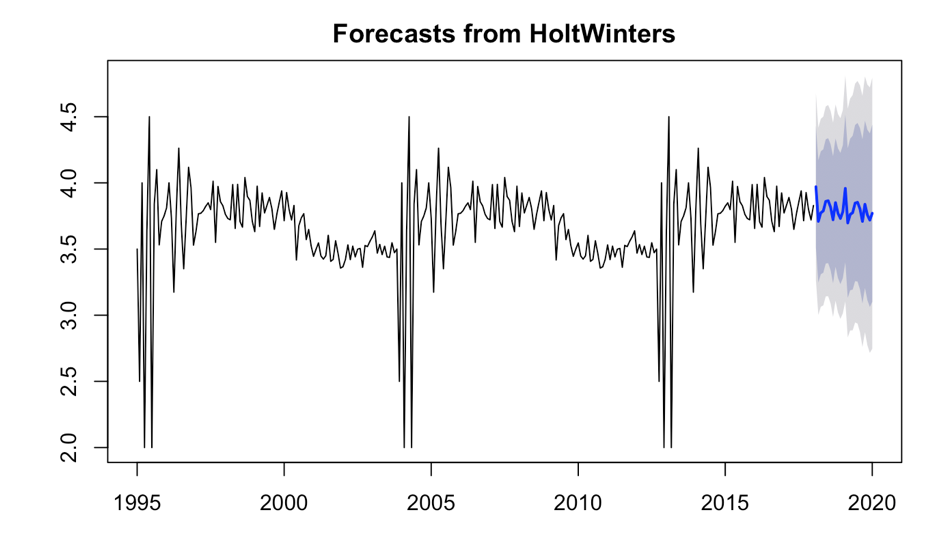


Here we are predicting for years till 2020 and we see that majority of ratings are between 3.5-4.0

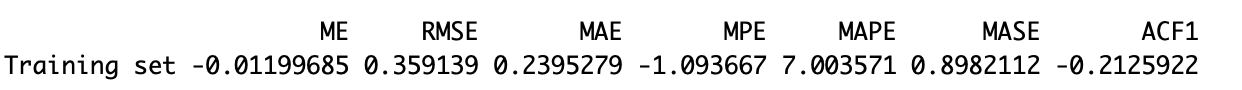
And we check for accuracy using RMSE (room mean square error)

Here we have RMSE of 0.27

Predicting Rating using Holt Winters Model:



Using the HoltWinter model we see that the predicting rating are closer to 4.0 and when we check RMSE for HoltWinter Model it is 0.35



**Conclusion**

In the given project, we created 2 models named ARIMA and HoltWinter, we then checked the accuracy based on the RMSE value. RMSE should be as low as possible. So after plotting both the models, we came to the conclusion that the accuracy of the ARIMA model is much better than the HoltWinter model. In both the model we got rating prediction between 3.5-4.0