BUSINESS REPORT

Airline Forecasting and Sentiment Analysis

Insight Report

2023

PREPARED BY:

Bobbi McDermott

CONTENTS

INTRODUCTION

The dataset is moderate sized comprising of 14640 observations and 15 features. It contains information on airlines in the US and a collection of tweets about the service, collected across 9 days. The objective is to evaluate the frequency of the tweets over a period of time to forecast future tweets. I'll look at sentiment analysis of the tweets to get a more accurate analysis of the Airline feedback was often neutral.

I'm going to perform both time series and text analysis.

Time Series is beneficial in business as it analyses, trends, seasonality, cyclicity and irregularity (Research Otimus, 2023) Behaviour changes in customer sentiment is important to the aviation industry as it can recognize habbits that can drive aircraft purchasing, flight routes, added extras and environmental changes (Events, 2022)

Text sentiment is crucial for business today with everything being digital. Airlines need to utilize this feaure to shape sales, marking, social media, brand stregth and services ("What Is Sentiment Analysis?," n.d.)

SECTION 1: BUSINESS UNDERSTANDING

The company that has commissioned this report is PlaneSimple, a new airline wishing to gain insights into the customer engagement and feedback from data collected on existing airlines. As social media is a high influencer in terms of retention, upsell and acquisition, it can also determine your brand and your audience. Therefore it is an excellent approach in the marketing launch strategies

SECTION 2: DATA UNDERSTANDING

The Data required cleaning as there was some missing cells (28.2%) in the overall data, Figure 1

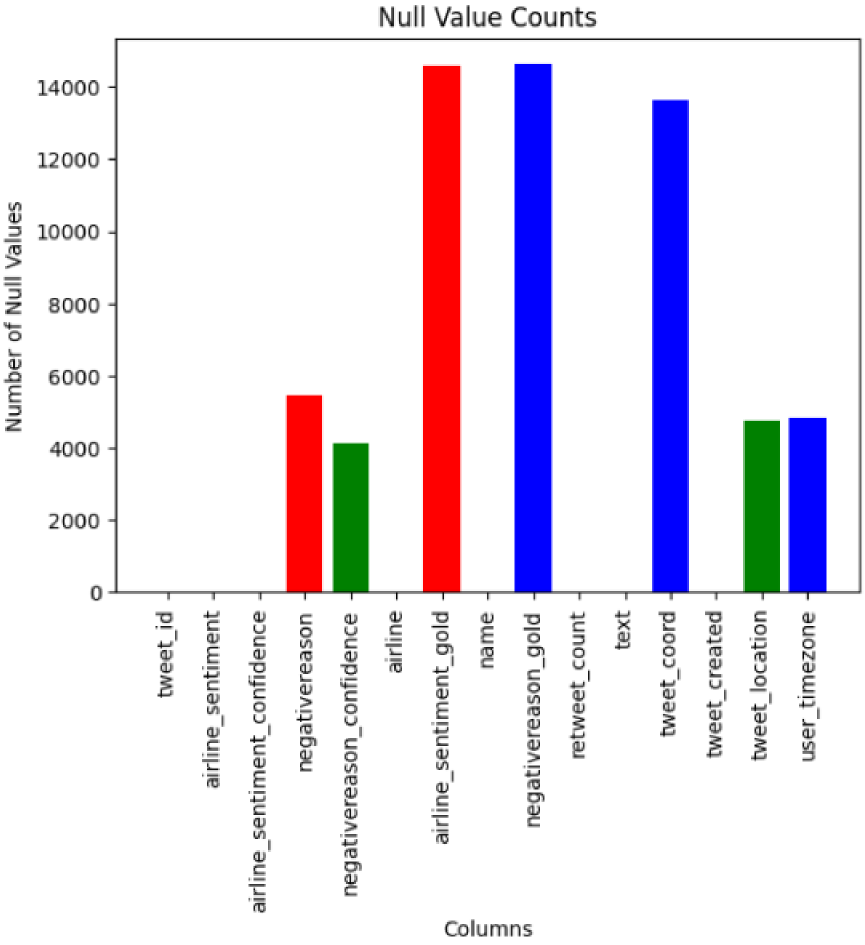


Figure 1

It encompassed global reviews. There were neutral reviews which do not provide much information.

SECTION 3: DATA PREPARATION / DESCRIPTION

Overall Preparation

I removed the Null values that did not affect our interpretation goal. The main outliers were in the retweet_count, however the IQR and box plots picked up all values as outliers so this is incorrect, as context clearly required all values. Figure 2

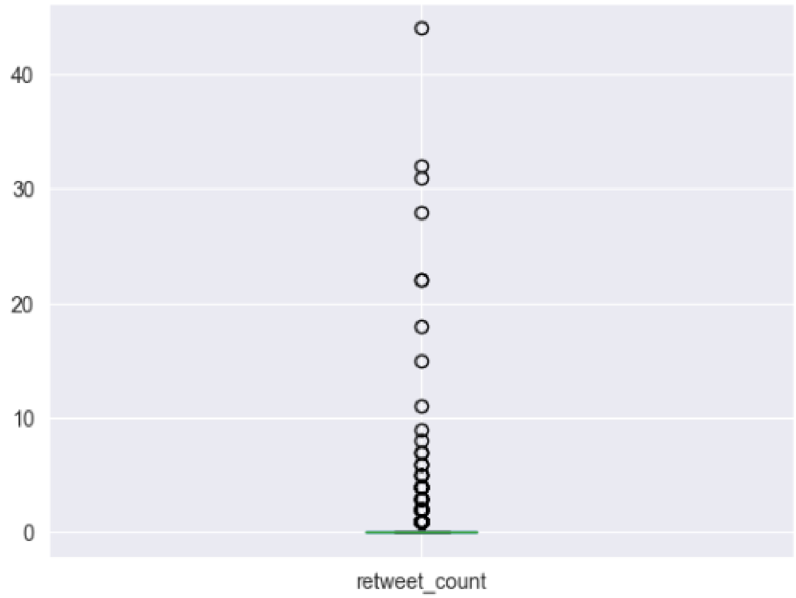


Figure 2

I populated missing tweets with those that were referenced and removed the observations with none.

Preparation for Time Series

For preparation of the model, I defined the time series dataframe of ‘tweet_created’ and ‘retweet_count’. I indexed the ‘tweet_created’ and sampled the data on an hourly basis over the 9 days and summed the ‘retweet_count’

SECTION 4: TIME SERIES: RETWEET COUNT

I wanted to get an overall view of how the time series looked per re_retweet count and established that the days the highest tweets occurred were Wednesday, Friday, Sunday and Monday, Figure 3

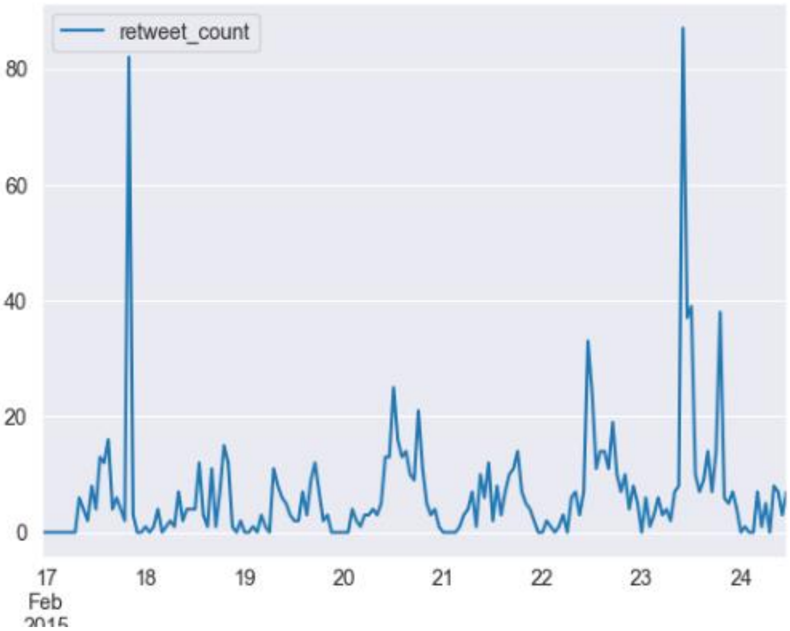


Figure 3

Autocorrelation and Partial Autocorrelation

Here I saw a few statistical anomalies but that the series was likely to be stationary, Figure 4, Figure 5

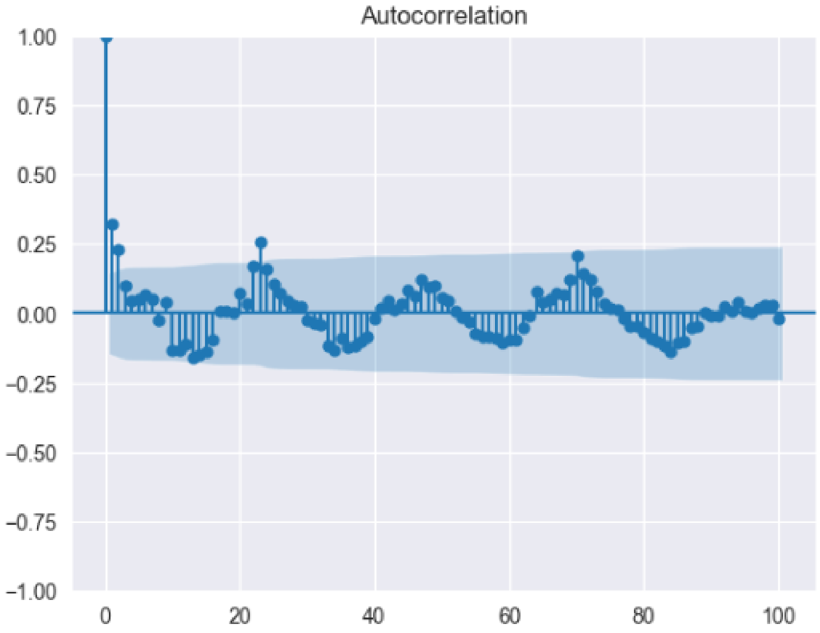


Figure 4

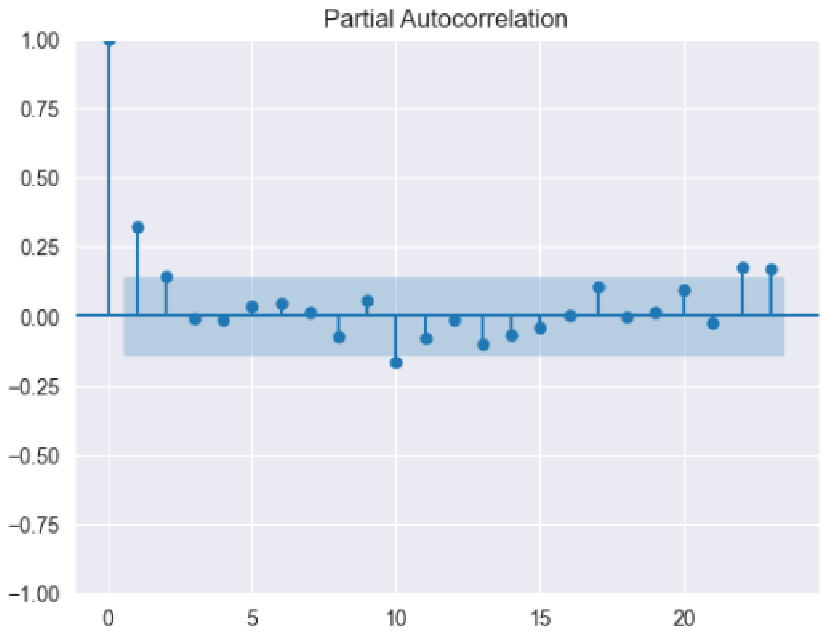


Figure 5

Augmented Dickey_Fuller Test to Determine Stationarity

It’s important to know stationarity of a series before defining a model as the statistical elements are different and the wrong model can return incorrect predictions (Rasheed, 2020)

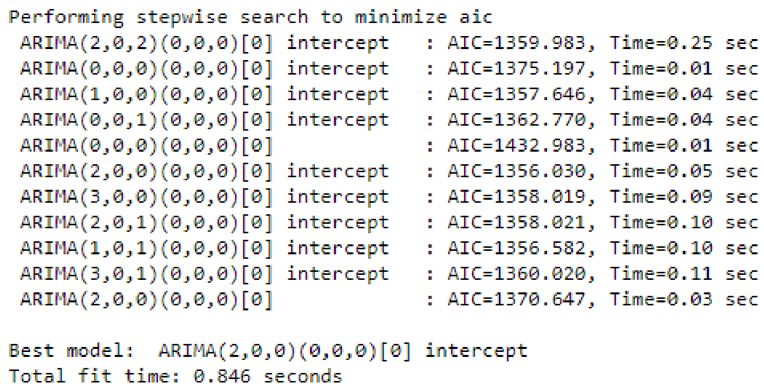
The results: the statistic (-6.735140719699589): The time series' withdrawal from stationarity is clear by this negative number. The evidence against the null hypothesis of non-stationarity is higher the more negative the value.. The p-value (3.221724499893661e-09): Here I can see the likelihood that the test statistic will be seen if the null hypothesis—that the time series is not stationary. This is correct. The null hypothesis can be rejected and the series can be determined to be stationary again (G, 2021)

Optimal Arima (Auto Regressive Inegrated Moving Average) Parameters

I will utilize The Akaike Information Critera (AIC) as it’s a widely used measure of a statistical model. It basically quantifies 1) the goodness of fit, and 2) the simplicity/parsimony, of the model into a single statistic.(Keshvani, 2013)

I use pmdarima, as it is a sophisticated and encompasses analysis capabilities including a collection of statistical tests of stationarity and seasonality(pypi.org, 2023)

This determined the best model fit and my p,d,q values, Figure 6



```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=1359.983, Time=0.25 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=1375.197, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=1357.646, Time=0.04 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=1362.770, Time=0.04 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=1432.983, Time=0.01 sec
 ARIMA(2,0,0)(0,0,0)[0] intercept   : AIC=1356.030, Time=0.05 sec
 ARIMA(3,0,0)(0,0,0)[0] intercept   : AIC=1358.019, Time=0.09 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=1358.021, Time=0.10 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=1356.582, Time=0.10 sec
 ARIMA(3,0,1)(0,0,0)[0] intercept   : AIC=1360.020, Time=0.11 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=1370.647, Time=0.03 sec

Best model:  ARIMA(2,0,0)(0,0,0)[0] intercept
Total fit time: 0.846 seconds
```

Figure 6

In the p, d, q model, p denotes the order of the autoregressive element, d the amount of initial differencing that is involved, and q the order of the moving average part (Hyndman and Athanasopoulos, 2018)

Train Test Split

Train test split is different in time series as it’s not random and follows a linear pattern to forecast (Radečić, 2022), Figure 7

I ran an 80/20 Split

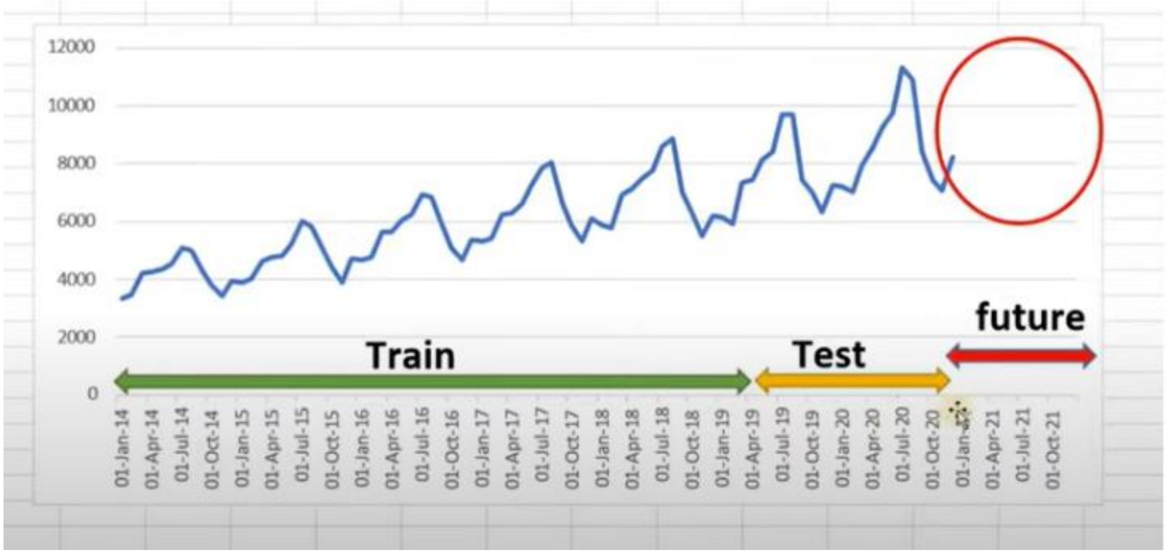


Figure 7

Arima Model

The model itself performed well showing a consistency with the original data, Figure 8, Figure 9

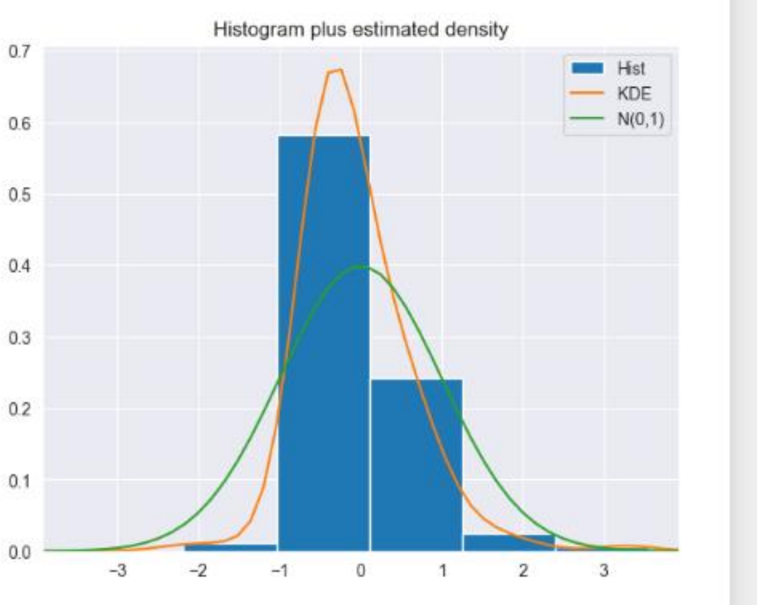


Figure 8

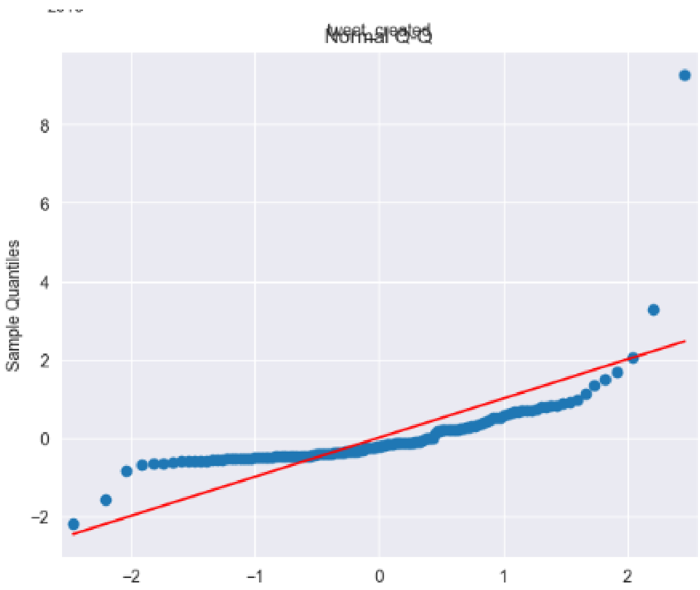


Figure 9

This plot is consistent with our previous observations there suggests the model is performing well overall. The future predictions are in alignment with the true values indicating it's showing an accurate slightly increasing trend, Figure 10

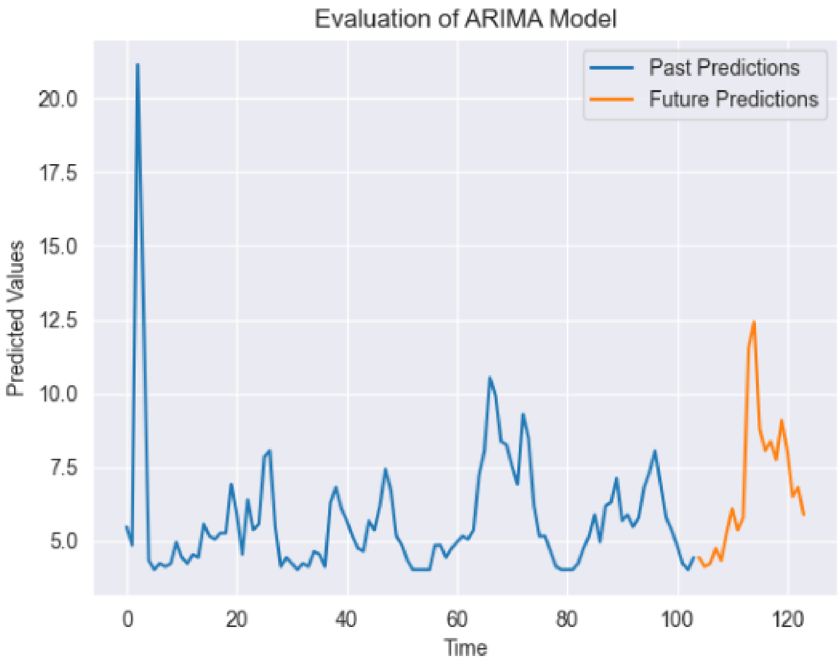


Figure 10

SECTION 5: TEXT / SENTIMENT ANALYSIS

Text Analysis is used to extract meaning from unstructured data to provide insights and allow it to be comprehended (Milward, 2023)

Tokenizing

Tokenization in machine learning is a process of splitting text into meaningful parts that can be represented as vectors (Kosar, 2022)

Airline Sentiment

With text / sentiment analysis I wanted to view the Positive/Negative in the airline_sentiment as it can provide Plane Simple with data on what works well and what requires better business decisions (Podolsky, 2022) Figure 11, Figure 12, Figure 13, Figure 14, Figure 15

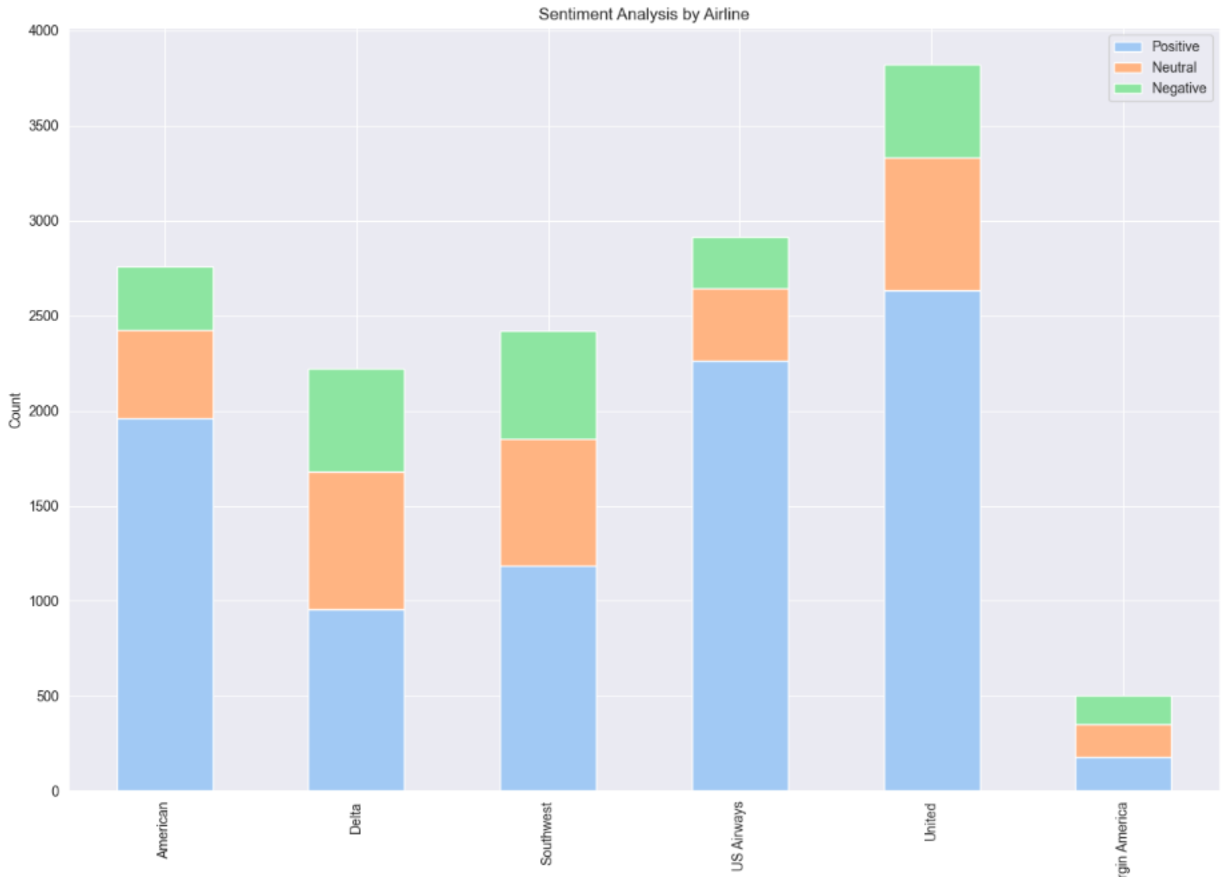

Sentiment Analysis by Airline
Count
0
500
1000
1500
2000
2500
3000
3500
4000
Positive
Neutral
Negative
American
Delta
Southwest
US Airways
United
rgin America

Figure 11

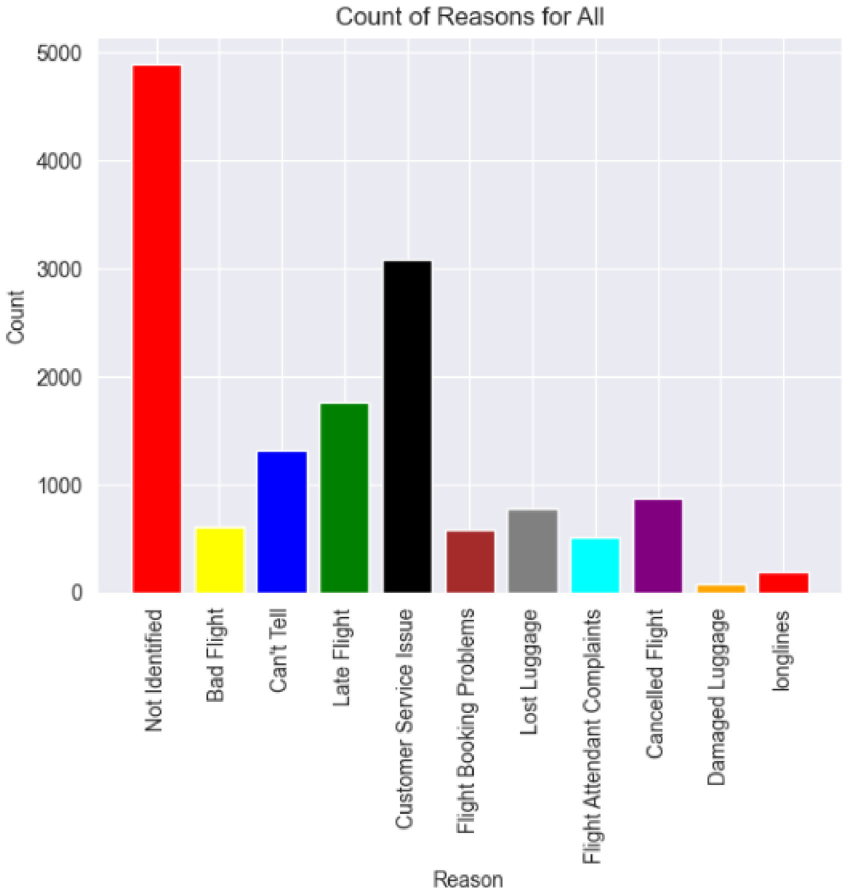

Count of Reasons for All
Count
0
1000
2000
3000
4000
5000
Not Identified
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines
Reason

Figure 12

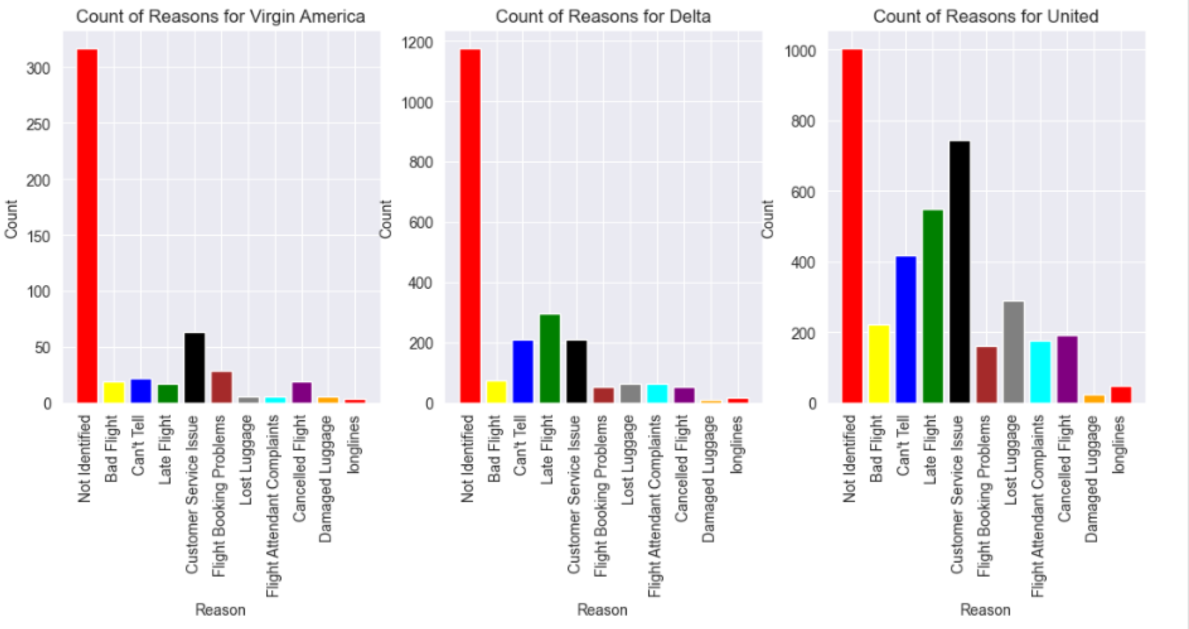

Count of Reasons for Virgin America
Count of Reasons for Delta
Count of Reasons for United
Count
Reason
0
50
100
150
200
250
300
400
600
800
1000
1200
Not Identified
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines

Figure 13

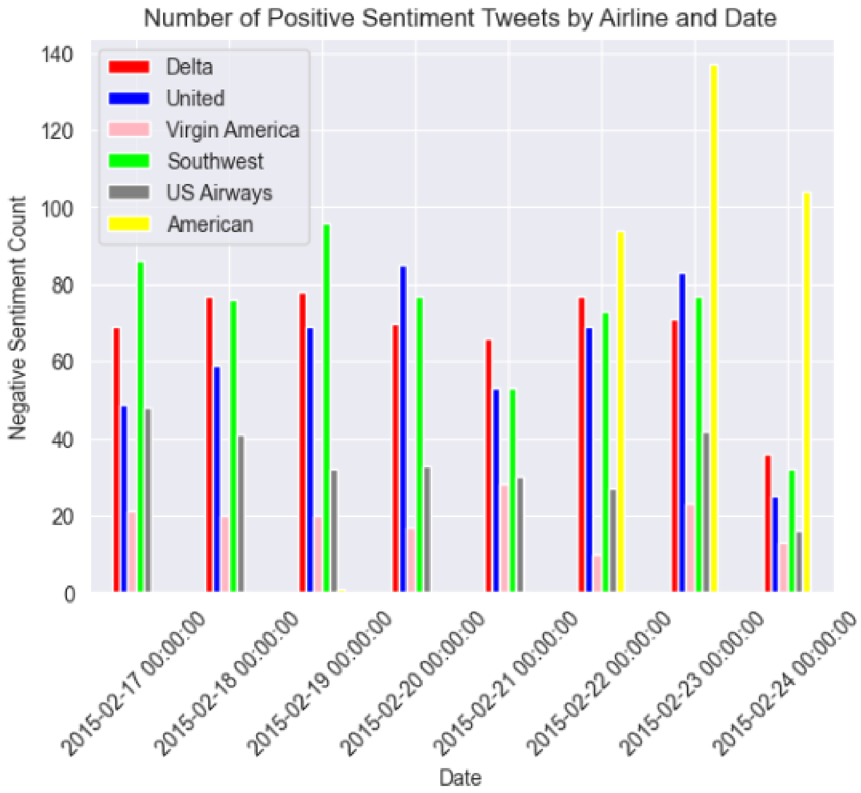

Number of Positive Sentiment Tweets by Airline and Date
Delta
United
Virgin America
Southwest
US Airways
American
Negative Sentiment Count
0
20
40
60
80
100
120
140
2015-02-17 00:00:00
2015-02-18 00:00:00
2015-02-19 00:00:00
2015-02-20 00:00:00
2015-02-21 00:00:00
2015-02-22 00:00:00
2015-02-23 00:00:00
2015-02-24 00:00:00
Date

Figure 14

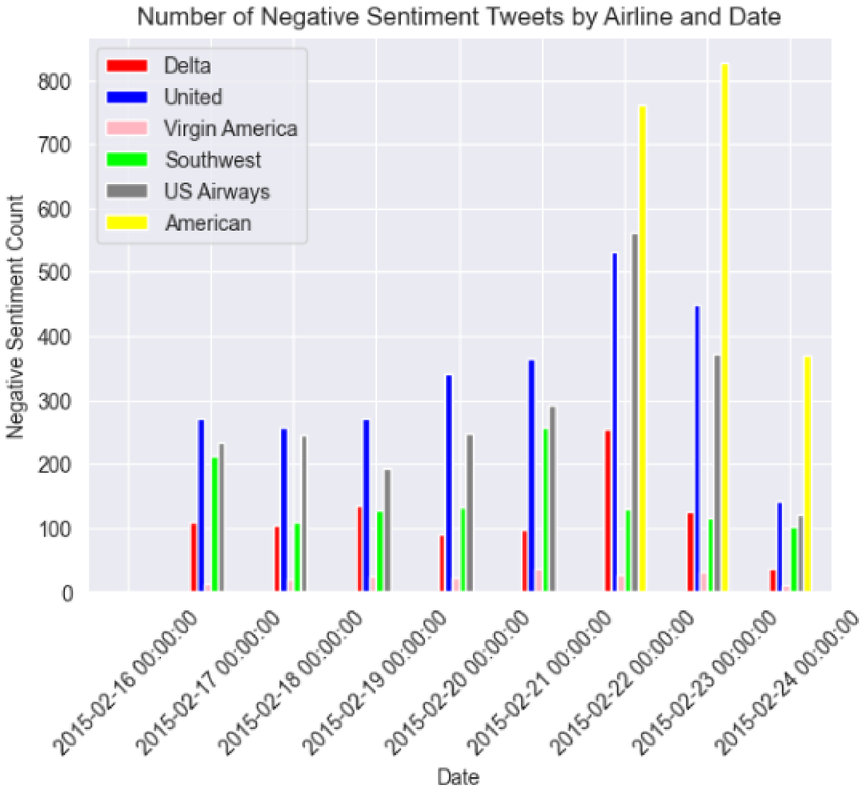


Figure 15

Compound Score

A compound score is the aggregate of the score of a word, or precisely, the sum of all words in the Vadar lexicon, normalized between -1 and 1 (Shah, 2021)

It determines the sentiment score for each collection of text

Multinormial Naieve Baes

This algorithm frequently serves as a starting point. The basic idea of Naive Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes (Smetanin, 2018)

The Model performed well with an Accuracy of 0.89 and an F1 Score of 0.818. The accuracy measures overall correctness while F1 measures precision and recall, Figure 16

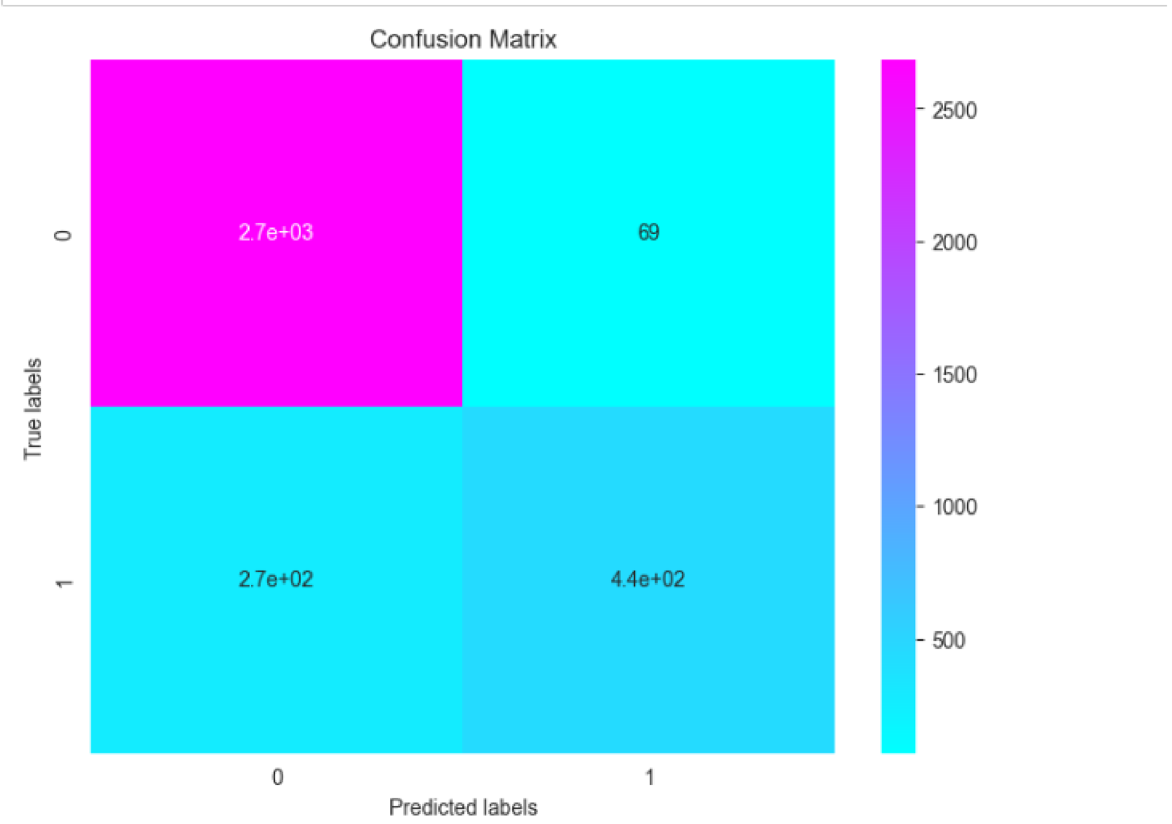


Figure 16

TensorFlow was implemented to determine dimensionality of embedded space and the architecture of the model (Brownlee, 2017), Figure 17, Figure 18

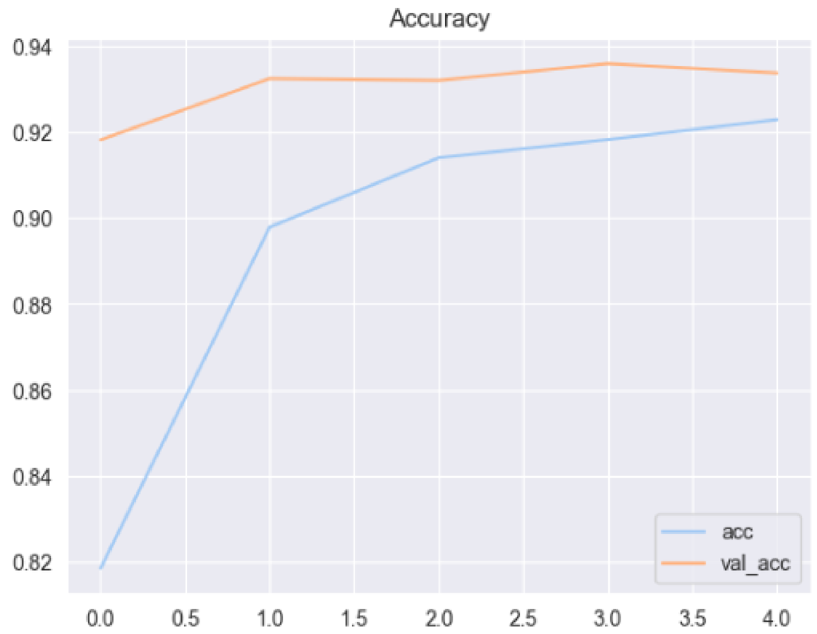


Figure 17

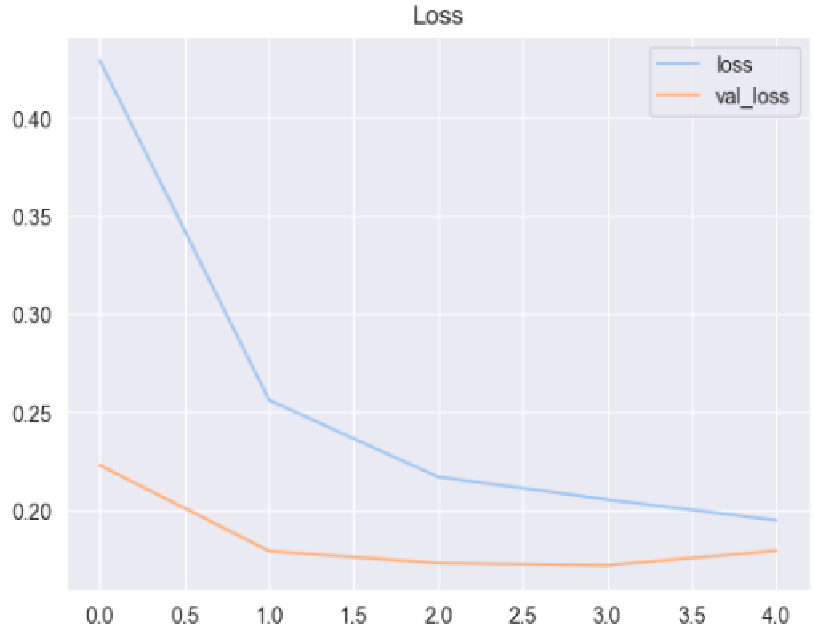


Figure 18

SECTION 6: CHALLENGE: TIME SERIES POSITIVE SENTIMENT

The data was smaller as I decided to take the compound scores that were calculated in text analysis to create a new dataframe, only containing the positive values. The dataframe was smaller with 5784 observations.

The forecast appeared to follow a similar patter to the retweet_count as was also stationary and steady in the it's pattern and slightly increasing, Figure 19

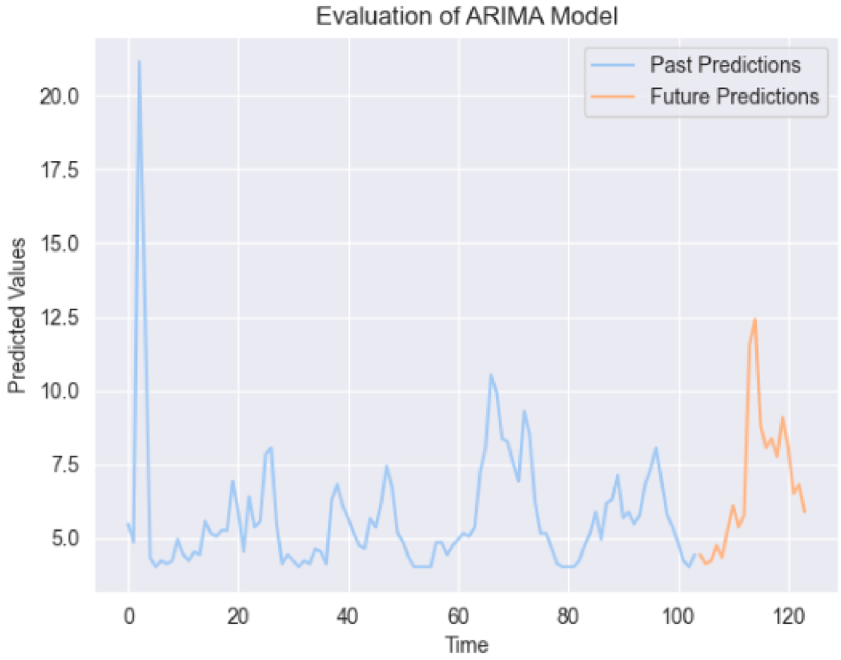


Figure 19

SECTION 7: TIME SERIES LOCATION

I did time series on location as it is interesting to see if the tend deviated from that of the retweet count or positive tweet sentiment

I was glad I conducted this analysis as it provided vital insight to the Airline. The time series was markedly different in that it forecasted a dramatic increase, Figure 20

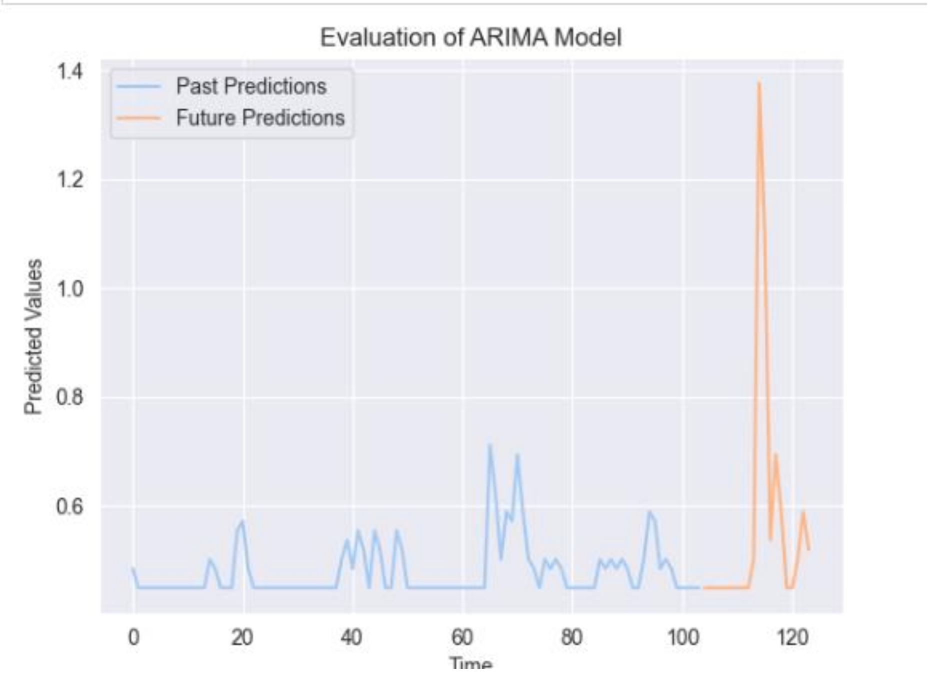


Figure 20

SECTION 8: RESULTS, CONCLUSIONS AND LIMITATIONS

Both results in time series and text analysis were successful. It shows how machine learning tools are extremely beneficial to business.

The time series indicated that a more precise breakdown per location would be warranted as there are different results per region. Text Analysis showed that he accuracy in determining the positive and negative tweets is advantageous in this digital age

The limitations were that the data had holes in it and it would have needed further analysis

BIBLIOGRAPHY

Brownlee, J., 2017. Deep Convolutional Neural Network for Sentiment Analysis (Text Classification). MachineLearningMastery.com. URL https://machinelearningmastery.com/develop-word-embedding-model-predicting-movie-review-sentiment/ (accessed 5.30.23).

Events, Uk.M.&, 2022. Creating mass consumer behavior change in the travel industry. Passenger Terminal Today. URL https://www.passengerterminaltoday.com/opinion/creating-mass-consumer-behavior-change-in-the-travel-industry.html (accessed 5.30.23).

G, V.K., 2021. Statistical Tests to Check Stationarity in Time Series. Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/ (accessed 5.29.23).

Hyndman, R., Athanasopoulos, G., 2018. 8.5 Non-seasonal ARIMA models | Forecasting: Principles and Practice (2nd ed).

Keshvani, A., 2013. Using AIC to Test ARIMA Models. CoolStatsBlog. URL https://coolstatsblog.com/2013/08/14/using-aic-to-test-arima-models-2/ (accessed 5.30.23).

Kosar, V., 2022. Tokenization in Machine Learning Explained [WWW Document]. URL https://vaclavkosar.com/ml/Tokenization-in-Machine-Learning-Explained (accessed 5.30.23).

Milward, D., 2023. What is Text Mining, Text Analytics and Natural Language Processing? Linguamatics [WWW Document]. URL https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing (accessed 5.30.23).

Podolsky, M., 2022. Council Post: Why Negative Reviews Can Help Your Business Improve [WWW Document]. Forbes. URL https://www.forbes.com/sites/forbesbusinesscouncil/2022/09/06/why-negative-reviews-can-help-your-business-improve/ (accessed 5.21.23).

pypi.org, pmdarima, 2023. pmdarima: Python’s forecast::auto.arima equivalent.

Radečić, D., 2022. Time Series From Scratch — Train/Test Splits and Evaluation Metrics [WWW Document]. Medium. URL https://towardsdatascience.com/time-series-from-scratch-train-test-splits-and-evaluation-metrics-4fd654de1b37 (accessed 5.30.23).

Rasheed, R., 2020. Why Does Stationarity Matter in Time Series Analysis? [WWW Document]. Medium. URL https://towardsdatascience.com/why-does-stationarity-matter-in-time-series-analysis-e2fb7be74454 (accessed 5.30.23).

Research Otimus, 2023. Time Series Analysis for Better Decision Making in Business [WWW Document]. URL https://www.researchoptimus.com/article/what-is-time-series-analysis.php (accessed 5.30.23).

Shah, R., 2021. Different Methods for Calculating Sentiment of Text. Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2021/12/different-methods-for-calculating-sentiment-score-of-text/ (accessed 5.30.23).

Smetanin, S., 2018. Sentiment Analysis of Tweets using Multinomial Naive Bayes | by Sergey Smetanin | Towards Data Science [WWW Document]. URL https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b (accessed 5.30.23).

What Is Sentiment Analysis? [WWW Document], n.d. . Business News Daily. URL https://www.businessnewsdaily.com/10018-sentiment-analysis-improve-business.html (accessed 5.30.23).