Customer Time-series Data Analysis and Revenue Prediction with ARIMA and CNN-LSTM Deep Learning Neural Networks

Table of Contents

1. Introduction

The purpose of this study is to conduct customer data analysis and to predict revenue in the future 30 days for company AAVAIL. AAVAIL is a video service company similar to NETFLIX. To increase the competitiveness, AAVAIL launched a tiered, subscription-based service which had showed promise in the USA. Therefore, the experiment testing the new approach was carried out outside of the US and there are now a couple of years of data with a few thousand active users. The data are transaction-level purchases across 38 different countries and are invoiced in batches. Management has nearly decided to make the switch to the new model, but they find it difficult to predict monthly revenue.

The aim of this project is first to do the customer data analysis on the current database and then build a model that, at any point in time, to predict the revenue for the following month especially for the top ten countries with the most revenue.

In addition, for easy applying by other teams with less IT skills like the marketing team, the management needs the API to automate the whole process, from extracting relevant data from multiple data sources to model deployment for revenue prediction.

2. Data acquisition and quality assurance

2.1 Data sources

The data for this case study comes from the online retail data set and are available through the UCI Machine Learning Repository. The original data were published as a study that used RFM model to explore customer segmentation in the data set. The data presented in this study are derived from this data set with simulated features and re-named columns to align with the AAVAIL case study. The datasets are a couple of jason files with details of the transaction from the customer since 2017.

2.2 Data cleaning for quality assurance

Data downloaded from multiple files were combined into one table. There were a lot of missing values due to the inconsistent column name which had been fixed and the customer data with extreme values (for instance the product price is below zero) had been removed. The data also have been sorted in a chronological order to carry out predictive or forecasting analysis.

Table 1. First five rows of the cleaned dataset

	country	customer_id	invoice	price	stream_id	times_viewed	year	month	day	invoice_date
1	United Kingdom	13085.0	489434	6.75	79323W	12.0	2017	11	28	2017-11-28
2	United Kingdom	13085.0	489434	2.10	22041	21.0	2017	11	28	2017-11-28
3	United Kingdom	13085.0	489434	1.25	21232	5.0	2017	11	28	2017-11-28
4	United Kingdom	13085.0	489434	1.65	22064	17.0	2017	11	28	2017-11-28
5	United Kingdom	13085.0	489434	1.25	21871	14.0	2017	11	28	2017-11-28

3. Methodology

We will need to state the ideal data to address the business opportunity and clarify the rationale for needing specific data. The ideal data would contain a feature set based on which the revenue of the AAVAIL could be predicted such as the number of subscribers, types of subscription, location etc and the target variable would be the revenue. This will help in building a supervised learning pipeline which would be predicting the target variable "Revenue" based on the feature set or dependent variables.

An alternative scenario would be the application of time-series analysis where the revenue of AAVAIL is given along with the timestamp. Thus, a time-series forecasting technique (ARIMA and Deep Learning) could be applied to obtain the revenue prediction based on the historical data.

First, the exploratory analysis will be conducted to reveal general information:

1. Assimilate the business scenario and articulate testable hypotheses.
2. State the ideal data to address the business opportunity and clarify the rationale for needing specific data.
3. Investigate the relationship between the relevant data, the target and the business metric.
4. Create a function/python script to extract cleaned data for modelling from multiple data sources and automate the process of data ingestion.

The second step will focus on modelling:

1. To forecast with ARIMA with the generated time-series data
2. To create features for revenue forecasting with baseline model Random Forest
3. To develop a deep neural network with CNN-LSTMs

4. Exploratory Data Analysis

4.1 Top 10 countries with the most revenue

It is clear from Figure 1 that the UK dominates the market outside the US. EIRE, Germany, France and Norway are the second.

Figure1. Top 10 countries with the highest revenue

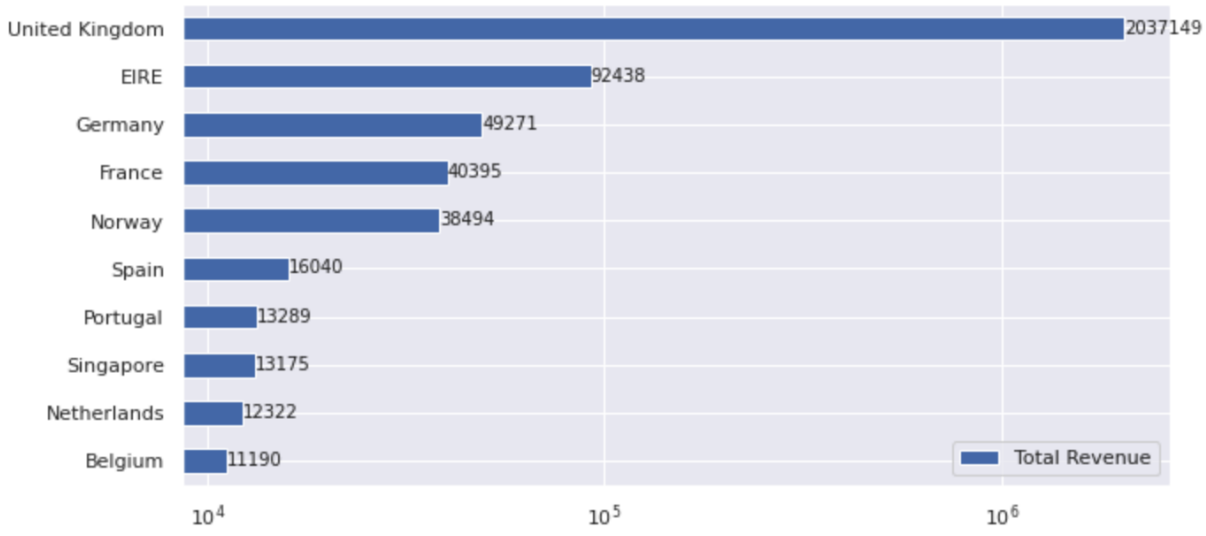
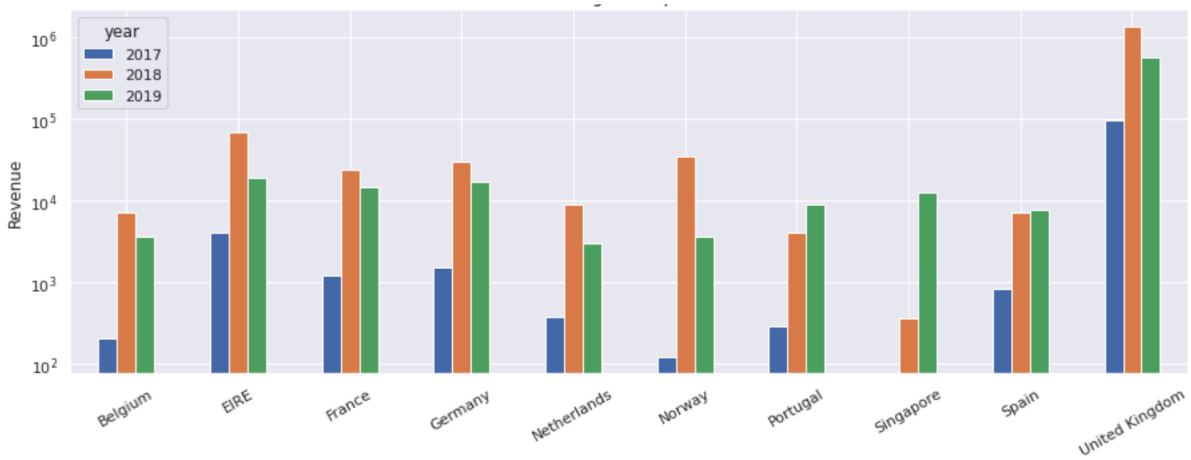
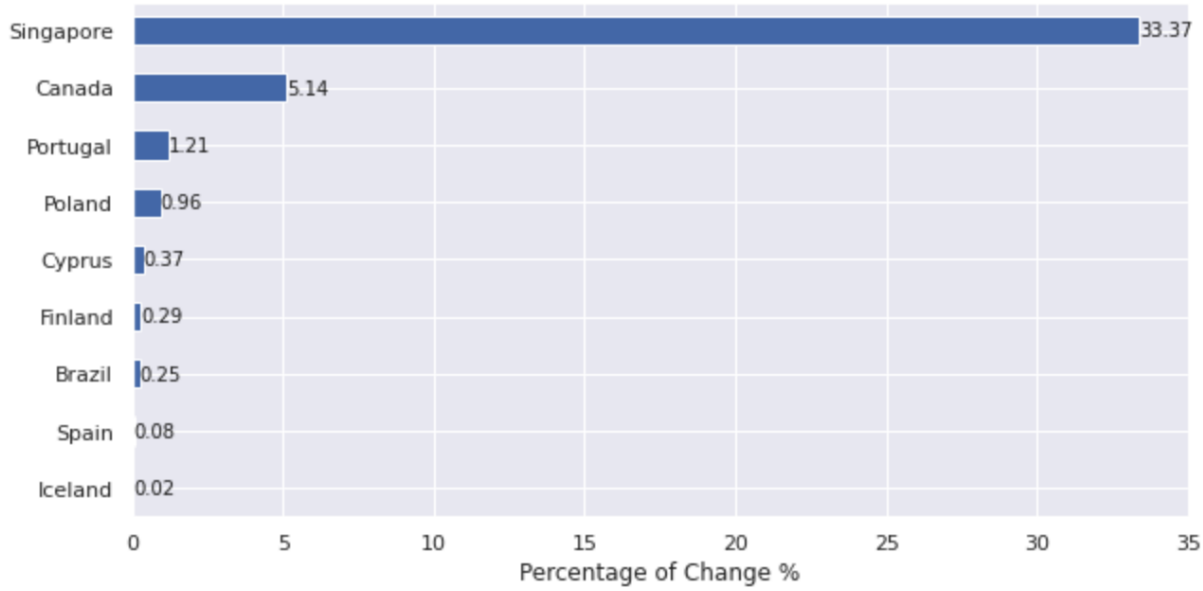


Figure 2. Revenue changes of Top 10 countries



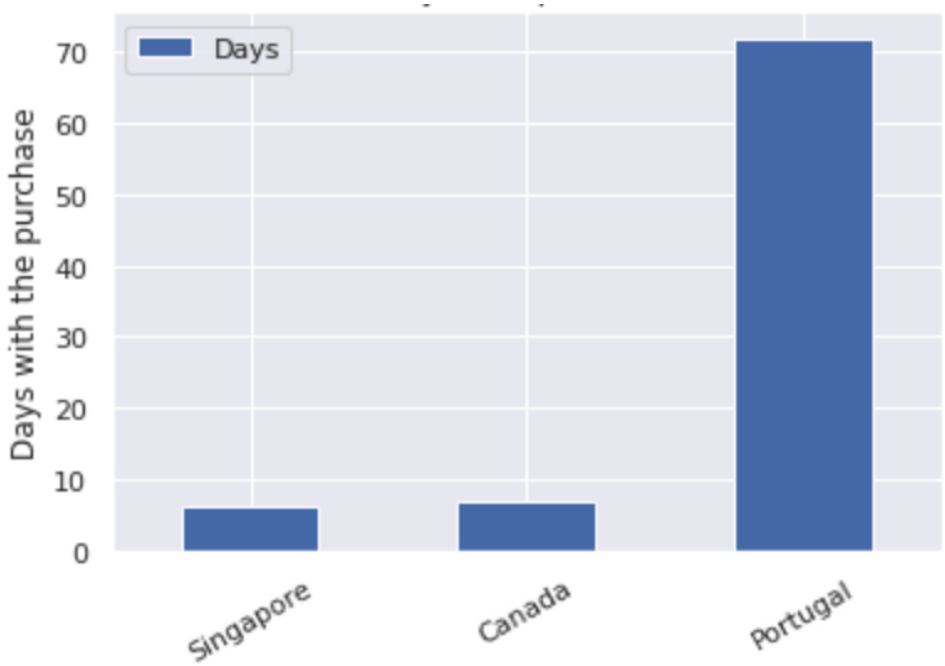
Interestingly, among the top 10 countries, the revenue was increased only in three countries. It is clearer from the figure below that there were 9 of total 34 countries managed to increase the revenue since 2018 and three of them are among the top 10 countries with the highest revenue which are Portugal, Singapore and Spain. Since Singapore performs overwhelmingly better (up more than 33%), it is worth to reveal the reasons.

Figure 3. Countries with increased revenue in 2019



As shown by Figure 4, it turns out Singapore and Canada have extremely fewer orders than most of the countries. There are less than 10 days with the active purchase therefore more data is required to determine the strategy performance in these countries.

Figure 4. Purchase history of the three best performance countries



4.2 Top 10 popular streams

Top 10 most viewed streams are captured and their price changes (moving average 120 days) are indicated by Figure 5. It is clear that the price of the majority of the streams is increased during the year 2019. As indicated by Figure 6, the increased price may be the main factor for the purchase. The customer is quite sensitive to the price since the purchase declined with the rise of the price except for the one stream. More details need to be obtained regarding the genres of the stream. The marketing team could recommend or advertise popular genres to the target countries.

Figure 5. Price changes of the top 10 popular streams

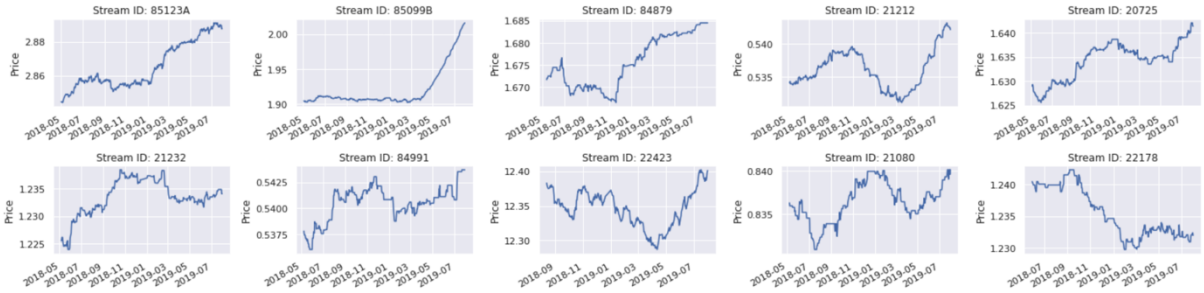
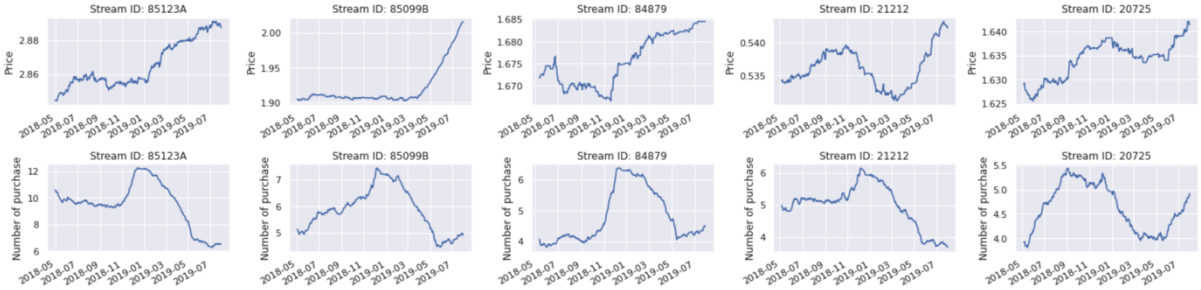


Figure 6. Stream price changes vs. Number of purchases of the top 5 popular streams



4.3 Time series data preparation and analysis

In order to carry out a time series analysis, record of each day should be considered, and the data frame should be in chronological order so that forecasting models can fit and provide revenue i.e price for the following month.

Let's start by aggregating the transactions by day. The time-series data has been generated in Table 2.

Table 2 First five rows of Time series data set

	date	num_customer	num_invoices	num_streams	total_views	revenue	weekday	month
0	2017-11-28	117	137	1063	13285.0	7541.01	1	2017-11
1	2017-11-29	98	120	1090	12111.0	6982.95	2	2017-11
2	2017-11-30	116	137	1093	14541.0	9214.00	3	2017-11
3	2017-12-01	82	93	1009	11319.0	5573.63	4	2017-12
4	2017-12-02	26	30	285	3310.0	1443.26	5	2017-12

It is interesting to see the revenue distribution by the weekday (Figure 7). Surprisingly, the revenue on Wednesday is higher than the weekend. Probably because people need to relax and fuel up to stay strong in the middle of the week. Obviously, there are outliers could affect the forecasting which needs to be removed. In terms of the distribution by month, the revenue rises gradually since August then declines slightly in December after reaching the peak in November. This means this time series has a seasonality trend which is important for modelling.

Figure 7. Revenue distribution by weekday

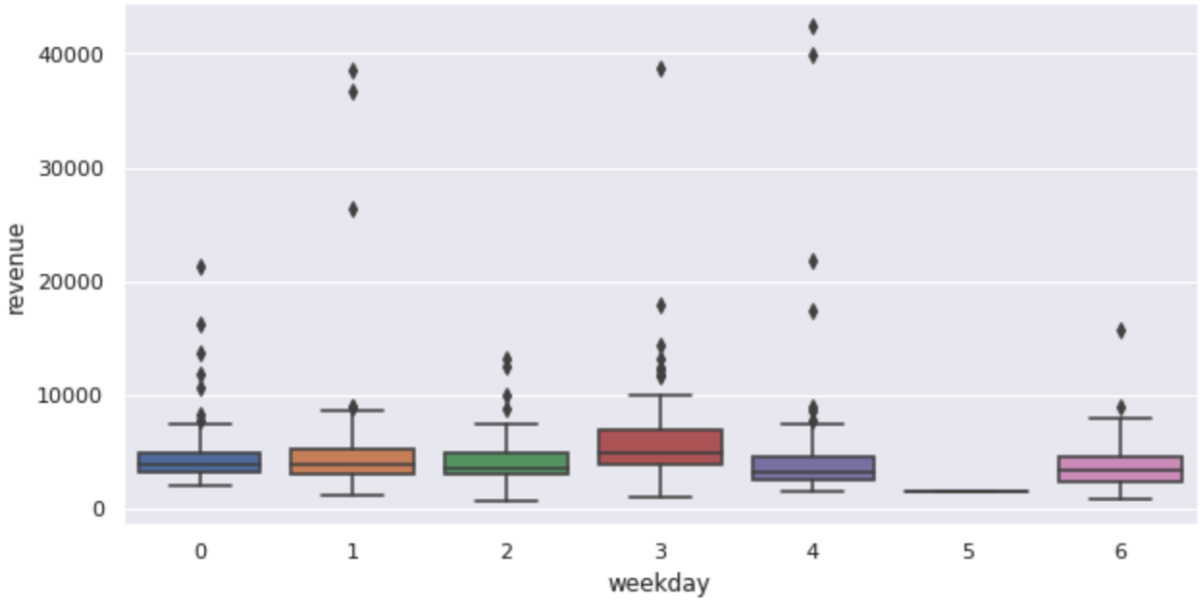
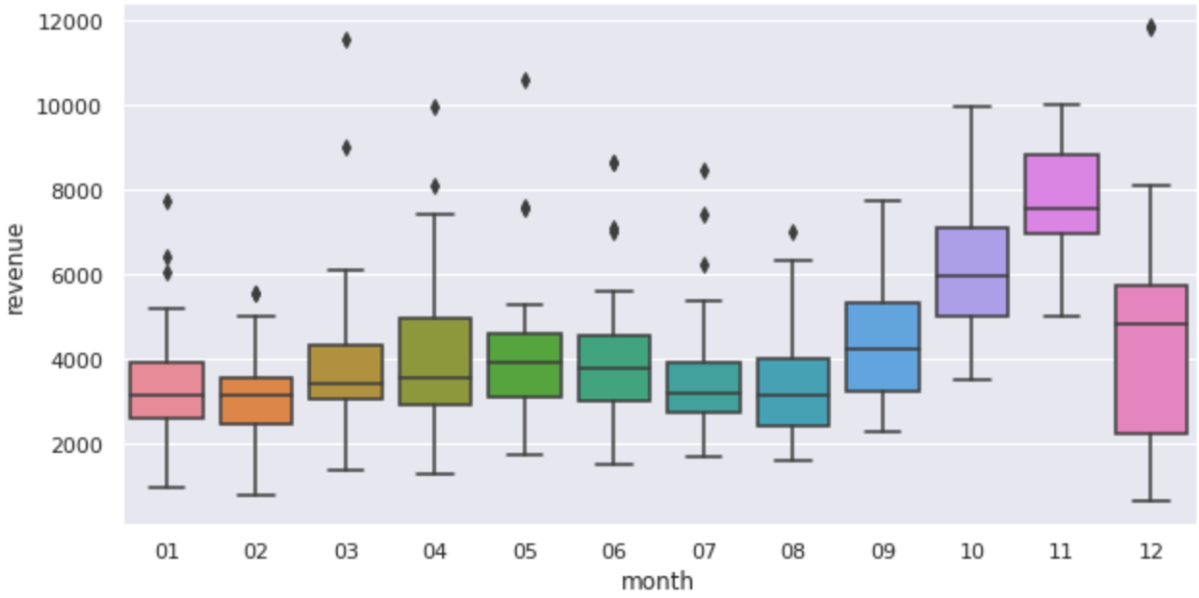
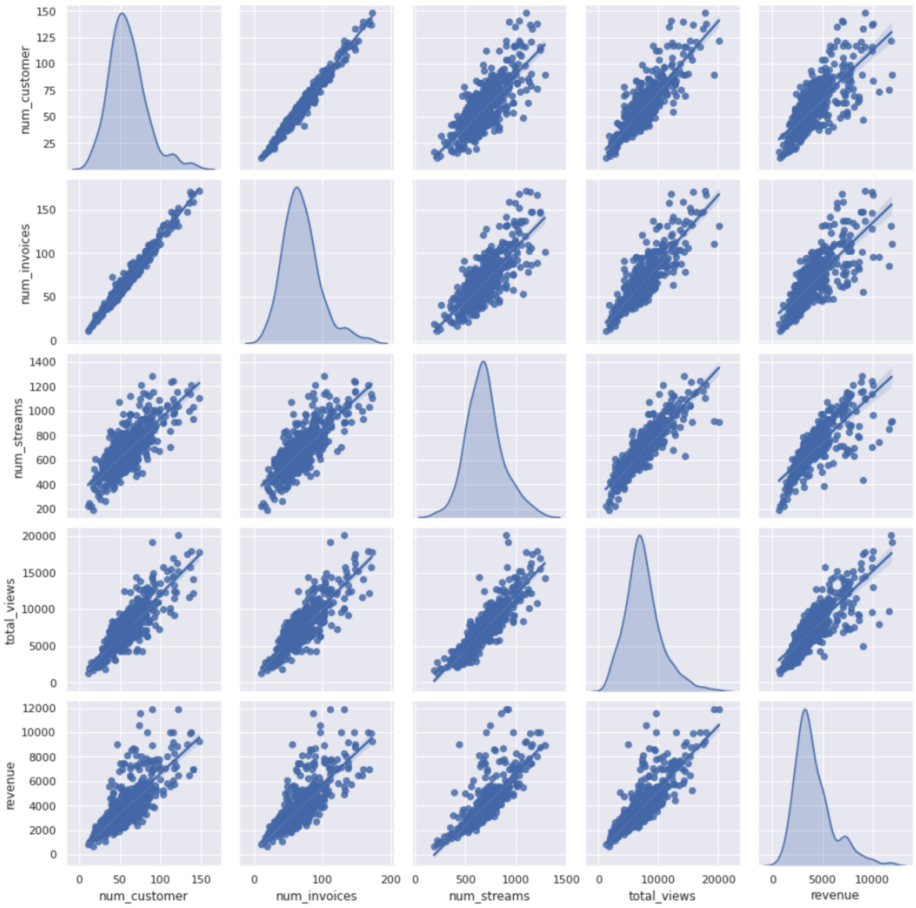


Figure 8. Revenue distribution by month after removal of outliers



The relationship between each feature is shown below. It is shown that there are positive linear relations between revenue and other features.

Figure 9. Relationship between revenue and other features



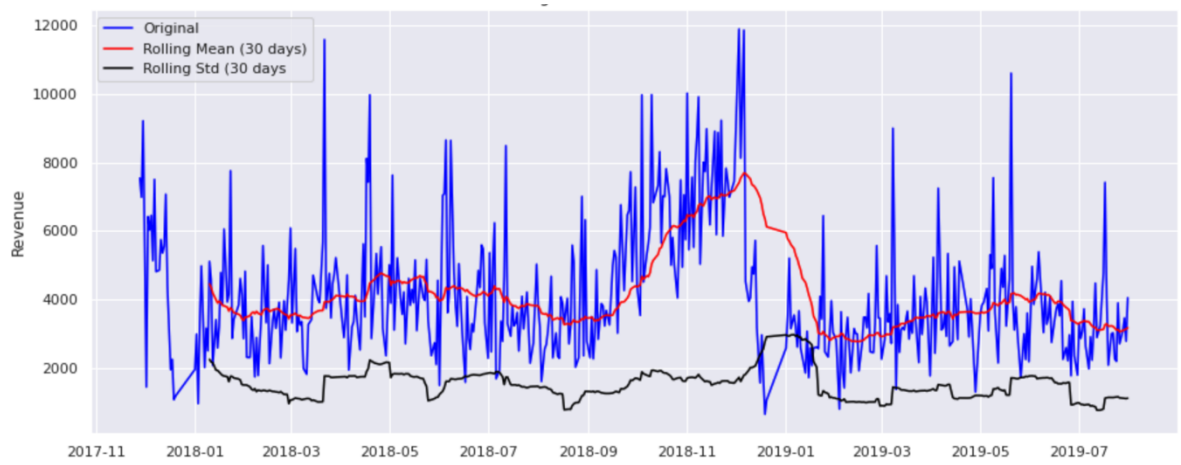
5. Revenue Forecasting with ARIMA

ARIMA, short for 'Auto-Regressive Integrated Moving Average' is a popular model that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

5.1 Stationarity check of a time-series

As the first step, the stationarity needs to be checked. There are mainly two ways to check the stationarity. The first is by looking at the data. By visualising the data, it should be easy to identify a changing mean or variation in the data. For a more accurate assessment, there is the Dickey-Fuller test. If the 'Test Statistic' is greater than the 'Critical Value' then the time series is stationary.

Figure 10. Revenue and it's rolling mean/standard deviation (30 days)

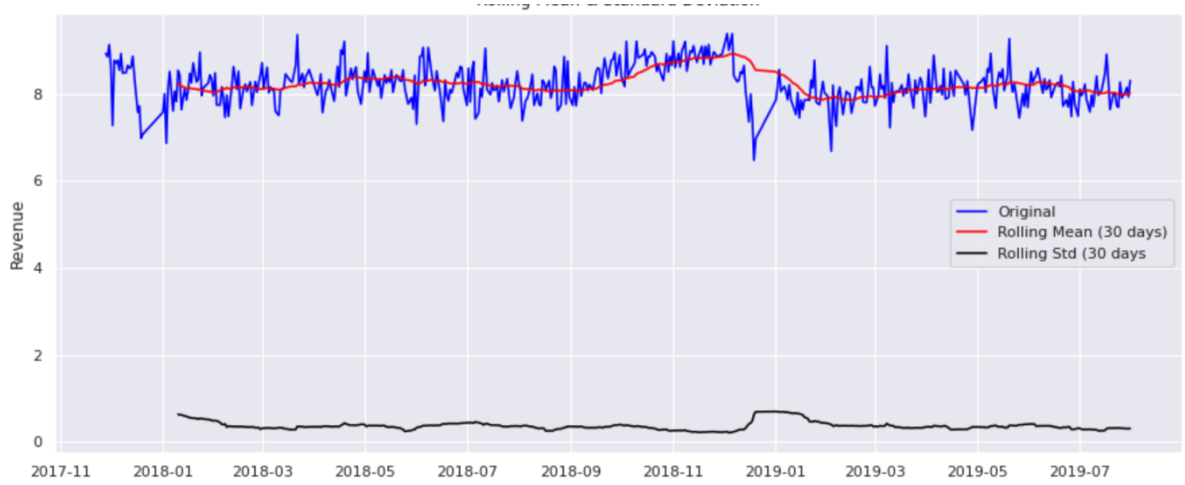



```
Results of Dickey-Fuller Test:
p-value = 0.0005. The series is likely stationary.
Test Statistic                  -4.261209
p-value                          0.000518
#Lags Used                       5.000000
Number of Observations Used    470.000000
Critical Value (1%)             -3.444340
Critical Value (5%)             -2.867709
Critical Value (10%)            -2.570056
dtype: float64
```

It is clear that the revenue crushed around December 2018. The p-value is 0.000518 which is small enough to confirm the stationarity. But since sometimes log-transformed and log-transformed-differencing data work better, these two

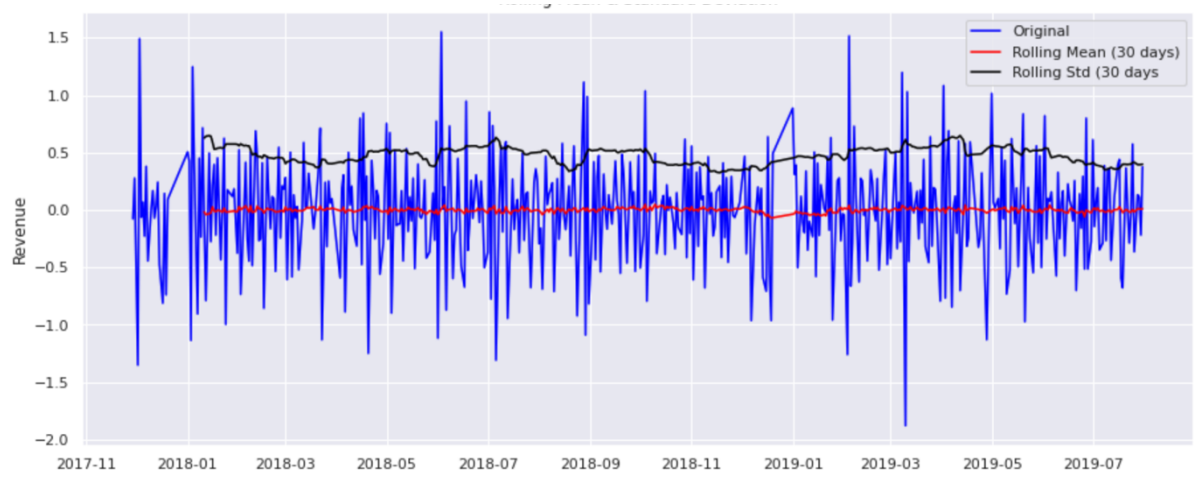
methods are also tested. Figure 11 and 12 shows the Log-transformed and Log-transformed-differencing revenue respectively. It clear that the Log-transformed-differencing performs the best since the p-value reaches around 2.5e-18. However, since an over differenced series may still be stationary, which in turn will affect the model parameters, the grading search techniques will be applied to determine which one is the most appropriate for modelling.

Figure 11. Log-transformed Revenue and it's rolling mean/standard deviation (30 days)




```
Results of Dickey-Fuller Test:
p-value = 0.0001. The series is likely stationary.
Test Statistic                  -4.815103
p-value                          0.000051
#Lags Used                       5.000000
Number of Observations Used    470.000000
Critical Value (1%)             -3.444340
Critical Value (5%)             -2.867709
Critical Value (10%)            -2.570056
dtype: float64
```

Figure 12. Log-transformed-differencing Revenue and it's rolling mean/standard deviation (30 days)



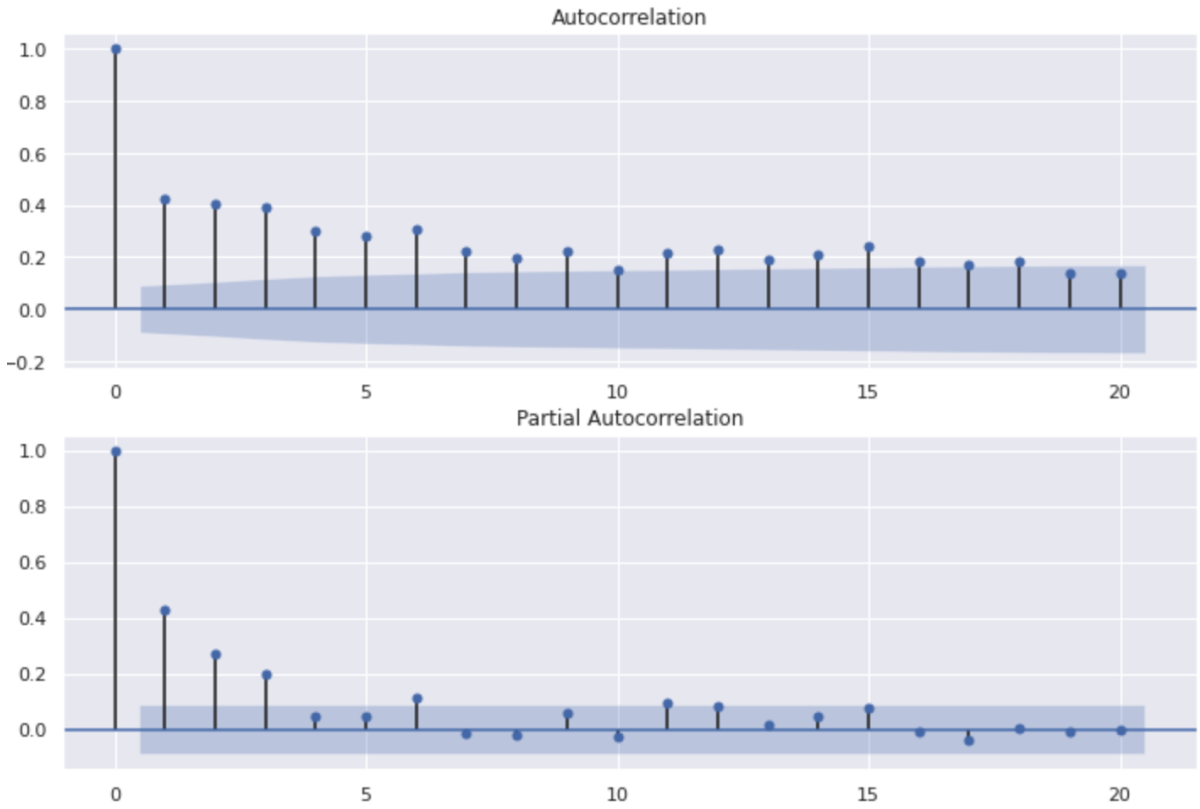

```
Results of Dickey-Fuller Test:
p-value = 0.0000. The series is likely stationary.
Test Statistic                -1.035194e+01
p-value                        2.523440e-18
#Lags Used                     1.000000e+01
Number of Observations Used    4.640000e+02
Critical Value (1%)           -3.444522e+00
Critical Value (5%)           -2.867789e+00
Critical Value (10%)          -2.570099e+00
dtype: float64
```

5.2 ACF and PACF

PACF (Partial Autocorrelation) plot can be applied to determine the order of the AR term (p). Partial autocorrelation can be imagined as the correlation between the series and its lag, after excluding the contributions from the intermediate lags. So, PACF sort of conveys the pure correlation between a lag and the series.

ACF plot can be applied to investigate the number of MA terms (q). An MA term is technical, the error of the lagged forecast. The ACF tells how many MA terms are required to remove any autocorrelation in the stationed series. Figure 13 shows the ACF and PACF plot for log-transformed data.

Figure 13. ACF and PACF plot for log-transformed data



Based on the plot, q could be 1 and p could be 1 or 2, d could be 0 or 1. After grid search testing for both original and log-transformed data, the final confirmed parameters turn out to be p is 2, d is 0 and q is 1 with the log-transformed data and the data does not need to be differenced.

5.3 Model validation

Figure 14. ARIMA Model validation with test data

The mean absolute error achieved $969.92 with the ARIMA model and the trend has been captured except for some extreme values.

6. Revenue Forecasting with Machine Learning

6.1 Feature Engineering

Since all original features are correlated, more features need to be created to be able to successfully apply machine learning. As shown by Figure 15, newly created features are the sum of revenue in the previous day, previous week, previous month, previous three months and previous year. The average number of invoices and views in recent 30 days has also been added as additional features. What needs to be pay attention here is the target is the sum of the revenue in the next 30 days.

Figure 15. Examples of created features and targets for machine learning

	previous_7	previous_14	previous_28	previous_70	previous_year	recent_invoices	recent_views	revenue in next 30 days
440	21427.37	40809.27	78000.33	203540.841	101251.99	67.916667	6865.208333	79273.020
441	21873.92	39964.27	78761.67	202746.391	101206.90	67.708333	6971.083333	74728.960
442	21380.42	42218.77	78053.24	202354.180	98954.07	66.782609	6970.478261	78497.780
443	20526.87	41778.94	78931.26	202173.720	102095.47	65.217391	6836.608696	78169.001
444	18996.36	39784.47	81145.12	199438.840	101554.02	64.500000	6735.791667	78150.191

6.2 Modelling with baseline model: Random Forest

Figure 16. Random Forest Model validation with test data

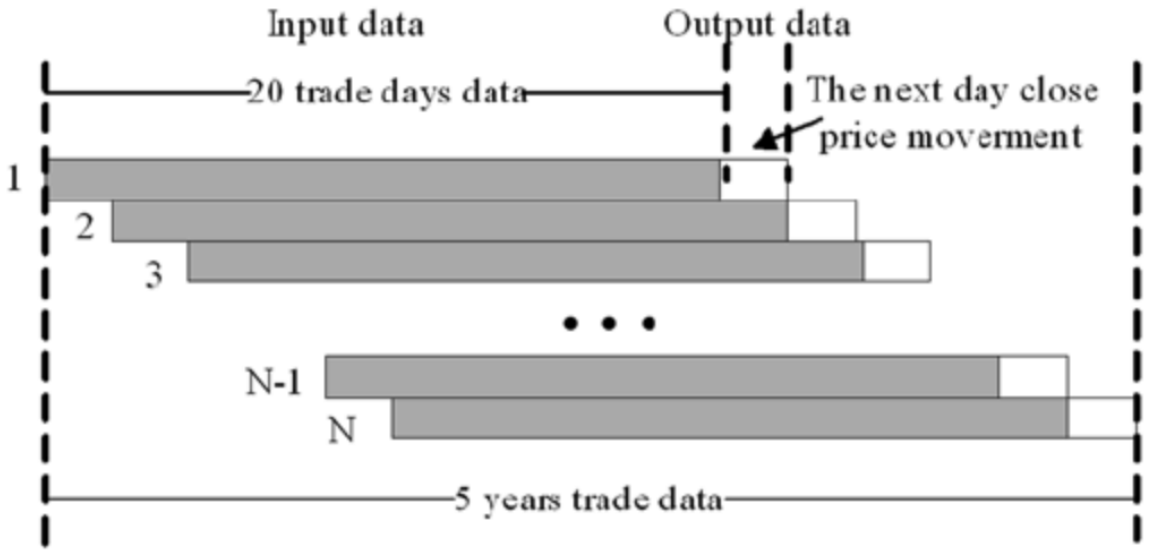
As shown by Figure 16, after the parameters tuning with grid search, Random Forest does not perform well with the mean absolute error of $8820. The trend has not been well captured.

6.3 Modelling with Deep Learning with Tensorflow

Time Series data can be highly erratic and complex. Deep Learning methods make no assumption about the underlying pattern in the data and are also more robust to noise (which is quite common in time series data), making them the top choice for time series analysis.

Data processing for deep learning is different. Before we move on to predicting, it is important to first process our data in a form that is understandable to a mathematical model. As shown by Figure 17, Time series data can be transformed into a supervised learning problem by using a sliding window to cut out datapoints. The expected output of each sliding window is then the timestep after the end of the window. Sliding window transformation as it is just like sliding a window across prior observations that are used as inputs to the model in order to predict the next value in the series.

Figure 17. Illustration of data processing for Deep Learning



To handle the large dataset, convolutional Layer (CNN) combined with Long Short-Term Memory (LSTM) Neural Network were developed. The test case was conducted first to determine the appropriate learning rate which should be 1e10-5 (Figure 18). Then the stochastic gradient descent and the mean absolute error loss function was applied for predictive modelling. At last, the relationship between the number of epochs and the loss has been plotted for confirmation of the appropriate number of epochs as shown by Figure 19.

Figure 18. The plot of the Learning rate vs. Loss for epoch

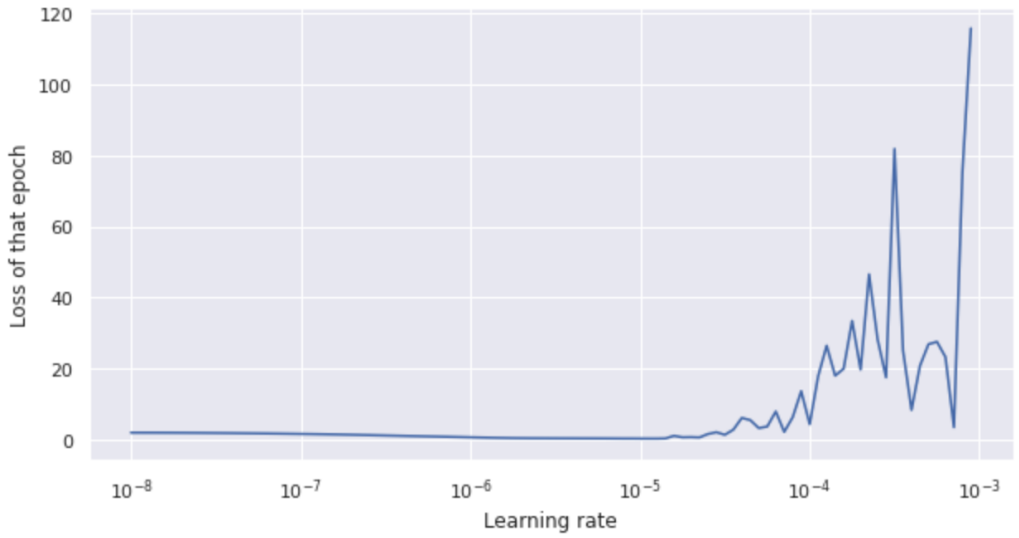
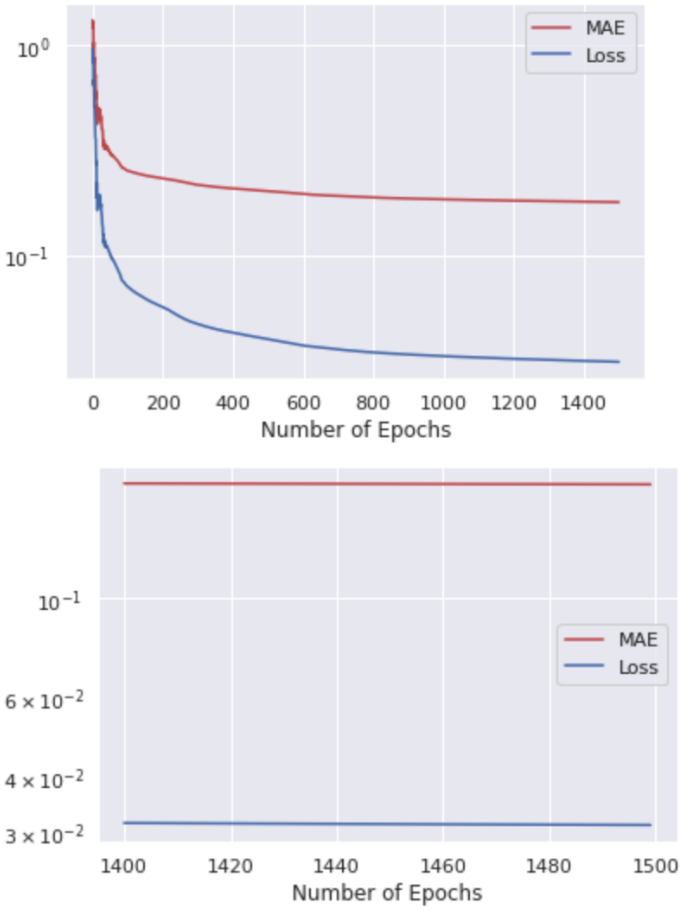
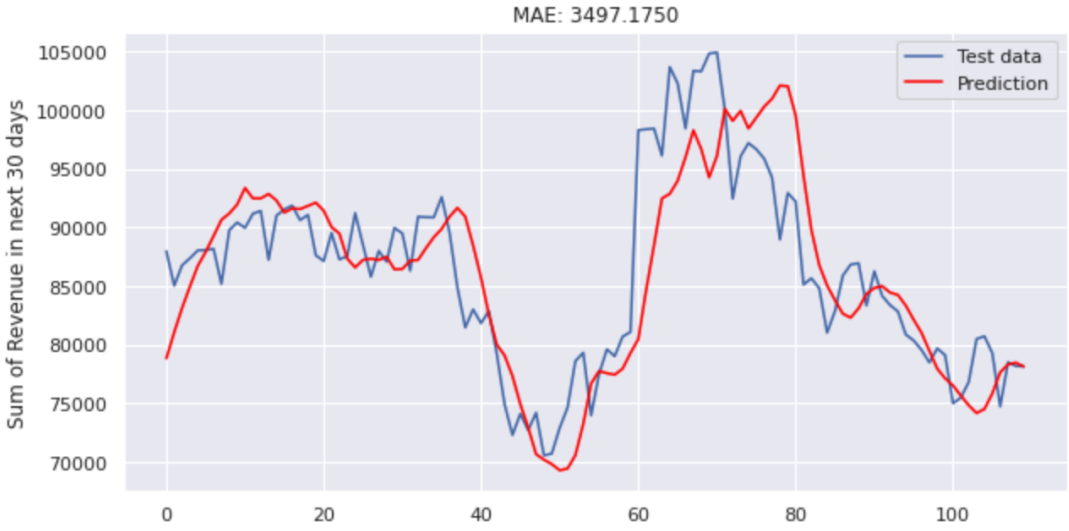


Figure 19. The plot of Number of Epochs vs. MAE and Loss



The results are indicated in Figure 20. The developed CNN-LSTM model was able to predict the future revenue with MAE of $3497 and captured the trend quite well. This model is much better than the baseline Random Forest Model.

Figure 20. CNN-LSTM Model validation with test data



7. Results and Discussion

In this study, the sales data has first been analysed. The revenue in the majority of the countries has declined in 2019 probably due to the rising price of the streams. The customers tend to order more on Wednesday during the week and when it is approaching November. The developed API could also reveal the sales details for each country. To forecasting the revenue, ARIMA, Random Forest and deep learning with CNN-LSTM have been developed. ARIMA was able to predict the revenue by day with the mean absolute error of less than $1000. Neural networks performed the best and were able to predict the 30 days revenue with a mean absolute error of only $3000. The developed models could be applied by the management team to predict the revenue in any duration. This information is valuable for the company to tailor the marketing strategy for each targeted country and to improve the current business models.

Models in this study mainly focused on the sales data. However, there are more factors affecting revenue. For example, the change of the customer tastes on videos and the local economy which are obviously more difficult to extract and quantify. But if they were able to be considered in the models, there would be significant improvements in predictions.