|  |
| --- |
| Revenue Forecast |
|  |

|  |  |
| --- | --- |
| August 31, 2020 | Examination and Forecasting of Revenues |

Using Python-based machine learning models, this report details the examination of the 2016-2020 revenues and a prediction of future revenues.

Revenue Forecast

Examination and Forecasting of Revenues

Contents

[1. ChALLENGE 2](#_Toc49591551)

[2. DATA 3](#_Toc49591552)

[3. Exploratory data analysis 4](#_Toc49591553)

[4. MODeling 6](#_Toc49591554)

[ARIMA Model 6](#_Toc49591555)

[FBProphet Model 6](#_Toc49591556)

[Covid-19 Effect 7](#_Toc49591557)

[Testing Data 7](#_Toc49591558)

[Model Results 7](#_Toc49591559)

[5. predictions 9](#_Toc49591560)

[6. conclusion 10](#_Toc49591561)

[7. Code 11](#_Toc49591562)

# ChALLENGE

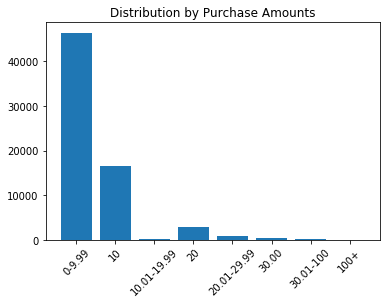
elementary OS provides a Linux based operating system that can be used in place of Windows and macOS. Its main revenues come from voluntary fees that are paid when downloads are made. There are several years' of revenue streams to examine. The challenge is to examine the streams and identify patterns and also be able to create a forecast of future revenues.

# DATA

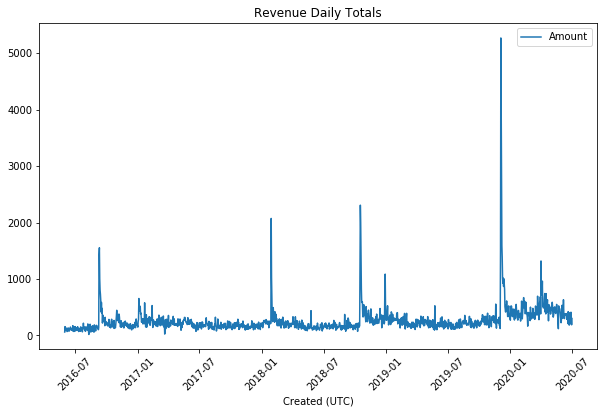
The data is availabe as CSV files that have been downloaded. The files contain all of the transactions for two separate accounts that receive the revenues. Contained within the files are fields that should be kept private (for example, email id of donor). The two files carry identical fields but unique data.

# Exploratory data analysis

On the elementary download page, there are three preset amount buttons: $10, $20, and $30 as well as a "custom" button which allows a purchaser to input any amount (including $0). A first look at the data was to see how the responses were distributed about these amounts.



When summed to a daily level, the revenue amounts stay relatively level except for some exceptionally high peaks. Most of these peaks are closely tied to new version releases.

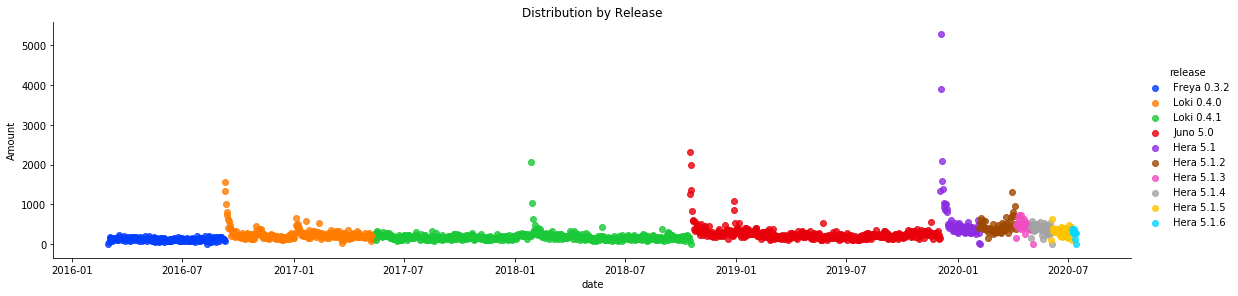


The peaks could be caused by higher than average purchases or an increased number of purchases, so both were examined:

|  |  |
| --- | --- |
|  |  |

While there were fluctuations in the mean daily amount, it appears that the actual number of purchases is what is driving the revenue peaks.

Finally, a look at how the different releases show over time:



# MODeling

This is a time series with a continuous numeric target. In order to be able to predict the future from history, the time series must be stationary. A KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test can check for the null hypothesis that x is level or trend stationary. A KPSS test was run against the daily revenue total series and returned a value of 0.01. Since it was not greater than 0.05, the hypothesis that the series was stationary had to be rejected. A second run against the log of the daily revenues, yielded .1. So, the target will be log(revenue) instead of revenue.

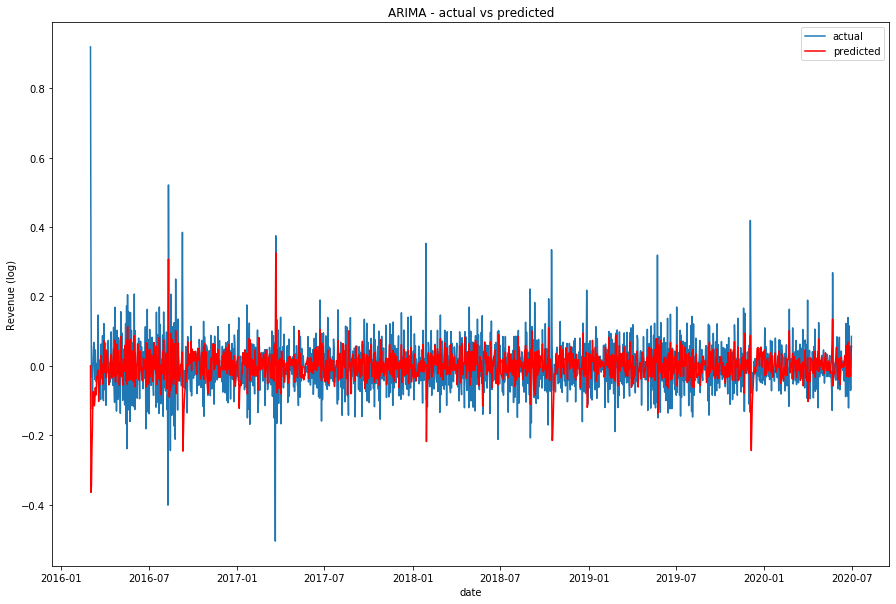
## ARIMA Model

The data was run using an Autoregressive Integrated Moving Average (ARIMA) model. ARIMA combines autoregression (time sequence as a linear function of the time series at previous times) with moving average (time sequence as a linear function of the residual mean errors of the previous times). This integration of methods help to make the time series stationary.

Even after tuning of the hyperparameters, the model was not a good fit. The scores from the final tuned model were:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **explained variance score** | **MAE** | **R2 score** |
| ARIMA | -0.3263541004611714 | 1.6760290448086197 | -306.8386569321416 |

The ARIMA model does not take seasonality into account, so the irregular peaks were difficult to fit.



## FBProphet Model

A search for a model that would accommodate irregular seasonality, produced the Facebook Prophet model. Prophet forecasts are based on an additive model that is fitted to yearly, weekly, and daily seasonality trends. Most importantly for this case, it also allows the incorporation of holidays which are irregular. These can be used to mark the release dates and other events that boost the downloads. Prophet works best with several seasons of historic data.

### COVID-19 Effect

In addition to the irregular peaks, an examination of the daily revenues shows that there was a noticeable increase in the number of downloads in 2020. This is probably caused by quarantining due to COVID-19. Prophet automaticly calculates changepoints in trending but also allows for manual identification. Several different changepoints were tested with the following results:

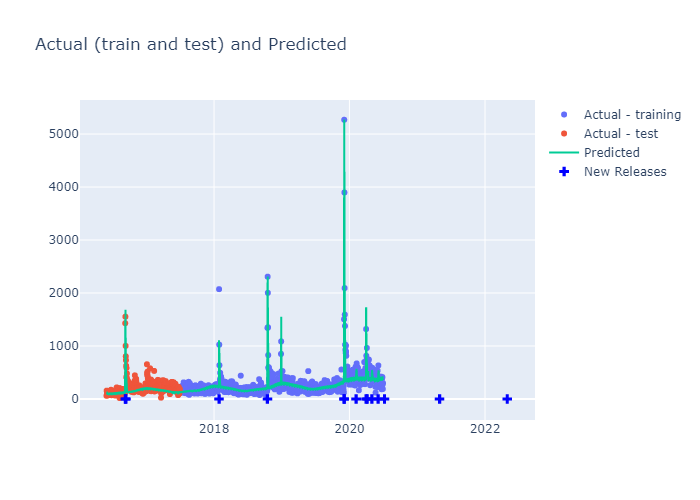
|  |  |  |  |
| --- | --- | --- | --- |
| **date** | **explained variance score** | **MAE** | **R2 score** |
| None | -0.03271216828489942 | 0.3826648469668074 | -0.20408094408184496 |
| 1/1/20 | 0.5550138787397201 | 0.2690427485044921 | 0.5378174854455271 |
| 2/1/20 | 0.5463961396525929 | 0.2726485724064258 | 0.5267105785147848 |
| 3/1/20 | 0.5352505851139009 | 0.27710073626397946 | 0.5127002805499307 |

The 1/1/20 changepoint provided the best fit and was used for modeling and forecasting.

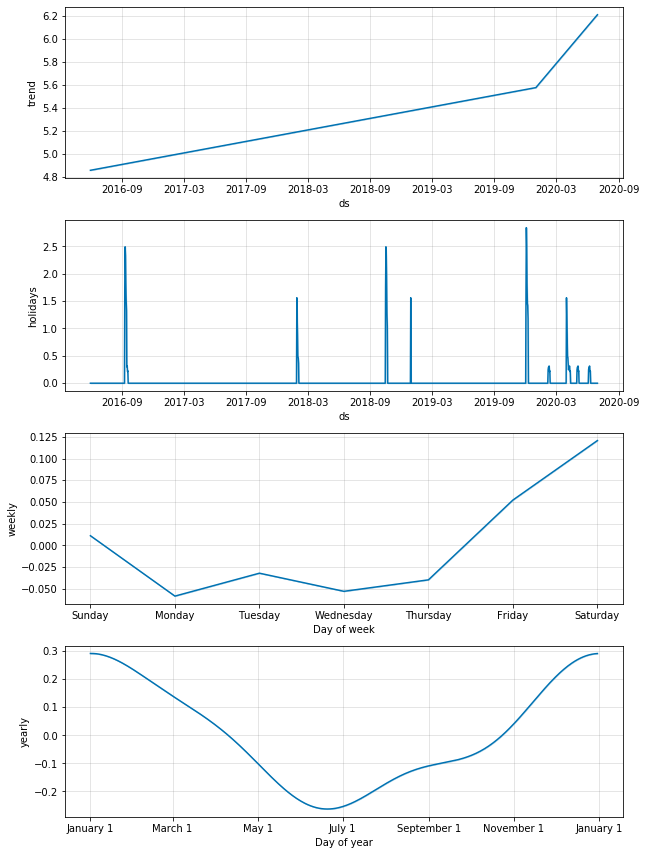
### Testing Data

Since the model is working with annual seasonality, it is recommended that the test data include at least a full year of data. The changepoint had to be included in the training data, so that meant that the test data had to be taken from the start of the series instead of the end. So, the first year's data was given to the test group and the final three years' went into training.

### Model Results

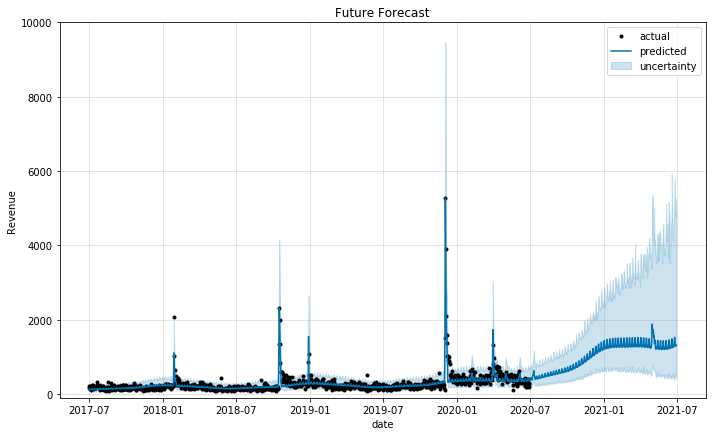


Here are the components of the forecast:



# predictions

Once the model was fitted and tested against the heldout test data, it could be used to predict future revenues. The model was run asking for an entire year's worth of forecast.



# conclusion

Future revenues are predicted to climb. However, given the trending change caused by Covid, it is recommended that this model be rerun once things get back to a more normal basis.

elementary might want to experiment with different preselected amounts on the download page. Since the majority of users have chosen the $10 preset button, perhaps raising this amount might encourage users to be more generous.

Not all of the revenue spikes have been identified. Most of the peaks were correlated with new releases. However, one very signficant boost occurred when elementary was featured very favorably in a Forbes article (December 2019). Further study of these peaks might help the company to understand what is driving the increases and encourage future occurrences.

# Code

Code for this project can be found at:

https://github.com/NancyKecso/Springboard/tree/master/Capstone 3 elementary OS revenue projection

Note: data files have not been stored on github in order to protect user privacy