CIVE 7100 Project Report

Solar Flares Forecasting

Harshada Sasturkar (NUID: 002750340)

Abstract:

Solar flares are astronomical phenomena where there are localized eruptions of electromagnetic radiation in Sun's atmosphere. When absorbed by the earth's atmosphere, the radiation can disrupt satellite communications and reduce their lifetime. Accurately predicting the flare activity beforehand can help us better prepare for such disturbances. Additionally, it can provide more insights into other associated solar features and similar electromagnetic events occurring on other stars in the universe. Over the years, NASA's RHESSI mission has been monitoring solar flares and recording their characteristics in a dataset. In this project, this dataset is analysed to explore the underlying patterns and trends. Furthermore, autoregressive machine learning models (ARIMA, SARIMA, SARIMAX) and neural network models (LSTM, TCN) are trained on the data to forecast monthly number of solar flares and their average duration. A comparative analysis is done between the models using RMSE as the performance metric. The results obtained show that neural network models were able to perform significantly better than autoregressive models for this data. LSTM and TCN both had similar performance with a little difference in their RMSE values, LSTM having a slight edge over TCN.

Introduction:

Study of solar flares has been an ongoing active area of research. A solar flare is an intense localized eruption of electromagnetic radiation in the Sun's atmosphere and are thought to occur when stored magnetic energy in the Sun's atmosphere accelerates charged particles in the surrounding plasma. This results in the emission of electromagnetic radiation across the electromagnetic spectrum [1]**.** NASA's Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI) is a solar observatory spacecraft tasked with detecting solar flares and collecting data from them [2]**.** The dataset is a time series of solar flare observations with their characteristics such as the duration, position, count of photons, count of peaks, energy produced by the flares, etc. The aim of this project is to first explore the time series, identify any underlying patterns and study the trends present in it. Secondly, train machine learning models to forecast the number of solar flares that could occur per month and their average duration.

The mechanism of solar flares is still not completely understood and as of now there aren't any reliable physical models which can accurately describe and forecast the same. Hence, majority of the related work has been done with a data driven approach and using machine learning models [3]. Some have used state of the art ensemble tree models like Random Forest and Light Gradient Boosting Machine (LightGBM) [4]. However, with the advent of various autoregressive and neural network models which are designed to work with time series data, some papers have focussed on a comparative analysis of their prediction accuracy and performance [5]. Taking reference from these papers, three Autoregressive – ARIMA, SARIMA and SARIMAX, and two neural network type models – LSTM and TCN, are implemented in this project and a comparative analysis of their performance is done. The forecasting problem can be divided into two subproblems depending on the target variable to be predicted – number of flares and duration. Both univariate and multivariate (if plausible) model trainings are done on these targets to understand how the addition of more features affects model performance. The final objective of the project is to understand the mechanism behind these models and analyse their accuracy to find out the one most suitable for making forecasts on the data in question. A discussion is also done on what data and model specific factors could have contributed to the forecasting success/failure.

Motivation:

Solar flares emit radiation across the electromagnetic spectrum at all wavelengths, from radio waves to gamma rays. Even though this radiation is absorbed by the earth's upper atmosphere, the ionization it causes can degrade satellite communications. Radiation also heats up the atmosphere causing it to expand increasing the drag on earth-orbiting satellites. This reduces their lifetime in orbit. Flares are speculated to be associated with a much more severe type of geomagnetic solar activity called coronal mass ejections (CMEs) which can significantly impact not only the satellites but also power

grids present on land [6]. If we can accurately predict flares, we could potentially predict CMEs as well, leading us to better prepare for the disruptions. It can shed more light on other solar features such as Sun's core, atmosphere, and the active regions in it called sunspots. Ultimately, these high energy processes are not unique to our star. Flares potentially occur on other stars in the universe as well (called stellar flares) and the ability to accurately forecast and therefore understand solar flares can provide insights into stellar ones too.

Methodology:

The approach used in the project consists of five major tasks.

1. Exploratory data analysis:

The Solar Flares Dataset [7] is openly available on Kaggle. It has 116k observations ranging from Feb 2002 to Feb 2018 and 18 columns with flare related features [Figure 11]. Firstly, the descriptive statistics of the dataset are printed out which includes its summary statistics and missing value information [Figure 12]. The data was already in a cleaned state. Then comes the data visualization part where univariate plots are made for the features (columns) of interest explaining their distribution in the data [Figure 1-4, 13-16]. Bivariate plots are plotted to shed light on the relationship between the target variable and specific features [Figure 17-18].

2. Data preparation and preprocessing:

In this step, the original data is aggregated into a monthly time series where each observation corresponds to a month of a year. This gives the total numbers of solar flares for each month and their respective average durations [Figure 19-20]. The time series is then split into train and test sets containing observations from 2002 to 2014 and 2015 to 2018 respectively. For each of the two targets to be forecasted – number of flares and duration, we perform time series decomposition [Figure 5-6]. The time series is made stationary by doing one seasonal differencing and one non-seasonal first order differencing. Based on the ACF – PACF plots, an order(p, d, q) of (2, 1, 1) is decided for number of flares and (1, 1, 1) is given for duration. Both the targets have seasonal order(P, D, Q, m) of (2, 1, 1, 12). The correlation plot, ANOVA testing and Granger causality tests [Figure 21-26] determine the features to be selected and help in deciding whether to train univariate models or both univariate and multivariate models.

3. Model training and forecasting:

A total of five models, three autoregressive and two neural networks, are trained on the train set and the forecasts are made on the test set [Figure 27-30, 32-38]. Following is a brief description of the models:

1) ARIMA: Combines autoregression (AR), differencing (I), and moving average (MA) components to capture different aspects of a time series.

2) SARIMA: Extends the ARIMA model to account for seasonal patterns in time series data. By incorporating seasonal components alongside the AR, I, and MA terms.

3) SARIMAX: Expands upon SARIMA by incorporating additional external or exogenous variables into the model. This integration allows for the inclusion of external factors that influence the time series.

4) LSTM: A type of recurrent neural network (RNN) designed to model sequential data by effectively capturing long-term dependencies.

5) TCN: Unlike traditional recurrent models, TCN utilizes convolutional operations with dilated convolutions to efficiently capture long-range dependencies in sequences.

4. Comparative analysis of model performances:

Finally, to compare the model performances, RMSE (Root Mean Squared Error) between the actual values and predicted values was used as the metric on both train and test data [Figure 31, 39-41, 7-10]. The results obtained are discussed in subsequent sections.

Results:

1. Exploratory data analysis:

Studies have shown that the solar flares activity has a trend cycle where the number of flares gradually increases and then decreases over a period of 11 years. Figures 1 and 2 show the same trend especially when focused on the years 2008 to 2018. Examining if the same trend is shown by the flare duration – it is found that the average duration does have a weak trend but when separated into individual energy bands, there doesn't seem to be a significant pattern.

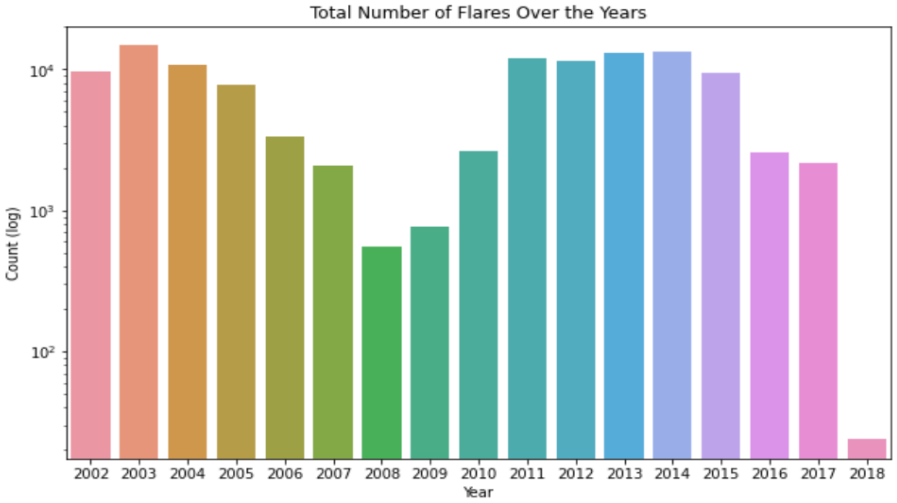


Figure 1: Distribution of number of solar flares over the years

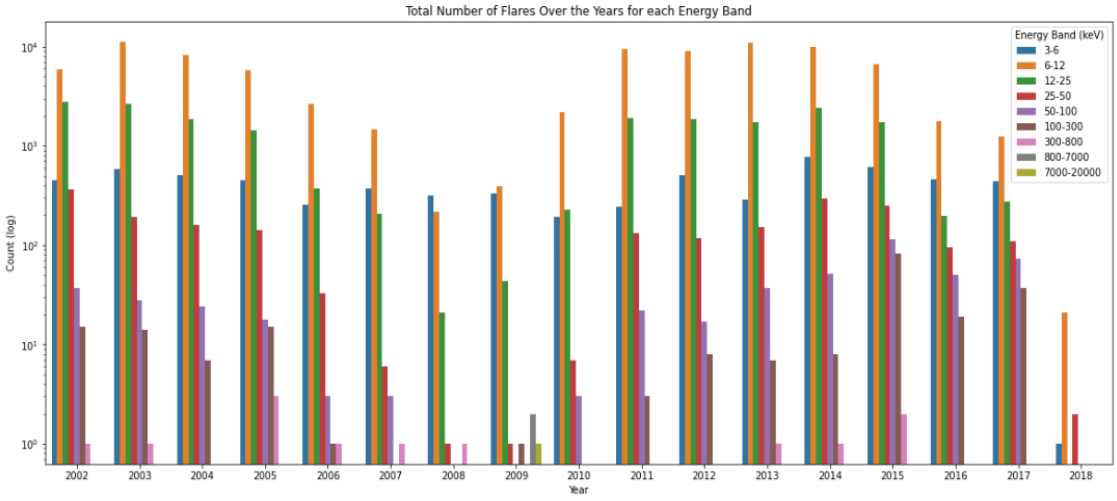


Figure 2: Distribution of number of solar flares in each energy band over the years

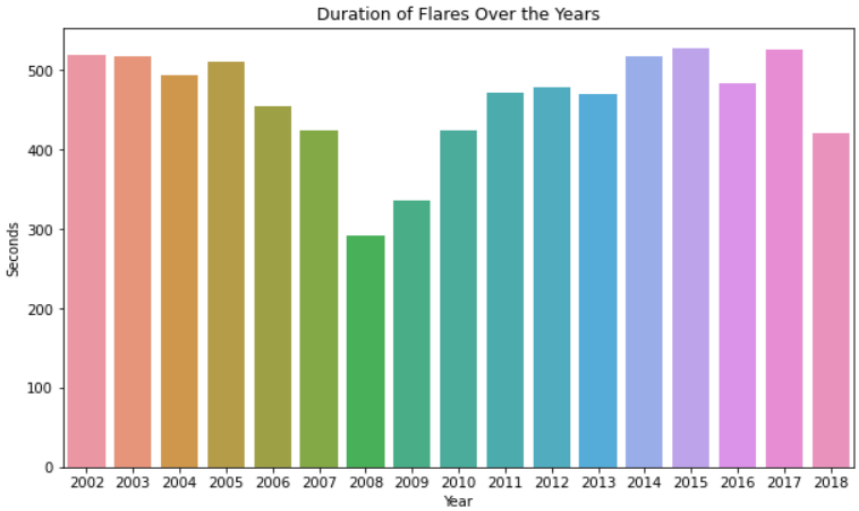


Figure 3: Distribution of duration over the years

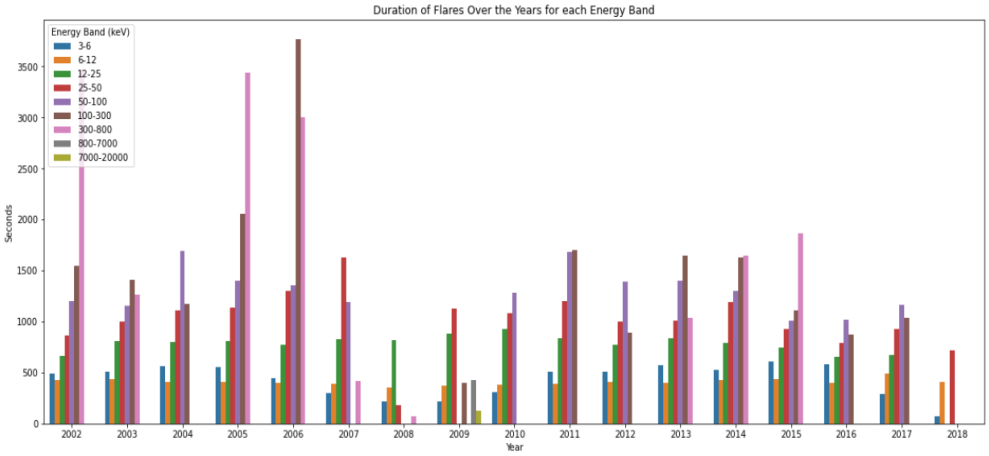


Figure 4: Distribution of duration in each energy band over the years

2. Data preparation and processing:

However, the trend and seasonal components for both targets become clear after time series decomposition.

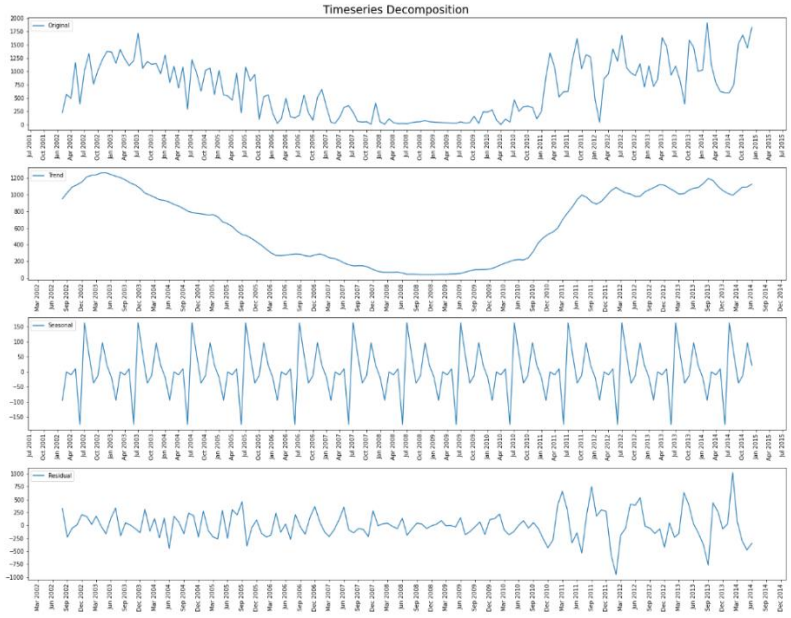


Figure 5: Timeseries decomposition for number of flares

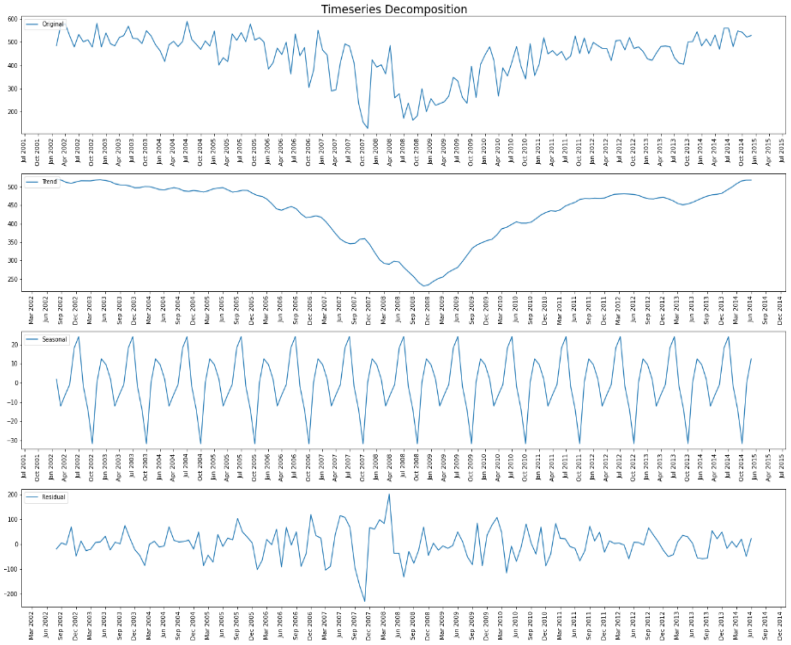


Figure 6: Timeseries decomposition for duration

3. Model training and forecasting:

Following are the Test RMSEs obtained from univariate modelling on Number and, both univariate and multivariate modelling on Duration.

1) Target- Number of flares:

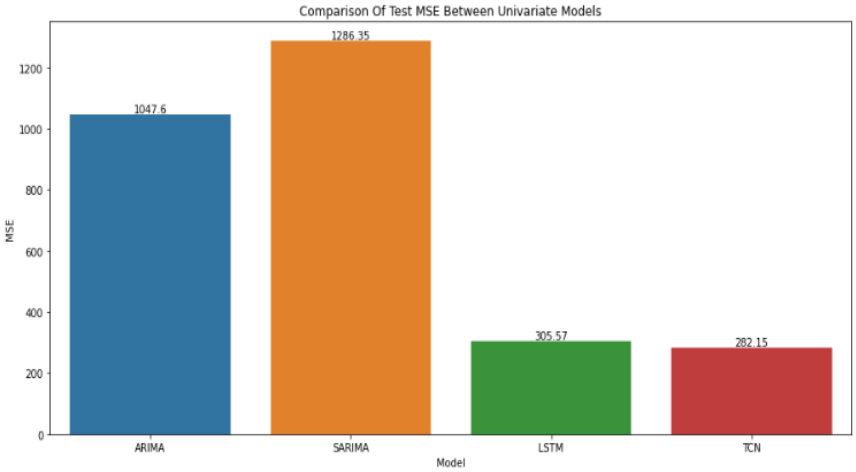


Figure 7: Test RMSE comparison for number of flares

2) Target – Duration:

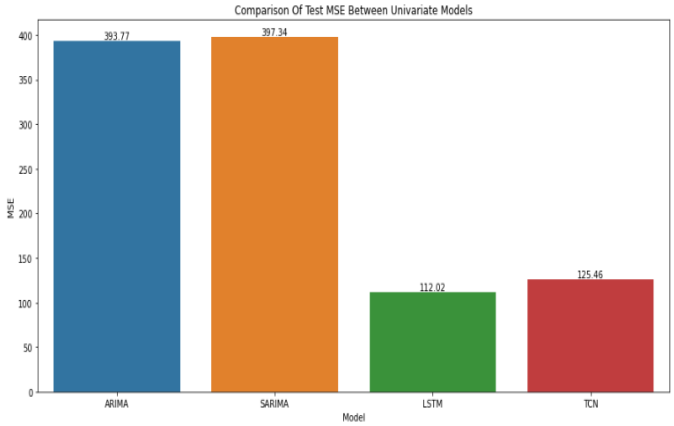


Figure 8: Test RMSE comparison of univariate models for duration

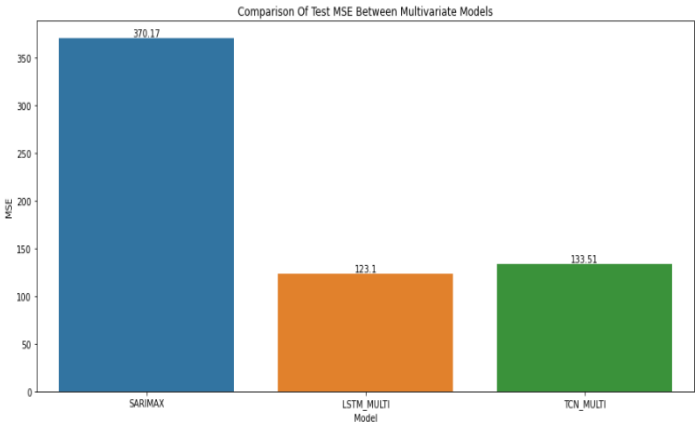


Figure 9: Test RMSE comparison of multivariate models for duration

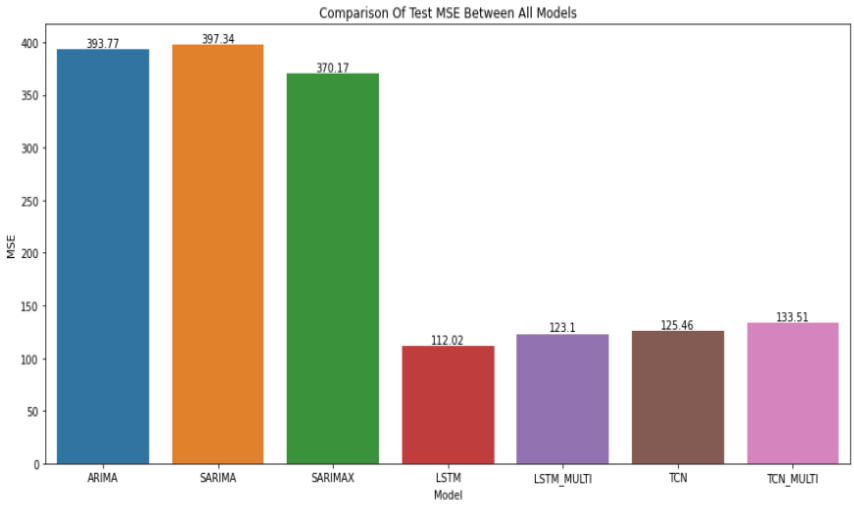


Figure 10: Test RMSE comparison of all models for duration

Conclusion:

The solar flares dataset was successfully analysed to understand its underlying patterns and trends. Multiple ml models were trained on the data for forecasting monthly solar flare numbers and their average duration. A comparative analysis of model performances was also carried out. The results show that neural network models (LSTM and TCN) performed better than autoregressive models (ARIMA, SARIMA and SARIMAX) for both target variables. For target Number, LSTM and TCN have RMSE 305 and 282 respectively while RMSEs of ARIMA and SARIMA re in the range of 1000s. Similarly, for target Duration neural network models have RMSE averaging around 125 while autoregressive models have around 390 in both univariate and multivariate models. SARIMAX does perform a little better compared to its univariate counterparts. Both LSTM and TCN had similar performance. It can be concluded that there isn't much difference between the performances of LSTM and TCN though it can be said that LSTM has a slight edge over TCN.

Discussion:

The performance of any ml model greatly depends on the configuration and size of the data. Autoregressive models are designed to work very well on time series, and they surpass the performance of neural networks as well **[ref]**. The results obtained in this project contradict this. It could mean that the data is significantly nonlinear as autoregressive are a type of linear models while neural networks are better at identifying nonlinear patterns. Models like LSTM and TCN are also more suitable for capturing long range dependencies which could be another reason for their success. The order and seasonal order parameters passed to autoregressive models are the base of the equations behind their working. The small size of the training dataset which was a challenge faced in this project could have contributed to the parameter values being not so accurate leading to poor performance. A much larger dataset with more recent observations could have been beneficial in improving prediction accuracy for all models. However, one advantage autoregressive models have is in terms of interpretability. The order parameters input in the models are data specific and determine how the data will be perceived by these models. Whereas in case of neural networks only the data is given as the input and there is no way to know how the model is fitting onto it.

Future scope:

The scope of this project can be extended to make forecasts for future time steps. Currently the forecasts are made on the test data with observations till 2018 only. Other than number of flares and duration, features like flare energy level, number of peaks and photon counts could be considered as targets for the models. The data has flag columns storing detector specific information which could be further analyzed. We can experiment with popular models like VAR, Prophet and mSSa. The models implemented in this project can also be tried out on a non-solar (stellar) flares dataset (if available).

References:

[1] Solar flares https://en.wikipedia.org/wiki/Solar_flare

[2] Reuven Ramaty High Energy Solar Spectroscopic Imager
https://en.wikipedia.org/wiki/Reuven_Ramaty_High_Energy_Solar_Spectroscopic_Imager

[3] Research Progress on Solar Flare Forecast Methods Based on Data-driven Models, 2023. [Ke Han et al]
https://iopscience.iop.org/article/10.1088/1674-4527/acca01

[4] Machine learning techniques applied to solar flares forecasting, 2021. [F. Ribeiro, A.L.S. Gradvohl]
https://doi.org/10.1016/j.ascom.2021.100468

[5] Comparative Analysis of Machine Learning Models for Predicting Travel Time, 2021. [Armstrong Aboah, Elizabeth Arthur] https://doi.org/10.48550/arXiv.2111.08226

[6] RHESSI - Impact of flares https://hesperia.gsfc.nasa.gov/rhessi3/mission/science/the-impact-of-flares/index.html

[7] Solar Flares Dataset https://www.kaggle.com/datasets/khsamaha/solar-flares-rhessi?select=hessi.solar.flare.UP_To_2018.csv

Appendix:

	flare	start.date	start.time	peak	end	duration.s	peak.c/s	total.counts	energy.kev	x.pos.asec	y.pos.asec	radial	active.region.ar	flag.1	flag.2
0	2021213	2002-02-12	21:29:56	21:33:38	21:41:48	712	136	167304.0	12-25	592	-358	692	0	A1	P1
1	2021228	2002-02-12	21:44:08	21:45:06	21:48:56	288	7	9504.0	6-12	604	-341	694	9811	A1	P1
2	2021332	2002-02-13	00:53:24	00:54:54	00:57:00	216	15	11448.0	6-12	-310	375	487	9825	A1	P1
3	2021308	2002-02-13	04:22:52	04:23:50	04:26:56	244	20	17400.0	12-25	-277	378	469	9822	A1	P1
4	2021310	2002-02-13	07:03:52	07:05:14	07:07:48	236	336	313392.0	25-50	-272	390	476	9825	A1	GS

Figure 11: Original solar flares data frame

Summary Statistics

	flare	duration.s	peak.c/s	total.counts	x.pos.asec	y.pos.asec	radial	active.region.ar
count	1.161430e+05	116143.000000	116143.000000	1.161430e+05	116143.000000	116143.000000	116143.000000	116143.000000
mean	1.099563e+07	493.643009	215.086617	3.768843e+05	-7.681625	-42.185495	687.896989	990.071550
std	9.888161e+06	434.131763	839.382841	3.048797e+06	755.773503	401.904509	511.364382	1351.853636
min	2.021213e+06	8.000000	0.000000	8.000000e+00	-10012.000000	-10005.000000	0.000000	0.000000
25%	4.112632e+06	212.000000	28.000000	2.284000e+04	-701.000000	-247.000000	467.000000	0.000000
50%	1.112297e+07	364.000000	56.000000	5.856000e+04	0.000000	-71.000000	759.000000	691.000000
75%	1.404222e+07	628.000000	144.000000	1.798080e+05	708.000000	198.000000	946.000000	1564.000000
max	1.711151e+08	4444.000000	113156.000000	4.355501e+08	1190.000000	1223.000000	14154.000000	9999.000000

Missing Values

	Missing Values	Data Types
flare	0	int64
start.date	0	object
start.time	0	object
peak	0	object
end	0	object
duration.s	0	int64
peak.c/s	0	int64
total.counts	0	float64
energy.kev	0	object
x.pos.asec	0	int64
y.pos.asec	0	int64
radial	0	int64
active.region.ar	0	int64
flag.1	0	object
flag.2	0	object
flag.3	19907	object
flag.4	20202	object
flag.5	61180	object

Figure 12: Data frame summary statistics and missing value information

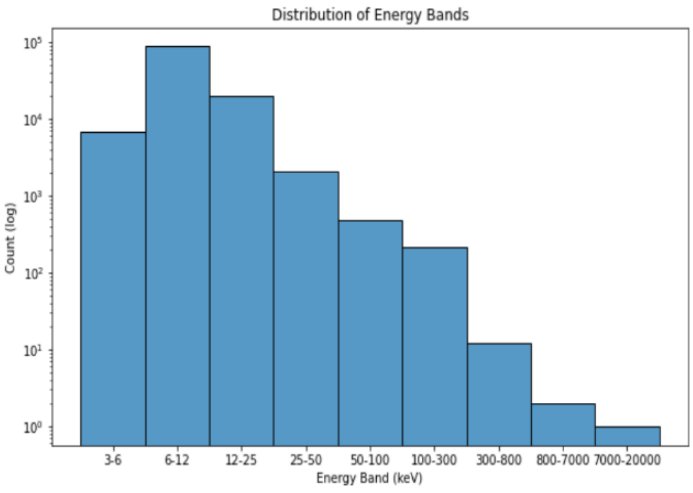


Figure 13: Distribution of energy bands

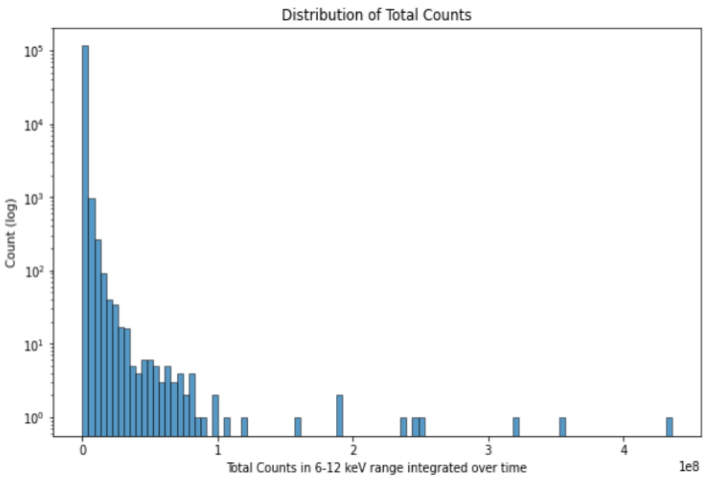


Figure 14: Distribution of total photon counts (6-12 keV)

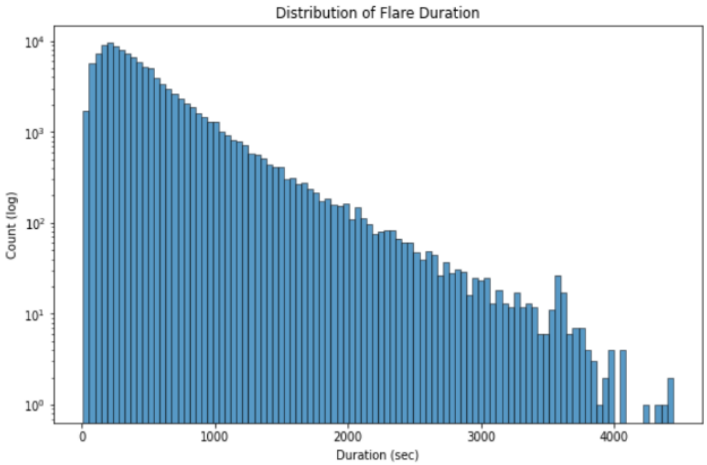


Figure 15: Distribution of duration

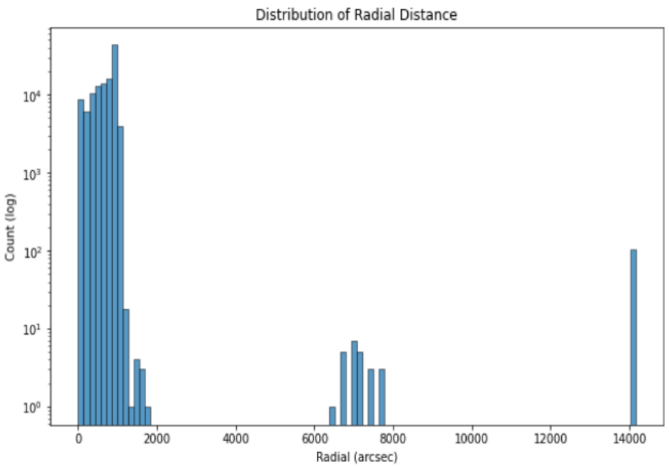


Figure 16: Distribution of radial distance

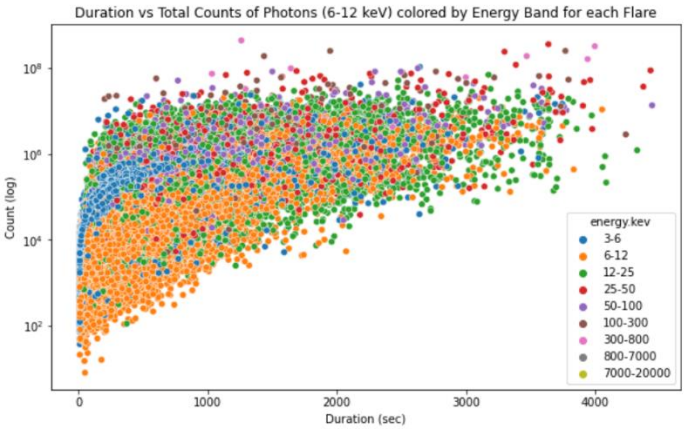


Figure 17: Duration VS total count of photons (6-12 keV) coloured by energy bands

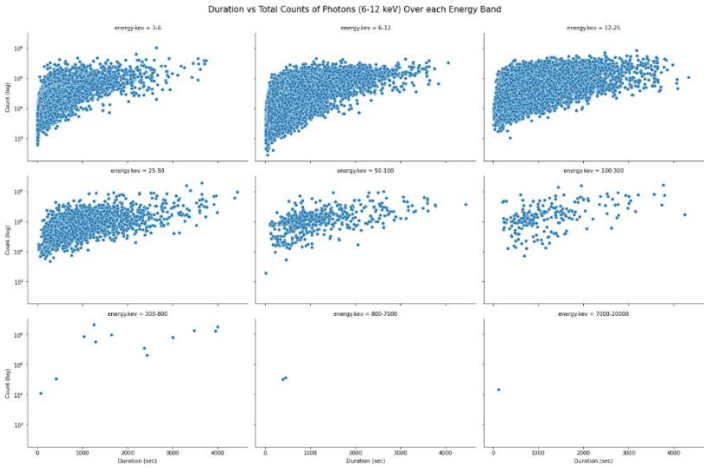


Figure 18: Duration VS total count of photons (6-12 keV) divided by energy bands

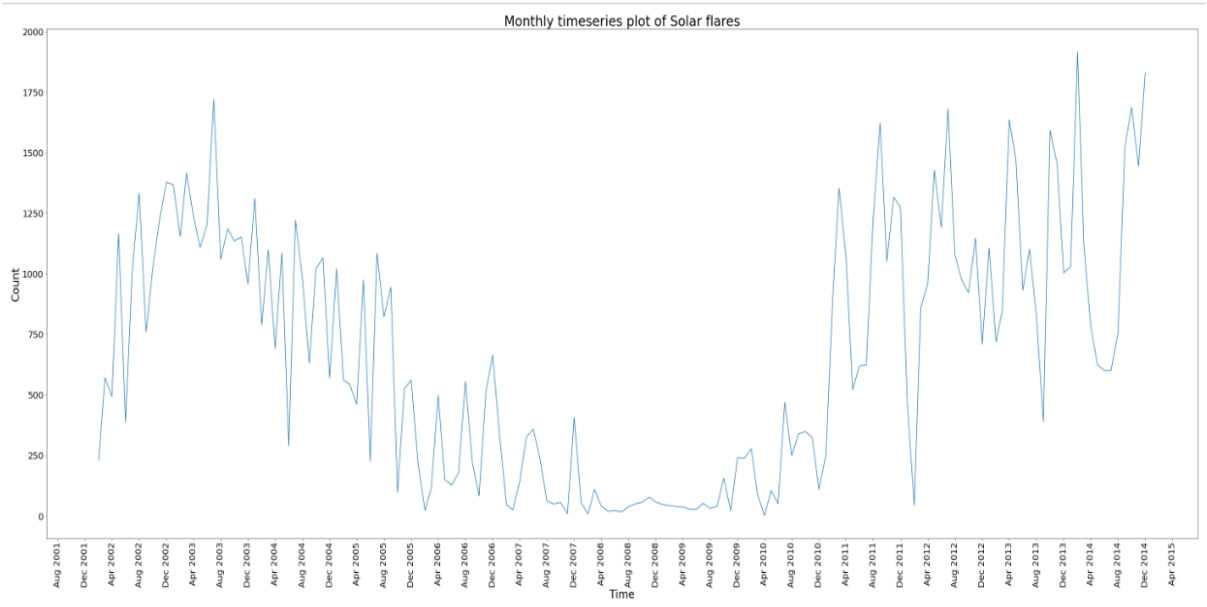


Figure 19: Timeseries of solar flares per month of each year

	year	month	Number	Duration	Energy	Radial	Total_counts	Peaks
2002-02-01	2002	2	231	484.987013	6-12	r1	3.060434e+05	157.480519
2002-03-01	2002	3	571	569.912434	6-12	r1	1.492297e+06	530.789842
2002-04-01	2002	4	493	578.547667	6-12	r1	8.919634e+05	361.298174
2002-05-01	2002	5	1165	526.908155	6-12	r1	6.305009e+05	329.327897
2002-06-01	2002	6	388	478.618557	6-12	r1	2.811340e+05	170.613402
...	...	...	...	...	...	...	...	...
2017-11-01	2017	11	166	292.650602	3-6	r1	2.329086e+05	285.343373
2017-12-01	2017	12	3	272.000000	6-12	r1	1.957500e+04	49.333333
2018-01-01	2018	1	6	581.333333	6-12	r1	3.284233e+04	83.333333
2018-02-01	2018	2	17	360.941176	6-12	r1	1.178812e+04	33.411765
2018-03-01	2018	3	1	492.000000	6-12	r1	1.581600e+04	42.000000

194 rows × 8 columns

Figure 20: Final timeseries data frame of solar flares

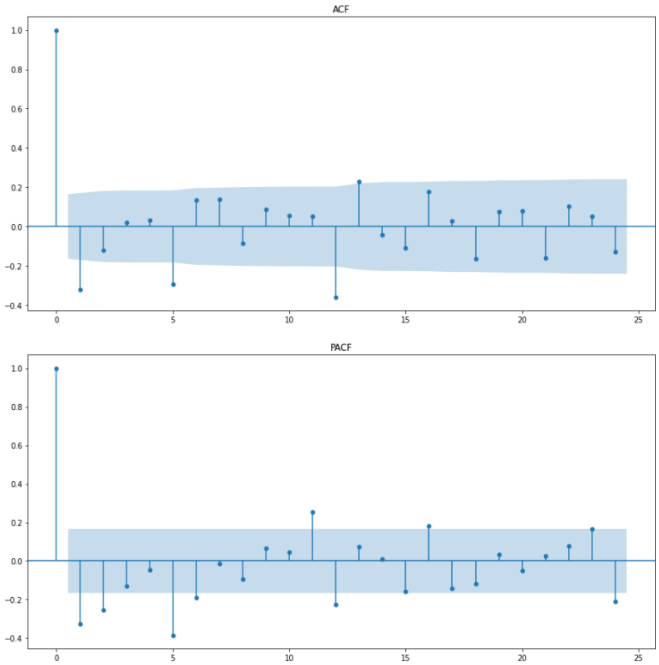


Figure 21: ACF - PACF plots for target number of solar flares

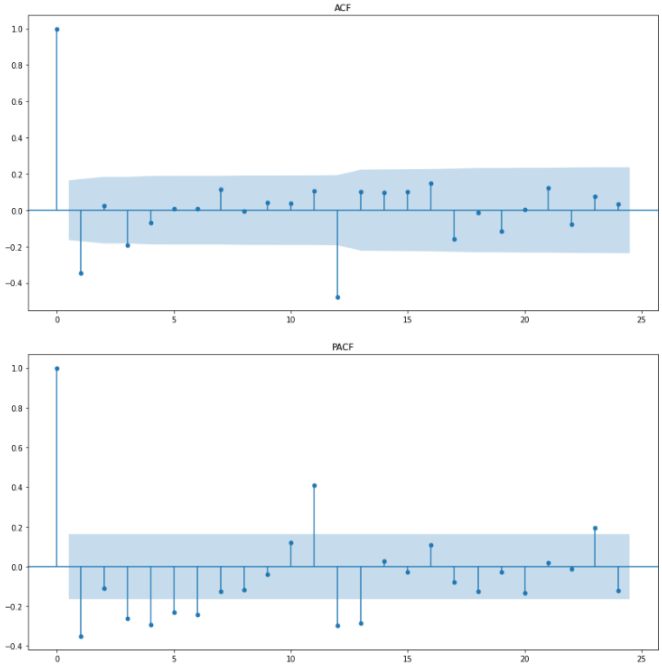


Figure 22: ACF - PACF plots for target duration

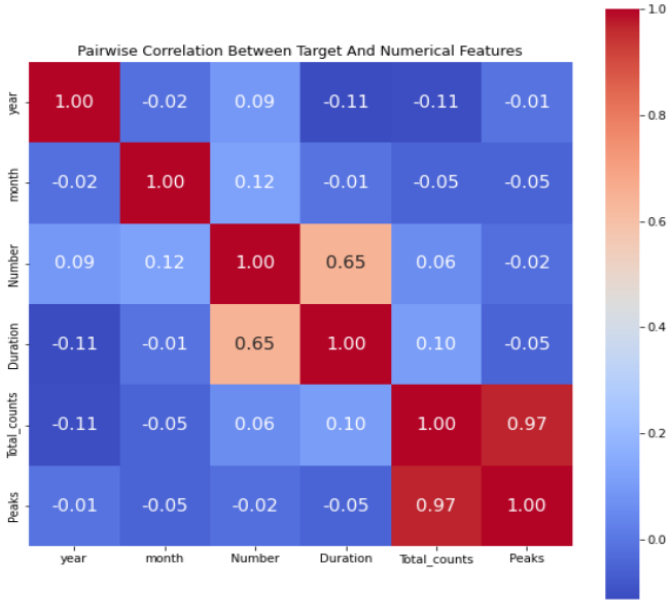


Figure 23: Correlation plot between numerical features

Target: Number of flares, Categorical feature: Energy

ANOVA test between target and categorical feature:

	sum_sq	df	F	PR(>F)
Energy	1.108451e+07	1.0	55.411183	6.615458e-12
Residual	3.060629e+07	153.0	NaN	NaN

Target: Duration, Categorical feature: Energy

ANOVA test between target and categorical feature:

	sum_sq	df	F	PR(>F)
Energy	1.006266e+06	1.0	277.040793	3.700322e-36
Residual	5.557260e+05	153.0	NaN	NaN

Figure 24: ANOVA test results between targets and categorical feature

Target: Number of flares

```
Granger Causality between Number and Duration:

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.2794  , p=0.5980  , df_denom=138, df_num=1
ssr based chi2 test:   chi2=0.2854  , p=0.5932  , df=1
likelihood ratio test: chi2=0.2852  , p=0.5933  , df=1
parameter F test:         F=0.2794  , p=0.5980  , df_denom=138, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.3647  , p=0.6951  , df_denom=135, df_num=2
ssr based chi2 test:   chi2=0.7564  , p=0.6851  , df=2
likelihood ratio test: chi2=0.7544  , p=0.6858  , df=2
parameter F test:         F=0.3647  , p=0.6951  , df_denom=135, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=1.1102  , p=0.3474  , df_denom=132, df_num=3
ssr based chi2 test:   chi2=3.5072  , p=0.3198  , df=3
likelihood ratio test: chi2=3.4636  , p=0.3255  , df=3
parameter F test:         F=1.1102  , p=0.3474  , df_denom=132, df_num=3
```

```
Granger Causality between Number and Energy:

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.1734  , p=0.6777  , df_denom=138, df_num=1
ssr based chi2 test:   chi2=0.1772  , p=0.6738  , df=1
likelihood ratio test: chi2=0.1771  , p=0.6739  , df=1
parameter F test:         F=0.1734  , p=0.6777  , df_denom=138, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.0158  , p=0.9843  , df_denom=135, df_num=2
ssr based chi2 test:   chi2=0.0328  , p=0.9838  , df=2
likelihood ratio test: chi2=0.0328  , p=0.9838  , df=2
parameter F test:         F=0.0158  , p=0.9843  , df_denom=135, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.1676  , p=0.9181  , df_denom=132, df_num=3
ssr based chi2 test:   chi2=0.5294  , p=0.9124  , df=3
likelihood ratio test: chi2=0.5284  , p=0.9126  , df=3
parameter F test:         F=0.1676  , p=0.9181  , df_denom=132, df_num=3
```

Figure 25: Granger causality test results between target number of flares and features duration, energy

Target: Duration

```
Granger Causality between Duration and Number:

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=2.9140  , p=0.0901  , df_denom=138, df_num=1
ssr based chi2 test:   chi2=2.9774  , p=0.0844  , df=1
likelihood ratio test: chi2=2.9464  , p=0.0861  , df=1
parameter F test:         F=2.9140  , p=0.0901  , df_denom=138, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=1.6325  , p=0.1993  , df_denom=135, df_num=2
ssr based chi2 test:   chi2=3.3860  , p=0.1840  , df=2
likelihood ratio test: chi2=3.3457  , p=0.1877  , df=2
parameter F test:         F=1.6325  , p=0.1993  , df_denom=135, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=1.3511  , p=0.2607  , df_denom=132, df_num=3
ssr based chi2 test:   chi2=4.2682  , p=0.2339  , df=3
likelihood ratio test: chi2=4.2040  , p=0.2403  , df=3
parameter F test:         F=1.3511  , p=0.2607  , df_denom=132, df_num=3
```

```
Granger Causality between Duration and Energy:

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=1.0314  , p=0.3116  , df_denom=138, df_num=1
ssr based chi2 test:   chi2=1.0538  , p=0.3046  , df=1
likelihood ratio test: chi2=1.0499  , p=0.3055  , df=1
parameter F test:         F=1.0314  , p=0.3116  , df_denom=138, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.6963  , p=0.5002  , df_denom=135, df_num=2
ssr based chi2 test:   chi2=1.4442  , p=0.4857  , df=2
likelihood ratio test: chi2=1.4368  , p=0.4875  , df=2
parameter F test:         F=0.6963  , p=0.5002  , df_denom=135, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=10.0781 , p=0.0000  , df_denom=132, df_num=3
ssr based chi2 test:   chi2=31.8375 , p=0.0000  , df=3
likelihood ratio test: chi2=28.6672 , p=0.0000  , df=3
parameter F test:         F=10.0781 , p=0.0000  , df_denom=132, df_num=3
```

Figure 26: Granger causality test results between target duration and features number of flares, energy

Target: Number of flares

1) Univariate modelling:

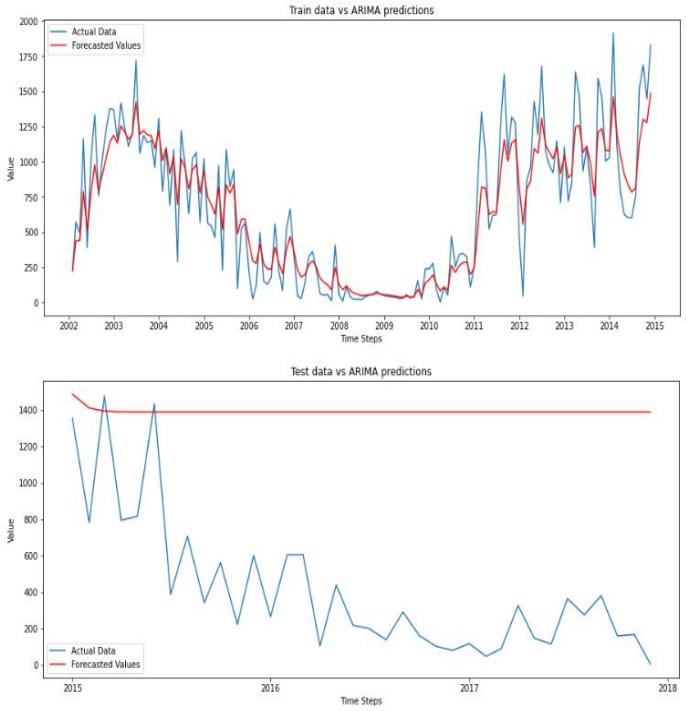


Figure 27: Actual values VS ARIMA predictions of number of flares

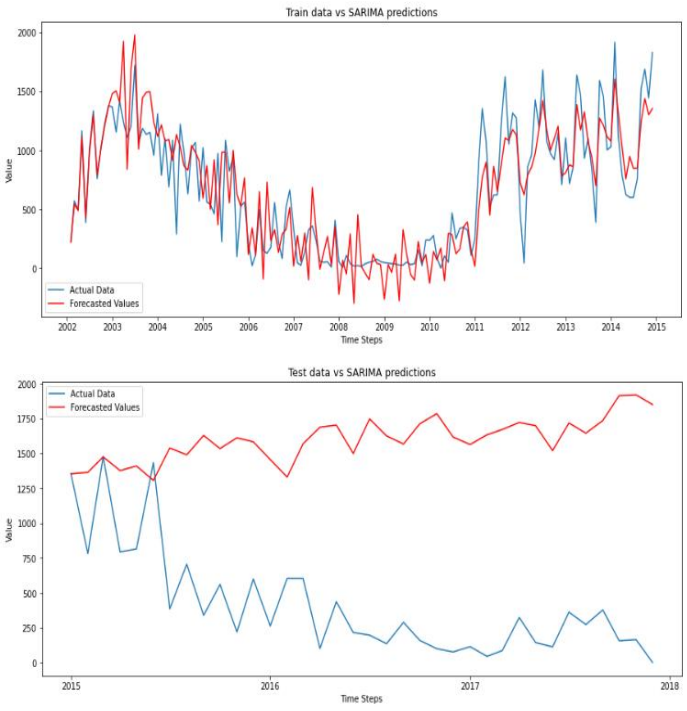


Figure 28: Actual values VS SARIMA predictions of number of flares

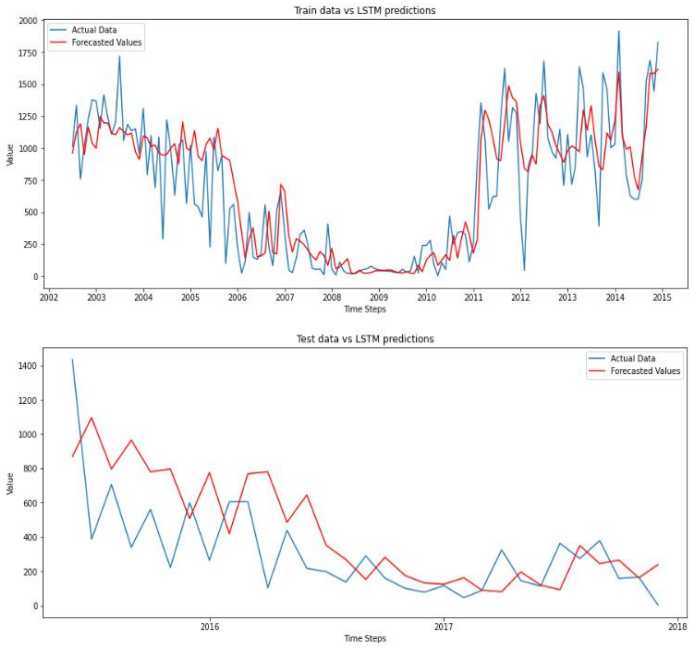


Figure 29: Actual values VS LSTM predictions of number of flares

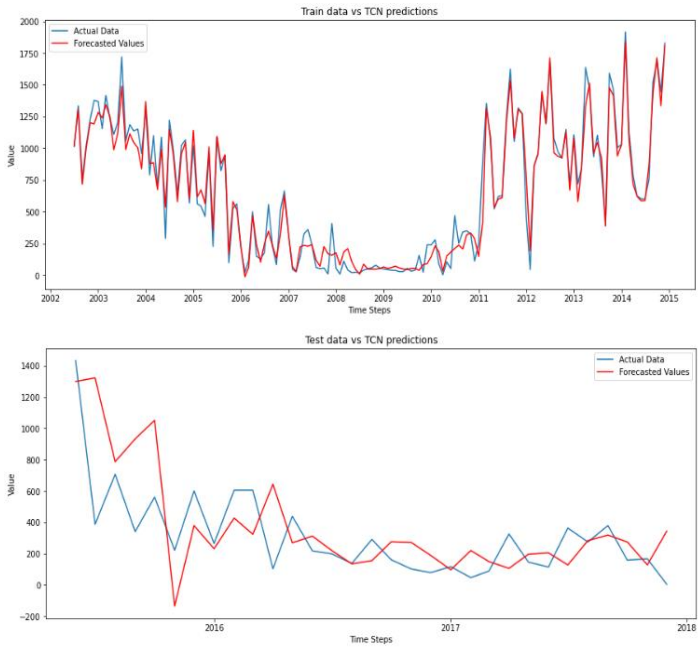


Figure 30: Actual values VS TCN predictions of number of flares

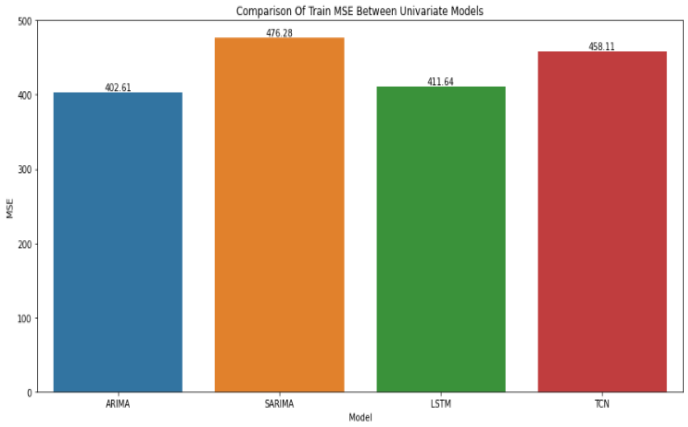
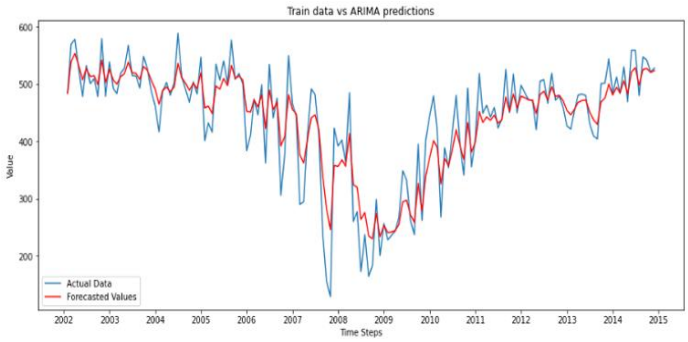


Figure 31: Train RMSE comparison for number of flares

Target: Duration

1) Univariate modelling:



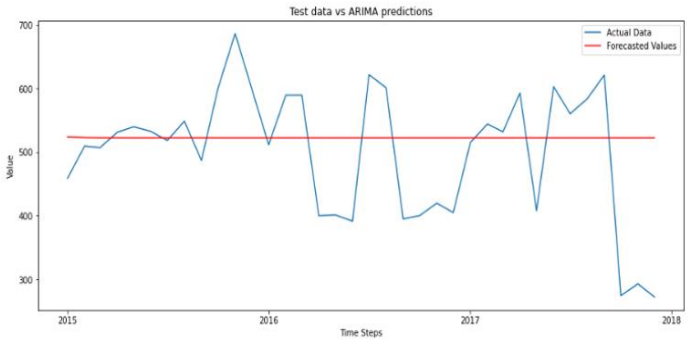
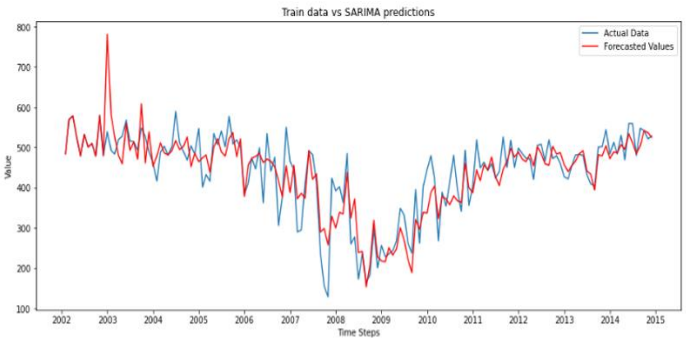


Figure 32: Actual values VS ARIMA predictions of duration



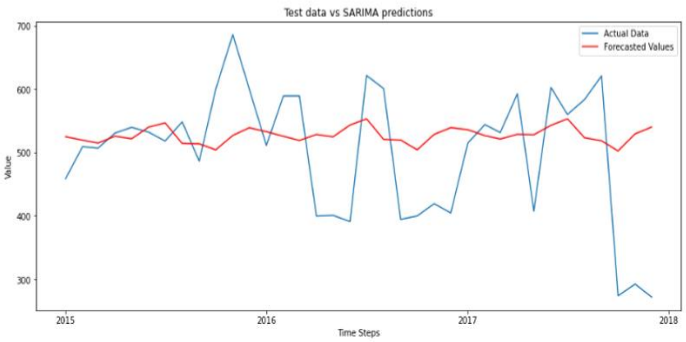


Figure 33: Actual values VS SARIMA predictions of duration

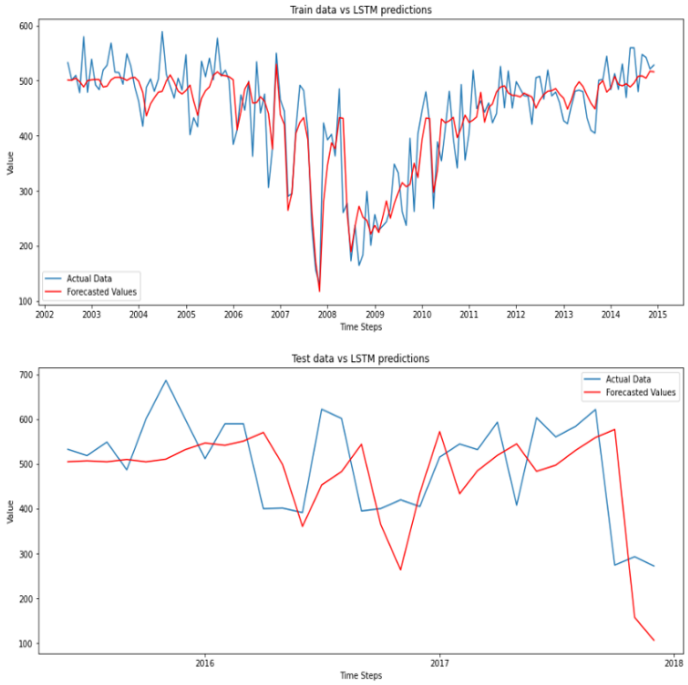


Figure 34: Actual values VS LSTM predictions of duration

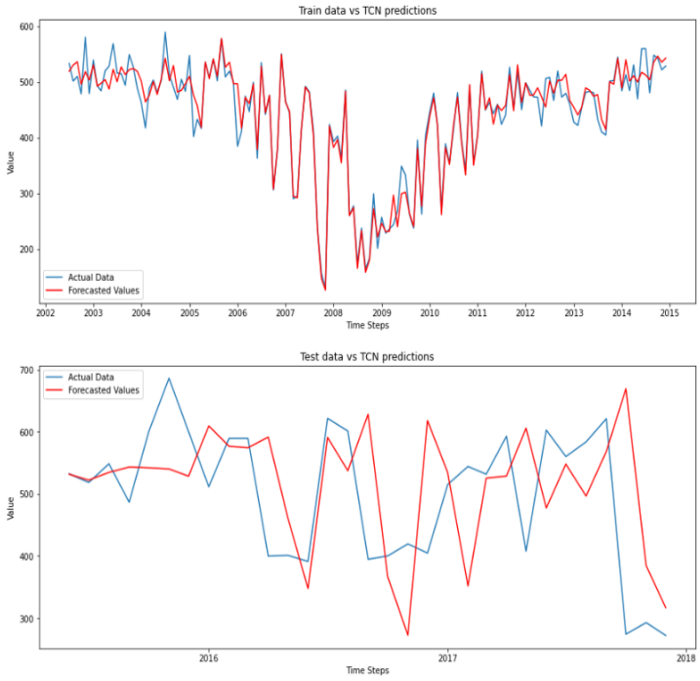


Figure 35: Actual values VS TCN predictions of duration

2) Multivariate modelling:

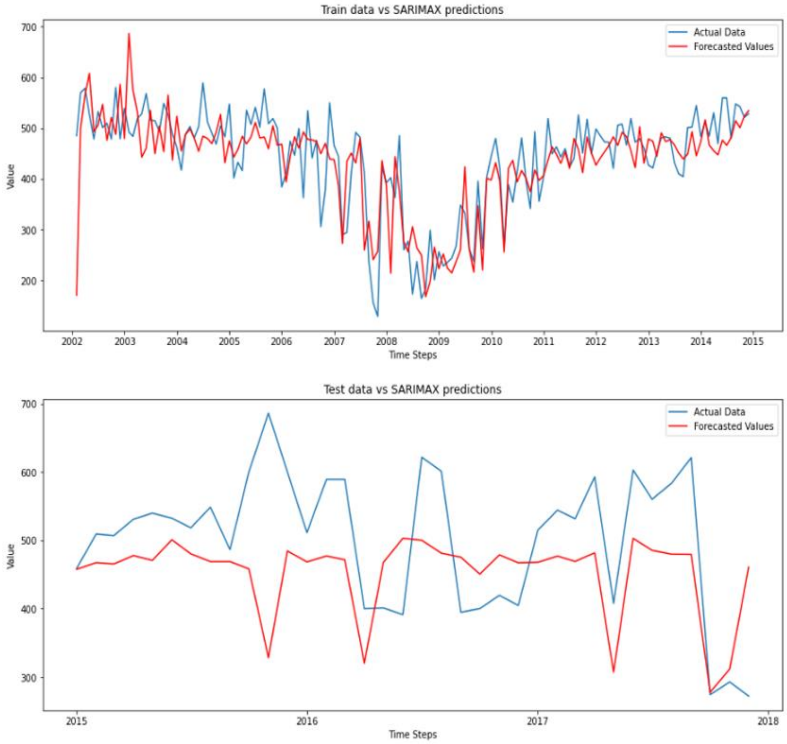


Figure 36: Actual values VS SARIMAX predictions of duration

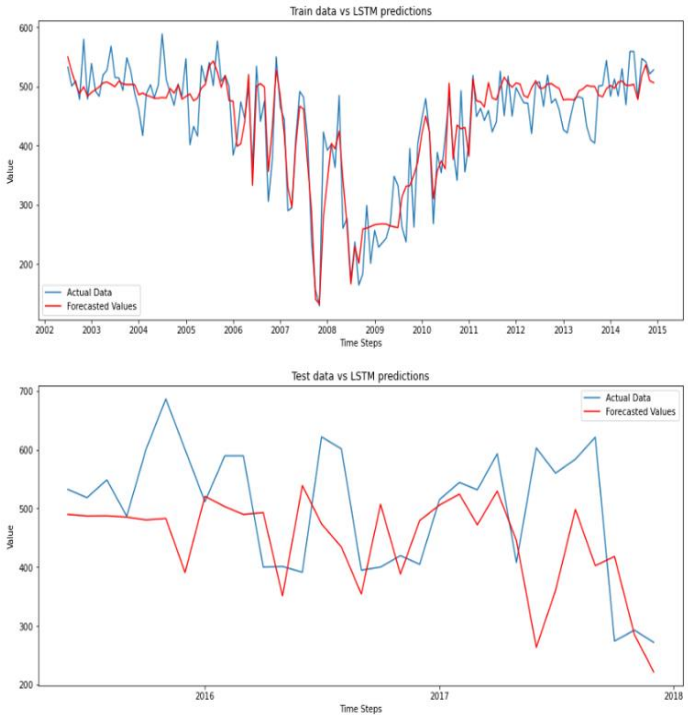


Figure 37: Actual values VS multivariate LSTM predictions of duration

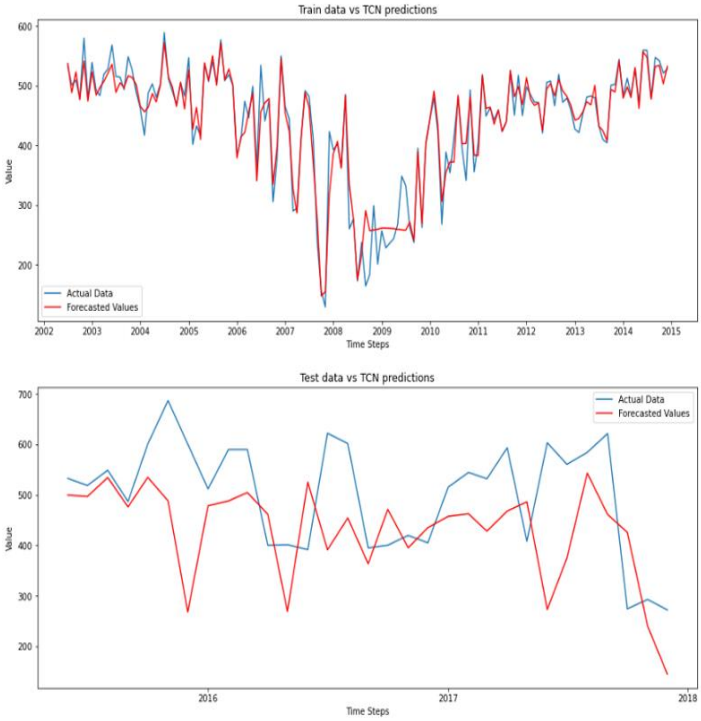


Figure 38: Actual values VS multivariate TCN predictions of duration

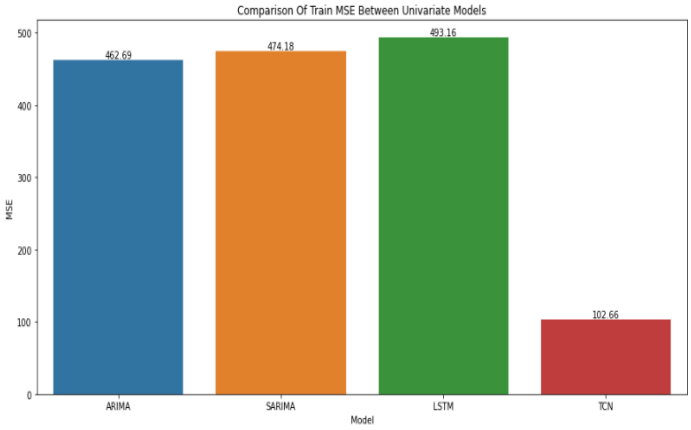


Figure 39: Train RMSE comparison of univariate models for duration

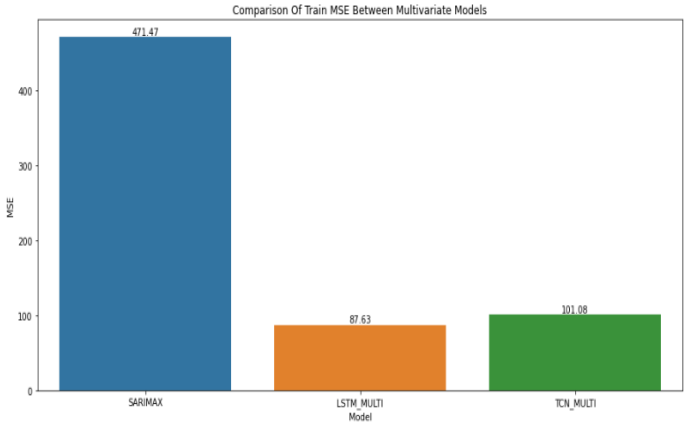


Figure 40: Train RMSE comparison of multivariate models for duration

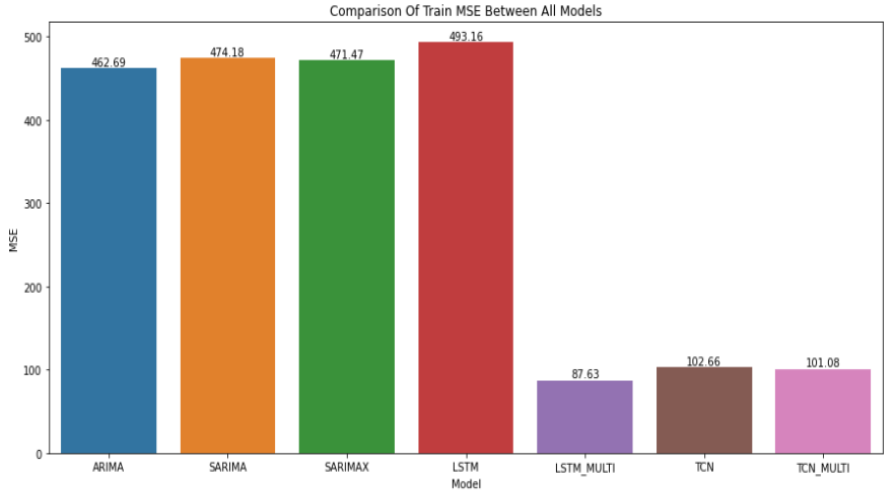


Figure 41: Train RMSE comparison of all models for duration