Comparative Analysis of Sunspot Number Predictions with Machine Learning Models

Argyro Talamagka

School of Electrical and Computer Engineering
Technical University of Crete

Supervisor: Dionysios Christopoulos

July 26, 2024

Table of Contents

Introduction

Motivation

Figure 1: Photo from my visit to the Royal Observatory in Belgium (ROB) during the "Space Weather" training course.

- Passion for Astronomy and Space Science.
- Interest to understand and investigate different forecasting models.

What are sunspots?

- Dark spots on the Sun's photosphere.
- Low temperature and intense magnetic activity.
- $\pm$ 11 year solar cycle (Schwabe cycle).
- Solar activity impacts Earth's space weather and technology by causing:
  - satellite disruptions and damage
  - power grid failures - pipeline corrosion
  - increased radiation exposure, particularly affecting aviation

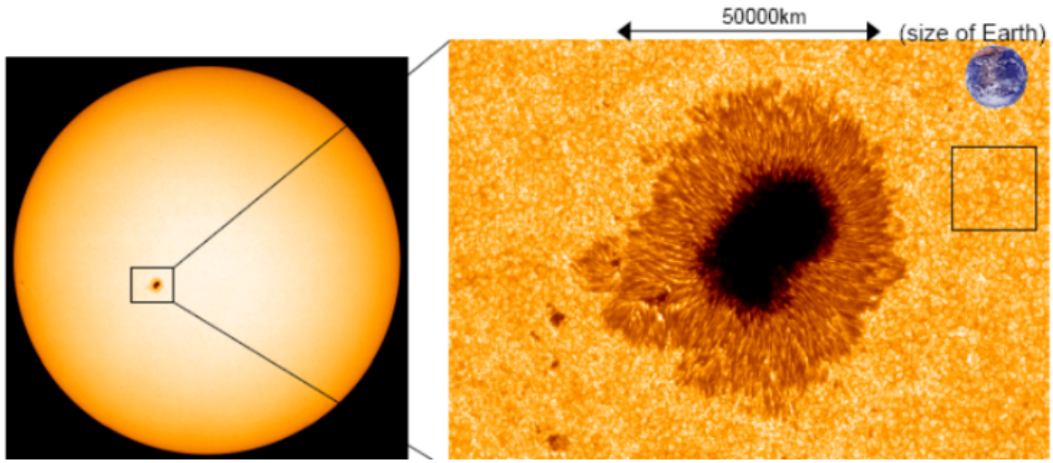
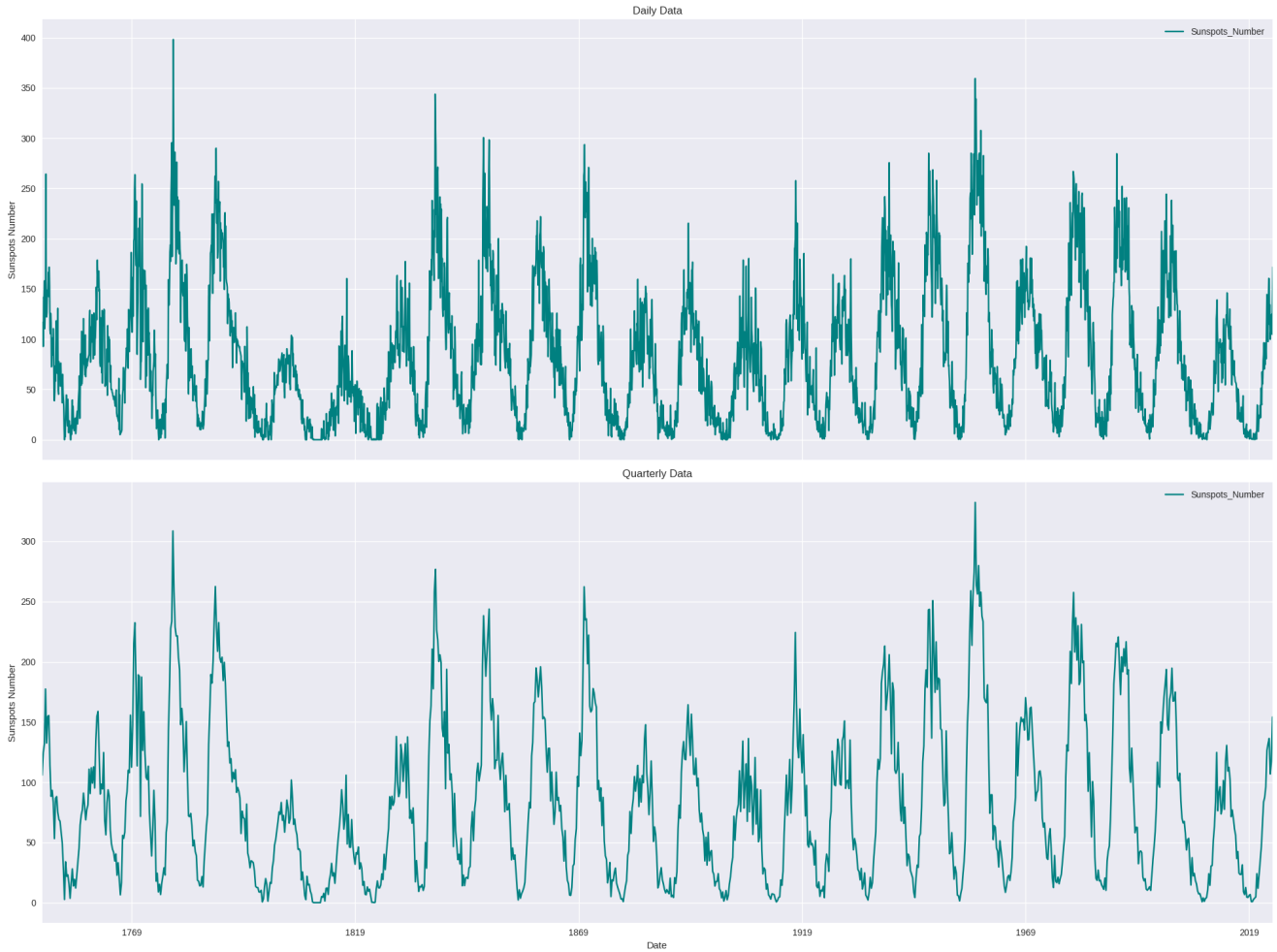


Figure 2: Close-up of a Sunspot.

Exploratory Data Analysis

Exploratory Data Analysis

- Monthly data from 1749 - 2024 retrieved from ROB's SIDC (3305 rows).
- Programming language used: **Python** - Environment: **Colab**



Stationarity Test and Basic Plots

Augmented Dickey-Fuller Results

Metric	**Value**
Test Statistic	-10.61
p-value	5.8×10^{-19}
Lags Used	28
Observations Used	3,276
Critical Value (1%)	-3.43
Critical Value (5%)	-2.86
Critical Value (10%)	-2.57

- **Null Hypothesis (H_0)**: The time series has a unit root, meaning it is non-stationary.
- **Interpretation**: Since the p-value is very small, we reject the null hypothesis and conclude that *the series is stationary*.
- **Critical Values**: Test statistic is more negative than the critical values at 1%, 5%, and 10% levels, further supporting stationarity.

SARIMA Modeling

ACF and PACF Plots

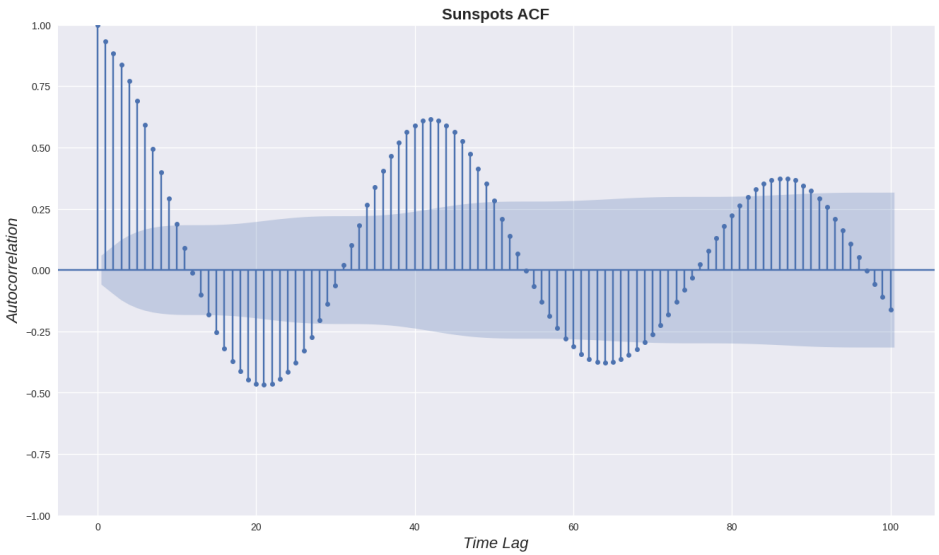


Figure 4: Auto-Correlation Function

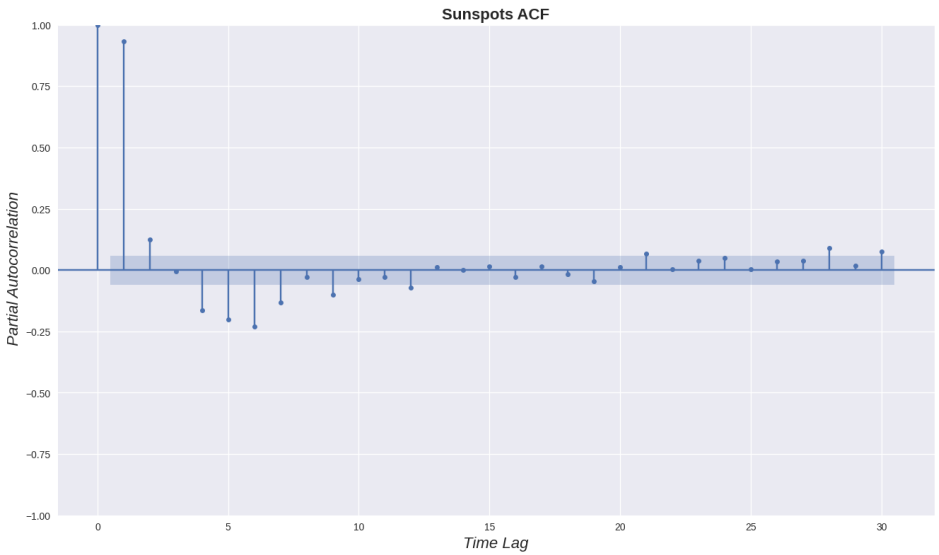


Figure 5: Partial ACF

Selecting SARIMA parameters

- From ACF we identify the seasonal pattern (44 quarters = 11 years).
- PACF indicates that we should use p=6 or 7 for the AR component.

SARIMA Results

SARIMA(p, d, q)x(P, D, Q, S)
SARIMA(7, 0, 1)x(1, 0, 1, 43)

- Out-of-sample prediction on 10% test set does not perform well.
- One step-ahead predictions (rolling forecast) is a much better option.
- For one step-ahead, we train the model for the N values, we make a prediction and then we use N+1 (actual) values for the next prediction as history.

Metric	**Value**
MAE	11.744
R^2	0.910

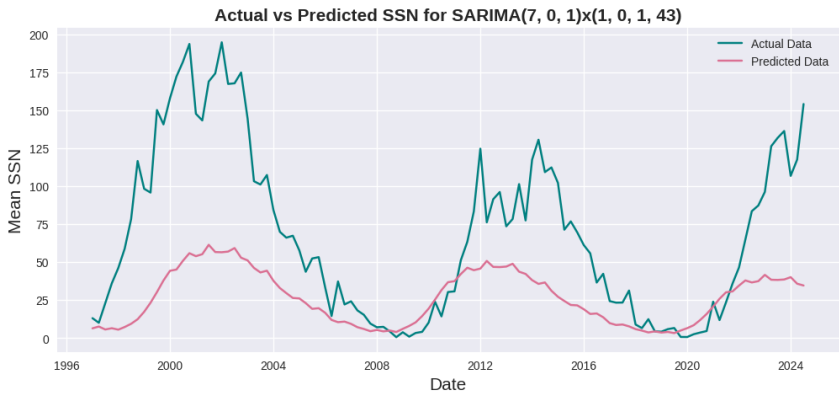


Figure 6: Static Forecast

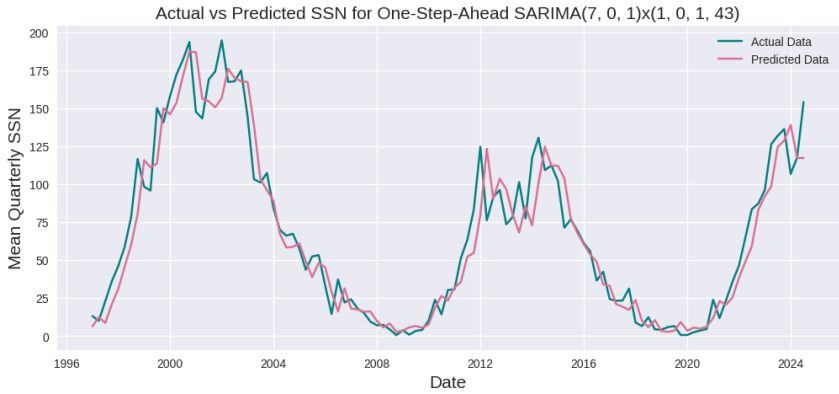
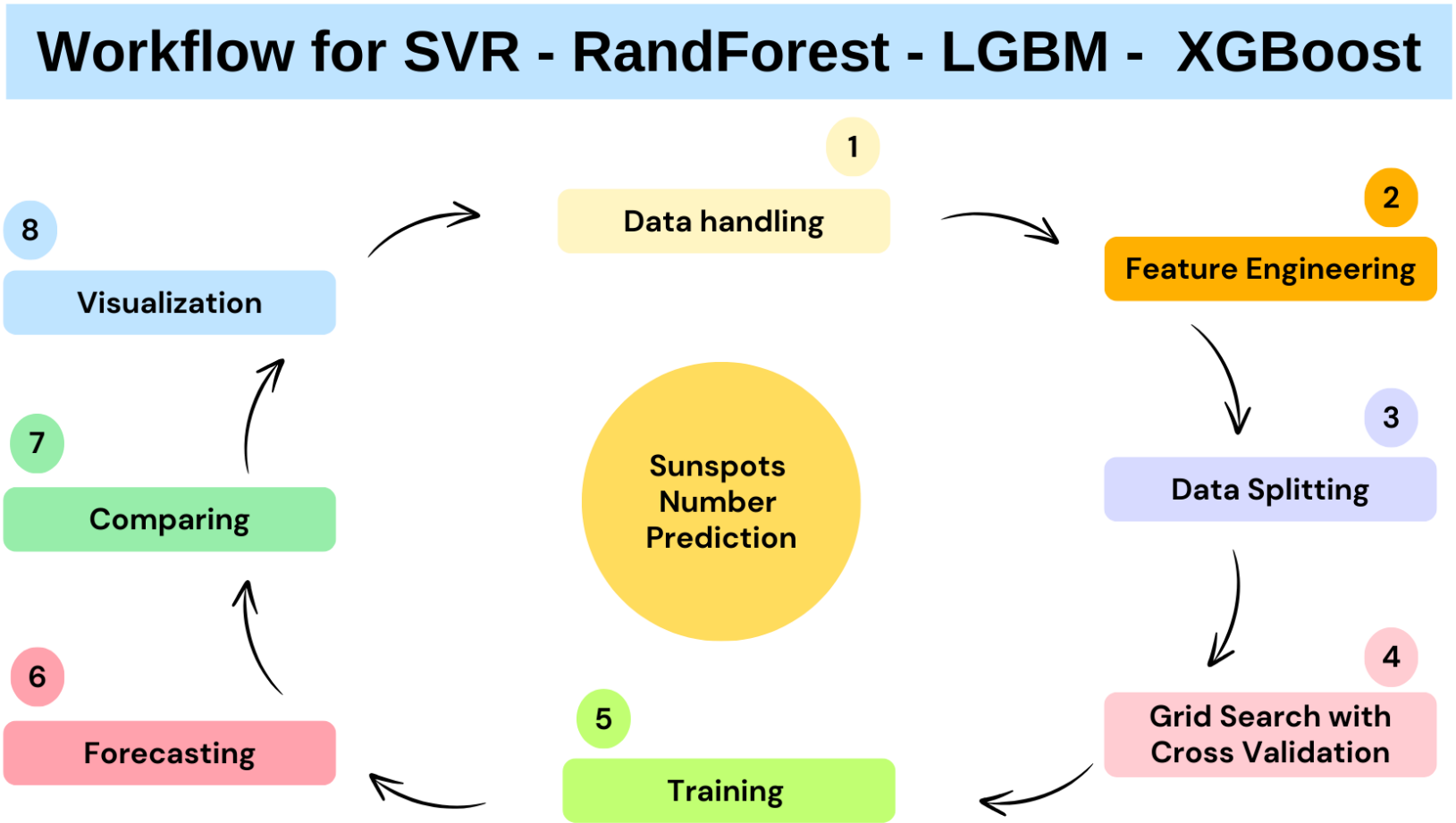


Figure 7: Rolling Forecast

Machine Learning Models
(SVR - Random Forest - LGBM - XGBoost)

Steps Followed for the Machine Learning Models



SVR - Random Forest

Lagged Values as Features (1, 2, 3 , 4 ,5 ,6)

- **SVR**(Lags(1,2))
  MAE: 13.742
- **Random Forest**
  MAE: 12.046

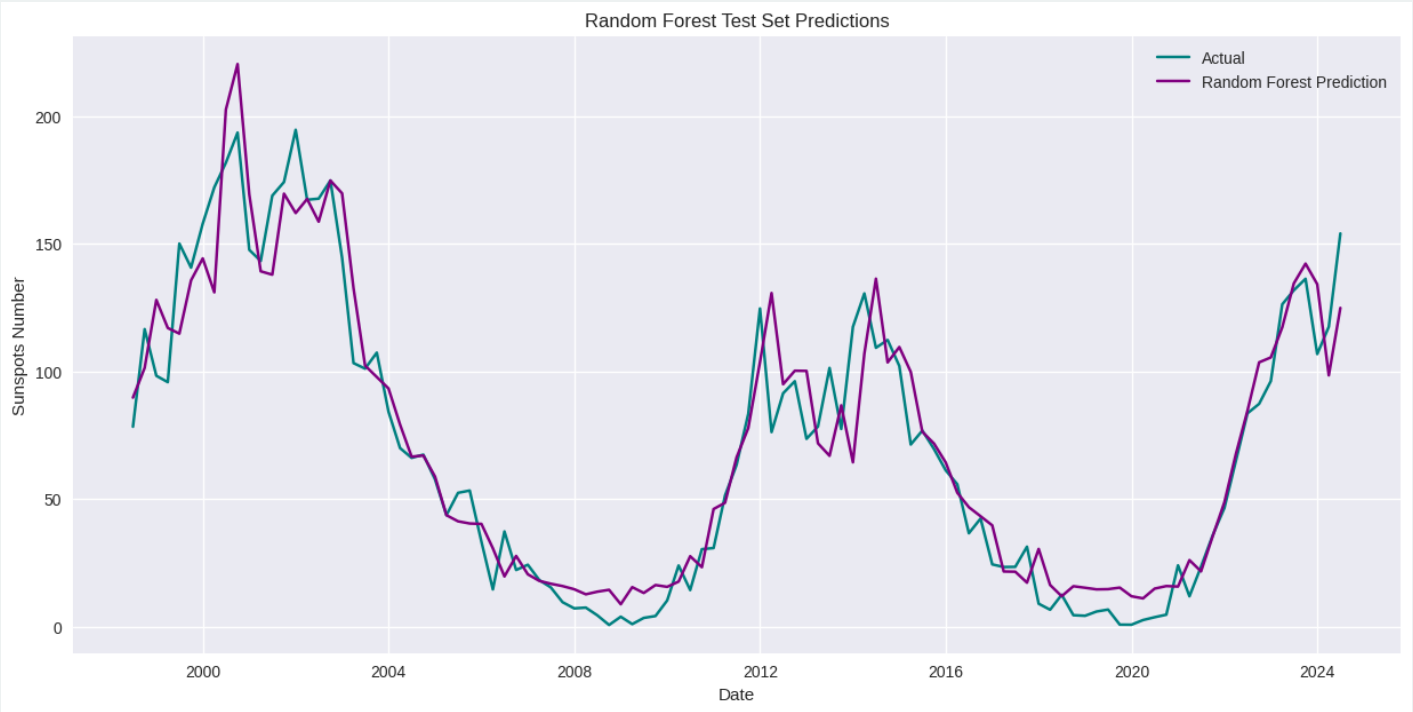


Figure 8: Predictions for the same test set (last 10%)

LGBM vs XGBoost

Lagged Values as Features (1, 2, 3, 4, 5, 6)

- **XGBoost**
  MAE: 13.394
- **LGBM**
  MAE: 13.208

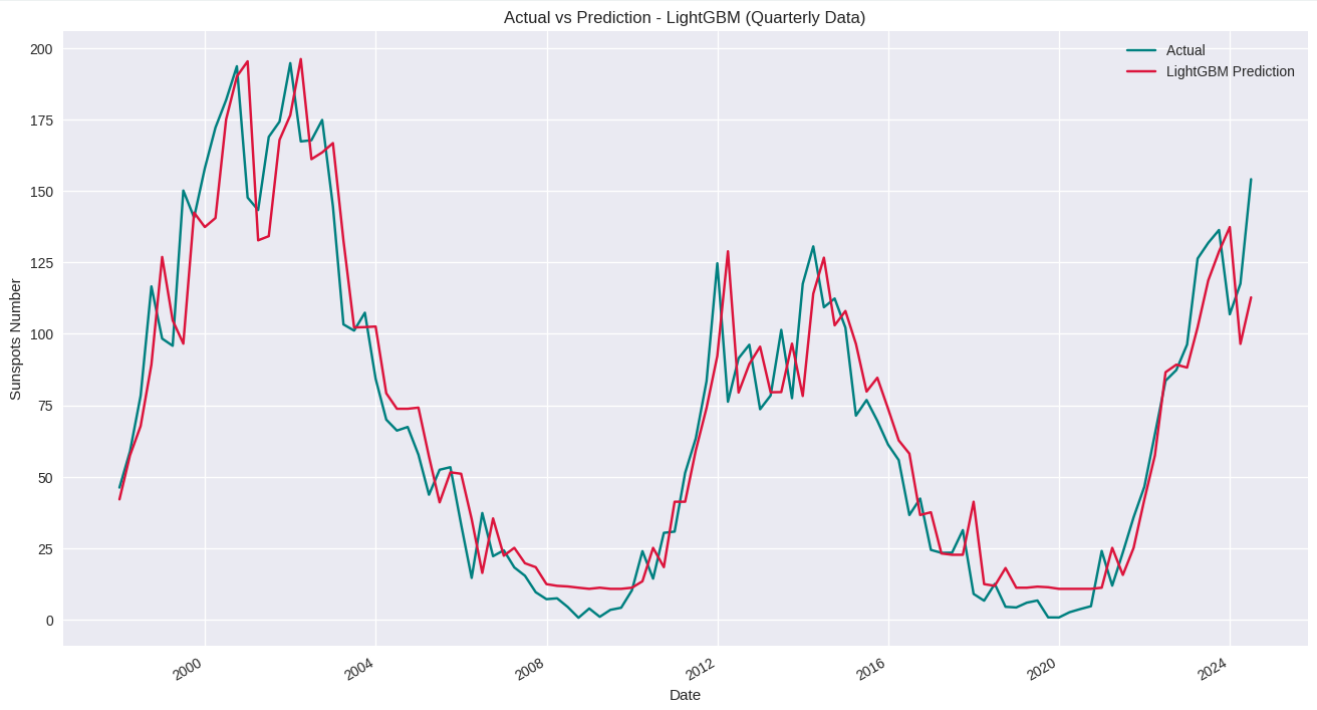


Figure 9: LGBM Predictions for the same test set(last 10%)

Improving Machine Learning Models with Extra Features

- **Feature Importance:** Adding relevant features can enhance model performance.
- **Idea:** Exploring the impact of related features such as geomagnetic indices **Kp, Ap and solar radio flux (10.7 cm wavelength)** on predictive accuracy.
- Those indices are key indicators in space weather.

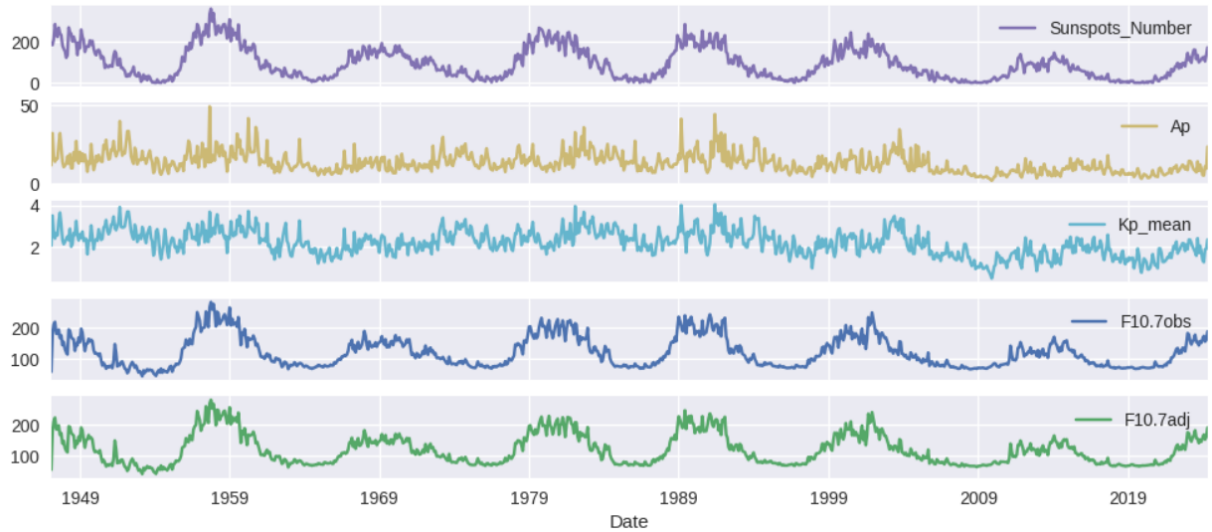
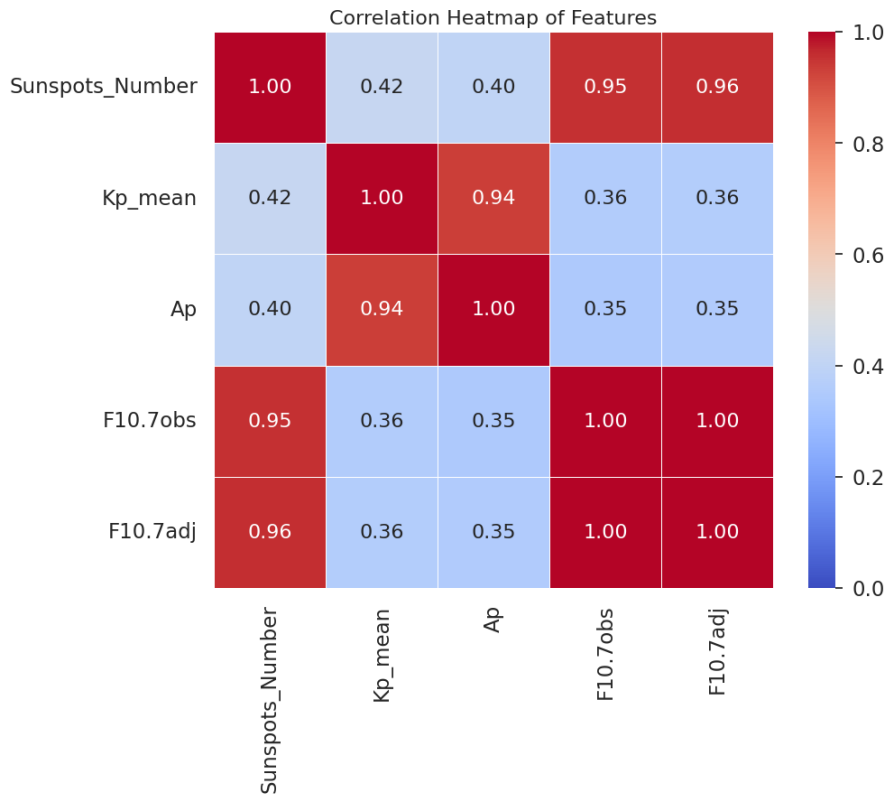


Figure 10: Subplot of Sunspots number with other features.

Preprocessing the additional features

- The new data (33.791 rows) were retrieved from Geomagnetic Observatory Niemegk, GFZ.
- Daily data from 1932 until today.
- Resampled to mean monthly Sunspot Number, Kp and Solar Flux (10.7) and that gave me 1109 rows, but only 928 rows had data for F10.7cm.
- Will integrating these features into the predictive models, help with our predictions?



Adding features to Machine Learning Models

- Yes, it helps.
- Comparing the models with just the **time lags** vs the **time lags + extra features** on the same data set.
- We see a small improvement in the accuracy when we add the "extra" feature.

Comparison of performance between only Lagged and Extra Features

Model	Lagged Features			Full Features		
	MAE	RMSE	R^2	MAE	RMSE	R^2
LightGBM	12.00	16.1621	0.8864	11.56	15.2522	0.8988
XGBoost	12.92	16.587	0.867	11.55	15.2766	0.8985
RandomForest	12.17	16.798	0.864	11.99	16.586	0.867
SVR	15.77	18.1624	0.8566	17.41	19.8919	0.8279

Recurrent Neural Network (LSTM)

Recurrent Neural Network: LSTM

- Designed to handle sequential data. Memory cells, capture long-term dependencies
- Use gates (input gate, forget gate, and output gate) to regulate the flow of information.
- Capture non-linear patterns in data, making them powerful for modeling complex real-world phenomena.

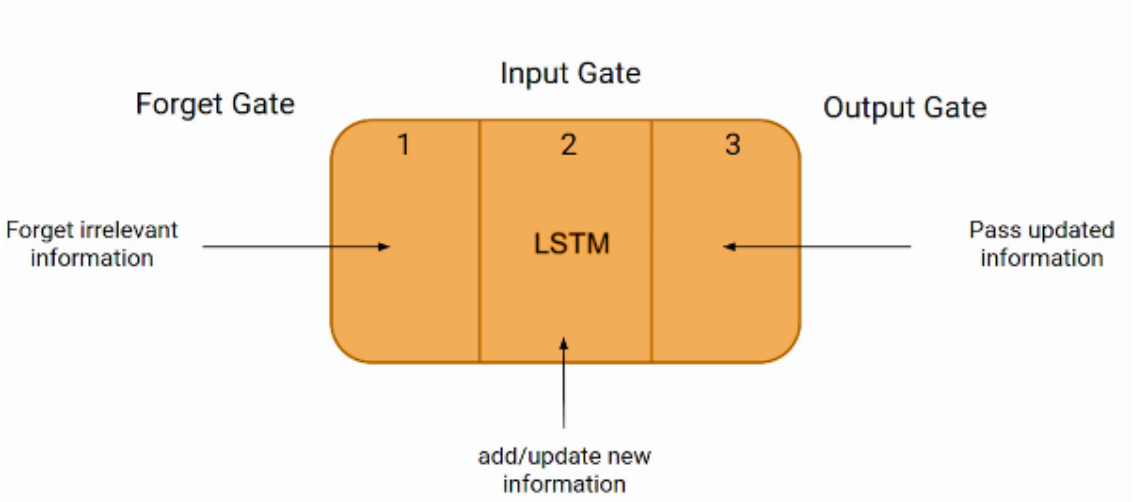


Figure 11: Simple Representation of how LSTM works

Training LSTM

- Grid Search with 5-fold Cross Validation for hyperparameter tuning, with early stopping criterion.
- Trained using sequences of 44 quarters..
- Parameters: **Batch Size** = 16/32, **Epochs** = 50/100, **Optimizer** = adam/rmsprop, **Units** = 50/100/200

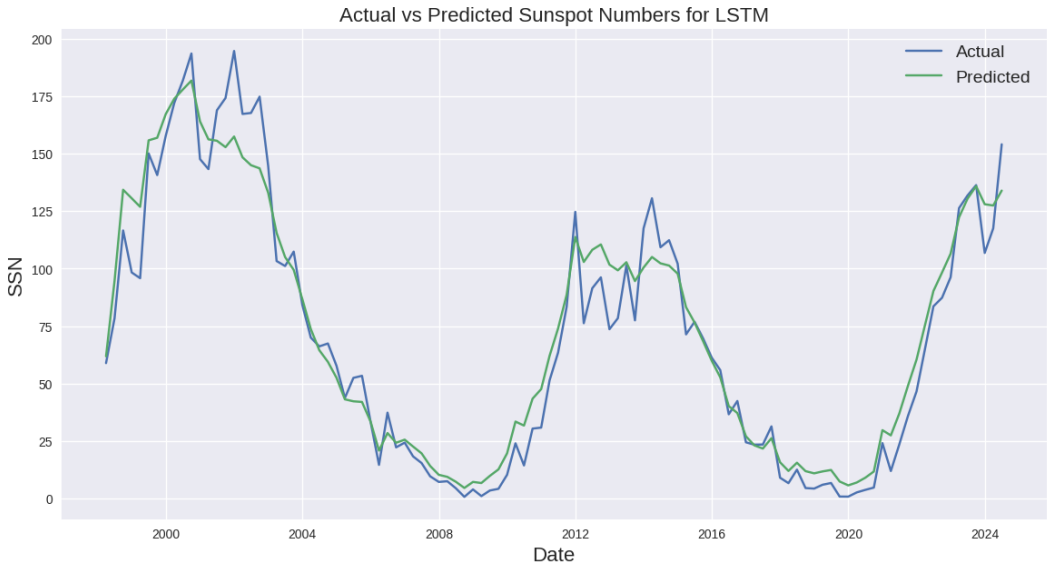


Figure 12: LSTM Predictions on Test Set (last 10%)

Model Comparison

Model Performance

- The model with the best performance based on MAE, RMSE and R^2 is the LSTM.

Comparison of Test Metrics for Different Models and Feature Sets

Model	MAE	RMSE	R^2
LSTM	9.46	12.34	0.95
SARIMA(7, 0, 1)x(1, 0, 1, 43)	11.74	16.43	0.91
LightGBM	13.21	17.26	0.90
XGBoost	13.39	17.38	0.90
RandomForest	12.05	16.38	0.91
SVR	13.74	17.72	0.90

Conclusions

Key Findings

- **Best Performing Model:** LSTM with MAE of 9.46 and R^2 of 0.95.
- **Feature Enhancement:** Adding extra features (Kp, Ap, Solar Flux) improved model accuracy, but only slightly.
- **SARIMA vs. Machine Learning:** While SARIMA provided reasonable predictions, ML models, especially LSTM, showed superior performance.

- Classical models are easier to interpret and assess prediction uncertainty but may not perform as well, while more complex models, though potentially more accurate, might complicate reliability estimation.

Future Work

How I would proceed with the analysis

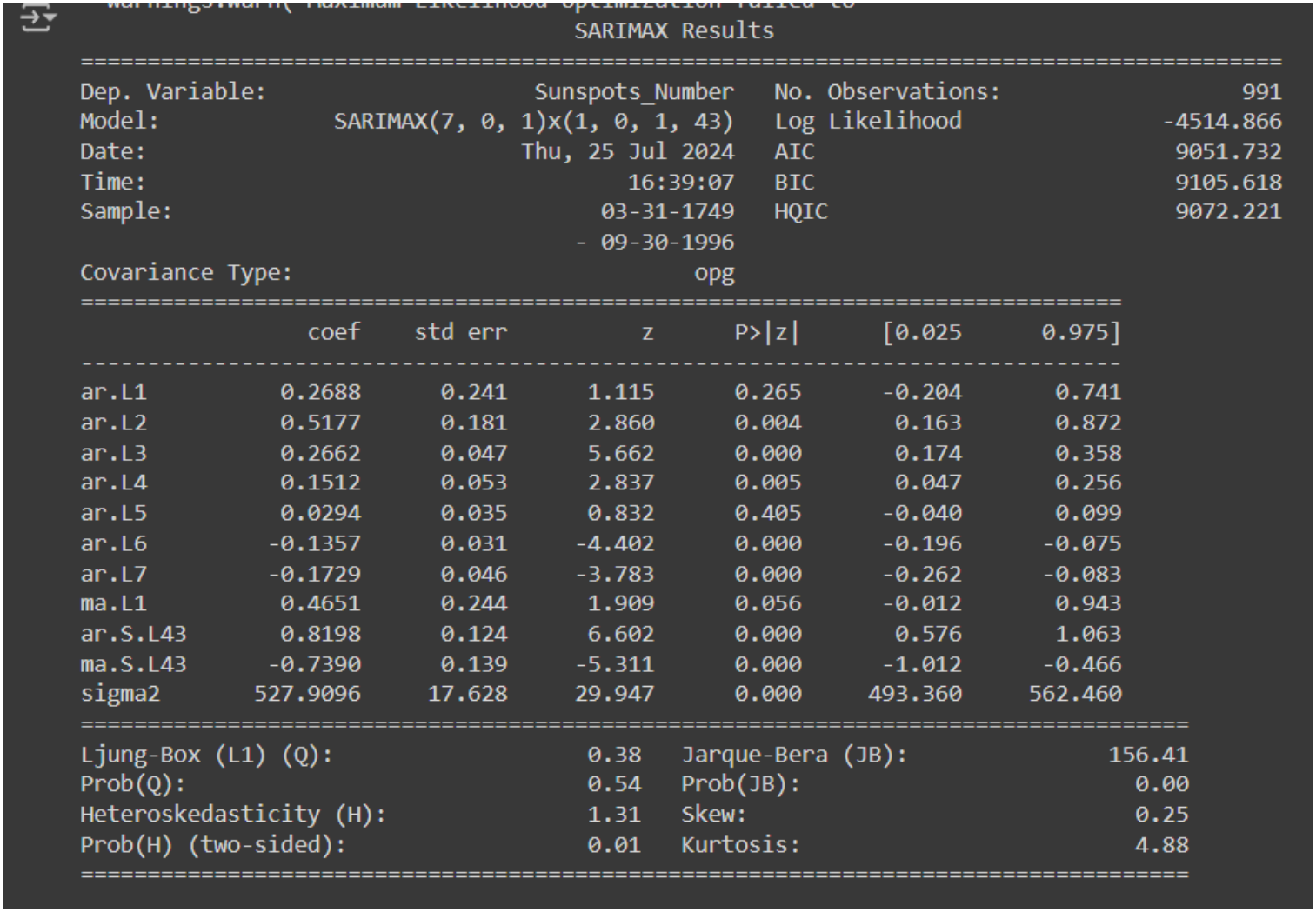
- **Enhance feature engineering:** Adding cyclical features(sin-cos), rolling mean etc.
- **Enhanced Predictive Models:** With more computational power, more complex models could be implemented for **daily predictions** and **larger grid searches**.
- **Incorporate Additional Data:** Utilize more comprehensive datasets, including sunspot area and solar magnetic activity, for all the candidate models.
- **Real-Time Forecasting:** Develop systems to generate real-time sunspot forecasts with improved accuracy.

References

- **Analytics Vidhya (2021)**. *Introduction to Long Short-Term Memory (LSTM)*. Retrieved from `https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/`
- **Box, G. E. P., Jenkins, G. M., Reinsel, G. C. (2015)**. *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.
- **Breiman, L. (2001)**. *Random Forests*. Machine Learning, 45(1), 5-32.
- **Clette, F., Lefèvre, L.(2016)** The New Sunspot Number: Assembling All Corrections. Sol Phys 291, 2629–2651 . https://doi.org/10.1007/s11207-016-1014-y
- **Chen, T., Guestrin, C. (2016)**. *XGBoost: A Scalable Tree Boosting System*. In KDD 2016. `https://arxiv.org/abs/1603.02754`
- **Dang, Y., Chen, Z., Li, H., Shu, H. (2022)**. A Comparative Study of non-deep Learning, Deep Learning, and Ensemble Learning Methods for Sunspot Number Prediction. Applied Artificial Intelligence, 36(1).
- **Hathaway, D. H. (2010)**. *The solar cycle*. Living Reviews in Solar Physics, 7(1), 1.
- **Helmholtz Centre Potsdam GFZ (n.d.)**. *Kp Index Data*. `https://kp.gfz-potsdam.de/en/data`
- **Hochreiter, S., Schmidhuber, J. (1997)**. *Long Short-Term Memory*. Neural Computation, 9(8), 1735-1780.
- **Matzka, J., Stolle, C., Yamazaki, Y., Bronkalla, O. and Morschhauser, A., 2021.** The geomagnetic Kp index and derived indices of geomagnetic activity. Space Weather, https://doi.org/10.1029/2020SW002641
- **Ng, K (2016)**. Prediction Methods in Solar Sunspots Cycles. Sci Rep 6, 21028.
- **Uccle Solar Equatorial Table (USET)**. `http://www.sidc.be/uset/`


THANK YOU!

Extra



SARIMAX Results

Dep. Variable:	Sunspots_Number	No. Observations:	991
Model:	SARIMAX(7, 0, 1)x(1, 0, 1, 43)	Log Likelihood	-4514.866
Date:	Thu, 25 Jul 2024	AIC	9051.732
Time:	16:39:07	BIC	9105.618
Sample:	03-31-1749	HQIC	9072.221
	- 09-30-1996		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2688	0.241	1.115	0.265	-0.204	0.741
ar.L2	0.5177	0.181	2.860	0.004	0.163	0.872
ar.L3	0.2662	0.047	5.662	0.000	0.174	0.358
ar.L4	0.1512	0.053	2.837	0.005	0.047	0.256
ar.L5	0.0294	0.035	0.832	0.405	-0.040	0.099
ar.L6	-0.1357	0.031	-4.402	0.000	-0.196	-0.075
ar.L7	-0.1729	0.046	-3.783	0.000	-0.262	-0.083
ma.L1	0.4651	0.244	1.909	0.056	-0.012	0.943
ar.S.L43	0.8198	0.124	6.602	0.000	0.576	1.063
ma.S.L43	-0.7390	0.139	-5.311	0.000	-1.012	-0.466
sigma2	527.9096	17.628	29.947	0.000	493.360	562.460

Ljung-Box (L1) (Q):	0.38	Jarque-Bera (JB):	156.41
Prob(Q):	0.54	Prob(JB):	0.00
Heteroskedasticity (H):	1.31	Skew:	0.25
Prob(H) (two-sided):	0.01	Kurtosis:	4.88

Figure 13: SARIMA statistics

Extra

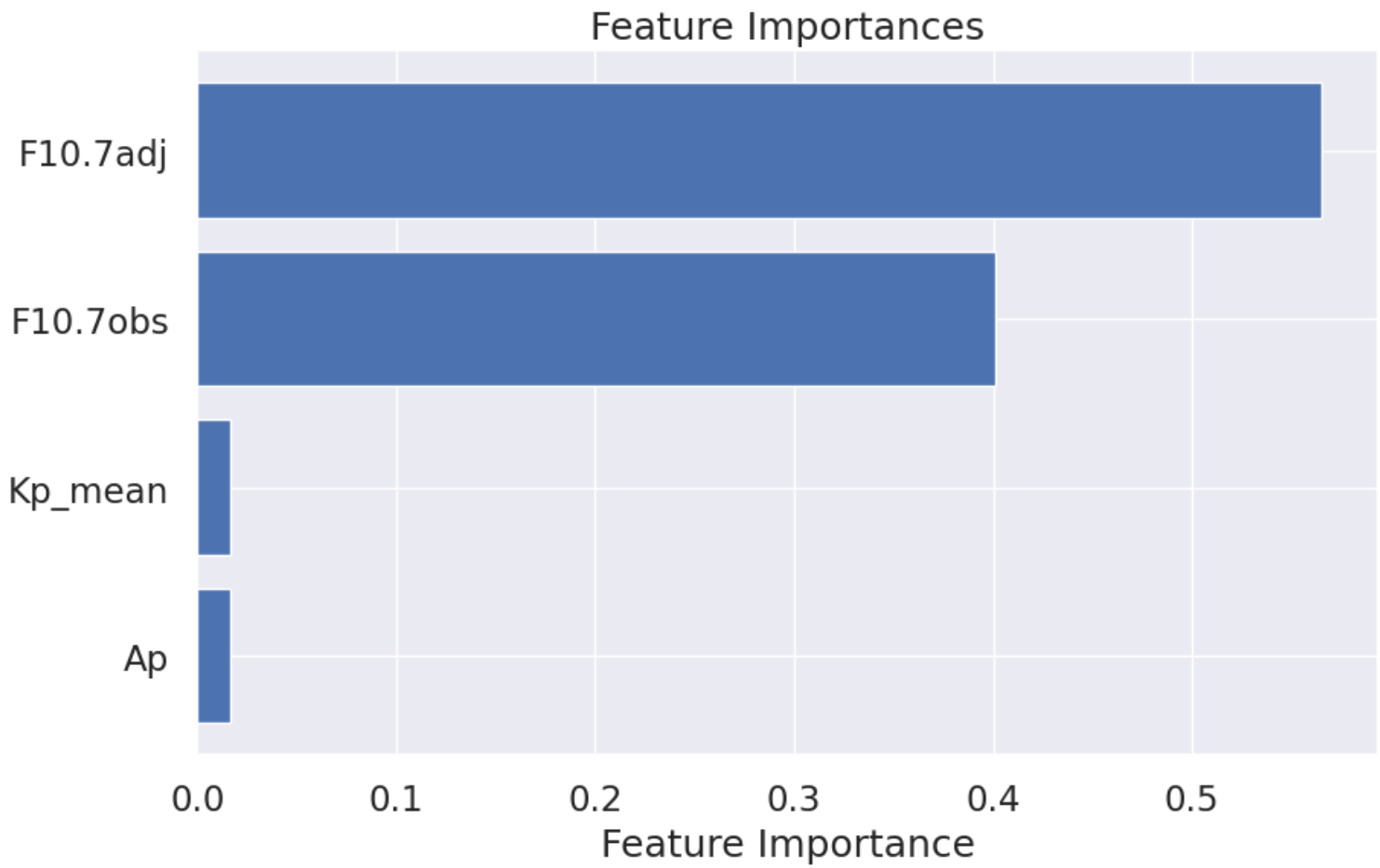


Figure 14: Feature Importance for Random Forest

Extra

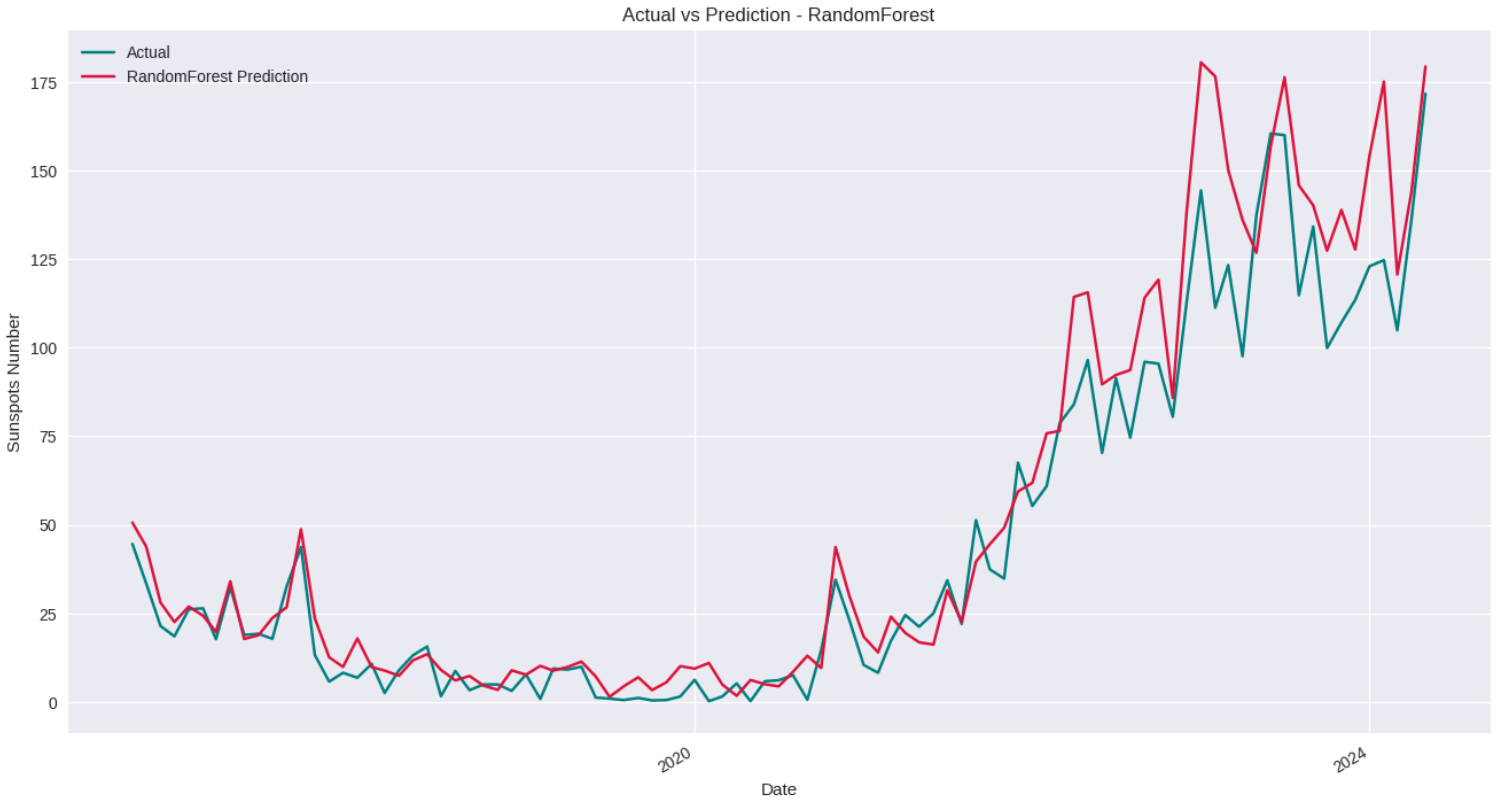


Figure 15: Random Forest's Predictions with extra features