Time Series Analysis

ARIMA-UCM-ML MODELS

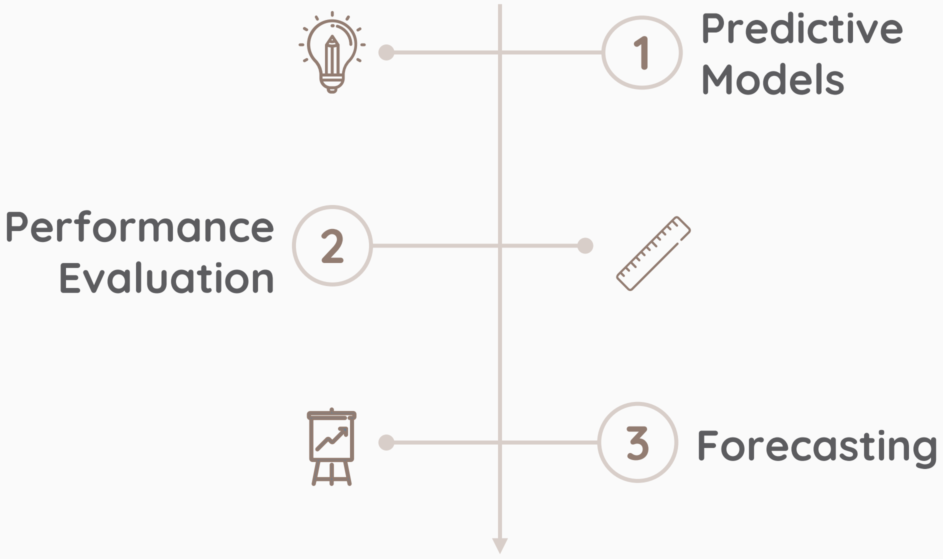
Luca Porcelli 853189

Data & Goal

Data

The dataset consists of three columns: date (yyyy-mm-dd), weekday, and ave_days. It contains 3009 observations, each representing a day starting from 2007-01-04.

Goal



1 Predictive Models

2 Performance Evaluation

3 Forecasting

Outline

1

Pre-processing

General data analysis, checking for null values and outliers.

2

Stationary Analysis & Transformation

Verification with two tests and logarithmic transformation.

3

Models

SARIMA, UCM, ML models and optimization with optimal parameter search.

4

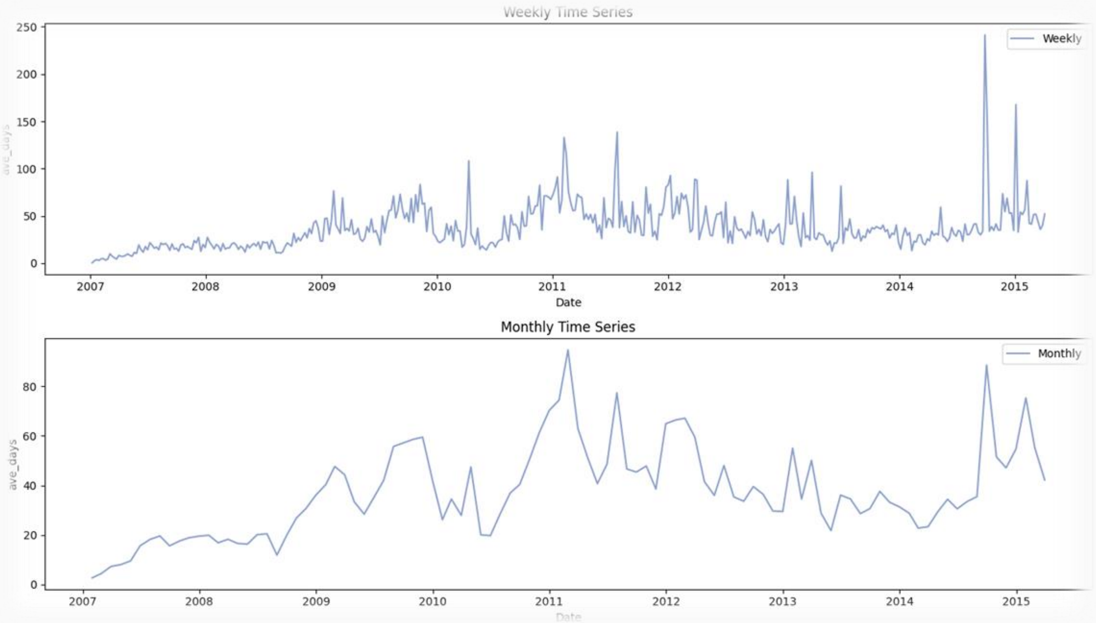
Forecasting

Forecasts from 2015-04-01 to 2015-11-07.

01 Pre-processing

General data analysis, checking for null values and outliers

Exploratory Analysis



Trend

An increasing trend from 2008 to 2012, followed by a decreasing trend from 2012 to 2014, with a recovery in 2015.

Outliers

Presence of high outliers that deviate from the mean.

Missing values & Outliers

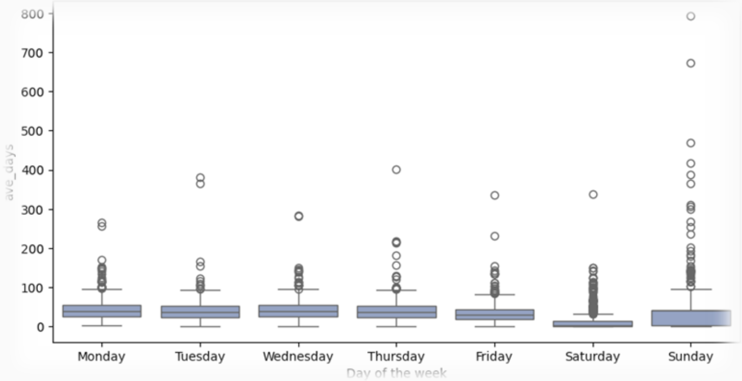
Missing values

Presence of 202 null values out of 3009 (6.71%):

- Monday: 4.65%
- Thursday: 0.93%
- Wednesday: 0.00%
- Tuesday: 0.23%
- Friday: 0.93%
- Saturday: 1.40%
- Sunday: 38.84%

Outliers

The outliers represent 3.39% of the entire dataset. To address the outlier issue, it was decided to set a maximum threshold for the ave_days value at 100.



02 Stationary Analysis & Transformation

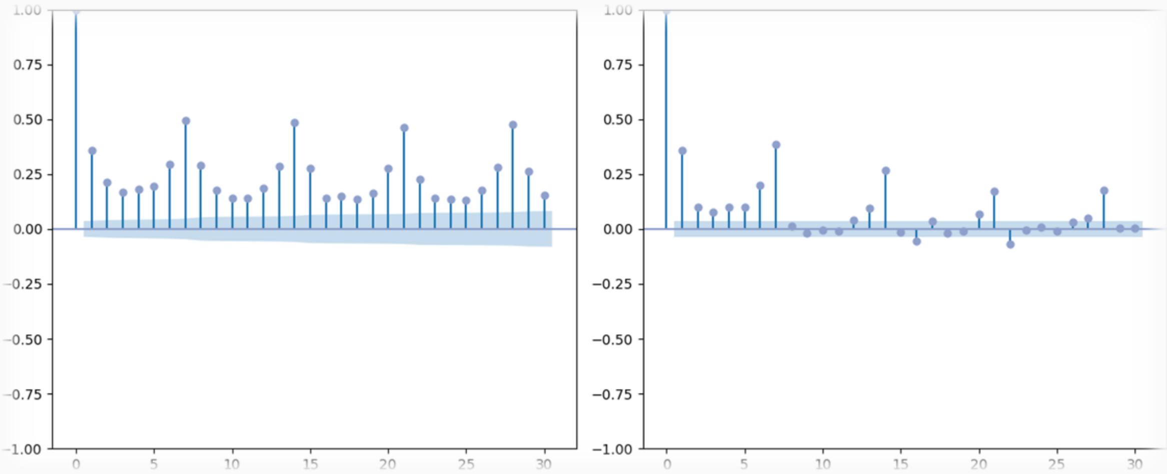
Verification with two tests and logarithmic transformation.

Stationary Analysis

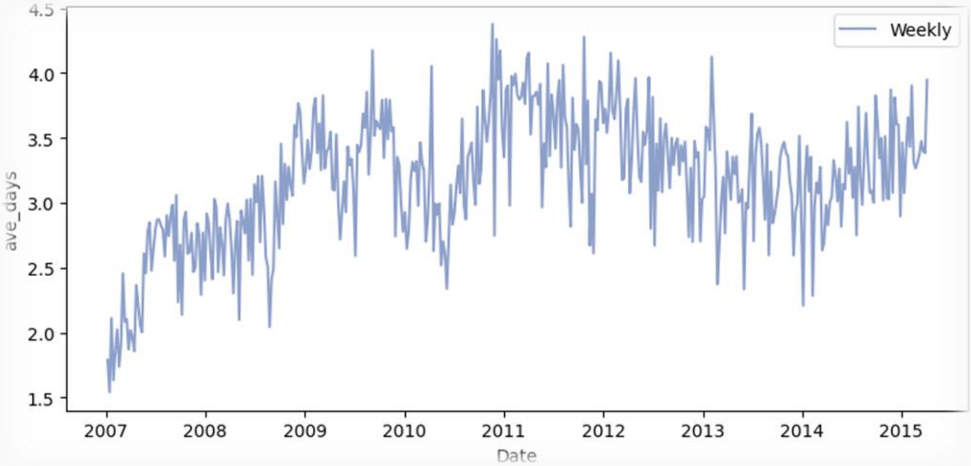
Two tests were applied, ADF and KPSS:

- The ADF test obtained a p-value of 0.0049, rejecting the null hypothesis.
- The KPSS test obtained a p-value of 0.01, accepting the null hypothesis

Presence of weekly seasonality.



Transformation



- To reduce the variance of the time series and improve outlier handling.
- The ADF and KPSS tests showed identical values to those obtained with the original time series.
- The logarithmic transformation did not resolve the issue of seasonality.
- There are no longer any significant spikes, and the overall variability of the series has been reduced.

03 Models

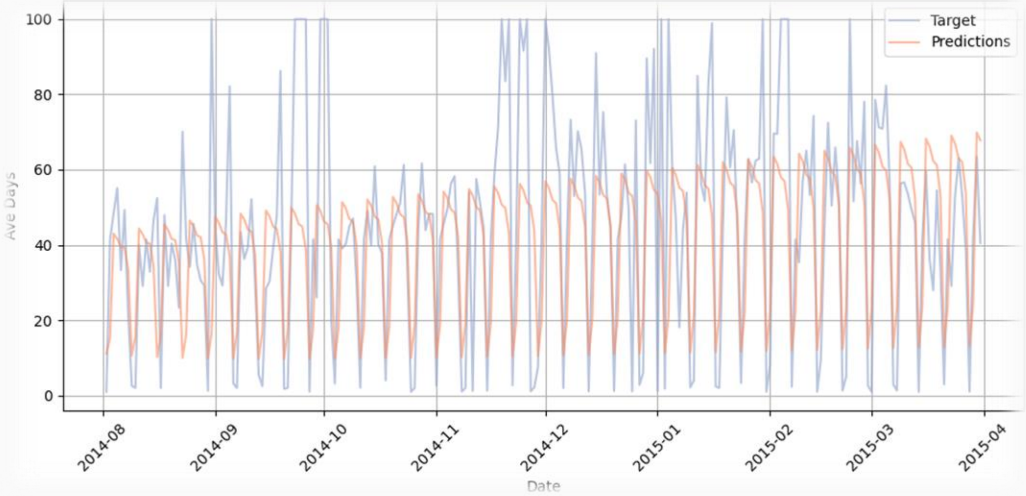
SARIMA, UCM, ML models and optimization with optimal parameter search

SARIMA

To optimize the model's performance using a grid search technique. All possible combinations of these parameters ranging from 0 to 3.

- SARIMA(1,1,0)(1,1,2)[7]
- MAE: 16.97

Tends to rise correctly but has difficulty predicting extreme values.

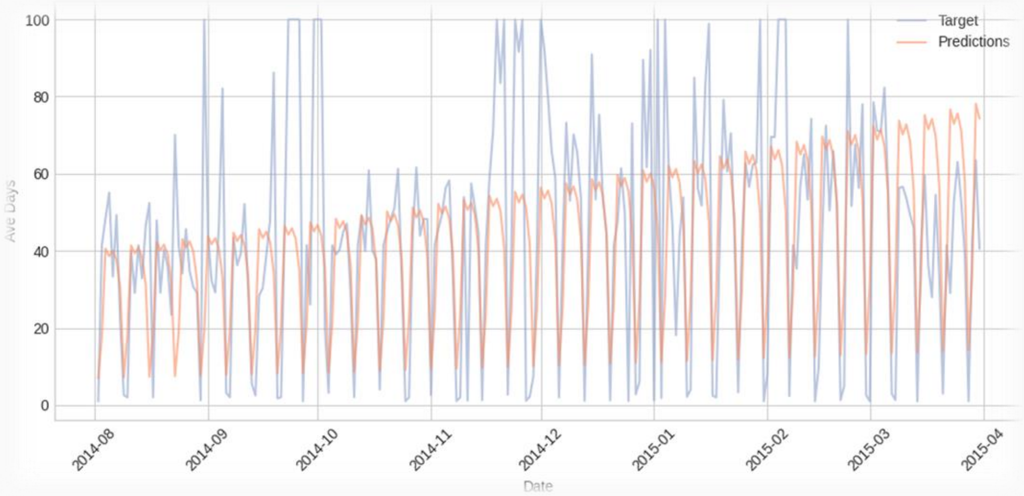


UCM

The combination of all components is performed: level, trend (deciding whether to use it or not), seasonal component (7), and stochastic properties, including level, trend, and seasonality.

- Smooth trend component and 7 sinusoids
- MAE: 16.94

The obtained values are higher compared to the forecasts of the SARIMA model.

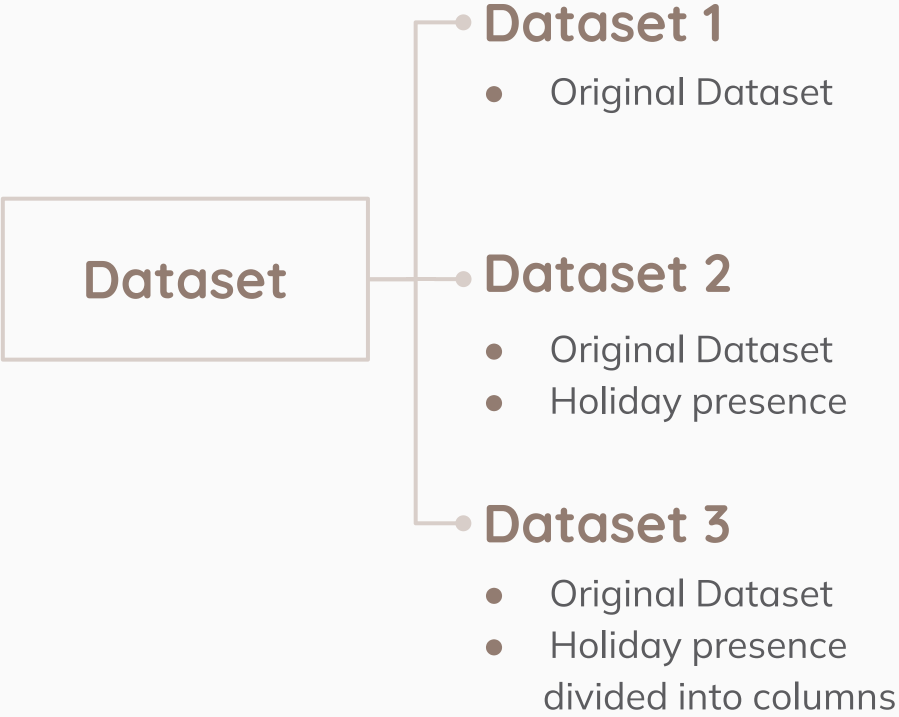


More information

Data

Additional information, including Italian holidays.

Holidays:
New Year’s Day, Epiphany, Liberation Day, Labour Day, Republic Day, mid-August, All Saints’ Day, Immaculate, Christmas Day, and St. Stephen’s Day.



Dataset

Dataset 1

- Original Dataset

Dataset 2

- Original Dataset
- Holiday presence

Dataset 3

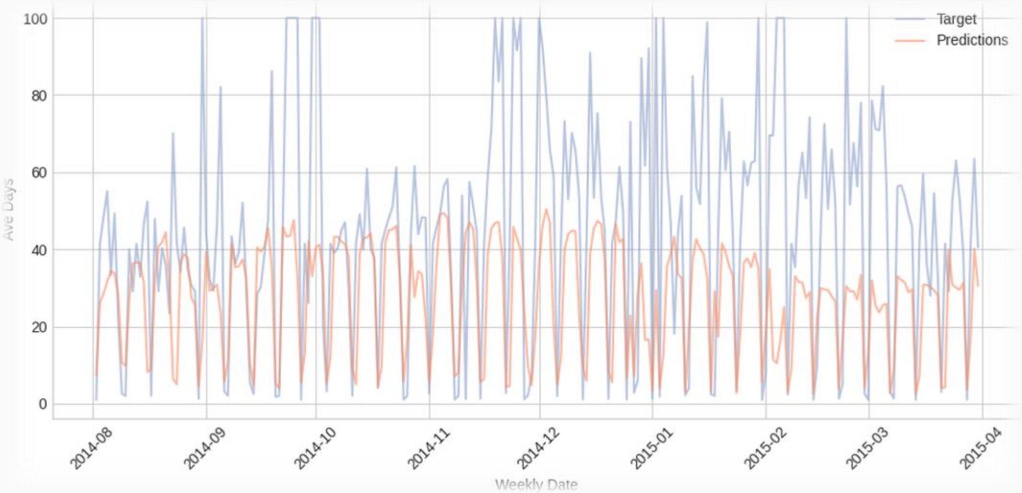
- Original Dataset
- Holiday presence divided into columns

ML

To optimize the model's performance using a grid search technique.

- Xgboost
- **Random Forest**
- SVR

	Dataset1	Dataset2	Dataset3
Xgboost	22.44	21.80	22.13
RF	21.97	21.94	**21.78**
SVR	22.85	22.74	22.80



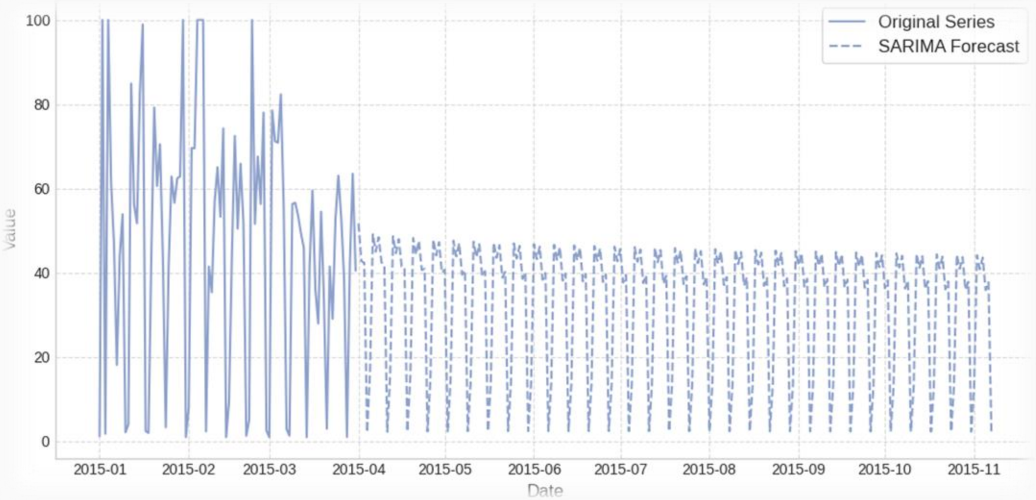
04 Forecasting

Forecasts from 2015-04-01 to 2015-11-07

SARIMA Forecast

Best model of SARIMA

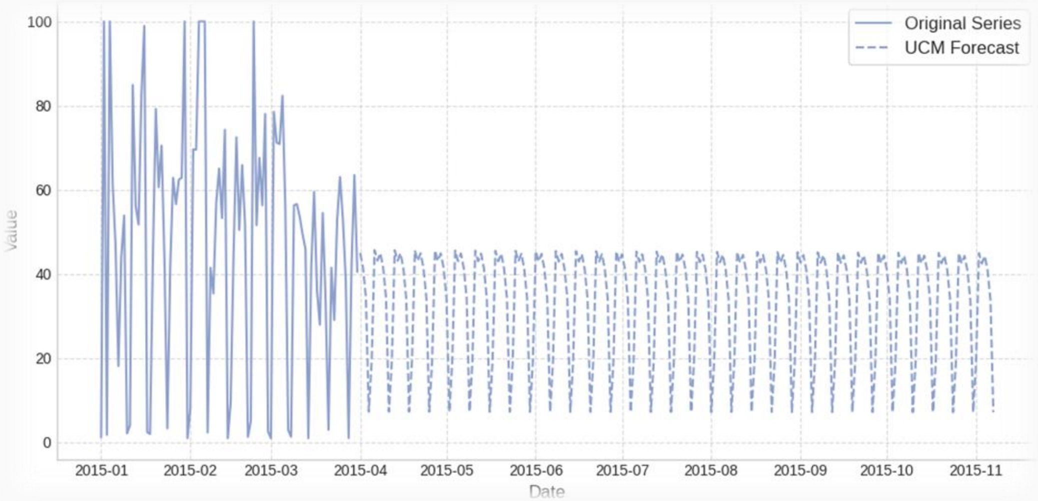
Trend: Decrease



UCM Forecast

Best model of UCM

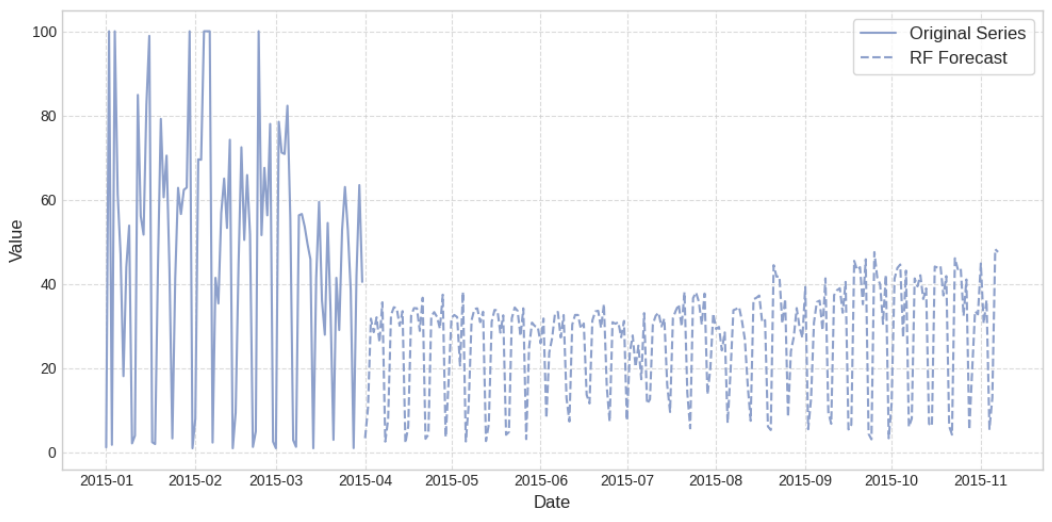
Trend: Stable



RF Forecast

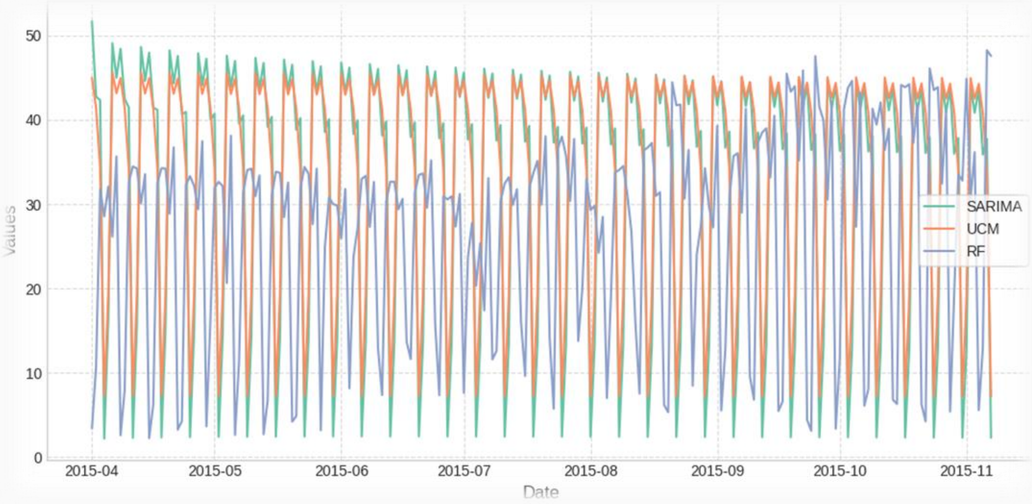
Best model of RF

Trend: Stable and slight increase



Forecast

Best model for each family



Thanks!

Do you have any questions?