Big Data Analytics

Group-C

Group C

1. Dhairya Kathpalia[2021H1540832P]
2. Himani Singh[2021H1540822P]
3. Naman Dhameja[2021H1540829P]
4. Reshma Gadde[2021H1540820P]
5. Vishva Bhalodiya[2021H1540833P]
6. Vaibhav[2021H1540846P]

Content:

- Introduction
- Data Engineering
- Approaches we followed
- Final Model
- Visualization
- Conclusion

Introduction

We are using household power consumption Dataset which consists of measurements gathered in a house located in Paris, France between December 2006 and November 2010 (47 months).

The documentation identifies the following variables of interest:

date: Date in format dd/mm/yyyy

time: time in format hh:mm:ss

global_active_power: The total active power consumed by the household (kilowatts).

global_reactive_power: The total reactive power consumed by the household (kilowatts).

voltage: Average voltage (volts).

global_intensity: Average current intensity (amps).

sub_metering_1: Active energy for kitchen (watt-hours of active energy).

sub_metering_2: Active energy for laundry (watt-hours of active energy).

sub_metering_3: Active energy for climate control systems (watt-hours of active energy).

Data Engineering

1.Load the data using Spark API format as a pyspark.sql.dataframe.DataFrame
2.Splitting the Date and time as separate columns for date as day,month,year and for time hour,minute
3.Casting the datatypes in appropriate format
3.Handling missing values
4.Data transformation using Vector Assembler
5.Standard Scaling of data for each feature or variable to unit variance

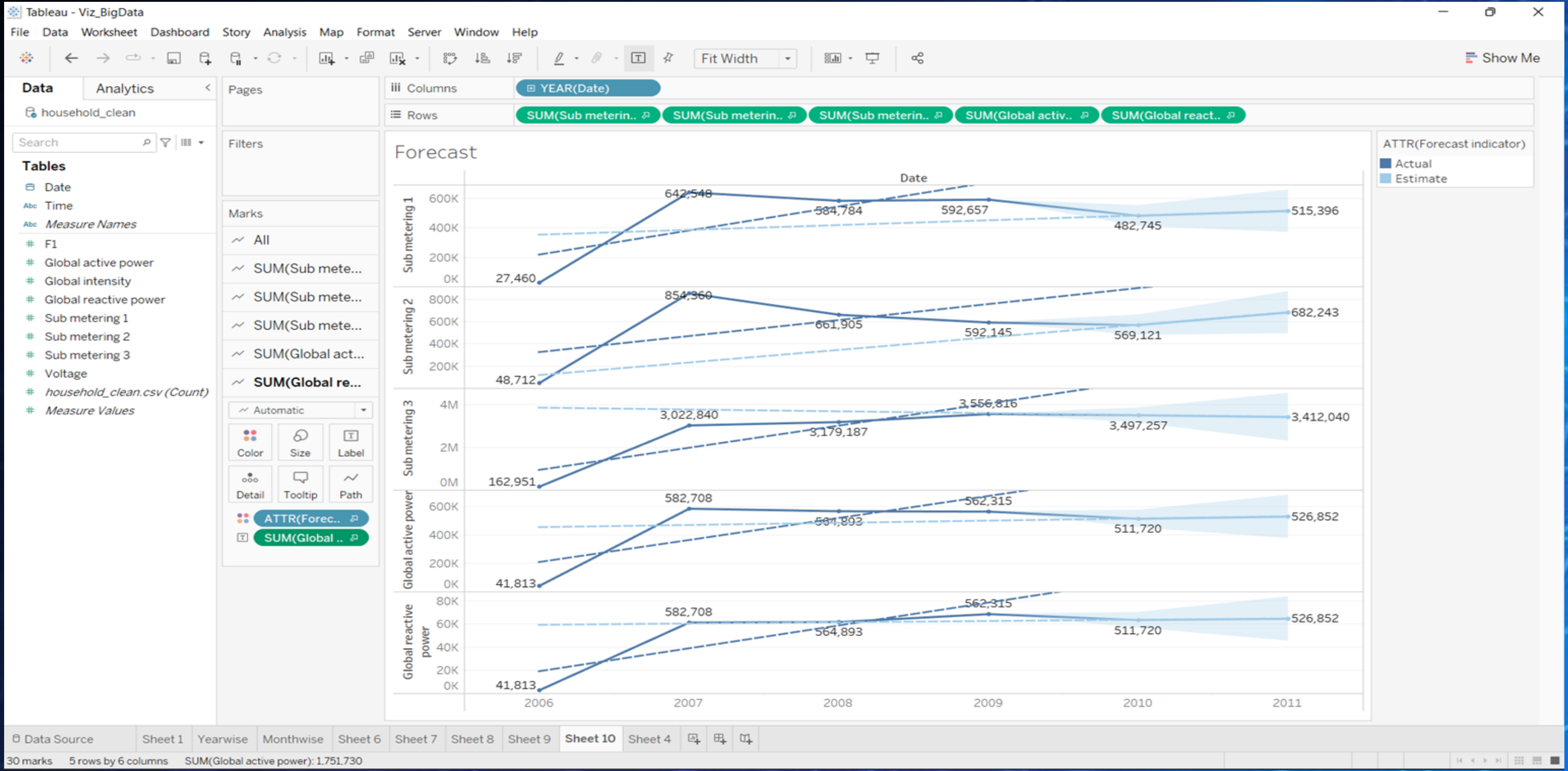
Approaches we Followed

1.Initially as the data is time series data we approach to apply 3,5,7,9 period moving average as the data is highly volatile we switched to regressor methods

2.We have used four kind of Regression methods

- Random Forest Regressor
- Linear regression
- Decision tree Regression
- GBT

Visualizations



Columns: YEAR(Date)

Rows: SUM(Sub meterin.. SUM(Sub meterin.. SUM(Sub meterin.. SUM(Global activ.. SUM(Global react..

30 marks 5 rows by 6 columns SUM(Global active power): 1,751,730

Conclusion:

Out of which Random Forest Regressor is best

Model	RMSE
Linear Regression	0.22
Random Forest Regressor	0.040
Decision Tree Regressor	0.12
GBT Regressor	0.13

Thank You