London Weather Prediction Using Apache Spark

COURSE:

CSYE 7200: BIG-DATA ENGINEERING USING SCALA

PROFESSOR:

ROBIN HILLYARD

TEAM 9

RENTENG HUANG, SHUANGSHUANG XU, BALAJI MUDALIYAR

Goals of the project

- To predict the future weather condition of London city(1 week).
- To develop Apache - Spark Scala code to clean, train, model the data.
- To use Apache - Spark Scala MLib(Machine Learning Library) to predict the weather.
- We will be implementing the UI using play framework.
- Collaborative learning and knowledge sharing.
- Delivering each milestone on time

Use cases

Model input

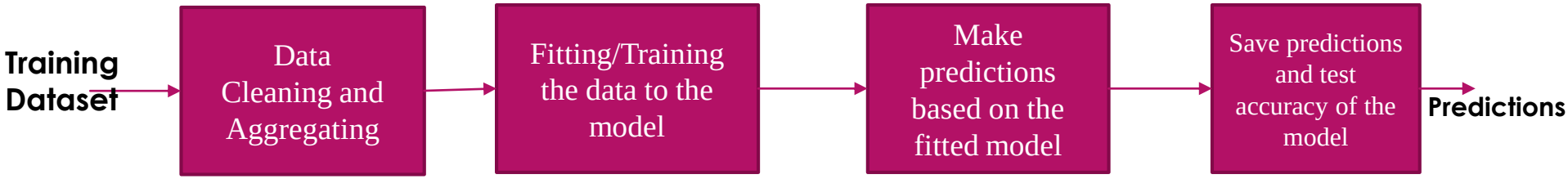
- Date

Model Output

- Weather prediction (Sunny, Cloudy, Rainy, Foggy, Clear)
- Temperature, Pressure

Methodology

- The cleaning of the training dataset
- Fitting of the data to the model
- Making predictions based on fitted model
- Calculating accuracy of the model
- Save predictions



Data source

- **Kaggle : (https://www.kaggle.com/jeanmidev/smart-meters-in-london/data)**
- **21000 record dataset containing hourly weather information of London city**

visibility	windBearing	temperature	time	dewPoint	pressure	apparentTemperature	windSpeed	precipType	icon	humidity	summary
5.97	104	10.24	2011-11-11 00:00:00	8.86	1016.76	10.24	2.77	rain	partly-cloudy-night	0.91	Partly Cloudy
4.88	99	9.76	2011-11-11 01:00:00	8.83	1016.63	8.24	2.95	rain	partly-cloudy-night	0.94	Partly Cloudy
3.7	98	9.46	2011-11-11 02:00:00	8.79	1016.36	7.76	3.17	rain	partly-cloudy-night	0.96	Partly Cloudy
3.12	99	9.23	2011-11-11 03:00:00	8.63	1016.28	7.44	3.25	rain	fog	0.96	Foggy
1.85	111	9.26	2011-11-11 04:00:00	9.21	1015.98	7.24	3.7	rain	fog	1.0	Foggy
1.96	115	9.33	2011-11-11 05:00:00	8.87	1015.91	7.19	3.97	rain	fog	0.97	Foggy
1.3	118	9.31	2011-11-11 06:00:00	8.82	1015.7	7.1	4.1	rain	fog	0.97	Foggy
1.22	114	8.85	2011-11-11 07:00:00	8.69	1016.08	6.48	4.23	rain	fog	0.99	Foggy
1.4	120	9.13	2011-11-11 08:00:00	8.75	1016.33	6.84	4.2	rain	fog	0.97	Foggy
1.38	121	9.23	2011-11-11 09:00:00	8.7	1016.57	7.07	3.96	rain	fog	0.97	Foggy
1.35	115	9.21	2011-11-11 10:00:00	8.76	1016.26	6.96	4.16	rain	fog	0.97	Foggy
1.72	127	9.78	2011-11-11 11:00:00	9.23	1016.17	7.68	4.14	rain	fog	0.96	Foggy
1.83	129	9.91	2011-11-11 12:00:00	9.34	1015.92	7.91	3.97	rain	fog	0.96	Foggy
2.53	120	10.22	2011-11-11 13:00:00	9.73	1015.49	10.22	4.25	rain	fog	0.97	Foggy
3.67	122	10.63	2011-11-11 14:00:00	9.73	1015.1	10.63	4.28	rain	partly-cloudy-day	0.94	Mostly Cloudy
3.77	126	10.34	2011-11-11 15:00:00	9.7	1015.32	10.34	4.3	rain	partly-cloudy-day	0.96	Mostly Cloudy
3.99	121	10.31	2011-11-11 16:00:00	9.66	1015.13	10.31	4.8	rain	partly-cloudy-day	0.96	Mostly Cloudy
4.02	129	10.39	2011-11-11 17:00:00	9.73	1015.41	10.39	4.13	rain	partly-cloudy-night	0.96	Mostly Cloudy
4.17	130	10.79	2011-11-11 18:00:00	9.79	1015.62	10.79	4.41	rain	partly-cloudy-night	0.94	Mostly Cloudy

PROJECT DETAILS

DATE	MILESTONE
15-Mar	Project Start Planning & Data work
23-Mar	Data finding & Cleaning(complete)
24-Mar	Spark Self-Learning, Mlib study
1-Apr	Coding
5-Apr	Implementation
9-Apr	Testing
13-Apr	Final Presentation & Documentation
15-Apr	Project End

Milestones/sprints

Programming in Scala and code repository

- **Most part of the project will be programmed in Scala including**
  - Cleaning
  - Splitting(Training& Testing)
  - Fitting/Training Data to Model
  - Predictions
  - Accuracy Calculation
- **Code repository : GitHub**

https://github.com/001239511ShuangShuangXu/csye7200-spring2018-group9

Acceptance criteria

- **The accuracy of the model predicting weather will be correct 4/7(day/day).**
- **Target a Root Mean Square Percentage Error (RMSPE) of 0.40**

End