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

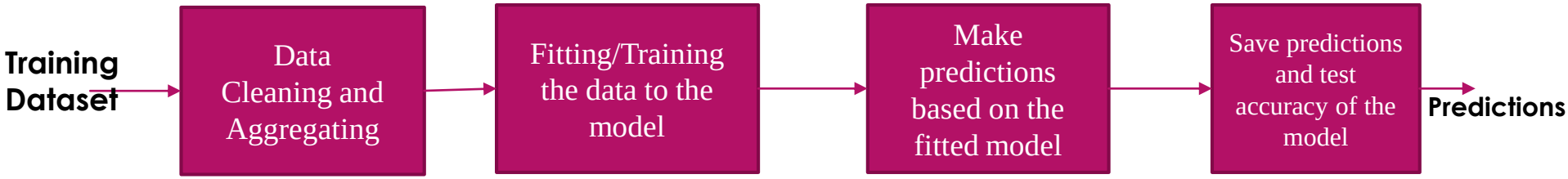
- Date

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

- **Weather Underground : (http://api.wunderground.com/api/API_KEY/history_DATA/q/UK/London.json)**
- **3700 record dataset containing weather information of London city**

date	meantem	maxtempr	mintempr	meanpres	maxpressu	minpressu	meanhum	maxhumic	minhumid	rain
1/1/2010	1	3	-1	1005.76	1010	1002	83	93	70	0
1/2/2010	2	5	-1	1014.83	1019	1010	81	93	70	1
1/3/2010	0	2	-2	1022.69	1024	1019	89	93	81	1
1/4/2010	-1	2	-4	1018.05	1023	1012	87	100	70	0
1/5/2010	-1	2	-4	1004.19	1011	1000	86	100	65	0
1/6/2010	0	2	-3	1003.11	1007	1000	91	100	75	0
1/7/2010	-1	1	-3	1009.64	1016	1007	89	93	86	0
1/8/2010	-2	0	-4	1021.53	1024	1017	94	100	87	0
1/9/2010	-1	1	-3	1021.45	1023	1018	90	100	75	0
1/10/2010	1	1	1	1017.12	1018	1017	93	93	93	1
1/11/2010	1	2	0	1018.79	1019	1018	92	100	87	1
1/12/2010	2	2	1	1009.15	1019	1002	86	93	75	0
1/13/2010	0	1	0	1000.69	1002	999	88	93	81	1
1/14/2010	2	4	0	1004.97	1013	1001	97	100	93	1
1/15/2010	4	6	2	1017.97	1020	1013	93	100	87	0

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