David P. Stonehouse

University of Maryland University College

DATA 650 – Big Data Analytics

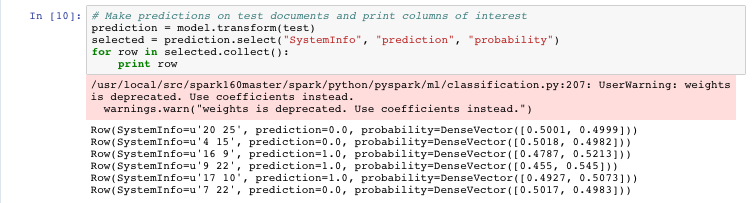
Fall 2016, Section 9040, Professors Gortcheva and Woo

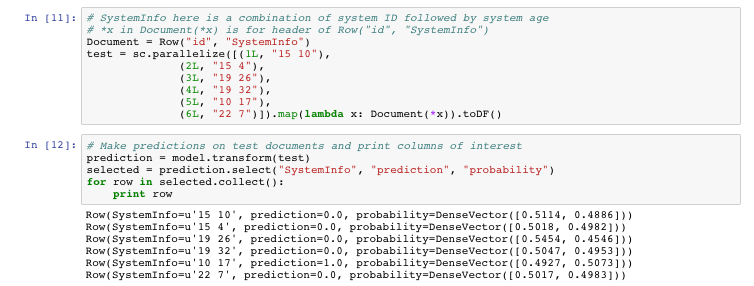
Assignment 4: Spark Analysis of HVAC systems

## HVAC Model Results

The analysis for this question is based on heating, ventilation and air conditioning (HVAC) data. Details of the data are at Mehrotra (2016). The variables in the data are Date, Time, TargetTemp, ActualTemp, System, SystemAge, and BuildingID. The goal of analyzing the data is to develop a model

Using the IBM Bluemix Spark application, data was read in.





Logistic regression is suitable for this case. Per Statistics Solutions (2016), logistic regression assumes a number of things. One is that, for binary logistic regression, the target variable is binary. It assumes that P(y=1) is the probability of the event happening. Only meaningful variables should be included and all meaningful variables should be included. Also, logistic regression requires a large number of samples relative to the number of independent variables. The minimum is 10 cases per independent variable, but often the recommendation is more like 30. The data in this case meets all of these criteria. There are plenty of records (8000) relative to the number of independent variables (2). The target variable is binary, and does represent the label of 1 to mean the event is happening.

An alternative algorithm appropriate for this case is support vector machines (SVMs).

## In-memory processing

.

## Spark, HDFS and Object Storage

Difference between HDFS and Object Storage (mention network latency) - there is an option for no replciations.

Relationships between Spark and MapReduce – talk about shuffle

Spark and HDFS (Tachyon)

## Programming Languages for Spark

Convenience in coding

Code size

Performance

.

## References

Chambers, B., 2015, “Spark MLLib - Predict Store Sales with ML Pipelines”, retrieved from <http://www.sparktutorials.net/Spark+MLLib+-+Predict+Store+Sales+with+ML+Pipelines>

Ganesan, K., 2014. “All About Stop Words for Text Mining and Information Retrieval”. Retrieved from <http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html>

Mehrotra, N., 2016. “Use Apache Spark to build machine learning applications on HDInsight”, retrieved from <https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-ipython-notebook-machine-learning/>

Statistics Solutions, 2016, “Assumptions of Logistic Regression”, retrieved from <http://www.statisticssolutions.com/assumptions-of-logistic-regression/>

Zotti, R., 2013, “How does a Support Vector Machine (SVM) work?” retrieved from <http://stats.stackexchange.com/questions/23391/how-does-a-support-vector-machine-svm-work>

## Appendix: Supporting Information

**Selected logs from R code:**

**R source code:**

Now try with tf-idf

kfit <- DoKMeans(m.tf.idf.transpose2,8)