311-Service-Request Data Analysis using Apache Spark - SOEN 691 Project

Apoorv Semwal Hareesh Kavumkulath Loveshant Grewal

Abstract

Recent advances in the field of Big Data Analytics and Machine Learning have introduced a plethora of open-source tools and technologies for both, academia and the growing data analyst community. In this project we try leveraging one such popular distributed data processing framework Apache Spark to analyse 311 - Service Request Data for the city of New-York. Being updated almost on a daily basis for the last 10 years, massive size of this dataset makes it a suitable candidate for analysis using a distributed data processing framework like Spark. Making use of Spark Ecosystem libraries like Spark SQL and Spark ML, on this dataset, enables us to derive some interesting insights, which might drive better resource planning within the city. Identifying the 3 primary goals for this project we first try answering a few statistical questions like "most frequent complaints reported(across entire city/borough wise)", "Average time to resolve the request (category/department wise)", "mostly used source for making request(borough wise)" and "most busy days/months in terms of request volumes". Arriving at these statistical figures involve making extensive use of Spark SQLs Dataframe API. Secondly we generate a model for predicting the closure time for any new incoming service request, after comparing performance of a set of selected supervised learning algorithms available in Spark ML. As part of our last goal we would be applying K-Means clustering over a selected set of features dividing the dataset into clusters to further analyse them for identifying any underlying service request patterns.

1. Introduction

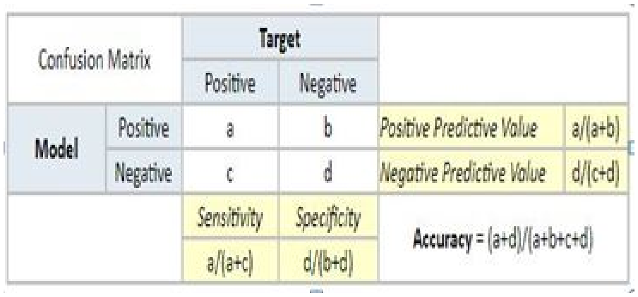
2. Section 2

1. Sample 1: Its some random text:

- Sample Item 1 ***Sample Italics*** in the dataset.

- Sample Item 1 ***Italics***,

2. Sample 2:



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	Specificity	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Figure 1. Sample Confusion Matrix.[1]

Sample Table	
Sample1	Sample_1, Sample_2, Sample_3_1
Sample1	Sample_1, Sample_2, Sample_3_1
Sample1	Sample_1, Sample_2, Sample_3_1

Sample Reference of a figure in text. Figure 1.

Multicol Table				
ID	Sample2	Sample3	Sample4	Sample5
1	74.5	64.5	63.5	66.0
2	78.41	81.61	81.81	82.95

$$\frac{\partial SampleEquation_{y_i}}{\partial X} \tag{1}$$

3. Conclusions

References

[1] TAVISH SRIVASTAVA. Analytics vidhya. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/, Aug 2019. 1