# Machine Learning for Customer churn prediction using big data analytics

YATIN SHARMA, DIKSHA YADAV[1,*]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: yatins@indiana.edu, yadavd@iu.edu

project-001, April 30, 2017

**This project involves use of machine learning algorithms to identify customers who are most likely to discontinue using the service or product.**

© 2017 https://creativecommons.org/licenses/. The authors verify that the text is not plagiarized.

**Keywords:** Prediction, Bigdata, Apache Spark, MLlib, Hadoop, Analytics

https://github.com/cloudmesh/sp17-i524/tree/master/project/S17-IR-P016/report/report.pdf

## CONTENTS



## 1. INTRODUCTION

We will use Apache Spark[1] machine learning library for fitting a predictive model on a massive dataset. Detailed analysis and modeling will be carried out in Python Programming language.

## 2. EXECUTION SUMMARY

The tentative schedule for this project has been outlined below:

1. March 13-March 19, 2017: Create virtual machines on Chameleon, FutureSystems and Jetstream clouds
2. March 13-March 19, 2017: Deploy Hadoop cluster to the clouds and install the required software packages to the clusters and also finalize data.
3. March 20-March 26, 2017: Data Preprocessing and applying transformation to extract features from the data.
4. March 27-April 09, 2017: Use MLlib to train and evaluate various machine learning algorithms and choose best based on various performance metrics.
5. April 10 - April 16, 2017: Create deployable software packages in Python.
6. April 17-April 23, 2017: Complete Project Report.

## 3. WORKFLOW

The project will make use of the following four components.

1. Apache Spark
2. Hadoop
3. Spark MLlib

## 4. DEPLOYMENT

We will deploy our application using Ansible[2] playbook. Deployment of Master/slave nodes will be done hadoop/spark distributed cluster environment. Different cloud systems that will be used in the project include Chameleon,FutureSystems and JetStream.

## 5. BENCHMARKING

Performance of the Hadoop/Spark clusters deployed on different clouds will compared for benchmarking.

## 6. CONCLUSION

TBD

## 7. ACKNOWLEDGEMENT

TBD

## REFERENCES

[1] A. S. Foundation, “Overview - spark 2.1.0 documentation,” Web Page, accessed: 03-12-2017. [Online]. Available: http://spark.apache.org/docs/latest/index.html

[2] “Ansible Documentation,” Web Page. [Online]. Available: http://docs.ansible.com/ansible/index.html