Sarthak Kothari

📱 (+1) 617-840-3450 | ✉ kothari.sar@husky.neu.edu | 🏠 sarthakkothari.me | 🐙 sarthakkothari | 💼 sarthakkothari | Boston, MA | Availability: Dec. 2019

Education

Northeastern University *Boston, MA*
MS IN DATA SCIENCE *Sep. 2017 - Exp. Dec. 2019*
- Relevant Courses: Supervised Machine Learning, Unsupervised Machine Learning and Data Mining, Large-Scale Parallel Data Processing, Causality in Machine Learning, Algorithms

K. J. Somaiya College of Engineering *Mumbai, India*
BACHELOR OF ENGINEERING IN COMPUTER SCIENCE *Jun. 2016*

Technical Knowledge

Specialties	Predictive Modeling, Natural Language Processing, Deep Learning, Distributed Computing, Data Visualization.
Programming Languages	Python (Numpy, Pandas, Scikit, SciPy, Tensorflow, etc.), R, Java, etc.
Databases	Hadoop, Spark, Neo4j, MongoDB, SQL Server, MySQL, Postgres, Oracle.
Data Visualization	Tableau, PowerBI, D3.js, SSRS, Telerik, Excel, etc.
Other Skills	AWS, Linux, Jupyter Notebook, RStudio, Eclipse, Git

Work Experience

Center for Complex Network Research - Northeastern University *Boston, MA*
RESEARCHER *Jun. 2019 - Present*
- Developing high dimensional graph embedding methodologies using metapath2vec in python to quantify the success and impact in scientific careers of researchers.
- Developing models to use these embedding for predicting success, recommendation venues and identifying the author of the papers.

Staples Inc. *Framingham, MA*
DATA SCIENCE INTERN *Jan. 2019 - Present*
- Forecasting the demands for 550K+ Staples SKU's for the upcoming week by developing an ensemble of models which offered a 23% improvement to the existing model.
- Visualizing the delivery footprint for Staples operations to recognize key areas of interest using Python and D3.js.
- Building data pipelines to process more than 200K+ weekly vendor records using Python and presented related visual stories to executive management.

Fidelis Cybersecurity *Bethesda, MD*
DATA SCIENCE INTERN *Jul. 2018 - Aug. 2018*
- Migrated millions of records to Neo4j, a graph database, from MongoDB and Spark using Python in order to increase efficiency and data retrieval time for complex queries by 40%.
- Executed clustering analysis on the malware alert data harnessing the intelligence obtained by analyzing relationships between nodes in the graph database to explain how different malware files are associated with each other using Python.

Hansa Cequity *Mumbai, IN*
ASSOCIATE ANALYST *Aug. 2016 - Jun. 2017*
- Modeled a logistic regression for lead scoring which would rate a lead's likelihood of becoming a customer for a client with accuracy of 90%.
- Developed efficient data ingestion pipelines using SSIS for 100K+ records and build descriptive dashboards in SSRS enabling clients to evaluate product and campaign performance, boosting sales by 15%.

Projects

Recommender Bias, Northeastern University, Boston MA *May. 2019 - Present*
- Using different classes of Bandit Algorithms to model and solve the problem of recommender bias under the guidance Prof. Baeza-Yates.
- Testing these models on various datasets like MovieLens, Coupon Purchase Prediction, etc.

Matrix Multiplication in Distributed Environment, Northeastern University, Boston, MA *Nov 2018 – Dec 2018*
- Performed comparative study between Cannon's vs Simple Block Partitioning for Distributed Matrix Multiplication algorithm.
- Implemented the algorithms using MapReduce benchmarked them on AWS EMR cluster using matrices of different sizes.

Predicting Student's Portuguese Grade, Northeastern University, Boston, MA *Oct 2018 – Dec 2018*
- Predicted student's final grade for Portuguese language and Math course using demographical, socio-cultural and academic features.
- Built an ensemble model which first predicts the midterm performance to then predict the final grade using Python.

Twitter Follower 2-Hop Neighbors, Northeastern University, Boston, MA *Sept 2018*
- Implemented replicated and reduce side joins in MapReduce as well as Spark to find 2-hop followers for the user.
- Executed the algorithms for 11M+ users in the Twitter Edges dataset using AWS EMR cluster.