

# Raka Dalal

ASPIRING DATA SCIENTIST/ MACHINE LEARNING ENGINEER

☎ (443)515-7912 | ✉ rak2@umbc.edu | 🏠 rakadalal1409.wordpress.com | 📷 RakaDalal | 🌐 rakadalal

## Education

### University of Maryland Baltimore County- Baltimore, MD

Aug 2016 - Aug. 2018

MASTERS IN COMPUTER SCIENCE

GPA: 4.0/4.0

### Jadavpur University - Kolkata, India

Aug 2012 - May 2016

BACHELOR OF ENGINEERING IN COMPUTER SCIENCE AND ENGINEERING

GPA: 8.16/10.0

### Online Certified Courses (UDEMY)

Data Science A-Z, Deep Learning A-Z, Hive for Processing Big Data, Complete & Practical SAS, AWS Concepts, AWS Essentials, Apache Spark

## Technical Skills

<b>Languages</b>	Python(Most Experienced), C, C++, C#, Java, MATLAB, Octave
<b>Tools/Frameworks</b>	PyTorch, Keras, TensorFlow, Scikit-learn, NumPy, Pandas, NLTK, Greta, <b>ETL Tools:</b> SAS, SSIS
<b>Big Data Frameworks</b>	Hadoop, Hive, <b>Apache Spark:</b> RDD, DataFrames, SparkSQL, MLlib, Spark Streaming, GraphX
<b>Data Visualization</b>	Tableau, Microsoft Excel, <b>Web Visualization:</b> Flask
<b>Database Concepts</b>	SQL, PL/SQL, PostgreSQL
<b>AWS Services</b>	IAM, VPC, S3, EC2, RDS, SNS, CloudWatch, ELB, Auto Scaling, Route 53, Lambda

## Experience

### UMBC

Baltimore, MD

GRADUATE RESEARCH ASSISTANT

Jan 2017 - May 2018

- Created an end-to-end pipeline that uses a semi-supervised bootstrap learning model to extract different relations from large-scale cybersecurity text dataset with limited training samples and populate a knowledge graph. We evaluated our model on the CVE dataset and achieved high accuracy. [\[Github\]](#)
- Created an end-to-end pipeline to automate the process of extracting semantic relation from sensor input by formalizing knowledge as digital twin models coming from sensors in industrial production lines and by introducing a semantic query mechanism. We published this work at IKG 2017. [\[Publication\]](#)

### General Electric

GRC, Niskayuna, NY

RESEARCH FELLOW INTERN

May 2017 - Aug 2017

- I built a semi-supervised bootstrap learning based approach to extract relations from huge unstructured text dataset with an iterative client feedback loop. Evaluations over diverse datasets, including aircraft engine maintenance records and a Google relation extraction corpus, showed promising results. I also designed a user Interface using Python Flask to showcase my work and facilitate users to give feedback.

### Samsung Research & Development

Bangalore, India

SUMMER RESEARCH INTERN

May 2015 - July 2015

- Detected the level of stress of user with high accuracy using raw accelerometer and gyroscope data. Filtering and Principal Component Analysis were applied on the data to get the resultant signal. Fast Fourier Transform of this final data gave the heart rate which was used to infer the stress level.

### IIT Kharagpur

Kharagpur, India

SUMMER RESEARCH INTERN

May 2014 - July 2014

- The project involved analysis of GPS data of trucks to detect hotspots based on stopping time. We clustered the data using Density-Based Spatial Clustering (DBSCAN) and subsequently typecast the data with respect to stopping time distribution, busyness distribution and finally visualized them on Indian roads.

## Major Projects

### Auto-Completion System [\[Github\]](#)[\[Blog Post\]](#)

Aug 2018 - Sept 2018

- The project involves building an auto-completion system for customer service representatives by suggesting sentence completions. For this I developed the trigram Katz Backoff model from end to end and used it for prediction. Finally, this system was wrapped in a local HTTP server using Flask.

### Analysis of Insurance Data for Safe Driver Prediction [\[Github\]](#) [\[Kernel\]](#)

Aug 2018 - Sept 2018

- Analyzed a large Insurance Company dataset in Kaggle comprising of 595212 datapoints and 59 variables. I handled imbalanced data, missing values, performed extensive feature engineering and finally trained an Artificial Neural Network to predict if a driver will file an insurance claim.

### House Price Prediction [\[Github\]](#) [\[Kernel\]](#)

Aug 2018 - Sept 2018

- I worked on the Ames Housing dataset which consisted of 79 explanatory variables describing every aspect of residential homes in Ames, Iowa. The training data has 1461 datapoints of 81 variables. After extensive feature engineering I predicted the final price of each home by using XGB Regressor.

### Prediction of academic references for Wikipedia articles [\[Github\]](#)

Jan 2015 - Aug 2016

- Crawled wikipedia for academic CS articles starting from Wikipedia Books to gather interesting statistics about the references in wikipedia articles till 2012. The data was cleaned, analyzed and finally, a statistical model was built using N-grams to predict references that will be added in the future.

### Intelligent Cybersecurity Recommender (UMBC)

Feb 2017 - Apr 2017

- Designed an effective recommender system that analyzes the probable vulnerabilities and recommends the least vulnerable products. It works on question-answer(Q/A) based model where the system admin is expected to ask product vulnerability questions to which the system suggests a right option.

### Detecting Stance in Tweets (UMBC)

Feb 2017 - Apr 2017

- Reproduced the work by Saif M. Mohammad et. al. in the paper SemEval-2016 Task 6: Detecting Stance in Tweets. Designed a system that for any given tweet and target entity, determines whether the tweeter is in favor of the given target, against the given target or neutral to it.