SOUVIK DUTTA

sdutta9@illinois.edu, 217-735-0789

1814 12th Ave S, Seattle, WA 98144, USA GitHub:/TheKivs LinkedIn:/souvik-dutta

- Ph.D. candidate with 4+ years experience in Statistical Analysis and Machine Learning with BigData (AWS Certified)
- 4+ years experience with Machine Learning libraries on **Python** like **Scikit-learn**, **Pandas**, **Numpy**, **Matplotlib**, **XGBoost**, **NLTK**, **TensorFlow**, **Keras**, **PyTorch**

EDUCATION

Ph.D. candidate, University of Illinois Urbana-Champaign, USA **Sept 2020**

- **Field of research**: Machine Learning in Computational Physics **GPA:** 3.92/4.0
- **Relevant Courses:** Machine Learning Theory, Deep Learning, Statistical Learning II, Natural Language Processing

B.Tech., Indian Institute of Technology Bombay, India Jul 2009 - Apr 2013

- **Relevant Courses:** Data Structures & Algorithms, Optimization, Data Analysis & Interpretation, Linear Algebra
- **Topic**: "Data analysis using novel clustering algorithms", received "Undergraduate Research Award"

WORK EXPERIENCE

Graduate Research Fellow, University of Illinois Urbana-Champaign, USA Aug 2015 - present

- Developed **2** novel machine learning algorithms in **Python** for complex classification and discrimination tasks
- Collaborated with Big Data engineers in performing **ETL** on terabyte-scale data from electronic signals using **Spark**
- Devised **3** feature-selection methods based on **Random Forests**, **PCA** and **XGBoost** to reduce input space by **27%**
- Improved **outlier detection** methods with **KNN** (PyOD); automated pipeline to distribute data after batch processing

Data Scientist Intern, Flipkart Online Retail, India May - Jul 2019

- Performed feature selection & trained **Logistic Regression** model for binary classification of subscription lapse (Y/N)
- Used **Random Forests** to identify distinct segments for lapsing customers to enable personalized targeting strategies

Teaching Assistant, University of Illinois Urbana-Champaign, USA Sep 2013 - Jul 2017

- Designed **40** tutorial sessions for Statistical Data Analysis (STAT 542) in **Python** for a class of 90 graduate students
- Instructed **coding sessions** for the course on Statistical Methods; won the Dean's award for **"Outstanding TA"** twice

Data Scientist Intern, Machine Learning Group, CERN, Switzerland May - Aug 2012

- Created a data cleaning framework in **Python** for outlier detection, leading to **14%** faster optimization convergence
- Developed and deployed a novel clustering algorithm for **20%** faster processing of big data from electronic sensors

RECENT PROJECTS

Tumor detection from MRI images, Gardner Neuroscience Institute, Cincinnati OH [*remote*] Jun - Aug 2020

- Led team of 6 interns to develop a multi-class classification model with **OpenCV** to detect tumors at **0.83** F1-score
- Implemented an AlexNet **CNN** using **Keras** to segment images for targeted study of tumor location, size and growth

Generating image captions using ML, Siebel Center for Computer Science, Urbana IL Mar - May 2020

- Trained an attention-based model in **TensorFlow** for generating captions for MS-COCO and ImageNet datasets
- Achieved **7%** higher accuracy (better grammar, less word repetition) at caption generation compared to VAE models

Sentiment Analysis for Recommender Systems, Amobee Innovation Center, Champaign IL Jan - Apr 2020

- Analyzed review sentiments via **TF-IDF** with POS-tags and lemmatisation (using **NLTK**) on client's e-comm website
- Tailored page-ranking algorithm to category display pages using web-traffic data; led to **4%** rise in add-to-cart rate

TECHNICAL SKILLS

Programming: Python, R, C++, SQL, Microsoft Office, MATLAB, SAS, BigQuery (GCP)

Libraries: Scikit-learn, Pandas, NumPy, SciPy, Statsmodels, Matplotlib, PyTorch, Keras, XGBoost, NLTK

Machine Learning: Bayesian classification, Linear & Logistic Regression, KNN, K-means, Decision Trees, SVM, Random Forests, PCA, Recommender Systems, Natural Language Processing, Convolutional Neural Nets

Mathematics: Linear algebra, Probability theory, Multivariate statistics, Optimization