HARSHEEL SINGH SOIN

hss2148@columbia.edu | 500 Riverside Drive, New York, NY 10027 | (646) 726-2969
linkedin.com/in/harsheelsoin | github.com/harsheelsoin

EDUCATION

Columbia University | Data Science Institute | New York, NY **Dec 2018**
M.S. in Data Science, GPA: 3.68/4.00
Relevant Courses: Algorithms, Statistical Inference, Machine Learning, Exploratory Data Visualization, Applied Deep Learning

Indian Institute of Technology Kharagpur | Department of Biotechnology | Kharagpur, IN **May 2017**
B. Tech (Honors) and M. Tech in Biotechnology and Biochemical Engineering, GPA: 8.21/10.00
Micro-Specialization in Biomedical Devices and Instrumentation, GPA: 8.67/10.00
Relevant Courses: Data Structures, Biostatistics, Bioinformatics, Computational Biophysics, Computational Phylogenetics

LANGUAGES AND TECHNOLOGIES

- Python, R, SQL, MATLAB, C++, D3.js, MongoDB, Bash, HTML, LaTeX, Markdown, CSS
- Git, AWS, GCP, Apache Spark, Tableau, Trifacta, Travis CI, OpenCV, MS Office, LabVIEW

EXPERIENCE

AUDIBLE, INC | *Research Assistant – Data Science* | *Newark, NJ*
Predicting First 6-Month Sales of Audiobooks **Sep 2018 – Dec 2018**

- Built XGBoost based regressors and classifiers to predict the first 6-month sales of audiobooks and their sales categories (ratings)
- Determined optimal book release period for modeling by accounting for adjustment of sales over time and effect of release year
- Developed impactful features to represent book publisher and genre in addition to using topic, writing and narration-style features
- Implemented class grouping strategies aligning closer with business interests to improve predictability of highly successful books
- Calibrated models' predicted probabilities and identified optimal probability thresholds leading to huge improvements in f1-scores

Studying Interaction between Topics and Genres; Predicting Content Quality Metrics **May 2018 – Aug 2018**

- Predicted content quality metrics (Listening Velocity and Completion Rate) using tuned XGBoost regressors based on writing-style and topic features, with the goal of identifying high-quality content for marketing, acquisition and recommendation purposes
- Investigated predictive power and direction-specific impact of features on targets via relative importance and partial-dependence
- Built models to predict audiobook genres from topic representations via multi-class and multi-label classification approaches
- Interpreted model coefficients and identified top topics per genre; Derived topic-space based genre representations using tuned DNNs with several business-oriented applications, including computing similarity across genres, topics and audiobooks

GOLDMAN SACHS | *Columbia Data Science Institute – Capstone Project* | *New York, NY*
Assessing Pharmaceutical R&D Similarity through FDA Clinical Trials **Sep 2018 – Dec 2018**

- Inferred similarities across clinical trials using representations extracted from trial metadata, via TF-iDF and Word2Vec models
- Developed manual validation strategy for comparison and benchmarking of unsupervised clinical trial similarity models
- Transformed trial-level similarity matrices to company-level and created interactive interface for ensembling them using Bokeh

DATA SCIENCE PROJECTS

Analyzing New York City Motor Vehicle Collisions **Mar 2018 – Apr 2018**
COLUMBIA UNIVERSITY | Final Project – Exploratory Data Analysis and Visualization | New York, NY

- Examined NYC accident data to answer key questions around the location, time and contributing factors of motor vehicle collisions
- Developed static and interactive visualizations including choropleths, map-based scatter plots and heat-maps using R and D3.js

Predicting Optimal Lineup for Daily Fantasy Football **Sep 2017 – Nov 2017**
COLUMBIA UNIVERISTY | Term Project – Projects for Data Science | New York, NY

- Developed GBM based regression model (R^2: 0.68) to predict player ownership using features derived from web-scraping and Twitter-sentiment; Performed mixed integer linear programming based multi-objective optimization on fantasy points (virtual scores) and projected ownership percentages for players; Packages used: Scikit-Learn, Selenium, PuLP, Tweepy, Beautiful Soup
- Built web dashboard with Python's Flask, featuring Tableau visualizations, to streamline weekly prediction of optimal NFL lineup

Developing Robotic Exoskeleton for Stroke/Amputee Patient Rehabilitation using Electroencephalography **Aug 2016 – May 2017**
IIT KHARAGPUR | Master's Thesis – School of Medical Science and Technology | Kharagpur, IN

- Implemented object recognition with distance and dimension estimation from live camera feed using Python's OpenCV
- Acquired EEG data for motor imagery and actual limb movement, pre-processed for artifact removal and extracted features for EEG classification including Hjorth, power-periodogram and auto-regressive parameters using MATLAB and EEGLab
- Performed classification to label EEG data with associated hand or finger using SVM, kNN and RF classifiers from Scikit-Learn

ACHIEVEMENTS AND LEADERSHIP

- Selected as Board Member, Columbia Data Science Society (2017-18); Organized Hackathon, panel sessions, technical workshops
- Awarded prestigious J. N. Tata Endowment Scholarship (2017) towards Masters in Data Science at Columbia University
- Appointed as Vice President, Association of Biotechnologists (2015-16); Conducted webinars, field trips, in-house competitions
- Spearheaded IIT Kharagpur's first-ever participation in top-tier global synthetic biology competition (iGEM 2015) at MIT, MA