Udip Bohara , Data Scientist

Links : Email | LinkedIn | Portfolio | Github

Technical Skills

Languages: Python/PySpark, R, JavaScript, MySQL, Mongo Query Language, Cypher/GraphQL, Bash, HTML/CSS
Platforms/Frameworks: Flask, Django, Dash, d3.js, jQuery, Spark, Dask, Hadoop, Hive, MongoDB Atlas, Neo4j
Tools: Git, Databricks, Google Cloud Platform, Azure, AWS, Jupyter, SPSS, Weka, Rapidminer, ArcMap, Tableau
Libraries: Pandas, NumPy, Matplotlib, Plotly, Scikit-learn, Statsmodels, PyTorch, spaCy, gensim, Tensorflow, Tidyverse

Experience

Graduate Research Assistant — Aug. 2019 – Present

Department of Computer Science, Mercyhurst University — *Erie, PA*

- Identified clusters (K-means, K-prototype, GMMs) in healthcare survey data to automate patients segmentation.
- Improved runtime of cyber Instrusion/Anomaly Detection Systems with Principal Component Analysis (PCA).
- Modeled cognitive decision-making in Twitter with Natural Language Processing (NLP) and Deep Learning (DL).
- Deployed an interactive GUI application for intra-department students to explore/analyze web-browser history.
- Boosted prospect students enrollment with Regression, Decision Trees, Clustering and Random Forest models.
- **Teaching Assistant**: Conducted lectures and graded exams for CIS-200 Linear Data Structures (70+ students).
- Wrote instructional material for Experimental Design, Hypothesis/AB testing, Linear Algebra in Jupyter.

Student Research Analyst — Feb. 2020 – May 2020

Johns Hopkins Applied Physics Lab — *Remote*

- Worked in John Hopkins program of Forecasting Counterfactuals in Uncontrolled Settings (FOCUS).
- Provided analysis/recommendations in simulation based intelligence/conflict strategies using hypothesis-testing.
- Effectively communicated and improved results by regularly collaborating with teammates and supervisors.

Data Scientist — Jan. 2019 – May 2020

Department of Institutional Effectiveness, Mercyhurst University — *Erie, PA*

- Translated broad business/academic problems to interpretable data-oriented predictive and prescriptive solutions.
- Cleaned and migrated data pipelines (ETL) from Ellucian Colleague to Google Cloud Platform (BigQuery).
- Regularly wrote PySpark and SQL scripts for effective querying, ad-hoc analysis, KPI identification and modeling.
- Facilitated data-driven budgeting by automating financial forecasts with econometric time-series modeling.
- Deployed a Retention/Churn Logistic Regression model using Python and Flask to identify at-risk students.
- Wrangled, identified and visualized novel Grade-Inflation (19000 students, over 500,000 records) using Python.
- Regularly communicated key findings using Matplotlib, Seaborn, Google Data Studio, Plotly, Tableau, etc.
- Maintained consultation with key stakeholders: Provosts and Deans to continuously build intervention models.

Highlighted Projects | Link to More Projects

ArXiv Papers Recommendation System | *PySpark, Neo4j, Gephi, Graphframes, Google Cloud Platform, sigma.js*
Built end-to-end scalable ETL Natural Language Processing (topic modeling and semantic/cosine similarities) search-engine Recommendation System pipeline in Graph Database (Neo4j) with Python for arXiv papers. Github

Electricity Demand Forecasting | *Python, Dash, Keras, Heroku, MongoDB*
Modeled advanced time-series Econometric forecasting models such as SARIMAX and Prophet and Deep learning methods: Dilated-CNN and LSTM to forecast Electricity Demand in all regions of USA. Deployed a Dash Application to display interactive and live results. Github

Information Extraction from Documents | *Python, PyTorch, Tesseract, OpenCV*
Researched and developed scalable end-to-end extraction of information from receipts/text documents using Optical Character Recognition (OCR) and semi-supervised deep learning with Graph Convolutional Networks (GCNs). Github

Education

Mercyhurst University — Erie, PA

Master of Science in Data Science | GPA 4.0/4.0 | Awarded Full Tuition Waiver — *Jan. 2019 – Dec. 2020*

Mercyhurst University — Erie, PA

Bachelor of Science in Biostatistics/Public Health — *Aug. 2013 – Dec 2017*

Relevant Courses

Probability and Statistics, Algorithm Development, Data Structures and Algorithms, Data Wrangling, Relational and Non-Relational Databases, Healthcare Analytics, Big Data Analytics, Machine Learning, Text Mining, Research Methods, Research Project, Data Visualization, Geospatial Data Analytics, Biostatistics I and II, Principles of Epidemiology I and II