Baipeng Gong

New York, NY, 10011 | (646)877-8364 | bg1622@nyu.edu | LinkedIn Profile

EDUCATION

New York University September 2021 – May 2023
Master of Science in Data Science **GPA: 4.0/4.0**

- **Anticipated Coursework:** Introduction to Data Science (A/B Testing), Machine Learning, Big Data, Natural Language Processing (NLP), Deep Learning, Probability and Statistics, Optimization and Linear Algebra, Time Series Analysis

New York University September 2016 – May 2020
Bachelor of Science in Data Science, Finance **GPA: 3.71/4.0**

- **Honors:** Cum Laude, University Honors Scholar, Founders Day Award, Dean's List
- **Coursework:** Data Structures, Database Systems, Information Visualization, Numerical Analysis, Econometrics

SKILLS

Programming Languages: SQL, Python (NumPy, SciPy, Pandas, Matplotlib, Scikit-learn, PyTorch, TensorFlow), MATLAB
Tools: Microsoft Office, Tableau, Git, Hadoop, Spark, HPC, Stata, LaTeX, iMovie

PROFESSIONAL EXPERIENCE

Morgan Stanley **New York, NY**
ESG Data and Analytics Summer Analyst June 2022 – August 2022

- Tried multiple data manipulation tricks, Python NLP libraries (including spaCy and NLTK), and evaluation metrics to match the company names between REITs dataset and Green Building Certification datasets **(over 100K records)**
- Built a language model using CountVectorizer for stop word detection and removal, N-gram for tokenization, and TF-IDF for embedding to find the best match between two lists of company names by comparing cosine similarity
- Established **a new scheme** that applied fuzzy string matching to compare text similarity by fuzzy ratios based on the Levenshtein Distance algorithm, **simplified the model and made it easier to understand for non-technical audiences**
- Leveraged the waterfall methodology to combine the above two schemes, **increased the number of matches by 80% and the matching accuracy to around 90%**
- Proposed using set theory and numerical analysis thinking to adjust the filtering threshold, **improved the flexibility and reliability of filtering than heuristic threshold selection**
- Analyzed the differences between the matching results of 3 Green Building Certification datasets, **obtained new findings and presented them to the Vice President, laid the foundation for the follow-up factor investing**
- Visualized Green Buildings on Google Map and drew choropleth maps using Google APIs and Python libraries such as Bokeh and GeoPandas

Analysys International **Beijing, China**
Data Analytics Intern June 2020 – December 2020

- Quantified business for **10+ clients** from various industries, determined performance metrics for each client
- Deployed web analytics tags for data collection, participated in building pipelines using SQL and Python, **sped up data processing time by 25%**
- Delivered data reports and dashboards with a special focus on Event Trends, Conversion Funnel, and Retention Analytics
- Provided clients with suggestions for improvement, monitored A/B Testing, **achieved significant growth in Daily Active Users, Page Views, Clickthrough Rate, Gross Merchandise Value, etc.** for different clients
- Wrote articles summarizing the optimization strategies for webpages/apps, posted them on company's WeChat Official Account, received **3000+ views** for some articles

RESEARCH & SELECTED PROJECTS

Improving Numerical Reasoning Skills for Financial QA February 2022 – May 2022

- Performed intermediate training with MathQA dataset to improve the automations of financial report analysis, **achieved accuracy 10% higher than general crowd performance**
- Utilized GenBERT and TASE-BERT encoders as drop-in replacement for BERT-base and RoBERTa-large models in the Financial QA architecture, **improved the accuracy by 1.5% than the baseline framework**
- Conducted qualitative analysis and investigated the errors, found that our model performs better for table-only questions and questions that require less than 3 operations

Evaluating COVID-19 Vaccine Hesitancy among People September 2021 – December 2021

- Cleaned **1.5M+ user data** with **300+ features**, performed feature engineering with forward selection and one-hot encoding
- Constructed Logistic Regression model to identify the people who are most likely to be hesitant to get the COVID-19 vaccine, reached a **0.83 AUC score**
- Applied PCA and K-Means to cluster users into 6 groups, analyzed potential reasons people are not receiving vaccination

Baosen Luo

(510)–5705655 | bl3243@nyu.edu | linkedin.com/in/baosenluo

Education

New York University, NY — Sep 2021 – May 2023(Expected)
Master of Science in Data Science — GPA: 4.0 /4.0
University of California, Berkeley, CA — Aug 2017 – May 2021
Bachelor of Arts in Data Science, Economics — GPA: 3.8/4.0

Languages and Skills

- Python, MySQL, Pytorch, Hadoop, Spark, Tableau, Latex, Shell, Git
- Statistical Modeling, Deep Learning, Machine Learning, Visualization
- Convex Optimization, Hypothesis Testing, Time Series Analysis, Big Data,

Related Experiences

Data Scientist Intern | Roblox — San Mateo CA, May 2022 – Aug 2022

- Created a user level metric, Avatar Uniqueness, based on 2d thumbnails using DBSCAN clustering and Resnet-50 embedding; evaluated the metric by survey and correlation analysis; built a data pipeline to monitor the temporal evolution of the metric; proposed actionable product ideas
- Deployed an NLP framework based on Online Latent Dirichlet Allocation to automatically capture trending topics and detect emerging topics from help center tickets; analyzed users' frictions and pain points using the model results; identified a game engine bug and several glitches during the 2-month experimentation

Business Analyst Intern | Tencent Cloud — Palo Alto CA, Apr 2020 – Aug 2020

- Built a pipeline for analyzing and visualizing sales data extracted from Salesforce via Python and Tableau
- Conducted target market analysis and produced market research reports; contributed to Tencent America's 'Go-China' strategy by identifying 500+ potential clients in North America and successfully secured 102 Proof of Concept accounts and 26 Contract accounts

Projects

Movie Recommendation System — March 2022 – May 2022

- Built a collaborative-filtering based movie recommendation system using Apache Spark and high-performance computers; tuned hyper-parameters of Latent Factor Model using cross validation
- Assessed the quality of learned hidden representations of users and movies by visualizing the high dimensional representations in two-dimensional space using UMAP and t-SNE

A Model Comparison of ARIMA and GP in NBA popularity Forecasting — Oct 2021 – Dec 2021

- Performed Box-Jenkins's methodology and cross validation to fit an ARIMA model that captures the underlying trending, seasonal, and periodic structure within the data; achieved RMSE of 6.5
- Designed the kernel function and fine-tuned the parameters for the Gaussian Process Regression that consists of a combination of RBF, ExpiSineSquared, and Rational-Quadratic kernels; achieved RMSE of 5.0

Hypothesis Testing for Movie Ratings — Feb 2020 – Mar 2020

- Applied Welch-T, Mann Whitney, KS, and Bootstrapping tests to study research questions like whether people with different demographics view the same movie differently; are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?

Varsha Vattikonda

vv2116@nyu.edu | +1-805-743-9755 |Linkedin

Education

M.S. in Data Science, New York University, USA Sep 2021 - *May 2023
Coursework : Natural Language Processing, Deep Learning, Big Data (Spark), Probability and Statistics, Linear Algebra

BTech & MTech in Electrical Engineering, Indian Institute Of Technology Madras (IITM) Aug 2011 - May 2016

Skills

Programming Languages	Python, SQL, C# and R
Data Science	Statistical Models, Machine Learning Models, BI Models, Identify KPIs
Software Development	ETL , Object Oriented Design, Data Structures and Algorithms, Git
Cloud Services	Databricks, Azure (Data Factory, Analysis Services), GCP (Vertex AI, Earth Engine, BigQuery)
Reporting	Power BI, Splunk and Salesforce

Work Experience

Data Science Intern | MSCI Inc. ESG NYC, USA May 2022 - Aug 2022
Developed a PoC evaluating the capabilities of **Google Cloud Platform** and **Google's Earth Engine** in Spatial Finance that has a potential of **$1M** in annual recurring revenue

- Delivered an asset level flood forecast model using **KMeans Clustering** and **Random Forest**
- Deployed and automated the model to update daily flood probability in BigQuery using **Vertex AI**

Designed an enterprise architecture facilitating **Data Catalog** for around **10GB** data flowing in daily from around 2000 vendors

- Delivered a prototype for pluggable and extensible pipeline for data quality assurance using **Great Expectations** and **Airflow**
- Implemented a custom integration that pushes the metadata, lineage and data quality metrics of Delta tables to **Datahub** using **Spark** and **Databricks**

Data Science and Machine Learning Engineer |MSCI Inc., Mumbai, India Jun 2016 - July 2021
Multifaceted role in building a centralized Data Mart and pipelines for mining actionable insights from client's usage data. This project has earned an additional **$1.2M** of annual recurring revenue

- Achieved MAE of **5mins** wrangling around 10M rows of data to predict the duration of a simulation on the clients' portfolio using **K-Means Clustering and Linear Regression**
- Built a **Unigram (NLP)** model on 10K email conversations between the clients and consultants assessing the sentiments using **SentiWordNet**
- Built a **Churn Prediction** model using 10-15 features like magnitude of spikes, dips and trends of Client Engagement KPIs for 4 different product lines with around 5000 clients each using **Ensembled decision tree algorithms**
- Reduced churn rate by **10%** using **Statistical Data Modeling** and categorizing the users as DAU/WAU/MAU and thereby underlining the dips in usage
- **Collaborated** with Product teams to identify KPIs for products' usage and built a **Product Recommendation Model** underlining the potential cross-sells using **PCA** followed by K-Means Clustering
- Designed **ETL and QA Framework (C# API)** that extracts around **10GB** data each day from **10** different sources (eg. Splunk, Salesforce) and runs quality checks against self-learned **dynamic thresholds** using **Azure cloud services**
- Increased Client Insights Reporting platform adoption by **40%** by creating an intuitive and actionable **PowerBI** reporting solution in addition to firm-level trainings on the same

Projects

- **Precipitation Nowcasting** - A deep learning model based on **UNet (Pytorch)** using Negative Gaussian log-likelihood as the loss function with the model generating two output images - Mean and Standard deviation
- **Movie Recommendation Engine**- Comparative study on the performances of Movie Recommendation Engine using **Latent Factorisation** model (Alternating Least Squares for optimization) using **SparkML** (distributed system) vs Lenskit(single machine) for different data sizes
- A study on **The Evolving Sentiment of Work From Home** based on 3M samples of daily tweets over the last 3 years using **Naive Bayes Topic Model** to understand the topics of concern related to WFH and how the popularity and polarity of the topics have changed over time

David Trakhtenberg

847-331-8165 | davidtrakh@gmail.com | 567 6th St, Brooklyn, NY 11215

EXPERIENCE

Inspire11 — **Chicago, IL/Remote**
Senior Consultant, Data Engineering — July 2022 - Present
Consultant, Data Engineering — August 2020 - July 2022

- Modeled ten tables to create star schema which ingested 20M rows of Qualtrics survey data
- Translated five Alteryx workflows into 27 SQL (Snowflake) and two Python scripts
- Collaborated on development of a Spark architecture on top of Kubernetes which required parsing through active GitHub pull requests to develop novel solutions
- Consulted on the following clients: science research, government services nonprofit, trading firm

Clarity Insights — **Chicago, IL**
Senior Associate Consultant — December 2018 - March 2020
Associate Consultant — June 2017 - December 2018

- Developed PySpark code that met source to target requirements and created ETL pipelines, unit tested ETL code, and strengthened platform's common functions in code base
- Analyzed various industry's datasets by cleaning, transforming, and modeling (i.e. linear models)
- Utilized the following technologies: Hadoop ecosystem to manage file system, Microsoft Azure Cloud and TFS to build and release code, and SQL to validate data
- Documented best practices relating to useful Hadoop/Git commands, Spark error handling, a creative Git branching strategy, effective Python functions, and general knowledge transfer
- Conducted several stakeholder interviews to identify data consumption needs of the client and proposed appropriate solutions to sign new statements of work
- Consulted on the following clients: insurance, healthcare insurance, financial services

Zillow/NYU Capstone — **New York, New York**
Graduate Research Scientist — September 2021 - December 2021

- Implemented a novel algorithm (LayoutLMv2) to perform error analysis in the document understanding space
- Trained three modal baselines—BERT for text, ResNeXt-FPN for visual, and LayoutLMv2 for layout with the downstream task of document classification
- Learned various practical deep learning skills—working with Hugging Face, tuning deep learning models (using WandB), and communicating/presenting on complex data science subject matter

EDUCATION

New York University — **New York, New York**
Data Science, M.S. — September 2020 - December 2022
Center for Data Science — GPA: 3.67

University of Illinois Urbana-Champaign — **Champaign, Illinois**
Aerospace Engineering, B.S. — August 2012 - December 2016
James Scholar Honors Program — GPA: 3.65

SKILLS

Technologies: Git, ETL, Hadoop, Apache Spark, AWS, Microsoft Azure, UNIX, Linux, Tableau, Talend, Alteryx, several databases (Oracle, MongoDB, Snowflake etc.), Jira, Docker

Languages: Python (3 Years exp.), SQL (4 Years exp.), R (1 Year exp.)

NYU coursework: Machine Learning, Probabilistic Time Series Analysis, Text As Data, Natural Language Processing, Linear Algebra, Big Data, Responsible Data Science, Deep Learning, Fundamental Algorithms

Gordon Chan

thegordonchan@gmail.com | +1-734-834-8486 | New York, New York

EDUCATION

NEW YORK UNIVERSITY — **New York, New York**
Master of Science in Data Science — **Expected Graduation: May 2023**

- GPA: 3.83/4.00
- GRE Quantitative Reasoning: 169/170; Verbal Reasoning: 166/170; Analytical Writing: 5/6

UNIVERSITY OF MICHIGAN — **Ann Arbor, Michigan**
Bachelor of Science, *with Distinction* — **Sep 2017 – Aug 2021**
Major: Economics; Minor: Computer Science

- Cumulative GPA: 3.83/4.00; Major GPA: 3.89/4.00

PROFESSIONAL EXPERIENCE

BW SOLAR — **Remote**
Data Analyst Intern — **Jun 2022 – Aug 2022**
A subsidiary of BW Group focused on the development of solar power generation and energy storage

- Planned and oversaw the transfer of the company's local MySQL server to a cloud-based solution (Azure), allowing for the automation of data pulling and uploading, as well as providing easier user access and flexibility for future scaling
- Developed an app in R that matched the visual grid data provided by Homeland Security to the physical properties of the transmission operators model, creating a visualization of the transmission system used to identify unique opportunities for potential new projects

PHOENIX CAPITAL (INTERNATIONAL) LIMITED — **Hong Kong**
Quantitative Developer Intern — **Jun 2020 – May 2022**
An asset management firm with a unique focus on technology and quantitative investment management strategies

- Developed a predictive classifier model in Python that can recognize common chart patterns in real-time price action candlestick data by utilizing computer vision and deep learning, to streamline and assist volume-based trading activities.
- Designed a custom Convolutional Neural Network with Tensorflow and OpenCV to achieve a training accuracy of over 95% on the predictive classifier model for chart pattern recognition.
- Produced a robust training set by combining comprehensive historical NASDAQ one-minute data to generate over 500,000 individual candlestick charts.
- Created a Slack app that would provide real-time price summaries of stocks via slash commands, as well as automatically generate regular market profiles for various key sectors during trading hours using price data streamed from DTN IQFeed.

LONGEVITY DESIGN HOUSE — **Hong Kong**
Project Intern — **Jul 2019 – Aug 2019**
A social enterprise providing interior design solutions to improve the 'aging at home' experience for senior residents

- Partnered with an NGO to lead an elderly home development project on a remote Hong Kong island, providing over 50 households with free home safety products to improve their quality of life and keep them safe.
- Conducted market research and collaborated with the design team to produce a material library that helped to streamline the design phase, thus allowing 20-25% faster turnaround to clients.
- Developed and presented a viability study for converting the material library into an online platform by conducting analysis on past projects.
- Strengthened and created English versions of all sales and marketing materials to increase service awareness and reach.

SERAYA ENERGY — **Singapore**
Product & Business Development Team Intern — **Aug 2018**
The retail arm of a leading electricity and utility provider for Singapore's new Open Electricity Market

- Refined and promoted the new 'Geneco' brand's electricity retail product offerings to households and commercial institutions for the soft launch of Singapore's Open Electricity Market.
- Collaborated with the sales, design, and technology teams to develop and implement a referral program via cold calling and promotional products, that leverages existing Seraya Energy customers to increase market share and reach.

ADDITIONAL

- **Multilingual** – Fluent in English, Cantonese, and Mandarin, working-level Spanish
- **Technical** – Proficient in C++, Python, SQL, and Microsoft Office, familiar with MATLAB, R, Stata, JavaScript, and Git
- Citizen of Hong Kong, France, and Canada

Xu Han

xu.han@nyu.edu | xh852 | Xu Han

EDUCATION

New York University, Center for Data Science, New York, NY — Expected 2023
Master of Science in Data Science

New York University, College of Art and Science, New York, NY — May 2020
Bachelor of Arts in Mathematics and Economics

SKILLS

Programming Languages/Tools: Python (NumPy, Pandas, Scikit-Learn, Jupyter notebook, Geopandas, Plotly), **SQL** (MySQL, PostgreSQL), Dash, R, Tableau, MATLAB, C++, Java, Excel, Google sheets, Git.

Relevant Courses: Machine Learning for Healthcare, **Natural Language Processing**, Natural Language Understanding, **Machine Learning**, Big Data, Probability and Statistics for Data Science, Optimization and Linear Algebra for Data Science, Introduction to Data Science, Introduction to Computer Science.

EXPERIENCE

Data Science Research Intern, **Marron Institute of Urban Management**, Brooklyn, NY — May. 2022 – present

- Collaborated with 2 other interns and the Illinois Department of Correction to build a job recommendation and a job application system for **20239 parolees** across **104 prisons** in Illinois.
- Generated an interactive and geographical data visualization dashboard to present distribution of parolees and prisoners by counties in Illinois and neighborhoods in Chicago with libraries such as Geopandas and Plotly.
- Analyzed parole and prison population datasets to allow policy makers at Department of Correction to implement better policy for helping parolees and to prevent crimes in the future.

Top 50 teams, **Humana-Mays Healthcare Analytics Case Competition**, New York, NY — Sep. 2021 – Oct. 2021

- Performed feature engineering on **367 features** of **10 million** Humana customers with resampling, one-hot encoding and PCA.
- Implemented ML models including XGBoost, SVM and Random Forest to identify customers' hesitant towards COVID vaccine.
- Categorized vaccine resisters and explored reasons (pre-existing conditions, access, misinformation, lack of trust) for the hesitance.
- Recommended policies for helping Humana to encourage customers to receive vaccines.

Universal Intern, **Aflac**, New York, NY — June 2020 - Aug. 2020

- Rotated through managerial, sales, marketing and financial operations departments and collaborated with 10 interns.
- Developed actionable recommendations, such as podcasts, to increase company's marketing effectiveness by **10%**.
- Analyzed and visualized Aflac's financial data by using Excel dashboard and presented to program supervisor to showcase insights.

PROJECTS

User/ item representation algorithm – Capstone @ Zillow | *neural network, big data, Python* — Sep. 2022 – present

- Working with Zillow AI team and constructed user/item representations for short- and long-term prediction with transformers4rec library
- Modifying the transformers4rec library to create and evaluate model on Zillow's dataset
- Improving the model by testing additional hypotheses, varying variables such as loss function, variables, sampling method, dataset

Data Analysis on Among Us Dataset | *Hypothesis Testing, Logistic Regression, Multiple Linear Regression, Python* — Dec. 2021

- Conducted hypothesis testing on approx. **2000 Among Us players** with **14 features**; suggested winning strategy for crewmates.
- Created a new dataset regarding crewmates' game data to predict player's imposter win rate by multiple linear regression.

Breast Cancer Prediction | *Scikit-Learn, Logistic Regression, Random Forest Classifier, KNN, Python* — May 2021

- Cleaned and turned **146.8+ KB** of numerical and text data into training and test sets.
- Constructed **3 machine learning models** to predict beingness of breast lumps based on **30 clinical parameters** of **569 patients**.
- Conducted exploratory data analysis and visualized information from dataset by using histograms, scatter plots, and heatmaps.
- Tuned 4 hyperparameters to increase accuracy of Random Forest Classifier model from 96% to 97%.

Khevna Parikh

Khevnaparikh96@gmail.com | LinkedIn | GitHub

EXPERIENCE

Data Science Intern June 2022 - present
Fox Corporation – Tubi, Inc.

- Liaise with product managers to address ad-hoc requests, including evaluating the content viewership decay of Tubi's trending 1K titles
- Conduct content analysis between Tubi and 153 streaming platforms to motivate executive decisions in content acquisition using PySpark in Databricks
- Develop dashboards to determine and track Tubi KPIs over time using SQL

Legislative Fiscal Analyst Jan. 2019 – Aug. 2021
New York State Assembly

- Implemented an income tax surcharge on high-income earners to increase State revenue by $4 billion, which eliminated budget deficits and continued funding of state programs
- Developed econometric models in SAS to forecast New York State tax revenue using a data file consisting of 10.5 million personal income tax returns
- Analyzed and executed budget recommendations on all personal income tax bills and effectively communicated insights on related bills to legislators and advocates
- Monitored $55 billion daily State personal income tax revenues and expenditures

Data Analyst Aug. 2018 – Dec. 2018
Hofstra Northwell Graduate School of Nursing

- Performed data entry procedures to process clinical placements for 200 nursing students
- Designed and assessed student course evaluations to enhance future iterations of class
- Recruited preceptors and co-facilitated onboarding and preceptor development sessions

PROJECTS

Emotions of War in Ukraine: Analyzing Public Sentiment from Twitter Data

- Collected 440K tweets from the Russia-Ukraine region using Twitter API and preformed preprocessing measures such word stemming tweets in Cyrillic alphabet
- Trained a Latent Dirichlet Allocation (LDA) topic model to discover the natural groupings of words or topics relevant to emotions of individuals of the ongoing Russian-Ukrainian conflict

Big Data: Movie Recommender System

- Deployed an Alternating Least Squares (ALS) model for implicit feedback in PySpark on MovieLens 25M dataset; Tuned various parameters including the number of latent factors to deal with the issues of scalability and sparseness

SKILLS

Python, Spark, SQL, R Studio, Git and Version Control, AWS Services, and familiarity with SAS
Regression, Classification, Clustering, and other Supervised & Unsupervised Machine Learning
Exceptional verbal knowledge of English, Hindi, and Gujarati; Intermediate knowledge of French

EDUCATION

Master of Science in Data Science May 2023
New York University

- *Courses*: Big Data, Machine Learning, Text as Data, ML in Finance, Linear Algebra
- *Leadership*: Vice President: Graduate Community Group, Public Finance Fellow, Grader

Bachelor of Science in Applied Mathematics & Statistics and Psychology May 2018
Stony Brook University

- *Leadership*: Teaching Assistant, Research Assistant, Academic Peer Advisor

Qin Yang

+1 (201) 565-5998 | New York, NY | qy692@nyu.edu

New grad with broad-based experience in building data-intensive applications, overcoming complex architectural, and scalability issues in diverse industries. Proficient in predictive modeling, data processing, and machine learning algorithms, as well as scripting languages and version control tool, including python, SQL, and Git. Capable of creating, developing, testing, and deploying highly adaptive diverse services to translate business and functional qualifications into substantial deliverables.

EDUCATION

New York University, Center for Data Science **New York, NY**
Master in Data Science Expected May 2023
GPA: 3.6/4.0
Relevant Coursework: Applied ML in Finance, NLP and Computational Semantics, Deep Learning, Machine Learning, Computer Vision, Time Series Analysis, Probability and Statistics, Big data

Shandong University, School of Management Science and Engineering **China**
Bachelor in Management Science and Engineering/Master in Econ Sep 2014 – Jul 2021
GPA: 3.8/4.0
Relevant Coursework: Advanced Micro and Macro Economics, Econometrics, Finance, Calculus, Linear Algebra, Data Principles and Applications, Python and Data Analysis, Database, C++

EXPERIENCE

Data Science Department, Campana Schott Inc. May – Sep 2022
Position: Data Science Intern

- Bayer: Product Sales Forecasting *(with NLP + Time Series Models)*

- Identify key drivers for a downshift in product sales by novel analysis of both consumer market dynamics and mobility data from internal and external sources, determined if sales can be expected to normalize over time.
- Analyzed >2m mobility data points over 3 years as a proxy for consumers' social interactions, scraped >5m Tweets to measure and plot the sentiment of products and analyze statistical relationship to company sales, and implemented NLP and feature importance techniques to understand the explanatory value of large-scale product reviews and opinions of target consumers.
- Ran the Facebook Prophet time series forecasting model to capture the general growth trend, seasonalities, and holiday effects of our products as benchmark results; adopted historical sales data to train deep learning model N-HiTS for capturing the variation of sales and making forecasts at different horizons and achieved 84% explanatory power of product sales when forecasting at a 24-week horizon.
- Present solutions to core clients, leadership teams, and forum audiences.

- Radius Health: Extraction of Customer Satisfaction Topics Regarding Product Effectiveness *(with NLP)*

- Utilized Zero-shot learning and Few-shot learning with Hugging Face pre-trained models in extracting consumers' post-purchase reactions from unstructured texts, assessed opportunities for service or product upgrade.
- Developed a real-world dataset by crawling 360,000+ users' reviews from internal and external websites, analyzed and annotated a set of reviews on a sentence level with different labels (e.g., beneficial effects vs. adverse effects).
- Fine-tuned pre-trained sentence classification models from Hugging Face (e.g., BART-large) on the manually labeled dataset, achieved 90% accuracy and delivered our research outcome in the form of a 20-page report and presentation to core stakeholders.

R&D Department, IFLYTEK Jun – Sep 2019
Position: Data Analyst

- Developed a nationwide retention program with Python, SQL, and Excel, saved 1000+ hours of labor per year.
- Identified procedural areas of improvement through 10+ million customer data, queried and analyzed data from department database system using SQL, and created 20 dashboards and presented findings to 10 stakeholders.
- Trained Linear Regression model, predicted repair costs for vehicles on the market, and increased the profitability of a nationwide retention program by 8% (~100 million).

PROJECT

Improve Speech Recognition Performance with Unpaired Audio and Text Data *(with Hugging Face Transformer and Pytorch, ongoing)*

- Fine-tune Hugging Face pre-trained ASR models on 360+ hours of audio dataset to boost speech recognition performance on highly accented and disfluent data.
- Analyze large audio datasets, measure, and benchmark model performance, as well as present model accuracy with visualization tools (e.g., Tensorboard, W&B).

Adversarial Learning on Neural Network Models for Text Classification *(with Hugging Face Transformer)*
Link to the research paper:https://drive.google.com/file/d/1fIH3axJJcmxTTtmSDQtVNDMrm7_Hrbmi/view?usp=sharing

- Extended the idea of self-supervised learning and combined it with the fine-tuning approach for boosting text classification performance in an adversarial setting.
- Pre-trained a generative model to predict the representative of adversarial input and fine-tune it with one additional output layer for downstream NLP tasks (e.g., sentiment analysis and textual entailment).

- Constructed 10,000+ strong adversarial texts dynamically by attacking STOA Transformer models (e.g., BERT and RoBERT) on three datasets (IMDB, AG's News, and SNLI), generated another 200,000+ weakly augmented texts by following the implementation of SSMBA and NLPAug methods as a comparison with baseline results.
- Improved the accuracy of classification and entailment tasks on adversarial texts by 30% on average, compared to Hugging Face pre-trained STOA models.

Bitcoin Price Forecast with Time Series Models *(with ARIMA & GPR)*
- Predicted future value of Bitcoin using Time Series techniques (ARIMA and Gaussian Process Model) in a 4-year period.
- Applied stationary check (ADF & KPSS tests) and transformation techniques (e.g., power transformation and difference) to preprocess 10,000+ price records of Bitcoin; determined the values of autocorrelation and partial autocorrelation of bitcoin price series by plotting ACF and PACF.
- Made predictions and evaluated model performance by tracking diverse metrics (RMSE or MAE); demonstrated findings in the form of a presentation to 40+ audiences and a 20-page written report.

Recommendation System for MovieLens *(with Apache Spark and Hadoop)*
- Implemented Alternating Least Squares (ALS) and popularity baseline model for collaborative filtering in making movie recommendations.
- Partitioned the rating data (58,000 movies and 280,000 users) into stratified training, validation, and testing samples based on user ID; made prediction and evaluation based on predictions for the top 100 items for each user with ranking metrics (e.g., precision at k, NDCG at k, RMSE, and MAP).
- Executed hyperparameter tuning with a number of latent factors, regularization, and iteration epochs to produce observable differences in evaluation score; make comparison to single-machine implementations (e.g., lightfm and lenskit) and achieved a 20% accuracy boost.

Sentiment Analysis of Comment Texts Based on Bi-LSTM *(with Keras and Tensorflow)*
- Encoded 50,000 sentences with a word-level Bi-LSTM sentence encoder and trained a neural network classifier for sentiment analysis.
- Performed NLP-based tokenization, lemmatization, and vectorization in machine understandable language and applied pre-trained Glove embedding to initialize embedding layer of word representations.
- Added dropout layer, early stopping, and max-norm constraints as regularization methods, and max/avg pooling layers to reduce variance and computation complexity, trained deep neural network Bi-LSTM, Bag-of-Words Classifier, and LSTM, and achieved the highest accuracy (97%) with Bi-LSTM.

SKILL

Key Skills: Data Visualization, Predictive Analysis, Statistical Modeling, Clustering & Classification
Technical Skills:
- **Tools:** Python, Spark, Hadoop, Dask, SQL, Tableau, Pytorch, MapReduce, Linux, HPC
- **Packages**: Scikit-Learn, Numpy, Scipy, Pandas, NLTK, Matplotlib, Seaborn, Jupyter Notebook
- **Statistics/Machine Learning**: Statistical Analysis, Linear/Logistic Regression, Clustering, Decision Tree, GBM, Deep Learning, Natural Language Processing
- **Link to Github**: https://github.com/opal-1996.

Jess Bunnag

www.linkedin.com/in/jesstbunnag/ | tb2817@nyu.edu | (347)-425-6342

Education

New York University, Center for Data Science — **New York, NY**
Master of Science in Data Science — Expected May 2023

Columbia University — **New York, NY**
Bachelor of Science in Computer Science — Sep 2015 – May 2019

Work Experience

Verkada — **San Mateo, CA**
Data Engineer Intern, Data Platform — June – August 2022

- Published product usage analytics dashboard using Airflow, Athena, and Metabase for the admin page, which allows users to monitor actions logged by any of the company's security system products. Helps product teams make data-driven decisions.
- Built dashboard for sales team to monitor time from shipping to camera installation, improving post-sales customer service.
- Designed experiment ID logging to enable A/B experimentation and metrics calculation when launching new model features.

Agoda — **Bangkok, Thailand**
Software Engineer, Metasearch Ads Bidding — July 2019 – August 2021

- Enhanced and maintained the automated bidding framework for 300M+ bidding entities, which uses XGBoost and CatBoost to predict optimal cost per click for ads listings on metasearch engines, including Tripadvisor, Trivago and Google HPA.
- Optimized the bidding framework by linking/reading/writing to database files on Hadoop, speeding up the pipeline by 20x.
- Designed and incorporated 2 new features into a margin prediction model—increased annual revenue by approximately $1M.

Appian — **Reston, VA**
Software Engineer Intern — June – August 2018

- Enhanced the Connected System Template SDK, which allows users to integrate third-party services into applications built on Appian's platform, by implementing a visual interface to integrate with Google Cloud Vision.
- Improved SDK's implementation of OAuth 2.0 by persisting access tokens for unsaved Templates that require authorization.

Sun Trading LLC. — **New York, NY**
Software Development Intern — June – August 2017

- Developed a simulation algorithm in C++ that analyzes order books in real time and identifies market trends to improve trading strategies. Results were used with treasury order books, to aid in filtering market-making data.

Technical Projects

Latent Factor Movie Recommendation System — **New York University**

- Built a movie recommendation system that recommends top 100 items to users. Used latent factor model with Spark's ALS API to learn user/item representations via matrix factorization. Trained on MovieLens datasets (27M ratings on 58K movies).
- Evaluated accuracy of latent factor model using the NDCG metric (0.014, compared to baseline of 4e-6).
- Used UMAP to visualize movie clusters and understand model errors (overrepresentation of outliers, genre overlap).

Comparing Image Similarity Ratings of ResNet50 and Humans — **New York University**

- Compared DL models to humans by analyzing similarity ratings that ResNet50 and humans gave to toy/real animal images.
- Fine-tuned ResNet50 using PyTorch by freezing the last layer of the network and reinitializing a new fully-connected Linear layer with 90 output features (the number of animal classes in our dataset). Achieved 85.24% accuracy after 10 epochs.

Movie Ratings Multi-output Linear Regression Model — **New York University**

- Used multi-output linear regression to predict personality from movie preferences, given a dataset of 1097 personality survey answers and movie ratings. Explored effects of tuning regularization hyperparameters in ridge and lasso regression models.

Skills / Academics

Languages: Python, PostgreSQL, Java, Scala, HTML, CSS, Javascript, Flask | **Tools:** Spark, Hadoop, Apache Airflow

Courses: Machine Learning, Big Data, Deep Learning, Computational Linear Algebra, Probability and Statistics, Natural Language Processing, Artificial Intelligence, Advanced Web Design Studio, Data Structures and Algorithms

Teaching: TA for Intro to CS and Programming in Java (Columbia), TA for Fundamental Algorithms (NYU)

Qi Dong

Jersey City, New Jersey • 6468661808 • qd2046@nyu.edu
www.linkedin.com/in/qi-dong-6a8b15170/

EDUCATION

New York University **May 2023**
Master of Science in Data Science, GPA: 4.0/4.0
Relevant Coursework: Machine Learning, Big Data, Introduction to Data Science, Programming for Data Science, Optimization and Computational Linear Algebra, Computational Cognitive Modelling

Rensselaer Polytechnic Institute **Dec 2020**
Bachelor of Science in Industrial Engineering | Minor in Psychology, GPA: 3.97 / 4.0
Relevant Coursework: Computer Science, Information System, Modeling and Analysis of Uncertainty, Statistical Analysis, Introduction to C++, Operations Research Method, Discrete Event Simulation Modelling

PROFESSIONAL EXPERIENCE

Load Forecasting & Analytics Intern - National Grid, Hicksville, NY **May 2022 – Aug 2022**
- Performed Exploratory Data Analysis on day-ahead gas load forecasting data using Pandas, cleansed & transformed the raw data and visualized trends in data with Matplotlib, Seaborn, and Plotly in Python.
- Detected data anomalies and coordinated with data source provider to propose & implement anomaly corrections, corrected 2 aberrations in gas load data and 3 irregularities in weather data.
- Built neural network machine learning models using Keras and PyTorch in Python on gas load time-series data to forecast short-term gas loads, applied hyper-parameter tuning, compared model performance with existing regression models and concluded that neural network outperformed regression models in most of the regions.

Logistics Intern - BMW Group, Beijing, China **Sep 2019 – Dec 2019**
- Facilitated arrival & delivery of items inside the warehouse, organized inventories & monitored ~80% of incoming deliveries daily, maintained a log of all receipts and disbursements.
- Conducted cost benefit analysis by estimating flow-in and flow-out units, full-time equivalent labor cost, space utilization and facility cost to calculate net present value of the new warehouse design.

RESEARCH PROJECT

Generalized Wind Power Prediction Modelling | *R, Python, Regression* **Jan 2020 – May 2020**
- Cleansed & transformed raw windfarm data, imputed missing values & treated outliers using R in RStudio to create a data asset for predicting wind power output using machine learning algorithms.
- Explored the data & identified 11 target windfarms for power prediction study, built a data ETL & engineering pipeline in R to gather and process the raw data into metric like hourly average, curtail, etc. for windfarms.
- Build a benchmark model with variables like air density, wind speed, and swept area of turbine blades using Linear Regression as a foundation for further development, achieved MAPE of 11%.

ACADEMIC PROJECTS

Movie Recommender System | *Python, Spark, Recommender* **Jan 2022 – May 2022**
- Built a Baseline Popularity, Spark Alternating Least Square, and LightFM models for movie recommendations using 1.2M records from the MovieLens dataset.
- Implemented hyperparameter tuning & cross-validation, evaluated models using Mean Average Precision, achieved best MAP of 0.182 with LightFM.
- Assessed model performances on small & large datasets & recommended models suited to different situations.

Movie Ratings Analysis | *Python, PCA, Clustering, Classification, Hypothesis Tests* **Sep 2021 – Dec 2021**
- Analyzed the impact of movie categories on movie ratings with hypothesis testing in Python.
- Performed correlation analysis & linear regression to determine factors that influence movie recommendations with a testing error of 1.2%.
- Leveraged Principal Component Analysis for feature reduction & kMeans to identify 4 clusters in the data.
- Constructed Nearest Neighbor & Neural Network classifiers to predict movie ratings using the clusters, evaluated model performance with confusion matrix, achieved 98% precision and recall.

SKILLS & CERTIFICATIONS

Programming Languages: Python (NumPy, Pandas, Matplotlib, SciPy, Scikit-Learn), SQL, R, C++
Analytical Tools: MySQL, RStudio, Jupyter, MS Excel, MS Access, MATLAB, Minitab, Visual Studio, Arena, CAD
Certifications: Machine Learning | Coursera, Advanced SQL for Data Scientist | LinkedIn, Six Sigma Green Belt | IISE

Soomin Kim

New York, NY 10128 (929) 837-0434
soomin.kim@nyu.edu linkedin.com/in/soominkim231 github.com/soominkim231

EDUCATION

New York University **May 2023**
M.S. in Data Science. GPA: 3.76/4.0 *New York, NY*

- **Relevant Coursework***:* Intro to Data Science, Probability and Statistics, Linear Algebra, Programming for Data Science, Data Visualization, Big Data, Machine Learning, Natural Language Processing (ongoing), Ethics of Data Science (ongoing)
- **Leadership**: Center for Data Science Student Leadership (Executive), Women in Data Science (Executive)

Dartmouth College **Jun 2020**
B.A. in Quantitative Social Science and Music *Hanover, NH*

- **Honors**: College Honor List, Citations for Academic Excellence

Relevant Certification: IBM Data Science Professional Certificate

TECHNICAL SKILLS

Languages: Python (pandas, NumPy, SciPy, sklearn, Matplotlib, PyTorch, Tensorflow) | R | SQL | MATLAB | STATA
Tools: Spark | Hadoop | Git | Unix | BigQuery | Tableau | ArcGIS | Google Colab | Jupyter Notebook | Excel | AWS

EXPERIENCE

Attentive*, New York, NY* **Sep–Dec 2022**
Machine Learning Engineer Intern

- Deploy state-of-the-art **neural network** model for **product recommendations** that outperforms existing model's performance
- Query into **Snowflake relational data warehouse** of 40 million different products over 8,000 distinct brands using **SQL**
- Load data of 8 million+ records and perform **EDA** on **Jupyter Notebook** via **AWS SageMaker** Notebook Instance Environment

Sony Music Entertainment*, New York, NY* **Jun–Aug 2022**
Data Science Intern

- Collected and **preprocessed** data pulled from internal **cloud databases** and Chartmetric/Spotify **API**s using **Python** and **SQL**
- Built **multivariate linear regression** model to predict Spotify first day streams; fine-tuned model from R^2 0.51 to 0.92 whilst validating model assumptions, which provided actionable **metric recommendations** to use when targeting new artists to sign
- Identified songwriter groups by track popularity using **unsupervised clustering** algorithms with **TF-IDF** method on text data and applied **PCA** to reduce vector dimensionality for cluster **visualization**, delivering songwriter collaboration recommendations

The Ripolles Lab, New York University*, New York, NY* **Oct–Dec 2021**
Data Analyst

- Developed **Python** algorithm that detects and visualizes music-induced goosebumps from real-time video of skin captured by prototype wearable sensor of **Raspberry Pi** architecture; expanded prototype use to 300 people in NYC area
- Set up virtual environments via **Conda**/**Python** on **Raspberry Pi Zero** that ensured **Python** script runs smoothly
- Utilized **Git** and worked on separate branches to allow code collaboration and to prevent merge conflicts

Ernst & Young*, Seoul, South Korea* **Sep–Dec 2020**
Technology Solutions Consulting Intern

- Collaborated on client-facing project to benchmark 8 Engineering Procurement Construction (EPC) companies' global operations
- Interviewed 40+ experts (C-level executives) on 30-120 min calls, successfully narrowing team's focus to 3 EPC companies
- Pivoted and adapted quickly as client guidance evolved, whilst always keeping the ultimate project objective in mind
- Summarized research on organizational structure & strategy into **Excel**/**PowerPoint**, effectively reporting to senior management

PROJECTS

Clustering Analysis of Korean Restaurants in NYC using Machine Learning and Python **Apr–Jun 2021**
IBM Data Science Professional Certificate

- Pulled and **prepared geodata** of 10K+ venues and 306 NYC neighborhoods (via **Foursquare API**) using **Python**
- Performed comprehensive **EDA** and **visualized** relationship between neighborhoods and number of Korean restaurants
- Employed **Elbow Method** to identify optimal number of clusters and **K-Means Clustering** to cluster neighborhoods based on similar mean frequency of Korean restaurants, identifying optimal locations to start Korean restaurant business in NYC

Factors Determining Sentence Length and Severity for U.S. Inmates using R **Jan–Mar 2019**
Dartmouth College Government Department

- Worked closely with 3 other undergraduates to conduct **multivariate linear** and **logistic regression** using **R** to investigate the statistically significant predictors of sentence length and severity for U.S. inmates across State and Federal Prisons
- Visualized results showing expected values of receiving an extreme sentence as a function of various predictors

TIFFANY LIN

Data Scientist, Analytics

tl3493@nyu.edu
(407) 234-7299
New York, NY
linkedin.com/in/lint12
github.com/lint12

EDUCATION

M.S.
Data Science
New York University
May 2023
New York, NY
3.67

Relevant courses

- Machine Learning, Probability and Statistics, Big Data, Optimization and Computational Linear Algebra

B.S.
Computer Science
University of Central Florida
December 2020
Orlando, FL
3.72

SKILLS AND TOOLS

- Python
- SQL
- Pandas
- NumPy
- Sklearn
- C++
- Hadoop
- Spark
- Excel
- JavaScript
- CSS
- ReactJS

WORK EXPERIENCE

Data Analytics Intern

Spectrum Enterprise
June 2022 – August 2022 | New York, NY

- Analyzed unstructured data, performed segmentation to improve campaign efficiency & present robust business insights to leadership
- Reduced the process time complexity for reporting by 90% by building a customized dashboard automated with Alteryx and SQL to clean and expedite data migration from various data sources: Salesforce and Eloqua
- Developed custom Datorama widgets for data visualization of KPIs to promote and track overall goal of increasing marketing attributed revenue by 11%
- Executed quality assurance of Datorama dashboards and diagnosed any data mismatch by inspecting data streams and raw data with Alteryx and advanced MS-Excel
- Performed statistical and exploratory analysis on marketing spend to support justification of Q3 budget increase

Data Research Assistant at NYU Law (Part-time)

New York University
May 2022 – August 2022 | New York, NY

- Transformed over 11 versions of raw coding interactions with a total of 411 companies and 336 policy instances into digestible and meaningful CSVs to help answer key research questions on company policies using PostgreSQL, Python, and Pandas
- Extracted and cleaned for valid coder behavior, relationship between text and policies, and metrics from SQL dumps and JSON objects with inconsistencies to further and monitor the effectiveness of the research

CWEP at Lockheed Martin (Contract)

University of Central Florida
January 2019 – March 2020 | Orlando, FL

- Assisted with design, development, documentation, testing, and debugging of real-time embedded software for the Advanced Vehicle Defense System program (AVDS)
- Analyzed the system's performance through instrumentation tools and MATLAB scripting
  - Examined for any erratic CPU and memory usage of the embedded software after a regression test through plotting data points on MATLAB
- Troubleshot technical issues via Linux commands and hardware for embedded systems such as routers and switches

Web Design Intern

VGlobalTech
October 2018 – January 2019 | Orlando, FL

- Maintained and updated on average 5 ADA compliant websites for communities daily
- Collaborated with a team to design websites for businesses via Word Press
- Edited code databases to alter the layout of websites in order to fit the client's needs

PROJECTS

Discord Bot: Wikipedia Game

Developer
August 2020

- Recreated and hosted the Wikipedia Game using an open-sourced REST API, a Discord JavaScript library for developing bots, and Heroku
- Customized the bot to be able to account for user errors, track the player's score and time, and assign the winner with the champion role

ELLE Web Portal

Frontend Developer
April 2020

- Transformed the website of a language learning tool to be compatible with 3 different types of platforms using ReactJS and React libraries
- Adapted the website for all user types to create, modify, and reuse language modules for their platforms and analyze user performance
  - Generated bar graphs and pie charts to offer the users a visual depiction of how well they were learning new terms on each platform

Multiple File Transfer UI (Lockheed Martin)

Developer
August 2019

- Developed a user interface with Qt framework to transfer FPGA images for an embedded system (did not support existing software)
- Produced a reliant file transfer protocol by using checksums and headers

VANESSA (ZIWEI) XU

1-929-988-3639
vanessaziweixu@gmail.com
https://www.linkedin.com/in/vanessa-ziwei-xu/
https://github.com/VavaXu

EDUCATION

New York University, Center for Data Science — **New York, NY**
M.S. in Data Science — *May 2023*

- ***Relevant courses:*** *Probability & Statistics, Linear Algebra, Machine Learning, Big Data, Text as Data, Programming, Natural Language Processing, Capstone Project*
- *President – CDS Diversity Affinity Group; Mentor for Undergraduates - Data Science Club*

New York University, Steinhardt — **New York, NY**
B.Mus. in Music Business; Minor in Data Science, Performance Studies — *May 2021*

- ***Relevant courses:*** *Causal Inference, Database Design & Implementation*
- *Music Business Student Mentor*

SKILLS AND INTERESTS

- **Technical Skills:** Shell. Git, Python (PyTorch, NumPy, Pandas, Scikit-learn), R (quanteda), SQL (T-SQL, MySQL, SQLite, MySQL Workbench), NoSQL, MapReduce, Hadoop, Spark, Dask, Advanced Excel, Final Cut Pro
- **Languages:** Chinese (Fluent), French (Conversational)
- **Interests:** Singing (Final round auditions of Sing! China and performed at venues and theaters); Drama Writing, Director and Performer (LAMDA Gold Medal Distinction)

PROJECTS

Extracting causal political narratives from text **– Deep-Learning based NLP Capstone Project** — *Sep 2022 - Dec 2022*

- Constructing machine learning pipeline for causal relation extraction from text as well as causal structure discovery in order to automatically extract narratives from a large corpus of 1 million political news articles
- Implementing text preprocessing methods, apply and test pipeline on data, draw informative conclusions on its performance, and suggest potential improvements to explore how politically aligned media outlets propagate different narratives about the same set of facts, how selective reporting is structured, and the issue of polarization
- Utilizing Deep-Learning based NLP methods and their Python implementations including BERT, GPT-3, as well as Deep Learning libraries including PyTorch or TensorFlow

Using Human Rights Texts to Predict Country Income Category **– NLP Project (R)** — *Mar 2022 - May 2022*

- Explored how effective human rights texts can be in predicting monetary values – countries' income levels
- Applied text analysis methods:
  - Text pre-processing: stemming, removing punctuation, stopwords, and numbers, converting to lowercase, replacing apostrophes with empty string
  - Tokenized, filtered, and made features into document term matrix in order to measure features by word counts
- Built Naïve Bayes Multi-class Classification model with Laplace smoothing in R to train 10,000+ human rights text data files
- Made country income classification predictions and achieved 82.46% accuracy; and generated a 10-page report

Collaborative-filter Based Recommender System on MovieLens **–Group Project (Python & Spark)** — *Mar 2022 - May 2022*

- Built Popularity Baseline Model and Spark's Alternating Least Squares Model (ALS) after preprocessing and partitioning to predict the top 100 movie recommendations for each user from the MovieLens dataset with 280k rows
- My contribution:
  - Built Single Machine Implementation–LightFM model, an implicit feedback recommender, and further improved model performance with an absolute increase of 1.4% in precision_at_k and imrpved model efficiency by achieving less runtime
  - Used hyperparameter tuning on both ALS and LightFM model to achieve the best possible model result
- Implemented fast search algorithm with Annoy and established a brute force algorithm to further decrease runtime

WORK EXPERIENCE

New York University, Center for Data Science — **New York, NY**
Graduate Teaching Assistant, Causal Inference — *Jul 2022 - Aug 2022*

- Created course specific lesson plans and facilitated lab sessions in R language for Causal Inference
- Prepared weekly assignments, quizzes, exams and solutions for a class of 80 students
- Held Office Hours to help students work through complex questions and responded to inquiries and concerns
- Coordinated and released grading with the graders with careful consideration of confidential record

New York University, Center for Data Science — **New York, NY**
Graduate Assistant — *May 2022 - Aug 2022*

- Researched and generated detailed reports of alumni career placement stats with Python
- Reviewed and revised Career Development resources for undergraduate and graduate data science students
- Assisted with the development of internal undergraduate and graduate student portals, ensuring that website content was up-to-date and published content met guidelines
- Worked closely with the Academic & Student Affairs team on event logistics and student community building, including New Student Orientation and programming focused on diversity and inclusion

Xinge Hu

xh1051@nyu.edu • (617) 331-5722 • https://www.linkedin.com/in/xinge-hu-5336b5155/

EDUCATION

NEW YORK UNIVERSITY — **New York, NY, USA**
MS in Data Science **GPA: 3.76/4.00** — **Sep. 2021 – May 2023 (Expected)**
- Relevant Coursework: Machine Learning, Probability & Statistics, Linear Algebra, Big Data, NLP

BOSTON UNIVERSITY — **Boston, MA, USA**
BA in Statistics*, Minor in* ***Computer Science*** **GPA: 3.79/4.00** — **Sep. 2017 - Jan. 2021**

EXPERIENCE

Intellipro Group — **New York, NY, USA**
Data Analyst Intern — **Jun. 2022 - Aug. 2022**
- Performed descriptive analysis in **SQL** to analyze 1M+ students' information including major, gender, email, school location, etc, generated infographics using **Tableau,** and presented the data visualization results to company executives bi-weekly.
- Conducted data quality analysis and pre-processed modeling data by performing data cleaning and data transformation.
- Performed web scraping using **Python** Requests and Selenium packages, and expanded the company's database by 30,000+ records.

BETH ISRAEL DEACONESS MEDICAL CENTER — **Boston, MA, USA**
Data Science Research Intern — **Jul. 2020 - Nov. 2020**
- Collaborated with team's biologists, and performed statistical analysis to compare patients' protein data from different sources using **R**.
- Developed a Machine Learning classification model with Python package **scikit-learn** using **Logistic Regression** with L1 Regularization and Sequential Feature Selection with sigmoid SVM.
- Identified 20 most relevant proteins to surgery-caused delirium out of 1300 proteins, as well as signature proteins that distinguish different levels of delirium, and increased the overall **model prediction accuracy** from 70% to 89%.

BOSTON UNIVERSITY CENTER FOR CAREER DEVELOPMENT — **Boston, MA, USA**
Data Analyst Intern — **Jan. 2020 – May 2020**
- Conducted data cleaning and data transformations on survey results from 4000 college graduates.
- Developed a linear model that predicted college degree holders' salaries based on their demographics.
- Designed a User Interface to visualize salary prediction and distribution using **R Shiny** to provide better career consulting service.
- Presented data results to university executives and colleagues, and offered recommendations for data collection and survey design.

FINTECH GLOBAL — **London, UK**
Research Analyst Intern — **Feb. 2019 - Apr. 2019**
- Collected 12000+ global FinTech investments into MS **Excel**, doubling the size of the company's database.
- Analyzed investments and composed detailed analytical articles about FinTech subsectors on the company's research blog weekly.
- Collaborated with a 10-person team to organize a 3-day Global RegTech Summit attended by 500 industry professionals.

PROJECTS

Two-sided Market A/B Test Experiment Analysis (Metric Design, Experiment Evaluation) — **Jun. 2022**
- Performed **z-test** to evaluate key metric of Instacart shopper hiring funnel and validated the effectiveness of earlier background check.
- Analyzed the cost per acquisition(CPA) of multiple acquisition channels and provided marketing solutions for leads generation.
- Provided executable plans based on shopper hiring analysis to find bottlenecks and insights using **Python** and **Excel**.
- Recommended marketing strategy by analyzing the conversion rate and cost per acquisition of social media, referral, and job search.

Web User Activity Analysis (User Journey Analysis, Funnel Analysis) — **May 2022**
- Used **funnel analysis**, **cohort analysis** and **segmentation analysis** to acquire the reasons for decreased user email login rate.
- Analyzed Email campaign performance such as open rate and click through rate to diagnose user activity moving trend.
- Wrote **SQL queries** to impute retention rate and analyzed the moving pattern in the user engagement in **PostgreSQL**.
- Built **Tableau** dashboards to present the tendency in user engagement over time and breakdown by user type dynamically.
- Developed a methodology framework to provide practical recommendations as summarization.

Credit Card Users Segmentation (Unsupervised Learning, K-means) — **Feb. 2022**
- Conducted exploratory data analysis for geographic, demographic and transaction data from 10K+ credit card holders.
- Applied **PCA** and **normalization** for feature engineering, reducing 40+ features into 10+ features.
- Clustered users into five clusters with different models including **K-Means** and **RFM models**.
- Applied elbow method to identify the optimal number of clusters for K-Means models.
- Conducted profiling analysis for each cluster to interpret the characteristics and main driven features.

SKILLS

Data Analysis SQL | Python | Excel | R | Statistics
Data Visualization Tableau
Database PostgreSQL | MySQL
Machine Learning TensorFlow | PyTorch | Scikit-learn

YOU WANG

(201) 736-4733 | yw6127@nyu.edu | linkedin.com/in/you-wang-7094231a3

EDUCATION

New York University, New York, NY — Sep 2021 – May 2023 (Expected)
Master of Science in Data Science | **GPA**: 3.9/4.0
Relevant Courses*: Introduction to Data Science, Probability & Statistics, Optimization & Linear Algebra, Machine Learning, Big Data, Deep Learning, Natural Language Processing.*

Sun Yat-sen University, Guangzhou, China — Sep 2016 – Jun 2020
Bachelor of Science in Biotechnology | **GPA**: 3.8/4.0
Relevant Courses: *Database System, Computer Programing Language, Biostatistics, Machine Learning Fundamental*

TECHNICAL SKILLS

Tools: Python (NumPy, Pandas, Scikit-learn, Maplotlib) | R | SQL | LaTex | Tableau | Excel
Proficiencies: A/B Testing | Data Mining | Machine Learning | Big Data Analysis | EDA | Data Visualization| Statistics

PROFESSIONAL EXPERIENCE

Data Analyst Intern - **China Guangfa Bank Credit Card Center** — Aug 2020 – Nov 2020
- Extracted, cleansed & transformed 1.3M credit card transactions & customer demographics data using SQL queries, stored the processed data in a MySQL database daily, computed & visualized daily trends in 8 key KPIs with a Tableau dashboard.
- Applied feature engineering like One-Hot encoding, data binning & feature combination, augmented 25% new metrics & monitored daily key metrics of credit card operations data.
- Created 70+ interactive visuals, discovered data-driven insights with intuitive storylines & actionable recommendations daily.

Data Analyst Intern - **Datastory Information Technology** — Oct 2019 – Jan 2020
- Conducted exploratory data analysis for PepsiCo, Procter & Gamble and Bear Electric, identified & treated data anomalies using SQL & Python, segmented customers based on social media data, provided targeted audience profiles to brands.
- Tracked product upgrades & marketing campaigns performance with intuitive graphs & plots in PowerBI dashboards.
- Streamlined & automated 2 ETL processes, defined data cleaning & transformation rules, augmented data annotation with user portraits, and engineered 10+ new features from unstructured data, saved repetitive manual efforts by 60%.

Data Analyst Intern - **Kangmei Pharmaceuticals** — July 2019 – Sept 2019
- Implemented daily data ETL & processing pipeline for handling~400k sales records, produced daily & weekly reports with SQL & Excel for 10+ business users, saved ~20 hours per week of manual efforts.
- Analyzed sales variations with 9 other key product marketing metrics, performed correlation analysis & visualized trends with Python Plotly graphs & charts, empowered evidence-based decision making for product promotions.

PROJECTS

Diabetic Hospital Readmission Prediction | *Python, Machine Learning, Random Forest, Xgboost, Lasso Regression*
- Cleansed & processed 100+ features related to diabetes patient health & demographic data using Python Pandas & NumPy, performed correlation analysis for features selection & reduced to 44 significant features.
- Built & evaluated Random Forest, XGBoost and Logistic Regression classifiers with Sci-kit learn in Python to predict short-term & long-term readmission rates for diabetics, achieved an accuracy of 91% & AUC of 0.80.
- Constructed & optimized Lasso regression models to predict length patient stay in the hospital with MAPE of 12.6%.

CNN-Based Nucleic Acid Sequence Classifier | *Python, TensorFlow, Deep Learning, CNN*
- Processed HLA DNA sequence and 16s rRNA sequences into a 4 X 1 dimensional array using NumPy in Python.
- Designed a CNN with 6 convolutional, 6 pooling & 2 fully connected layers using TensorFlow in Python, achieved an accuracy of 93+% on all types of sequences.

Healthcare Provider Fraud Analysis | *Python, Machine Learning, Stacking, Random Forest, Xgboost, MLP*
- Cleansed & processed 160k+ rows of medical and insurance data from 5410 healthcare providers, studied and selected the features using Exploratory Data Analysis, created 25 new features
- Built a stacked model based on Random Forest, Xgboost and Multilayer Perceptron to predict the fraudulent behavior of healthcare providers, achieved an F1 score of 0.58 and an AUC score of 0.92

Citi Bike Network Flow Analysis | *Python, Machine Learning, RNN, Clustering*
- Preprocessed ~3M Citi Bike usage big data with 14 variables, incorporated multi-processing & multi-threading techniques to optimize date cleaning & preprocessing, decreased preprocessing time by 75%.
- Clustered 1000+ stations to 6 groups with K-Means algorithm in Python to reduce computational cost, transformed the data into 6X6 matrix to record trip counts within 2-hour intervals for each of the clusters.
- Built Random Forest, Graph Convolution Network & RNN models to predict the imbalancedness in bike distribution for each cluster within 2 hours, evaluated model performance & achieved MSE of 70.4 with RNN model.

Movie Recommendation System | *PySpark, Recommender System, Latent Factor Representation, ALS*
- Preprocessed 1.2M movie ratings from MovieLens, performed data cleaning & outlier treatment using Pandas in Python,
- Implemented collaborative filtering recommendation using Spark ALS method with latent factor representation of viewers & movies in PySpark, applied hyperparameter tuning with GridSearchCV, achieved MAP@100 of 0.001 & nDCG@100 of 0.011

Haizhu Wetland Ecosystem Service Value Analysis | *R, Python, Logistic Regression, Linear Regression*
- Applied Coarsened Exact Matching algorithm to deal with data imbalance, implemented Logistic Regression model applied backward elimination, quantified the impact of features on people's willingness to pay for ecosystem protection.