Heesuk Jang

(720) 201-7402 | Irvine, CA
heesuk.jang@gmail.com
https://www.linkedin.com/in/heesukjang/
https://www.heesukj.com/
https://github.com/heesukjang

PROFESSIONAL SUMMARY

Dynamic and results-driven Data Scientist and Analyst with over 7 years of expertise in data analytics, machine learning, and Natural Language Understanding (NLU). Passionate about Deep Learning and cutting-edge technologies, with a natural curiosity and love of learning that drive innovative solutions. Proven record of delivering high-quality code and completing complex projects on time, showcasing an ability to swiftly adopt new tools and methodologies. Experienced in developing novel approaches to energy market analytics. Strong leadership skills, providing mentorship, training, and actively supporting colleagues in adopting new technologies. Active contributor to S&P Global's Women in Technology initiative.

SKILLS AND TOOLS

- **Programming Languages:** Python (TensorFlow, Keras), SQL, R
- **Natural Language Processing (NLP):** BERT, BERTweet, SBERT, GPT
- **Deep Learning Models:** Convolutional Neural Networks (CNNs), Transfer Learning (ResNet, DenseNet, VGG, etc.), Long Short-Term Memory (LSTMs), Vision Transformer (ViT)
- **Machine Learning Techniques:** Regression, Classification, Gradient Boosting, Tree-based Algorithms, Clustering, Time Series Analysis (Seasonal ARIMA, STL Decomposition, Exponential Smoothing, and Prophet), Hyperparameter Tuning (Bayesian Optimization, Optuna, Random Search)
- **Statistical Methods:** Confidence Intervals, Bayesian Methods, Frequentist Methods for Hypothesis Testing (p-values)
- **Cloud Platforms:** AWS, GCP
- **Web Development:** Amazon EC2, AWS SageMaker, React Native, AWS Lambda, AWS Amplify, AppSync, S3, DynamoDB
- **Big Data Tools:** Azure Databricks, PySpark, Kafka, Hadoop, Presto, Git
- **Data Visualization:** Tableau, Plotly, Dash, Seaborn, Matplotlib, Databricks Dashboard
- **DevOps & Workflow Automation:** Docker, Airflow

EXPERIENCE

S&P Global Commodity Insights — **Irvine, CA (Remote)**
Data Analyst II — Sept. 2021 – Present

- Lead end-to-end projects monitoring natural gas pipeline flows, adapting throughput and specialized models to evolving geopolitical, weather, and economic shifts.
  - Oversee full throughput lifecycle - building, updating, and retiring groupings - to align with market dynamics, collaborating with cross-functional teams to streamline workflows.
  - Designed Databricks dashboards for anomaly detection using unsupervised ML (Isolation Forest, LOF, One-Class SVM) and GPT-4 to generate automated reports; improved detection accuracy by **95%** and speed by **200%**.
  - Built NLP models (Sentence-BERT, GPT-4) to automate post-merger gas meter alignment, reducing reconciliation time by **80%** and enhancing dataset accuracy.
- Defined 5 core metrics (bias, absolute/relative error, etc.) that improved oil & gas price forecasts for **20+ hedge funds**, driving more informed investment decisions.
  - Engineered ETL pipelines (Databricks, SQL, Python, PySpark, Airflow) to evaluate forecasts vs. actuals for **15 global benchmarks** (e.g., Brent, WTI), with automated weekly execution.
  - Automated comparative trend reporting in Tableau, eliminating **5+ hours weekly** of manual reporting.
- Trained/mentored junior analysts on Databricks, Python, and PySpark, cutting pipeline runtimes from weeks to minutes and raising team efficiency.
- Board member and speaker for **S&P Global Women in Technology**, promoting diversity and advancing machine learning adoption.

Austin Capital Data — **Austin, TX (Remote)**
Teaching Assistant in Python ML and NLP Bootcamp — Jan. 23, 2023 – Feb. 17, 2023

- Supported 75+ CDC data practitioners in live-stream training, answering in-class technical questions on Python ML and NLP.
- Led real-time breakout sessions, guiding trainees through coding exercises for ML and NLP projects.

S&P Global Market Intelligence — **Boulder, CO**
Senior Data Researcher — May 2018 - Sept. 2021

- Leveraged machine learning and data science to analyze global power generation across 10 regions and 190+ countries.
- Developed interactive Python Plotly Dash dashboards to analyze the impact of fossil fuels on the energy market.
- Engineered ETL pipelines for power and utility data, cutting processing time from days and weeks to under 10 minutes.

PROJECT

Please visit www.heesukj.com for more projects

WasteWizard with Computer Vision, UC Berkeley (Capstone Project) April, 2024
https://github.com/heesukjang/WasteWizardWithComputerVision
Developed a waste classification model (recyclable vs. non-recyclable) using advanced ML models, including ViT, CNNs, and Transfer Learning (ResNet, DenseNet, VGG, etc.), achieving 90% F1 score and 91% precision, with ViT as the top performer. Deployed to a web interface using Amazon EC2, React Native, AWS Lambda, AppSync, S3, and DynamoDB.

Flight Departure Delay Prediction, UC Berkeley December, 2023
https://github.com/heesukjang/PredictingFlightDelays
Predicted 15+ minute flight delays 2 hours in advance on GCP using PySpark, MLflow, MapReduce on Databricks with models like CNNs, Random Forest, MLP, and Bayesian-tuned XGBoost, achieving 69% F2 score, 90% precision, and 65% recall on 60M+ records , with Bayesian-tuned XGBoost as the top model.

Climate Change Dashboard, UC Berkeley August, 2023
https://github.com/heesukjang/ClimateChangeDashboard
Developed an interactive dashboard using Tableau, D3.js, and Vega-Altair to provide a comprehensive analysis of the global impact of climate change across four key performance indicators: global temperature changes, sea level rise, financial and risk metrics for a low-carbon economy, and the financial and human toll of climate-related disasters.

Improve English Essay with Artificial Intelligence, UC Berkeley May, 2023
https://github.com/heesukjang/ImproveEnglishEssayWithAI
Developed an Automated Essay Scoring (AES) system for grade 8-12 English Language Learners (ELLs) using BERT and BERTweet, evaluating essays on cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Achieved an MCRMSE loss score of 46%, providing tailored feedback to support ELLs.

Invasive Ductal Carcinoma (IDC) Breast Cancer Classification, UC Berkeley December, 2022
https://github.com/heesukjang/IDCBreastCancerClassification
Classified IDC Breast Cancer histopathology images as cancerous or non-cancerous using XGBoost and CNNs with Transfer Learning (VGG19, ResNet50, etc.). ResNet50 with the last layer unfrozen achieved a 0.76 test accuracy, 0.74 validation accuracy, and 0.92 AUC.

Global Energy Trend with Fossil Fuel Price Prediction, S&P Global Market Intelligence 2020
Developed an interactive web interface with Python Dash and Plotly to analyze global energy trends and predict fossil fuel prices using Seasonal ARIMA, STL Decomposition, Exponential Smoothing, and Prophet models.

EDUCATION

University of California, Berkeley – Master of Information and Data Science GPA 4.0/4.0 May 2024

Relevant Courses:

- Research Design and Applications for Data and Analysis
- Data Science Programming
- Statistics for Data Science
- Applied Machine Learning
- Data Engineering
- NLP with Deep Learning
- Data Visualization with Tableau, D3.js, and Vega-Altair
- ML at Scale (Handling Big Data with the MapReduce Framework in GCP, Azure, and Databricks)
- Capstone (WasteWizard with Computer Vision)

University of Colorado, Denver – Master of Business Administration (MBA) in Finance GPA 3.7/4.0
University of Colorado, Boulder – BS, Computer Science GPA 3.8/4.0