# Kunj Mehta

New York, NY | +1 848-437-1589 | kunjcmehta@gmail.com | linkedin.com/in/kunjmehta

### **TECHNICAL SKILLS**

- Languages: Python, Java, C#, SQL, JavaScript, Angular, Autosys (JIL), Unix
- Frameworks and Libraries: FastAPI, Pydantic, SQLAlchemy, PySpark, Airflow, Kafka, PyTorch, HuggingFace, pandas, matplotlib, scikit-learn, streamlit, NLTK, Flask
- Cloud and Engineering: AWS EMR, AWS RDS, AWS Sagemaker, Lambda, EC2, S3, Docker, Jenkins
- Databases and Visualization: TimescaleDB (Postgres), MySQL, MongoDB, SQLServer, AWS Redshift, Looker
- Certifications: <u>Financial Risk Manager Part I</u>, Cloud Practitioner (AWS), LookML Developer (Looker), Machine Learning Engineer (Udacity), Deep Learning Specialization (Coursera)

#### EXPERIENCE

**Software Engineer** | Barclays Services Corp | *New York, NY* 

- **Software Development:** *Led and mentored a team of two interns* working on a PoC to modernize backend reporting stack from SQLServer to a columnar store for cross-asset reporting
- Optimized existing queries and data models following best practices to make them on <u>average 2x faster on</u> <u>columnar store while consuming on average 75% less storage</u> compared to SQLServer
- Presented the PoC findings to stakeholders, securing buy-in for production use that will save \$200K+/year
- Automated generation of *four* Excel reports using Python and Autosys, thereby *decommissioning VBA application Built and enhanced 10+ API endpoints in MVC fashion* using C#, Python, FastAPI integrating the UI, orchestrator
- and backend of the IMA RFET project that facilitates calculation of risk-weighted assets (RWAs) by stakeholders
- Built CI/CD pipelines using Jenkins, Nolio to automate deployment of API endpoints across multiple environments
- Led the design of a two-step, batched ETL pipeline for ingesting raw trade data from multiple sources using Kafka and Autosys that *increased resiliency to network errors and eliminated the need for manual monitoring*
- Leveraged my risk regulatory expertise to coordinate user acceptance testing and enable prod deployment by *streamlining the product testing process and resolving 10+ bugs* in the product raised by stakeholders
- Devised unit testing strategies for the orchestration workflow that increased test coverage from 20% to 55%
- **Machine Learning:** Developing a generative text summarization framework that leverages upstream NER and uses knowledge distillation and LLMs to summarize financial news articles
- Designed a test suite and performed A/B tests to capture multiple relevant NER and summarization metrics for 10+ teacher-student model pairs
- Designing an active learning sampling strategy to continuously retrain the teacher summarization model and improve NER coverage in generated summaries

#### Data Scientist Intern | Eluvio | Berkeley, CA

#### Jun 15, 2022 – Aug 26, 2022

Sep 1, 2021 – May 1, 2023

Aug 14, 2023 – Present

- Worked on improving the media meta-tagging framework for movie distribution on the blockchain
- Logo Classification: Designed and restructured the logo classification pipeline in PyTorch from supervised to metric learning paradigm, elevating the model classification capacity 5x to 500 classes, achieving 67% accuracy
- Conducted 25+ A/B testing experiments using diverse base classification models and super-resolution blocks to select a champion model that reduced the *false positive rate of zero-shot classification from 10% to 2%*
- **NFT Recommender:** Pioneered and developed an *end-to-end ML pipeline for an NFT recommender system* and deployed an alpha version of the system using Docker *capable of handling 3,000 concurrent requests*

## Part-time Teaching Assistant | Rutgers University | New Brunswick, NJ

• Taught R, SQL and Amazon Redshift, and guided understanding of statistical concepts for 200 students

## **PERSONAL PROJECTS**

#### Food AI | Multi-modal Representation Learning (Python, PyTorch, Hugging Face, seaborn)

- Built a system for retrieval of food recipes using images of corresponding food items
- Surpassed the CCA baseline top-10 recall for recipe retrieval in the <u>im2recipe</u> paper by 20 percentage points using ResNet and BERT feature extractors and introducing cross-modality through a shared embedding layer [<u>Repo link</u>]
  New York Taxi Fare Prediction | Big Data (Python. pandas, matplotlib, PySpark, AWS EMR, AWS EC2)

## Performed feature engineering to focus on trip fares to and from airports and across different boroughs of NYC

Predicted taxi fares to a margin of +/- 2 (RMSE 4.28), training a Random Forest on the augmented data [Repo link]

**FPL Teammaker** | **Data Analysis** (*Python, NumPy, pandas, streamlit, matplotlib, PuLP*)

- Designed, and deployed on Streamlit, an algorithm that performs exploratory data analysis on the English soccer Fantasy Premier League (FPL) game data to suggest an optimal team to be entered into the game each week
- Consistently observing 50+ monthly active users. Team ranked top 2% out of 8.2 million players in 2020 [Repo link]

## **EDUCATION**

Master of Science in Computer and Information Sciences | Rutgers University, New Jersey | GPA – 3.95/4May 2023Computer Science Exchange Program | Princeton University, New JerseySep 2022 – Dec 2022Courses – Cloud Computing, Databases, Algorithms, Data Structures, Machine Learning, Computer Vision, AI, NLP