Kunal Kapur

408-239-3398 kunalkapur888@gmail.com https://www.linkedin.com/in/kunal-kapur1/ https://github.com/kunal-kapur

EDUCATION

Purdue University **December 2024**
B.S. in Computer Science and B.S. Data Science – GPA: 3.92 West Lafayette, IN
Relevant Coursework

- Foundations of Deep Learning
- Data Mining and Machine Learning
- Data Structures and Algorithms
- Computer Architecture
- Analysis of Algorithms
- Systems Programming
- Time Series
- Artificial Intelligence

SKILLS

Python | Java | C/C++ | Linux | Bash | JavaScript | AWS | Docker | Pandas | PyTorch | PySpark | SQL | Git | R

EXPERIENCE

State Farm PCM (Property and Casualty Underwriting Modeling) **May 2023 – August 2023**
Software Engineer Intern Remote
Tools: Python, AWS (S3 and Sagemaker), MLflow, Docker, Statsmodels
The PCM department develops statistical and machine learning models to assess the risk of insuring customers

- Reduced hours of manual labor by 70% and saved over a million dollars in EC2 costs by building a pipeline that automated the packing and unpacking of hundreds of trained general linear models
- Wrote Pytests to validate a given model throughout its final deployment phases, achieving a 98% test coverage
- Generated productivity gains by automating the processing of over 50 input and output data transformations for production models
- Created functionality for a Sagemaker processing script to facilitate the testing of models in a production environment

Presto Automation **May 2022 – August 2022**
Software Development and Data Analytics Intern San Carlos, CA.
Tools: JavaScript, Google Analytics, Pandas
Presto offers AI and machine learning-enabled restaurant automation to improve productivity and experience

- Developed tools and performed data analytics on thousands of data points pertaining to website data
- Created an ROI Calculator for the purchase of pay-at-table tablets that did calculations using 11 user-entered values
- Trained an ARIMA model that forecasted a production boost amongst fast food chains in light of a recession with the probability of error due to white noise being less than 0.1

The Data Mine, Purdue University **August 2021 – December 2022**
Undergraduate Researcher and TA West Lafayette IN.
Tools: Databricks, PySpark, Python, SQL

- Collaborated with Indiana University Health to classify 4 social determinants of obesity through clinical notes
- Made a module to preprocess over 6000 unique words and used topic modeling with 72 keywords to do classification
- Later served as a TA where I led a project involving forecasting hospital encounters following COVID where I coordinated work and provided technical guidance for PySpark and SQL for 9 students

PROJECTS

MojifyMe | *Python, PyTorch, Flask*

- Used a CNN (convolutional neural network) to predict which of 3 different emojis best matched a facial expression and hosted the model with flask for others to use
- Built the CNN with 3 hidden layers and a flattened input vector length of 800 and trained it using 3000 images from Kaggle
- Performed cross-validation to do hyperparameter tuning and achieve an 80% test accuracy

Guess the Song | *Python, Flask, Beautiful Soup, SQLAlchemy, JavaScript*

- Developed a front end with JavaScript that involved a user guessing a song title and artist based on lyrics they prompted for
- Used the Spotify API to query the top 50 Spotify songs and subsequently web scrape for their corresponding lyrics everyday, storing all the results on a SQLite database
- Used Flask to link the database and handle requests to access a random song

CMR Image Segmentation | *Python, PyTorch*

- Worked with 4 students to train an LSTM recurrent neural network that could segment the left ventricle using a dataset of 500 CMR images of patients with DMD (duchenne muscular dystrophy)
- Performed transformations to triple the size of the data set and add noise to increase the robustness of the model
- Achieved an average dice score (metric for segmentation accuracy) was 0.82 for the segementation

Twitter Figure Analysis | *Tweepy, NLP*

- Trained a sentiment model using a data set of 50 thousand tweets to identify the sentiment (positive or negative) towards a celebrity based on the requested tweets that mentioned them with a test accuracy of 0.91
- Used Tweepy (wrapper API) to request tweets that mentioned a chosen celebrity and provide the overall sentiment on said figure