# 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**
Bachelor of Science in Computer Science and Data Science – GPA: 3.92 West Lafayette, IN
*Relevant Coursework*

- Foundations of Deep Learning
- Computer Vision
- Machine Learning Theory
- Data Structure & Algorithms
- Analysis of Algorithms
- Artificial Intelligence
- Systems Programming
- Time Series
- Statistical Theory

## SKILLS

Python | Java | C/C++ | Bash | JavaScript | React | AWS | Docker | Pandas | PyTorch | TensorFlow | 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% by building a pipeline that automated the packing and unpacking of hundreds of trained **general linear models** to better facilitate their movement from experiment/evaluation phase to deployment
- Wrote **Pytests** to validate a given model throughout its final deployment phases, achieving a 98% test coverage
- Increased productivity by automating the processing of 50+ input/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
- Built a **JavaScript**-based ROI Calculator for pay-at-table tablets, processing 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 finding such residuals assuming they were white noise being 0.9, indicating good model performance

**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
- Served as a **TA** for 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 Assessment on DMD patients** | *Python, TensorFlow, Scikit-Learn*

- Worked with 4 students to develop a machine learning model that could segment the left ventricle of patients with DMD (duchenne muscular dystrophy) from CMR (cardiac magnetic resonance) imaging
- Wrote scripts to pre-process and augment data with different transformations and increase the data set size by a factor of 4
- Constructued a U-Net model that achieved an average dice score (metric judging segmentation) of 90% on test data

**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 76%
- Used **Tweepy** (wrapper API) to fetch tweets mentioning a chosen celebrity and analyze their overall sentiment