KISHORE KP, PH.D.

DATA SCIENTIST AND ML ENGINEER

Kishore KP, Ph.D

442-377-4717
kishorekp1818@gmail.com

Work Authorization:
US Citizen

Python – 10+ years
R – 10 years
SQL – 2 years
Matlab – 8 years

GitHub Profile
https://github.com/Kishore1818

EDUCATION:
   PhD, in Physics
   (Radar signal processing &
Data Analysis)
   MSC in Physics
   BSC in Phys, Math & Stat

Professional Summary

9+ Years Data Science and ML Engineer

I am an experienced data scientist with 9+ years of experience that leverages a deep understanding of machine learning, Artificial Intelligence, statistical, and mathematical techniques to propel business performance and extract maximum value from across several key domains.

- Experience in the application of Naïve Bayes, Regression analysis, Neural Networks/Deep Neural Networks, CNN, RNN, Natural Language Processing (NLP), Support Vector Machines, and Random Forest machine learning techniques.
- In-depth knowledge of statistical procedures that are applied in both Supervised and Unsupervised Machine Learning, Deep Learning, and Artificial Intelligence problems
- Experience in advanced statistical and predictive models on big data sets using cloud/cluster computing assets with AWS and Azure
- Experience in RapidMiner automodel for data exploration and modeling production.
- Creative thinking/strong ability to devise and propose innovative ways to look at problems by using business acumen, mathematical solutions, data models, and statistical analysis.
- Experience implementing industry standard analytics methods within specific domains and applying data science techniques to expand these methods, for example, using Natural Language Processing methods to aid in normalizing vendor names, implementing clustering algorithms, and deriving novel metrics.
- Experience in text processing (text detection, text recognition), Image processing, Image classification, labeling, DPI encoding for image quality optimization, blur correction, skewed text correction.

Technical Skills

Machine Learning Natural Language Processing & Understanding, Image Recognition, and Detection, Forecasting **Artificial Intelligence (AI)** Text understanding, classification, Pattern Recognition, computer vision, Large Language model (LLM).	**Deep Learning** Data Mining, Machine Learning Algorithms, Convolution Neural Networks (CNN), Recurrent Neural Network (RNN), Multilayer Perceptional Neural Network (MLPNN), Long Short-Term Memory (LSTM), Support Vector Machines (SVM), and Random Forest (RF).
Data Modeling Bayesian Analysis, Statistical Inference, Predictive Modelling, Linear Modelling, Probabilistic Modelling, Time-Series Analysis.	**Analysis Techniques** Classification and Regression Trees (CART), Gradient Boosting Machine (GBM), TensorFlow, PCA, Regression, Naïve Bayes.
Applied Data Science Natural Language Processing, Machine Learning, classification, Social Analytics, and text recognition.	**Analytic Development** Python, R, IDL, SAS, SQL, MatLab
IDE Jupyter Notebook, Spyder, RStudio, Google Colab, sublime.	**Version Control** GitHub, Git
Packages and Visualizations Numpy, Pandas, SciPy, TensorFlow, PyTorch, Keras, Theano, Caffe, Matplotlib, Seaborn, ggplot, Tableau, and Plotly.	**Cloud Data Systems** AWS (RDS, S3, EC2, Lambda), Azure, GCP
Deployment Azure, AWS, Docker, Kubernetes, and Jenkins	**NLP** NLTK, Spacy, Gensim, Apache, Bert, Elmo, Bert3
Analysis Methods Advanced Data Modelling, Forecasting, Statistical, Sentiment, Stochastic, Bayesian analysis, Regression analysis, Linear models, Multivariate analysis, Sentiment analysis, Big data, and clustering.	**Soft Skills** Excellent communication and presentation skills; ability to work wee with stakeholders to discern needs accurately; leadership, mentoring, and coaching

Professional Work Experience

Data Scientist
Qualcomm
San Diego, California

Jan 2022 – present

I joined as a Data Scientist in the QTL-TECH VAS team at Qualcomm. I worked with the data science team where we collected image data and built convolutional neural network architectures, audio classification and object detection models using TensorFlow and PyTorch libraries on the Linux GPU execution environment.

- In the PUBG mobile game, player landing, and several icon identification purposes I used a pre-trained MobileNet multi-class model and achieved nearly 97% accuracy.
- Images were collected from the Call of Duty (COD) mobile game and cropped to obtain bullet number images (0-300) with different color backgrounds, and roughly equal data was collected for each class.
- Generated a classification model using both TensorFlow and PyTorch and the model weights were saved for DLC conversion using snpe-net tools.
- Tested model accuracies on the system side and the target side using both weights. We found PyTorch model accuracies were similar on both the system and target sides.
- Identification of bullet icons and gun icons on screen were done using object detection with YOLOv5 and YOLOv7.
- Participated in audio data collection and cleaning of different types of explosions in COD game and the generated audio classification model was trained and tested.
- Working on signal strength prediction using RSRP values. Finding the wifi signal strength on devices in regions near San Diego. I collected the datasets and merged the GPS and signal strength data sets to the nearest GPS time.
- Endeavored multiple approaches for predicting signal strength in few seconds in advance using ARIMA, and statistical methods.

Data Scientist
Verizon, Basking Ridge
New Jersey (remote)

Jun 2022 – Dec 2022

Verizon provides communications, technology, information, entertainment products, and services to businesses. I worked as a data scientist in their computer vision development team. I extracted frames from game videos and worked on object detection models to identify the basketball and player locations on each frame.

- Worked on data extracting from videos, pre-processing, and cleaning the data to perform feature engineering and data imputation.
- Trained an object detection model built on YOLOv5 that scored a 96.4% test accuracy.
- I collected different sizes and colors of basketball images to increase amount of balanced data among different classes
- I retrained model and tested and validated with precision as I realized F1 score was a bad metric for ball predictions, as it did not curb the number of false positives as well as precision.
- My object detection model was realized ahead of the 6-month pipeline and performed with a 95% accuracy in production making it one of the state-of-the-art models used for computer vision in Verizon's entertainment sector

Senior Data Scientist
FORM.COM
Los Angeles, California (Remote)

Jun 2020 – May 2021

Headed the Data Science team at Form.Com. The team was in charge of developing a computer vision and evaluation application to identify products being sold in individual stores. I implemented a Convolutional Embedding system to identify brand name products and movements in the cooler after successive image captures.

- Develop a custom dataset for fine-tuning a deep neural network.
- Fine-tune a variety of image models with object-detection heads.
- Deploy finished model on edge devices using Tensorflow-Lite.
- Build various statistical models Statistical algorithms involving Time Series analysis, Survival Analysis, Multivariate Regression, Linear Regression, Logistic Regression, and PCA in financial projection
- Lead the development of the expected profit projection engine by applying machine learning with financial engineering and actuarial science.
- Build various statistical models: Statistical algorithms involving Time Series analysis, Survival Analysis, Multivariate Regression, Linear Regression, Logistic Regression, and PCA in financial projection.
- Perform in-force management, including survival analysis, churn analysis, and risk identification.
- Use pre-trained models to visualize feature maps in the intermediate layers and perform transfer learning.
- Use pre-trained models (VGG16, ResNets, U-Net, etc.) for transfer learning on small datasets.
- Lead various cross-department projects and work closely with internal stakeholders such as business teams, product managers, and engineering teams.
- Work on customer segmentation using an unsupervised learning technique clustering.

WATSON NLP Engineer
IBM
San Jose, California (Remote)

Aug 2019 - Apr 2020

Spearheaded IBM's Physical Documentation WATSON initiative. The project entailed processing thousands of scanned documents using a combination of Convolutional Neural Networks and Natural Language Processing, and worked with Word2Vec, BERT, LDA (Latent Dirichlet Analysis), ELMO (bidirectional LSTM), and Spacy (NER). This project entailed converting optically scanned documents into usable data for a later consumption by WATSON. The project involved using CNNs for document classification, AWS Textract, and Google Tesseract for OCR and NLP techniques for Named Entity Recognition (NER) and text document processing.

- Worked with a team to implement a high quality human interactive system NLP and other Deep Learning techniques.
- Performed classification on text data using NLP fundamental concepts, including tokenization, stemming, lemmatization, and padding using millions of documents.
- Constructed an NLP-based filter utilizing embedding and LSTM layers in Tensorflow and Keras.
- Used of a variety of NLP methods for information extraction or named entity recognition, topic modeling, and relationship extraction using NLTK and Spacy.
- Some of my NLP models using Python:

    o https://github.com/Kishore1818/Neural-Networks-and-Deep-Learning-models

- Collected the image/text data, identified the proper features in the data, image pre-processing, and data split into train and test for build the Deep Neural Network and Artificial Intelligence models.
- Designed and built document summarization models using Python and Tensorflow to scan large documents and extract/summarize key facts and items from the scanned documents.

Kishore KP, Ph.D.   949-231-7905 | kishorekp1818@gmail.com

Senior Data Scientist
The Weather Channel
Atlanta, Georgia

Sep 2015 - August 2019

Heavy rainfall prediction is a major problem for the meteorological department, as it is closely associated with the economy and daily life. For this project, I developed several time-series models to predict rainfall across several distinct regions using a large amount of historic data from 1901-2015. These predictions were then additionally used to generate advance warnings in advance for natural disasters like floods and drought across the globe.

- Served as Project Manager. Led five engineers and three data scientists in this NASA/UCI joint project.
- Built and maintained dashboard and reporting based on the statistical models to identify and track key metrics and risk indicators.
- Used Python and IDL to retrieve the historical hierarchical format (HDF5) and clean data prior to implementing and model training.
- Wrote functions to perform pre-processing to impute the missing values using a linear interpolation technique.
- Established normalization of features in the data to reduce noise and maximize signal-to-noise ratio.
- Feature selection/reduction using Principal Component Analysis (PCA) to avoid multi-collinearity issues.
- Worked on data cleaning and ensured data quality, consistency, and integrity using Pandas and Numpy.
- Data stationarity purpose: I removed the long-term trend, seasonality, and periodicity from each region/location of the time series.
- Data stationarity validated with the Dickey-Fuller test.
- Applied Multiple Linear Regression and ARIMA Machine Learning models for rainfall prediction.
- Strong seasonality favored ARIMA performance.
- Final prediction accuracy was 86.67% and the F-measure value was 0.88 for estimating the efficiency of the model.
- Used Machine Learning techniques to predict future long-term trends using historical datasets.

Kishore KP, Ph.D.   949-231-7905 | kishorekp1818@gmail.com

Data Scientist
US Bank (Personal Project)
Irvine, California

For this project a request evaluation system was developed to ingest customer requests from a diverse set of digital and handwritten sources, which were then filtered based on urgency and forwarded to the appropriate department. Handwritten sources were pre-processed using Optical Character Recognition (OCR) techniques, and then handed to the classification model in a hierarchical approach. Relevant departments were identified using metadata such as form and request type determined by the source. I used Python for developing Machine Learning models and Deep Learning models.

- Machine Learning models of Bayesian and KNN techniques along with Tesseract were compared for OCR applications based on model accuracy and speed.
- Models achieved 91%, 96%, and 97% accuracy, respectively.
- OCR model performance evaluated on textual and MNIST datasets.
- Tesseract provided the most consistent OCR results and was used in the productionized solution.
- Sorting subsequently done through the training and testing of an artificial neural network.
- Data Cleaning, Imputation, Tokenizing: used Python libraries (Pandas, NLTK, Numpy, Keras, Tensorflow, and PyTorch) to clean and prepare the data for analysis.
- Performed data integrity checks, data cleaning, exploratory analysis and feature engineering using R and Python.
- Performed machine learning, natural language processing (NLP), and statistical analysis methods such as clustering and classification.
- Urgency identified using a Natural Language Processing-based classifier.
- Final Deep Learning Bidirectional LSTM model achieved 85% test accuracy for identifying urgent vs. non-urgent.
- Production model deployed to a flask API for use by the business.
- Worked with a core-team of 4-6 individuals of Data Engineers, Data Analysts and Data Scientists.

Personal Project

Covid-19 CT Scan Analysis

Procured Covid Patient Lung CT Images and developed a CNN model to identify Covid Positive Lung Morphology. The idea was to develop a machine learning method to aid radiologists in detection. At the time, there were very few tests available. The Developed model could identify the illness correctly 78% of the time and had a recall of 0.9.

- CT scan image data analyzed using Deep learning models of Convolution Neural Networks (CNNs).
- Dataset consisted of an even split of 743 COVID and non-COVID images from the medical office.
- Improve patient care standards and reduce workloads by extracting essential information from literature, manuals, and research articles using Natural Language Processing (NLP).
- Generated some of COVID-19 models and visualizations:
  https://github.com/Kishore1818/Neural-Networks-and-Deep-Learning-models

  https://github.com/Kishore1818/Animations

- Generated Visualizations from SAS data using R:
  https://preethamvignesh57.wixsite.com/mysite/covid-19
- Generated Tableau visualization of Covid-19 Data:
  https://preethamvignesh57.wixsite.com/mysite/tableau

Education

Ph.D in Physics
Radar Signal Processing and Data Analysis
Sri Venkateswara University
Tirupati, Andhra Pradesh, India

Masters of Science in Physics
Sri Venkateswara University
Tirupati, Andhra Pradesh, India

Bachelor of Science in Mathematics, Physics, and Statistics
Sri Venkateswara University
Tirupati, Andhra Pradesh, India

Machine Learning, Deep Learning, Artificial Intelligence models, codes, and visualizations:
https://github.com/Kishore1818