Hari Kosaraju

614.246.1858 [| harinkosaraju@gmail.com](mailto:|%20harinkosaraju@gmail.com) | New Albany, OH, 43054

**Objective:**

Dedicated and highly skilled Machine Learning Engineer with a strong background in supervised and unsupervised machine learning, deep learning, data analysis, and model deployment. Seeking a challenging position to leverage my expertise in developing innovative solutions and contributing to data-driven decision-making.

**Status:** US Citizen

**EDUCATION**

* M.S. in Data Analytics, Franklin University, Columbus, OH
* Masters in Business Administration – Finance, Franklin University, Columbus, OH
* Bachelors in Computer Engineering, Pune University, Pune, Maharashtra, India

**CERTIFICATIONS**

* AWS certified in Machine Learning.
* Certified in Practice of Data Analytics (CPDA), The Ohio State University, Columbus, OH.
* Certified in Machine Learning.
* Certified in Deep Learning.
* Certified in Data Mining.
* Certified in Statistics.
  + Certified in Visual Analytics for Sense Making.
* Certified in Project Management Professional (PMP) (Current).
* Certified in ITL4 foundation.
* Certified in SAP HANA.

**Artificial Intelligence (AI) – Competitions/Awards/Presentations**

Participated in couple of Artificial Intelligence competitions on Drivendata.org.

* **Wind-dependent Variables (2021):** Predict Wind Speeds of Tropical Storms, in 2018, the NASA IMPACT team launched an experimental framework to investigate the applicability of deep learning-based models for estimating wind speeds in near-real time. Was ranked 48 and placed on leaderboard of the competition. Deep learning with pytorch lightning interface used for the competition.
* **Pump it Up (2021 – 2022):** Data mining the water table, using data from Taarifa and the Tanzanian Ministry of Water, a ML model needs to predict which pumps are functional, which need some repairs, and which don't work at all. Currently ranked 378/13600. Competition still live. Sci-kit learn frame work used to build Machine Learning models in Python..
* **Presentation (February, 2021)**: Gave presentation on Deep Learning/Computer Vision CNN to audience at my work place. This was part of my company’s innovation’s team effort to showcase, how it’s innovation team’s goals are aligned with knowledge of artificial intelligence.

## Trainings

* The Machine Learning Pipeline on AWS.
* SAP HANA – TZH300 – SAP HANA Implementation and Data Modeling.

## knowledge base

|  |  |
| --- | --- |
| * AWS ML cloud services * Computer Vision(CNN) * Data Mining | * Natural Language Processing * Sentiment Analysis - ML/DL * Legacy Systems |

## lANGUAGES

Python, R, Javascript, SQL/DB2, MongoDB (NoSQL), Django, Flask, Shiny .

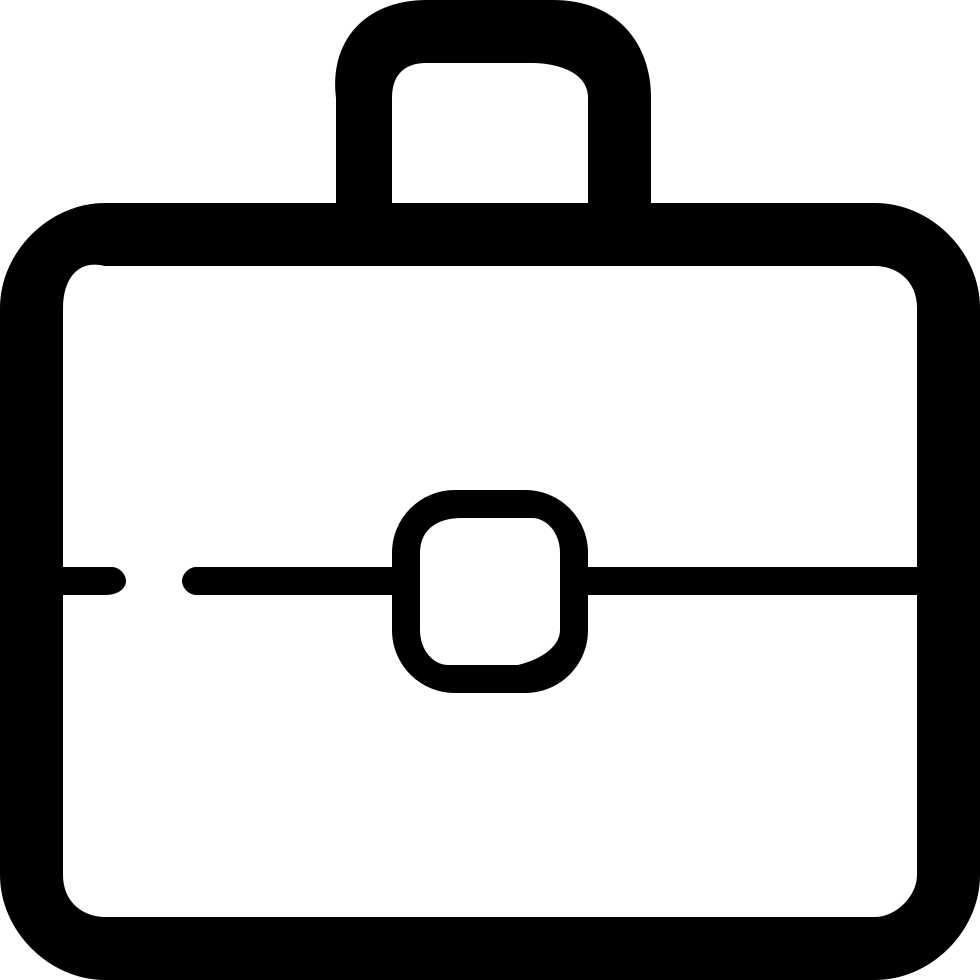
## FRAMEWORKS, packages and interfaces

Sci-kit learn, Pandas, Numpy, dplyr, Auto ML, Pytorch, Pytorch lightning, Tensorflow, Keras, Auto Keras, Keras tuner.

## EDITORS, VISUALIZATION Tools, and IDEs

Jupyter notebook, Jupyter lab, Google COLAB, SAS-Miner, SAS-Viya, SAS-Studio, Tableau, Matplotlib, Seaborn, ggplot2, R-studio, Spyder, Tensorboard, AWS Sagemaker, AWS Sagemaker Studio. **I**

## PROFESSIONAL EXPERIENCE



**Software Engineer at Department of Children and Youth**

State of Ohio (DCY), [Columbus, Ohio] [Nov, 2023 to Current]

Working as an AI Engineer for developing AI/LLMS solutions in Salesforce. Currently working on

Salesforce AI tools 1) Einstein Copilot, and 2) Einstein Bots.

Proper actions/prompts should be built in Einstein Copiolt to optimally use out of box LLMs given by

Salesforce. Also new LLMs are being developed to be used in Salesforce Copiolt as part of

‘bring your own model (BYO)’ concept.

Once proper actions and prompts are put in place, the development of Salesforce coding will

be made very efficient and fast.

Also, DCY being a very new department, I was given responsibility to develop predictive models. Currently we are in process of designing new programs for kids and collecting data for those programs.

**Data Scientist/Machine Learning Engineer**

State of Ohio (JFS/DOI), [Columbus - Ohio] | [May, 2020 – Nov, 2023]

Spearheaded the development of AI predictive models utilizing existing actuarial data, contributing to a range of actuarial applications using both Machine Learning and Deep Learning.

**Supervised Machine Learning Experience:**

* Developed Rgression models to predict healthcare expenditure for diabetic patients.
* Constructed Classification models to forecast customer choices between full coverage and no coverage.
* Utilized Generalized Linear Models (GLM) to predict frequency, severity, and loss ratios for insurance claims.
* Built GLM models to assess policy renewal likelihood and cross-sell opportunities for homeowners and auto insurance policies.
* Created predictive models using longitudinal and panel data for personal lines (e.g., homeowners and auto policies), commercial lines (e.g., trucks), insurance sales (territory-based policy sales prediction), and customer retention.
* Designed regression models to identify customers with severe diseases affecting their insurance policies, employing Generalized Additive Models (GAM) and kernel smoothing techniques.

**Unsupervised Learning Experience:**

* Applied various clustering techniques, including fuzzy clustering, k-means clustering, and hierarchical clustering, to build predictive models for detecting fraud and abuse in insurance claims.

Performed end-to-end life cycle Machine Learning tasks such as 1) Data Preprocessing, 2) Model Creation, 3) Model training, 4) Model Evaluation, and 5) Model Selection using AWS Sagemaker Studio.

* **Data Preprocessing:** Performed exploratory data analysis and data preprocessing for ML modeling using SageMaker Data Wrangler. Data was imported from Amazon S3, Amazon Athena, Amazon Redshift, and Snowflake.Using AWS Sagemaker Data wrangler, data was preprocessed which included treating for for missing values in both numerical and categorical data columns, bringing data to scale (min-max scaling,z-standardization), treated for outliers, treating data with various algorithms to remove skewness in each numerical data column. Also building a Feature Repository with SageMaker Feature Store which was used by other models. Pandas was used for data preprocessing.
* **Visualizations:** Generated data visualizations using Matplotlib and ggplot2 to illustrate various plots and insights.
* **Model Creation:** Implemented regression, classification, clustering, and ensemble machine learning algorithms such as randomforests, XGboost, k-means, KNN, decision trees, SVM, Bayesian, PCA, and Gaussian process.
* **Model Training:** Managed training jobs with SageMaker Experiments. SageMaker Experiments was used to track experiments among various training runs.
* **Model Evaluation:** Various model evaluation techniques implemented for Regression, Clasification, and Cclustering. Used ROC and Confusion matrix for evaluating classification models. Used least squares and

Maximum likehood estimation for evaluating regression models. Amazon SageMaker Clarify was used to discover underlying bias in the training data and model prediction and explain feature importance for an ML model. Created learning curves to learn how models were performing.

* **Model Selection:** Used gridsearch, meta tuners, stacking and other model selection techniques to select best model out of many models created.

**Development Environment: AWS Cloud**

Used Python, and R as development languages. Used various Python frameworks such as Scikit-Learn for Machine Learning and Keras, Tensorflow, and PyTorch for Deep Learning. Used Numpy, Pandas, SQL and Jupyter notebook instances to develop the applications. Using SageMaker Projects, pipelines, and templates implemented CI/CD ML pipeline and MLOps.

* Leveraged AWS Cloud infrastructure for end-to-end development.
* Utilized S3 storage as the primary data source.
* Employed Kinesis Streams to capture clickstream and application logs data.
* Performed real-time data analysis using Kinesis Analytics with SQL.
* Managed data storage and distribution using Kinesis Firehose, storing data in S3, Redshift, Elasticsearch, and Splunk.
* Configured Glue crawlers to automate schema inference and partition extraction for data stored in S3, Redshift, and RDS.
* Established a centralized Glue Data Catalog to store data tables, enabling automated schema inference and versioning.
* Integrated Glue Data Catalog with Redshift Spectrum, EMR, and Athena for enhanced data accessibility.
* Utilized Glue ETL for data cleaning, transformation, and enrichment before storage in S3, with applications in AWS SageMaker for model training.
* Set up AWS EMR clusters for ad hoc data querying using PySpark SQL.

**Data Scientist/Machine Learning Engineer**

Kore.ai, [Remote - Florida] | [May 2018 – April 2020]

* Developed AI Models for Asset Management for a bank where artificial intelligence was leveraged to optimize investment strategies and enhance portfolio performance within a given timeline. Timeseries MLforecasting Regression algorithms like Linear Regression, Regularized Linear Rregression, Random Forest Regressor, Gradient Boosting Regressor, and Decision tree Regressor algorithms were used to build AI models to predict the performance of portifolio/assets.
* AI models for fraud detection in banking applications related to credit cards were developed. Based on the behavioral input data of a customer, the AI model needs to predict if the customer can be issued a credit card or no. If model predicts that customer is good, then credit card can be issued.

**Big Data Development and AWS Ccloud Environment:**

**Data Ingestion:**

Set up data pipelines using technologies like Apache Kafka, and AWS Kinesis to ingest data from various sources.

Integrated data connectors and adapters to seamlessly collect data from databases, and APIs.

**Data Processing:**

Utilized batch processing frameworks like Apache Spark and Apache Flink for large-scale data transformation and analysis.

Employed stream processing with tools like Apache Kafka Streams and Apache Beam for real-time data processing.

**Data Orchestration and Workflow:**

Automated data workflows using tools like Apache Airflow and cloud-based solutions like AWS Step Functions.

Orchestrated complex data processing pipelines to ensure data quality, reliability, and timeliness.

**Data Storage Optimization:**

Implemented data compression techniques and columnar storage formats (e.g., Parquet, ORC) to optimize storage costs and query performance.

Managed data partitioning and indexing strategies to enhance data retrieval efficiency.

**Software Development Specialist - 3**

State of Ohio (SETS), [Columbus - Ohio] | [May 2012 – April 2018]

* Primary responsibilities included data & systems analysis, fraud detection, reporting and mitigation.
* Delivered significant financial savings thru in-depth data analysis of fradulent claims and build the protections within the processing of claims including targeting the channels sourcing the invalid claims data.
* Developed the data analysis platform to identify the unemployment data analysis and identified the County and local regions to target re-tooling programs to reduce the unemployment in the region.
* Built data reporting systems to recoupe the money paid in excess, track missing payments, daily payments to  custodial parent, daily payments received from absent parents etc.
* Responsible for daily ad-hoc reports, design and develop SETS reports for internal use, weekly, monthly, quarterly, and yearly federal reports.
* Software development in legacy systems (COBOL/DB2). Worked on SETS allocation, distribution, locate, and case management subsystems.