**Venkata Sai Phaneesha Chilaveni**

[venkatasaiphaneeshachilaveni@gmail.com](mailto:venkatasaiphaneesha@gmail.com) | [linkedin.com/in/phaneesha-chilaveni](http://linkedin.com/in/venkata-sai-phaneesha-chilaveni) | <https://github.com/phanee16> | +17164269247

**EXPERIENCE**

**Software Engineer (Data Science), MSR technologies, IL, USA.**   **Oct 2023-Present**

Projects/Client: Illinois Secretary of State

* **Proficient** in end-to-end data analysis and model development using **Python (NumPy, Pandas, Scikit-learn), Jupyter** **Notebook, SQL, and RESTful APIs** for data cleaning, feature engineering, model evaluation, and delivery of high-quality insights. Optimized data pipelines, reducing processing time by 25%.
* Designed and implemented interactive dashboards with **Power BI, Tableau, and SQL (Azure SQL Database)** for efficient data retrieval, management, and comprehensive reporting, translating complex data into actionable insights. Dashboards led to a 15% increase in data-driven decision-making.
* Led projects to clean and preprocess large datasets (up to 10 TB), enhancing data integrity and system efficiency by leveraging Python (Pandas), **Apache Spark on Azure Databrick**s, and **Azure Data Lake Storage** for efficient data processing and management. Reduced data processing time by 30%.
* Developed and optimized predictive models for customer behavior analysis using **Python (Scikit-learn, TensorFlow, Matplotlib**), ensuring accurate predictions (up to 90% accuracy) and user satisfaction (NPS score of 8.5/10).
* Designed and developed a predictive analytics project for customer churn prediction with **Python, SQL, IBM DB2, Docker,** and **Azure Kubernetes** Service for data integration, model deployment, performance, and scalability. Reduced customer churn by 20%.
* Collaborated with cross-functional teams of 8+ members, aligning data science projects with strategic goals through effective communication, **Agile methodologies (Jira),** and **CI/CD practices (Azure DevOps)** for continuous integration and deployment. Delivered projects 15% ahead of schedule.

**Data Scientist-Team Lead, Community Dreams Foundation, NY, USA**   **Apr 2023 – Oct 2023**

* **Led** a team of skilled data analysts in conducting in-depth analysis of energy data, utilizing advanced **Exploratory Data Analysis (EDA**) techniques to identify key insights and apply ML models for predictive analysis.
* Implemented an efficient and scalable database infrastructure, integrating cutting-edge web scraping technologies for real-time data acquisition and management.
* Employed **ad-hoc analysis** and reproducible analytical approaches to extract descriptive statistics from complex data sets. Applied advanced statistical models to analyze complex data sets, creating interactive dashboards in **Tableau** and **PowerBI** to visualize and communicate intricate patterns and trends effectively.
* Developed a customized chatbot for personalized user interactions, leveraging state-of-the-art **Large Language Models (LLMs)** like **Langchain** and **GPT-Turbo 3.5**, showcasing innovative AI applications including geospatial analyses and project management.
* Engineered a robust database solution on **Google Cloud Platform (GCP)**, utilizing **Vertex AI** for ML models and **BigQuery** for ad-hoc analysis, descriptive statistics, and geospatial analyses.
* Implemented **Apache Kafka** and **Apache Spark** for streaming and processing large volumes of data in real-time, significantly enhancing decision-making speed and accuracy in reporting key energy metrics and trends.
* Pioneered the integration of machine learning technologies like **TensorFlow** and **scikit-learn**, improving predictive models in energy data analysis through AI. Analyzed extensive datasets, boosting accuracy and efficiency.
* As a project management expert, I strategically led cross-functional teams, aligning initiatives with business goals. Utilizing Agile methodologies and collaborating with business users, I efficiently deployed data-driven solutions, presenting impactful results.

**Data Scientist Intern (Remote), Marvel Technology Solutions, NY, USA**  **Jun 2022 - Aug 2022**

* Developed and implemented a robust data science-driven product recommendation system, utilizing ML models and ad-hoc analysis. Presented results of reproducible analytical approaches, boosting click-through rates by **15%** and driving higher user engagement and increased sales.
* Leveraged **Google BigQuery** and GCP to streamline data processing, reducing processing time **by 40%,** leading to faster and more responsive product recommendations.
* Utilized **Apache Spark** and **TensorFlow** integrated with **GCP** to efficiently process and analyze 20% more data, enabling the system to scale with growing user demands.

**Data Scientist, Sarag Systems, Hyderabad, India.**    **May 2019 – July 2021**

* Proficient in **Python** programming, leveraging **NumPy, Pandas, and Scikit-learn** for data manipulation, analysis, feature engineering, and machine learning model development.
* Experienced in **SQL** querying and data management using **MySQL**, enabling efficient data extraction, transformation, and loading processes for large datasets.
* Applied customer segmentation techniques using **K-Means and Hierarchical** clustering algorithms from Scikit-learn, enabling targeted marketing campaigns and personalized product recommendations.
* Developed predictive models utilizing Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting algorithms for churn prediction, customer lifetime value estimation, and sales forecasting.
* Adept in data visualization using Tableau for effectively communicating insights through interactive dashboards, informative visualizations, and data storytelling.
* Conducted extensive feature selection and dimensionality reduction using techniques like PCA, Lasso, and Ridge Regression to enhance model performance and interpretability.
* Performed data preprocessing tasks, including handling missing data, outlier detection, and data normalization, to ensure high-quality inputs for machine learning models.
* Utilized cross-validation techniques, such as **k-fold and stratified cross-validation**, to evaluate model performance and prevent overfitting.
* Collaborated with cross-functional teams, effectively communicating complex data insights and model interpretations to stakeholders and business users.

**Research Intern, Indian Institute of Technology, Hyderabad, India**  **Feb 2017 - Nov 2019**

* Collaborated with a PhD Scholar to design and fabricate a **Passive Dynamic Walker** using **Solid Works** and **Laser beam machine**, achieving a remarkable 20% increase in efficiency through successful design analysis and optimization.
* Conducted a detailed simulation of the walker's performance on a 30 inclined slope using **MATLAB**, resulting in a highly optimized design.
* Presented the project at the **Connaisance** Conference, where the innovative design and efficiency improvement were showcased to an audience of over 50 professionals.

**SKILLS**

**Programming Languages*:*** Python, R, Java, C, SQL, HTML/CSS, MATLAB, Bash, COBOL, JCL, Julia, JavaScript

**Databases:** MySQL, MongoDB, IBM DB2

**Tools:** Tableau, Excel, Snowflake, Apache Spark, Jupyter Notebooks, GCP, Git, Docker, Kubernets, Apache Airflow, Talend

**GCP stack**(BigQuery, Looker ,Data Studio, AutoML, Datalab, Dataproc),

**AWS stack**( S3, Redshift, SageMaker, Glue, Lambda, EMR),

**Azure stack(**Data Factory, Databricks, Datalake, SQL Data Warehouse, PowerBI, COSMOS DB, Data Lake storage) ,

**IBM stack**( Watson studio, Watson Machine learning, Cloud Park, Db2, SPSS statistics, Cognos AI Analytics, Streams, InforSphere Information Server, Big SQL, DSX)

**Frameworks:** Django, Flask, Keras, TensorFlow, Pytorch, Matplotlib, Seaborn, Scikit-learn, Streamlit, Numpy, Pandas, Caret, Tidyverse, Shiny, RandomForest, XGBoost, Apache spark, Kafka, Hadoop, Flink, plotly, ggplot2, Bokeh, Tableau.

**EDUCATION**

**University at Buffalo, The State University of New York, MS in Data Science** **Aug 2021 – Feb 2023**

* Maintained a **4.0 GPA** while completing relevant coursework in Statistics, Data Analysis, Data Mining (R programming Language), Databases (SQL), Machine Learning, Numerical Analysis, and Data structures and Algorithms.

**JNTUH College of Engineering, Bachelors and Masters in Mechanical Engineering** **Aug 2015 – Jan 2021**

* Maintained a **3.9 GPA** while completing relevant coursework in Statistics, Higher Mathematics, Thermal Engineering, Product Design, Advanced manufacturing systems, Simulation, Modeling and Analysis, Optimization Techniques.

**RESEARCH PROJECTS**

**Travelling Salesman Problem (TSP)** [GitHub Link](https://github.com/phanee16/Traveling_Salesman_Problem_real-time)

***Tech Stack*:** Python (Scikit-Learn, Folium, Geo-Py, Image Io), OpenStreetMap API, Google Colab

* Developed a solution for the Traveling Salesman Problem, incorporating innovative technologies like the Nearest Neighbor algorithm, OpenStreetMap API, and advanced navigation techniques.
* Designed an interactive geographical map that visually illustrates optimized routes, simplifying complex optimization problems and offering seamless navigation for users.

**Forecasting Risk Gene discovery in Autism with Genome Scale Data** [GitHub Link](https://github.com/phanee16/-Comparative-Analysis-of-Boosting-Algorithms-for-Autism-Detection-using-Genome-Data-)

***Tech Stack:*** R (Caret, GGplot2, Random Forest, GBM, XGBoost, AdaBoost), RStudio

* Replicated and improved the analysis methodology from the research paper "Forecasting risk gene discovery in autism with machine learning and genome-scale data" by Brueggeman et al. (2018).
* Conducted a thorough comparative analysis of ensemble learning algorithms (e.g., XGBoost, AdaBoost) to enhance the prediction of autism risk genes, achieving an 84% AUC-ROC score and reducing the training error to 0.02795.

**Web application on NYC Collision Analysis**  [GitHub Link](https://github.com/phanee16/-Web-Application-for-Analyzing-NYC-Collision-Data-)

***Tech Stack:*** Python, Streamlit, PyDeck Google Cloud Platform, BigQuery, SQL, Looker Studio

* Extracted NYC collision data from NYC Open Data and analyzed it using GCP BigQuery. Visualized the results with Looker Studio and built a web application with Streamlit to share the insights.

**Revolutionizing Clothing Categorization with CNNs** [GitHub Link](https://github.com/phanee16/-Revolutionizing-Fashion-Classification-with-CNN-A-Deep-Learning-Approach-to-Enhance-Clothing-Categ)

***Tech Stack:*** Python (Pandas, NumPy, Matplotlib, TensorFlow, Keras)

* Created and trained Convolutional Neural Networks with TensorFlow on a Fashion-MNIST dataset. Initiated a base model with 3- layer neural network architecture and tuned the hyper-parameters for better results.

**Reinforcement Learning in Grid World: A SARSA Approach** [GitHub Link](https://github.com/phanee16/-Comparative-Analysis-of-Boosting-Algorithms-for-Autism-Detection-using-Genome-Data-)

***Tech Stack:*** Python (Numpy, gym, google\_colab)

* Implemented a flexible and adaptable reinforcement learning project that involved creating a Grid Environment class, SARSA\_Agent function, and render function to enhance the navigation and reward outcomes for an agent navigating through a 5x5 grid environment using SARSA algorithm.

**Highway Traffic Data Integration and Real-time Streaming Pipeline for Toll Plaza Analysis**

***Tech Stack***: Apache Airflow, Bash, Apache Kafka, Zookeeper, Simulators, Python,

* Developed and implemented a data pipeline using **Apache Airflow** to download, extract, transform, and consolidate data from various file formats (CSV, TSV, fixed width) with DAG definition, data extraction, transformation, and pipeline submission.
* Configured and managed a streaming data pipeline using Apache Kafka, including setting up **Zookeeper**, starting **Kafka** server, creating a topic, downloading, and configuring the Toll Traffic Simulator, and running the streaming data reader script and performed health checks to ensure the smooth functioning of the pipeline.

**ETL and Machine Learning (Edx)**

***Tech Stack***: *PySpark, Apache Spark, Elyra, IBM Watson , GitHub*

* Utilized the HMP dataset to develop a machine learning model, leveraging the open source CLAIMED library for data extraction, transformation, and loading; model creation was facilitated by **Apache Spark** and stored to Cloud Object Store.
* Employed the **Elyra JupyterLab** extension for editing notebooks and pipeline design, showcasing the utility of **IBM's Watson Studio Orchestration Flow** tool for cloud-based, end-to-end data science workflows.

**CERTIFICATIONS**

* **IBM Data Engineering Professional Certificate.**
* **IBM Data Analyst Professional Certificate.**
* **Google Data Analytics Professional Certificate.**
* **Snowflake – Data Engineer workshop.**
* Data Visualization with Tableau Specialization.
* Prompt Engineering for ChatGPT.