

Sreenidhi Hayagreevan

Data Scientist Intern

Santa Clara, CA, United States | [LinkedIn](#) | [Portfolio](#) | [GitHub](#)

Data scientist and machine learning engineer with expertise in LLM systems, RAG architectures, and big data processing. Currently pursuing a Master's in Applied Data Intelligence at SJSU.

EXPERIENCE

Chatgpt Brand Ambassador

OpenAI

Aug 2025 — Present

San Francisco Bay Area

- Selected as a ChatGPT Ambassador through the CSU × OpenAI partnership, supporting the rollout of ChatGPT Edu to 460,000+ students across 23 CSU campuses.
- Planned and conducted “Campus Pulse,” an interactive on-campus Q&A session with students focused on real-world AI usage, challenges, and expectations.
- Collected actionable feedback and student reviews on AI performance, usability, and academic impact to inform responsible AI adoption.
- Educated students on ethical, healthy, and safe AI use within an academic environment.

Instructional Student Assistant – Machine Learning & Database Technologies

San José State University

Aug 2025 — May 2026

United States

Evaluated over 100 graduate-level assignments on ML algorithms, Python, and SQL for technical correctness and code quality. Provided feedback on feature engineering and model selection. Reviewed database design including ER/EER models, normalization, and advanced SQL concepts like window functions.

Data Scientist Intern

Spiritual Data

Jul 2025 — Sep 2025

United States

Designed a hybrid RAG architecture combining Pinecone vector search and Neo4j knowledge graphs with evaluation metrics for scalable chatbot responses. Collaborated with the RLHF team to align model outputs with human feedback and created workflow visualizations that reduced onboarding time by 30%.

Data Analyst - SEO

The Hindu

Jun 2020 — Dec 2021

Chennai, Tamil Nadu, India

Conducted keyword analysis, competitor benchmarking, and audience segmentation to support content strategy.

Implemented on-page and off-page optimization strategies, improving mobile page speed from 15 to 65 and desktop from 25 to 75. Reduced page load time from 4 seconds to 1 second and cut bounce rate by 60%.

Data Analyst Trainee (SEO)

The Hindu

Dec 2018 — Jan 2019

Chennai, Tamil Nadu, India

- Worked closely with the development team, and helped extract all the urls from the main website using python
- Learned to use Google Analytics tools, and ways of data extraction from it
- Learned client server relationship and http client
- Learned to use various SEO tools, plugins and data extraction methods

EDUCATION

San José State University

Master of Science - MS in Applied Data Intelligence

GPA: 3.73/4.0

Aug 2024 — May 2026

University of Michigan

Certification in Python Programming

Grade: A

May 2024 — Jul 2024

Wharton Online

Specialization Certification in Advanced Digital Marketing and Growth Strategies

Advanced Digital Marketing course offered by Wharton Online, helped me gain a profound understanding of digital marketing strategies and tools. This comprehensive program covered strategic planning, content marketing, SEO, social media marketing, email marketing, PPC advertising, analytics, and conversion rate optimization. I also learned to stay up-to-date with the latest industry trends. Real-world case studies provided practical insights, ensuring a well-rounded education. Overall, the course has equipped me with the skills and knowledge to excel in the dynamic field of digital marketing, from crafting effective strategies to implementing data-driven decision-making.

Nov 2023

Thiagarajar College of Engineering

Bachelor of Technology - BTech in Information Technology

Grade: 3.4

Activities and societies: Smart India Hackathon, Comparing for various college functions , Student president of information Technology Department, Industrial Visit coordinator, Cultural club leader.

PROJECTS

NLP pipeline for extracting vehicle attributes from text descriptions, Named Entity Recognition for Vehicle Attributes Extraction

Built an NLP pipeline on the FindVehicle dataset with 42K descriptions and 21 entity classes. Converted CoNLL-style annotations to BIOES-style tagging and benchmarked RNN, LSTM, BiLSTM, and transformer architectures. Fine-tuned RoBERTa-base with AdamW, outperforming all baselines.

Wearable health data processing and risk segmentation using Spark, Big Data-Driven Health Risk Assessment

Processed 141M+ wearable health records totaling 17GB using PySpark on HDFS with Parquet format. Engineered risk features and identified key correlations across 127K+ users. Built K-Means segmentation and Linear Regression risk-scoring model deployed via Streamlit for real-time prediction.

Content and Behavior-Aware Movie Recommendation System

- Designed a full-stack web app for customized movie recommendations, leveraging subtitle data from IMDb API.
- Integrated both content-based and behavior-aware filtering to deliver tailored suggestions.
- Built an NLP pipeline to extract genre and summary metadata from subtitles.
- Analyzed implicit user behavior to support session-aware and dynamic content personalization
- Deployed the application on Amazon Amplify, enabling fast, real-time recommendations with 80% accuracy and maintaining computational efficiency.

Job Scam Detection in Gmail

- Built ML system to detect fraudulent job emails using Enron, survey, and synthetic data
- Engineered 15+ features (e.g., scam/urgency keywords, sender metadata) & trained models (LogReg, SVM, RF, XGBoost)
- Achieved 96.7% accuracy & 73.4% F1 with Random Forest; used SHAP for interpretability
- Applied SMOTE + Stratified K-Fold CV to address class imbalance

Retail Demand Forecasting System

- Developed an end-to-end demand forecasting pipeline using Snowflake for data storage, Apache Airflow for automation, and Superset for interactive dashboards, optimizing retail inventory and decision-making.
- Engineered advanced features from historical data and real-time sources like OpenWeatherMap, Census and FRED, using models such as SARIMA, XGBoost, and LSTM to predict demand with MAPE < 15%.
- Achieved a 20% reduction in inventory carrying costs, 50% fewer stockouts, and improved decision-making efficiency by 40%, empowering retailers to respond quickly to market trends.

Spotify Data Analytics: Trends and Forecasting

- Analyzed Spotify's top hits data to identify trends in track popularity, artist influence, and evolving music preferences over 3 decades.
- Designed an interactive Power BI and Tableau dashboard to visualize key metrics, including audio attributes and artist contributions, enabling actionable insights.
- Conducted EDA to uncover patterns in song characteristics, genre evolution, and their correlation with success on streaming platforms.
- Applied machine learning techniques, achieving 90% accuracy in forecasting factors influencing track popularity and success metrics.

Stock Price Forecasting Project

- Extracted 90 days of stock price data for two companies Netflix and Intuitive Surgical using the Alpha Vantage API.
- Designed and automated ETL process workflows using Airflow DAGs and stored the extracted data in Snowflake database.
- Developed an additional Airflow DAG for machine learning-based forecasting, predicting stock prices for the next 7 days with an accuracy of 96.01% for Netflix and 98.57% for Intuitive Surgical, and stored the predictions in a new Snowflake table.
- Used the last 90 days of historical data to perform calculations like RSI, moving average (7 days, 30 days), and price momentum using dbt and ran it as an Airflow DAG.
- Created in-depth visualizations in Preset based on the new calculated fields to analyze stock market performance for both companies.

Global Health Data Visualization & Analysis

- Extracted and analyzed real-time global data on critical diseases from the WHO, spanning from the early 1990s to 2024.
- Conducted in-depth exploratory data analysis (EDA) and data cleaning to ensure data quality and consistency.
- Designed and developed an interactive Tableau dashboard to visualize and explore the relationships between Tuberculosis and HIV.
- Identified key countries with effective disease detection and treatment strategies, providing valuable insights for global health initiatives.

Wharton Online Capstone Project, UpLIFTTS

- Developed an omnichannel digital marketing strategy for UpLIFTTS, a fictional social media marketing company
- Enhanced the social media presence, visibility and provided sentiment analysis, influencer suggestion and competitor

analysis by offering an advanced AI tech tool

--Curated a marketing channel mix using PESO model and created customer acquisition and retention strategies
--Created customer journey insights, buyer's persona, industry research report and implemented technology-driven approaches.

Web and Social Media Marketing Optimization for E-Commerce Company

--Leveraged data-driven insights through website and social media analytics and optimized digital marketing strategies for a newly launched E-Commerce company
--Analyzed traffic acquisition, user demographics, technology reports and derived user retention strategies for the website.
--Identified social media goal, prioritized best performing social media platform, scheduled posts and increased social media visibility

Omni-Channel and Customer-Centric Retail Strategy for Takeout & Delivery Business

--Designed an omni-channel retail strategy for a takeout & delivery business
--Incorporated new digital business models and created customer centric acquisition, retention strategies
--Enhanced customer journey by utilizing customer lifetime value insights, created effective acquisition, and retention strategies resulting in increased engagement, enhanced loyalty, wider reach, and improvised brand perception
--Built CAC, TAM reports for store location and using ROAS increased the advertising budget

Farmie (mobile app and web app)

- Ideated and built an app to enable farmers sell directly to their end consumers without the need for mediation
- Created a buyer seller portal, chatbot, weather prediction, plant disease predictor, seasonal crop suggestion, location based service, push notifications, and payment gateways
- Tools used: Eclipse, salesforce cloud, machine learning

Flight Pass Booking System

- Constructed a website to purchase flight pass
- Website was designed to allow the user to create an account, set multiple filters and make a purchase
- Retrieved various Airline data information via Google scholar
Tools used: Anaconda, Spyder. Algorithm used: Random forest

Multi-layer safety framework for LLM outputs with self-improving evaluation, GuardRail for LLM Safety Framework

Built a 4-layer guardrail system with input validation, RAG grounding, Chain-of-Debate multi-agent evaluation, and output verification. Processed samples through an Airflow ETL pipeline. Implemented output-stage verification for hallucination, knowledge base conflicts, and bias/privacy checks with S3-backed reflection memory and human feedback loops for threshold auto-tuning.

Custom speech synthesis model conditioned on emotion embeddings, Emotion-Conditioned Text-to-Speech using Tacotron2 & HiFi-GAN

Built a custom Tacotron2 encoder fusing text and emotion embeddings from LJSpeech and EmovDB datasets for emotion-conditioned speech synthesis. Achieved stable convergence of loss from 4.38 to 0.20. Pretrained and fine-tuned HiFi-GAN on Tacotron2-generated mels to reduce domain gap.

CERTIFICATIONS

Fundamentals Of Deep Learning , NVIDIA Deep Learning Institute	Jan 2026
Perform analytics in Power BI , Microsoft	Oct 2024
Python Specilization - Capstone project , University of Michigan	Jul 2024
Using Databases with Python , University of Michigan	Jul 2024
Using Python to Access Web Data , University of Michigan	May 2024
Python Data Structures , University of Michigan	May 2024
Programming for Everybody (Getting Started with Python) , University of Michigan	Mar 2024
Wharton Advanced Digital Marketing And Growth Strategies , Wharton Online	Nov 2023
Search Engine Optimization (SEO specialization) , University of California, Davis	Oct 2022
Google SEO Capstone Project , University of California, Davis	Oct 2022
Optimizing a Website for Google Search , University of California, Davis	Sep 2022
Google SEO Fundamentals , University of California, Davis	Sep 2022
Introduction to google SEO , University of California, Davis	Sep 2022

Advanced SEO Strategy , University of California, Davis	Aug 2022
Advanced Content and Social Tactics to Optimize SEO , University of California, Davis	Jul 2022
Digital Marketing Strategy and Planning , Digital Marketing Institute	Jun 2022
Digital Leadership and Digital Strategy Execution , Digital Marketing Institute	Jun 2022
The fundamentals of Digital Marketing , Google Digital Garage	Apr 2022
SEO fundamentals , Semrush	Apr 2020
Google Ads Display Certification , Google	

SKILLS

English · Hadoop · React.js · Amazon Web Services (AWS) · Node.js · HTML & CSS · BERT (Language Model) · C++ · Apache Superset · Microsoft Power BI · Pandas (Software) · Tableau · docker · Apache Airflow · Google Cloud Platform (GCP) · Core Java · HTML · Python (Programming Language) · Synthetic Data Generation · Data Build Tool (DBT) · Generative AI · Support Vector Machine (SVM) · Random Forest · Apache Spark · k-means clustering · Logistic Regression · Streamlit · EDA · Data Warehousing · Big Data Analytics · Natural Language Processing (NLP) · Machine Learning · Software Testing · Data Visualization · Large Language Models (LLM) · Machine Learning Algorithms · Reinforcement Learning · Artificial Intelligence (AI) · Demand Forecasting · Supply Chain Optimization · Database Management System (DBMS) · SHAP · SMOTE · K-Fold cross validation · TF-IDF · Evaluation Metrics · Teamwork · Easily Adaptable · Relationship Development · Team Motivation · Strategic Thinking · Skilled Multi-tasker · Negotiation · Communication · Problem Solving · Leadership · Analytical Skills · SQL · JavaScript · ETL/ELT Pipelines · PySpark · Snowflake · Pinecone · Neo4j · MySQL · NumPy · scikit-learn · PyTorch · Classification · Regression · Clustering · GANs · RNNs · LSTMs · NER · Transformers · RAG · LangChain · OpenAI API · Prompt Engineering · Chain-of-Debates · Multi-Agent Systems · Preset · R Studio · Matplotlib · Seaborn · Git · Kubernetes · Linux/Unix · Jupyter · VS Code · CI/CD · Object-Oriented Programming · Agile · Scrum