**Nevedtha Venkataraman**

**rishitha.23g@gmail.com**

**+1 2142998373**

**PROFESSIONAL SUMMARY:**

* **Overall 6 years of experience as Data Scientist and AI Engineer (4 years). Strong expertise in statistical modeling, data mining, data exploration, and visualization of structured and unstructured datasets, with a focus on AML/CFT transaction monitoring.**
* **Proficient in utilizing advanced language models such as GPT-3 and Hugging Face Transformers for text generation, summarization, sentiment analysis, and various other tasks.**
* **Demonstrated expertise in data sciences, mining, modeling, deep learning, CNN, clustering, boosting, classification, predictive modeling, LSTM, and LLMs.**
* **Machine Learning and Deep Learning:** Strong mathematical and statistical knowledge with hands-on experience in traditional and advanced algorithms like K-Nearest Neighbors, Logistic Regression, Linear Regression, Naive Bayes,SVM, Decision Trees, Random Forests, Gradient Boosted Decision Trees, Neural Networks, CNN, RNN, Autoencoders, Deep Generative Models, GANs, Computer Vision algorithms, Deep Reinforcement Learning.
* **Technical Proficiency:** Skilled in Python programming, Data Engineering, network security (VPC, Private/Public Subnet, NAT Gateway), microservices, Docker, Kubernetes, CloudFormation, ETL, and cloud computing (AWS Certified Cloud Practitioner). Experienced with AWS services like XGBoost, Keras, ResNet, VGG-16, Lambda, EMR, Hadoop, Spark, Kinesis (Streams, Firehose, Analytics), S3, Athena, Redshift, and SageMaker.
* **Model Deployment and API Integration:** Designed model endpoints using serverless functionality and API-gateway. Proficient in integrating AI models with RESTful APIs for seamless communication between web application components.
* **Team Collaboration and Leadership:** Led a team of offshore developers, designing algorithms and data processing pipelines, and managing IAM, API endpoints, and Lambda ETL.
* **Data Management and ETL:** Extensive experience in data modeling, collection, cleansing, ETL, BI, OLAP, data warehousing, and data lakes. Proficient in creating ETL mappings and scheduling jobs using Airflow for ETL batch processing into Snowflake, ensuring high data quality and integrity.
* **Big Data and Cloud Computing:** Hands-on experience with Hadoop, Apache Spark, Kafka, and cloud platforms (AWS, GCP, Azure). Proficient with EC2, S3, RDS, Lambda, Snowflake, and big data technologies.
* **Data Visualization and BI Tools:** Created interactive dashboards and reports using Tableau, Power BI, and D3.js to facilitate data-driven decision-making.
* **Programming and Scripting:** Proficient in Python, R,SQL, JavaScript, TypeScript, and shell scripting. Strong in data structures, algorithms, and object-oriented programming. Experienced in developing software applications using Python, Machine Learning, AI, Dialogflow CX, Generative AI, and Azure OpenAI.
* **Database Management:** Skilled in managing SQL and NoSQL databases (PostgreSQL, MySQL, Oracle, Teradata, MongoDB, Cassandra). Experienced in PL/SQL and database object creation.
* **Advanced Analytics and Statistical Analysis:** Expertise in hypothesis testing, ANOVA, PCA, time series analysis, correlation, and multivariate analysis. Proficient with scikit-learn, TensorFlow, and PyTorch. Strong in data preprocessing, Feature Engineering, and Model Evaluation Techniques.
* **DevOps and MLOps:** Implemented robust testing practices for ML Ops pipelines and ensured reliability and compliance with standards. Familiar with CI/CD pipelines and model deployment processes.
* **Text Mining and NLP:** Extensive experience in Text Mining , Topic Modeling, and Sentiment Analysis using TextBlob, NLTK, and SpaCy. Proficient in employing NLP techniques for tasks like Text Generation, Data Extraction, Named Entity Recognition, and Sentiment Analysis.
* **GenAI and LLMs**: Demonstrated familiarity with GenAI technology stack, including prompt engineering frameworks, guardrails for GenAI applications, and fine-tuning LLMs for optimal performance.
* **Frontend and Backend Development:** Proficient in front-end technologies (HTML, CSS, JavaScript) for developing interactive user interfaces and using Python and Flask for developing robust backend systems to support AI models.
* **Cloud and Containerization:** Experienced in using cloud platforms like AWS and Azure for deploying and scaling AI models. Familiar with Docker for containerizing AI applications and Kubernetes for orchestrating containerized deployments.
* **Network Security and Protocols:** Integrated PKI with enterprise applications for secure communication. Understanding of protocols like TCP/IP and proficiency in using Wireshark for network traffic analysis.
* **Documentation and Compliance:** Clear and comprehensive documentation of data sources, models, processes, and KPIs. Implemented data masking and encryption techniques to protect sensitive data, ensuring compliance with data privacy laws like GDPR.
* **Publications and Research:** Published research in top AI-related venues (NeurIPS, ICML, ICLR, AAAI, CVPR, MICCAI). Proven ability to troubleshoot and find solutions to complex issues, contributing to optimized model performance.
* **Professional Skills:** Strong analytical skills, attention to detail, critical thinking, problem-solving, excellent communication, and the ability to manage multiple priorities and meet deadlines. Skilled in storytelling and presenting complex ideas to non-technical audiences.
* **Healthcare AI Applications:** Developed AI solutions for medical imaging and healthcare data, improving diagnostic accuracy and patient outcomes. Experience with medical imaging data such as fundus photographs, OCTs, CT, and MRI.
* **Project Management and Agile Development:** Experienced in agile software development, project management, and leading multidisciplinary teams. Familiar with version control systems like Git and software development practices.
* **Adaptability and Continuous Learning:** Ability to quickly learn and adapt to new data analysis tools and techniques. Knowledge of popular LLMs (OpenAI GPT, LLaMA, PaLM) and using LangChain for GenAI applications.
* **Generative AI and Cloud Solutions:** Demonstrated the differentiation of Google Cloud through Generative AI-driven applications. Experience in using AWS services like S3, SageMaker for developing and deploying models at scale.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Category** | **Items** |
| **Databases** | PostgreSQL, cloud Bigtable, cloud Firestore, MongoDB, Cassandra, MySQL, HBase, Amazon Redshift, Snowflake, Neo4J, Microsoft SQL Server 2008/2010/2012, Oracle 11g/12c, DB2, Teradata, Netezza, Maria DB |
| **Machine Learning** | Linear Regression, Logistic Regression, Naive Bayes, AI/ML, LLMs, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Means Clustering, K-Nearest Neighbors (KNN), Gradient Boosting Trees, Ada Boosting, PCA, LDA, NLP, XGBoost, LightGBM, CatBoost |
| **Deep Learning** | Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Deep Learning on AWS, Keras API, Transformers, Generative Adversarial Networks (GANs) |
| **Python Libraries** | NumPy, Matplotlib, NLTK, Statsmodels, Scikit-learn/sklearn, SOAP, SciPy, Pandas, Gensim, Keras, TensorFlow, Pytorch |
| **Data Visualization** | Tableau, Python (Matplotlib, Seaborn), R (ggplot2), Power BI, QlikView, D3.js |
| **Languages** | Python (NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn), R, RTML, SQL, PL/SQL, MATLAB, Spark, Java, C#, HTML5, DHTML, WSDL, CSS3, C, C++, XML, Perl, MATLAB, Mathematica, FORTRAN, DTD, Schemas, Json, Ajax, JavaScript, Shell Scripting, TypeScript |
| **Operating Systems** | UNIX, Linux, Windows, Mac OS, UNIX, Macintosh HD, Sun Solaris |
| **Big Data/Hadoop Technologies** | MapReduce, Spark, Spark SQL, Azure, Spark Streaming, Kafka, PySpark, Pig, Hive, HBase, Flume, Yarn, Oozie, Zookeeper, Hue, Ambari Server |
| **Development Tools** | Microsoft SQL Studio, IntelliJ, Azure Databricks, Eclipse, NetBeans |
| **Public Cloud** | EC2, IAM, S3, Auto scaling, CloudWatch, Route53, EMR, RedShift, AWS ML, Lambda, Snowflake, Apache Airflow, Datadog, DataProc, GCP, Azure, Bigquery, Stackdriver, Firebase, Prometheus, Amazon Web Services, Google Cloud, Microsoft Azure, OpenStack, IBM Cloud, Oracle Cloud, Adobe Creative Cloud |
| **Development Methodologies** | Agile/Scrum, UML, Design Patterns, Waterfall |
| **Build Tools** | Jenkins, Toad, SQL Loader, PostgreSQL, Talend, Maven, ANT, RTC, RSA, Control-M, Oozie, Hue, SOAP UI |
| **Reporting Tools** | MS Office (Word/Excel/PowerPoint/Visio/Outlook), Crystal reports XI, SSRS, Cognos |
| **NoSQL** | Document databases, Graph databases, Key-value stores, Wide-column stores |
| **Data Warehouse** | MPP), Columnar data stores, ETL and ELT data integration, BI, AI, and machine learning |
| **Data Pipelines** | Instruction pipelines (classic RISC pipeline), Graphic pipelines units (GPU), Software pipelines (commands, program runs, tasks, threads, procedures), HTTP pipelining |
| **Data and Information Visualization** | Tableau, Power BI, Statistical Analysis, Data Cleaning and Preprocessing, Data Mining and Extraction, Data Modeling and Forecasting. |
| **Scripting** | Python and Ansible |

**EDUCATION**

* Bachelors - BTech in Electronics and Communication - SRM University - Chennai, 2018
* Masters - MBA in General Management- Bentley University- Boston, 2024

**PROFESSIONAL EXPERIENCE**

**Role: AI Engineer Aug 22 - Present**

**Client: Microsoft Corporation , WA**

**Project Description:**

In my role as an AI Engineer at Microsoft Corporation, I led the training and fine-tuning of large language models (LLMs) like GPT, BERT, and T5 for diverse NLP tasks, integrating Langchain with various APIs and databases. I contributed to a 30% reduction in fraud through advanced AI-driven detection and improved claims processing by 15%. I designed and trained transformer models for content generation, implemented cutting-edge techniques for model optimization, and utilized cloud-based GPU clusters for scalable training. Leveraging MLOps tools and cloud services, I ensured effective model deployment and lifecycle management. Additionally, I managed data processing workflows with Apache Spark and Kafka and applied DevOps practices for streamlined CI/CD.

**Responsibilities:**

* Trained and fine-tuned large language models (LLMs) like GPT,BERT, and T5 on diverse datasets for natural language processing (NLP) tasks such as text generation, summarization, and question-answering.
* Integrated Langchain with external APIs and services, such as document management systems, vector databases, retrieval systems and other databases like MongoDB, MySQL, and web scrapers. This integration enabled conversational AI systems to access and process diverse data sources.
* Helped Achieve 30% fraud prevention and mitigation through early detection of potentially fraudulent claims and realized a 15% improvement in overall claims processing and customer satisfaction.
* Designed Gen AI models to continuously analyze vast amounts of transactional data, user behavior patterns, and market trends, enabling the system to proactively identify and flag suspicious activities or outliers that might indicate fraud or other risks in real-time.
* Designed and trained transformer-based language models for open-ended text generation, creative writing, and content creation applications using self-supervised learning techniques. This included experimenting with different model architectures, training objectives, and data augmentation strategies to enhance model creativity and generative capabilities.
* Implemented techniques like contrastive learning, self-supervised pretraining, and transfer learning to improve the performance and generalization capabilities of generative AI models.
* Conducted error analysis and qualitative evaluation of LLM outputs to identify failure modes, biases, and areas for improvement. This involved analyzing model outputs, categorizing errors, and deriving insights to guide model refinement and dataset curation strategies.
* Planned strategic integration of Gen AI into the existing fraud detection system, aiming to further reduce false positives and enhance real-time anomaly detection capabilities.
* Explored the use of Gen AI's adaptive learning capabilities to create a dynamically evolving model, staying ahead of emerging fraud patterns and minimizing financial losses.
* Utilized Jupyter Notebook/Jupyter Lab for exploratory data analysis (EDA), leveraging its interactive capabilities to visualize data using tools like tools like Matplotlib, Seaborn, Tableau, PowerBI, perform statistical analysis, and iterate on data preprocessing and feature engineering tasks.
* Utilized MLOps tools like MLflow, Kubeflow, and TFX for model packaging, deployment, monitoring, and lifecycle management of generative AI models. This included setting up automated pipelines for model training and deployment, tracking experiment results, and managing model versions effectively.
* Implemented advanced techniques like prompt engineering, few-shot learning, and in-context learning to adapt large language models for specific tasks and domains.
* Leveraged cloud-based GPU clusters and distributed training frameworks like PyTorch Lightning and TensorFlow, Kera’s Extended for training and scaling of large generative AI models.
* Collaborated with subject matter experts and domain specialists to curate high-quality datasets and knowledge sources for training generative AI models.
* Designed and Implemented data processing workflows, optimizing performance, and ensuring data consistency and reliability Using Apache Spark and Apache Kafka for processing large-scale data streams and managing distributed data processing pipelines.
* Leveraged AWS services such as SageMaker for model training and deployment, EMR for big data processing, and Redshift for data warehousing. This included configuring and managing AWS resources, optimizing costs, and ensuring scalability and reliability of AI systems deployed on AWS infrastructure.
* Implemented DevOps practices continuous integration and continuous deployment (CI/CD) pipelines using Docker containers, automating the build, test, and deployment processes.
* Wrote complex SQL queries to extract, transform, and analyze data from diverse sources, ensuring data integrity and performance optimization. Applied SQL for data querying and manipulation in relational databases.
* Leveraged Python libraries such as Pandas, NumPy, spaCy, NLTK, and Scikit-learn for data preprocessing, analysis, and machine learning tasks. This included implementing data processing pipelines, feature engineering techniques, and machine learning algorithms to support AI model development.
* Utilized Docker and Kubernetes for containerization and orchestration of AI model deployment. This involved containerizing AI models, managing container deployments, and ensuring scalability and fault tolerance of deployed applications using Kubernetes orchestration.
* Worked on Git for continuous integration and continuous development (CI/CD) to maintaining code repositories, managing branches and merges, and facilitating collaborative development workflows to ensure code quality and project reproducibility.
* Utilized Kubeflow for managing machine learning workflows and pipelines for designing and orchestrating end-to-end ML pipelines, automating model training and deployment processes, and ensuring reproducibility and scalability of ML workflows.

**Environments:** Langchain, GPT, PyTorch, TensorFlow, Kera’s, TFX, MLflow, Apache Spark, Apache Kafka, AWS (SageMaker, EMR, Redshift), GCP (BigQuery, Vertex AI), SQL, Python, Pandas, NumPy, spaCy, NLTK, Scikit-learn, Hugging Face Transformers, Docker, Kubernetes, Git, Kubeflow

**Role: Sr Data Scientist/AI Engineer Jan’21 – Apr’22**

**Client: ZF Tech Centre, India**

**Project Description:**

At ZF Tech Centre India, I led projects as a Sr. Data Scientist/AI Engineer, collaborating with cross-functional teams to develop key metrics and perform cohort analysis, resulting in a 17% reduction in customer churn. I built and optimized churn prediction models using XGBoost, achieving 87% accuracy, and employed time series forecasting techniques, including ARIMA and FBProphet, to predict customer data usage and sales with up to 94% accuracy. I aggregated and analyzed device usage data, enhanced billing models, and created Tableau dashboards for business insights. Additionally, I developed data pipelines using Apache Pulsar and implemented robust model evaluation techniques to ensure performance and accuracy.

**Responsibilities:**

* Collaborated with other teams across Nielsen and developed key metrics to achieve business outcomes
* Collaborated with product managers to perform cohort analysis that identified an opportunity to reduce churn by 17% for a segment of customers
* Working closely with Sales and Marketing teams to understand the need for analytics
* Build customer churn prediction, partial churn prediction and propensity models using XG-BOOST with an accuracy of 87%
* Built Churn Prediction model using Jupyter Notebook, Python and SQL
* Aggregated daily device usage data for all the devices in a fleet to analyse fleet performance over time using Python and SQL
* Developed a forecasting model using time series analysis and machine learning algorithms to predict customer data usage.
* Explored techniques like ARIMA, LSTM, or Prophet to capture temporal dependencies and fluctuations.
* Utilize classification algorithms, such as logistic regression, decision trees, or ensemble methods, to achieve accurate churn predictions.
* Implemented various time series forecasting techniques to predict device usage, which helps in predicting churn.
* Used FBProphet to forecast sales and shipping at country and product level with 92% accuracy to avoid shipping delays
* Working closely with business and engineering teams to encourage statistical best practices with respect to experimental design, data capture and data analysis
* Build predictive models to predict possibility of data overage for IoT Customers
* Used FBProphet to forecast customers data usage with a 94% accuracy rate
* Built IoT data usage models using Jupyter Notebook, PySpark, SQL, Hadoop, and Hive
* Performed enhancements to the existing billing and payment models achieving 7% more accuracy and call volume reduction from 20.2M to 15.4M
* Created Tableau dashboards for quick reviews to be presented to business and end users
* Build data pipelines to extract the customer usage data from relational database and data streaming applications using Apache Pulsar
* Implemented rigorous model evaluation techniques, including cross-validation and metric analysis, to ensure model accuracy and generalizability and optimized model hyperparameters and configurations for improved performance.

**Environment:** ER Studio 9.7, Tableau 9.03, AWS, Teradata 15, MDM, AWS, GIT, Unix, Python 3.5.2,MLLib, SAS, regression, logistic regression, Hadoop, NoSQL, Teradata, OLTP, random forest, OLAP, HDFS, ODS, NLTK, SVM, JSON, XML, MapReduce, Google Dialog Flow.

**Role: Data Scientist Nov’19 – Jan’21**

**Client: Altran Technologies, Renault Nissan**

**Project Description:**

At Altran Technologies for Renault Nissan, I developed an intelligent recommendation system using advanced machine learning algorithms, including collaborative and content-based filtering, to personalize insurance policy suggestions. I led the creation and deployment of a state-of-the-art NLP-driven chatbot for customer support and claims processing, leveraging Google Dialog Flow for entity recognition and chatbot optimization. I developed solutions like “Insurance Insight GPT” and “Contract Negotiator GPT” for document summarization and data extraction. My efforts improved customer engagement by 17% and reduced response times and operational costs by 22%, while also enhancing data-driven decision-making by 25%.

**Responsibilities:**

* Developed an intelligent recommendation system utilizing advanced machine learning algorithms (collaborative filtering, content-based filtering etc.) to analyse customer profiles, historical data, and preferences for generating personalized insurance policy suggestions.
* Implemented a state-of-the-art chatbot driven by Natural Language Processing (NLP) to handle customer queries, provide information on policy details, and assist with claims processing.
* Lead the development and deployment of advanced machine learning models to power the Policy Recommendation System.
* Developed “Insurance Insight GPT” and Contract Negotiator GPT” solutions using LLMs for concise document summarization and Key-Value pair extraction.
* Conducted POCs and fine-tuning of LLMs such as GPT-3.5, LLama2, Flan-T5 to extract information from unstructured health insurance and financial documents.
* Utilized supervised learning techniques to analyse historical data and customer behaviour for accurate policy recommendations.
* Spearheaded the design and implementation of the NLP-driven chatbot, ensuring it understands natural language queries and provides context-aware responses.
* Collaborated with IT teams to seamlessly integrate the Policy Recommendation System and Chatbot with backend systems, including policy databases, CRM tools, and claims processing systems.
* Used Google Dialog Flow for entity recognition and machine learning which involves creating and configuring the chatbot or virtual agent within the Dialog Flow console
* Used the Dialog Flow console to test the chatbot by entering sample user inputs and evaluated how well it recognizes intents and extracts entities.
* Utilized the analytics and insights provided by Dialog flow to understand user interactions and identified areas for improvement and refinement of the chatbot based on user feedback and performance data.
* Elevated customer engagement by providing personalized policy recommendations, fostering a more positive and tailored experience which contributed to a 17% increase in customer satisfaction and engagement levels.
* Established a continuous improvement process for the machine learning models, involving regular retraining and optimization to adapt to changing customer preferences and market dynamics.
* Improved the efficiency of customer support operations by automating routine queries through the chatbot, reducing response times, and increasing customer satisfaction which lead to a 22% reduction in response times and operational costs.
* Worked on creating a conversational and natural interface for the chatbot to enhance user engagement and satisfaction.
* Overseen the design of an intuitive and user-friendly interface for customers interacting with the Policy Recommendation System.
* Incorporated sentiment analysis capabilities to gauge and respond to customer emotions effectively.
* Leveraged data-driven insights generated from customer interactions with the recommendation system and chatbot to inform strategic decision-making and enhance business operations which contributed to a 25% improvement in data-driven decision-making and overall business performance.

**Environment:** TensorFlow, scikit-learn, Apache Spark, Data Governance, CI/CD, SQL Server, ER Studio 9.7, Tableau 9.03, AWS, Teradata 15, ETL, MS Office Suite - Excel (Pivot, VLOOKUP), DB2, R, Python, Visio, HP ALM, Agile, Azure, Data Quality, Tableau and Reference Data Management.

**Role: Data Scientist Jul’18 - Nov’19**

**Client: SRM Technologies, Hitachi Automotive**

**Project Description:**

At SRM Technologies for Hitachi Automotive, I utilized a range of tools and technologies to develop comprehensive data solutions. I worked extensively with R and Python, leveraging libraries such as Pandas and Scikit-learn, and integrated CaffeDeep Learning Framework for machine learning tasks. I implemented end-to-end data analytics systems using AWS, Hadoop, and MongoDB, improving user engagement by 45% and tripling user conversations. My role involved data manipulation, aggregation, and visualization with tools like Power BI and QlikView. Additionally, I developed and optimized data models using Spark, Scala, and various machine learning methods, ensuring high data quality and effective data storage solutions.

**Responsibilities:**

* Worked with various R packages, including knitr, dplyr, SparkR, Causal Infer, and Space-Time, and coded R functions to interface with CaffeDeepLearningFramework.
* Proficiently used Python, using libraries like Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, SciPy, and NLTK to develop machine learning algorithms.
* Worked with CaffeDeep Learning Framework and also worked with different data formats like JSON, XML, and performed machine learning algorithms in Python.
* Performed data manipulation and aggregation from various sources using Nexus, Business Objects, Toad, Power BI, and Smart View.
* Implemented Agile Methodology for building internal applications and focused on integration overlap and Informatica's commitment to MDM.
* Set up storage and data analysis tools in Amazon Web Services cloud computing infrastructure.
* Implemented end-to-end systems for Data Analytics, Data Automation, and integrated with custom visualization tools using R, Mahout, Hadoop, and MongoDB.
* Worked with Data Architects and IT Architects to understand data movement and storage, using ER Studio 9.7.
* Utilized Spark, Scala, Hadoop, HBase, Cassandra, MongoDB, Kafka, Spark Streaming, MLLib, and various machine learning methods, increasing user lifetime by 45% and tripling user conversations.
* Extensively used Spark Data frames, Spark-SQL, Spark MLLib, and developed POCs using Scala, Spark SQL, and MLLib libraries.
* Applied Data Quality Validation techniques to validate Critical Data Elements (CDE) and identified anomalies.
* Worked extensively with Data Modeling tools like Erwin Data Modeler to design Data Models.
* Developed QlikView Data Models by extracting data from various sources, including files, DB2, Excel, Flat Files, and Big Data.
* Participated in all phases of Data Mining, including Data collection, Data cleaning, Developing Models, Validation, Visualization, and Gap Analysis.
* Implemented Classification using supervised algorithms like Logistic Regression, Decision trees, KNN, and Naive Bayes.
* Designed 3NF data models for ODS, OLTP systems, and Dimensional Data Models using Star and Snowflake Schemas.
* Updated Python scripts to match training data with the database stored in AWS Cloud Search for classification.
* Created SQL tables with referential integrity and developed queries using SQL, SQL Plus, and PL/SQL.
* Designed and developed Use Case, Activity Diagrams, Sequence Diagrams, and OOD using UML and Visio.
* Interacted with Business Analysts, SMEs, and other Data Architects to understand business needs and functionality for various project solutions.
* Identified and executed process improvements, hands-on in various technologies like Oracle, Informatica, and Business Objects.

**Environment:** AWS, R, Informatica, Python, HDFS, ODS, OLTP, Oracle 10g, Hive, OLAP, DB2, Metadata, MS Excel, Mainframes MS Vision, Map-Reduce, Rational Rose, SQL, and MongoDB.