**Venkatesh T**

**Data Scientist**

[**venkatesht.2702@gmail.com**](mailto:venkatesht.2702@gmail.com)

**PH: 8127784066**

**Professional Summary:**

* Data Scientist with 8+ Years of experience collaborating with cross-functional teams in data analysis, development, operations, and delivery of process innovation to achieve business goals in the fields of Machine Learning, Statistical Modeling, Predictive Modeling, Data Analytics, Data Modeling, Data Mining, and Business Intelligence.
* Proficient in managing the entire Data Science Life Cycle and actively involved in all the phases of the project including Data Acquisition, Data Cleaning, Data Engineering, Features Selection, Feature Engineering, Statistical Modeling, Testing, and Validation.
* Successfully developed Large-scale Data Platforms encompassing operational data stores, real-time metrics stores, attribution platforms, data warehouses, and data marts. Fostered team collaboration and maintained comprehensive system documentation.
* Extensive knowledge in Data Mining and Machine Learning with large informational datasets of Structured and Unstructured Data using programming languages like Python and R.
* Proficient in various Regression Algorithms such as Linear Regression, Ridge Regression, Lasso Regression, Principal Components Regression, Decision Trees, Random Forests, Ensemble Learning, and Support Vector Machines.
* Proficient in various Classification Algorithms such as Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, Decision Trees, Random Forests, Ensemble Learning, and Support Vector Machines.
* Proficient in various Ensemble Learning Algorithms such as Bagging, Random Forests, and Boosting Methods such as AdaBoost, Gradient Boosting, and XGBoost.
* Proficient in various Clustering Algorithms such as K-Means Clustering, DBSCAN, and Gaussian Mixture Models.
* Proficient in Dimensionality Reduction Algorithms such as Principal Component Analysis, Factor Analysis, and Random Projection.
* Proficient in creating stored procedures, importing, and exporting data using Relational Databases including MySQL, Microsoft SQL Server, PostgreSQL, SQLite, and Microsoft Access.
* Experienced in Data Visualization with Python libraries such as Matplotlib and Seaborn for creating static and dynamic visualizations and dashboards.
* Experienced in Data Manipulation with Python libraries such as NumPy, SciPy, and Pandas for data analysis and numerical computations.
* Experienced in Predictive Modeling with Python libraries such as Scikit-Learn for model development and performance evaluation.
* Experience with data transformation using Extract-Transform-Load/ETL Pipelines for Imputation, Outlier Handling, Transformation, Encoding and Scaling with Apache Airflow.
* Experience in Structured Query Language/SQL programming and developed stored procedures, triggers, user functions, subqueries, and joins for complex queries involving multiple tables with exception handling.
* Strong understanding of Statistical Modeling such as A/B Testing, Experiment Design, Hypothesis Testing, Significance Testing, and Correlation Methods.
* Proficient in data visualization tools such as Tableau, and Power BI to create visually powerful and actionable interactive reports and dashboards with customized parameters.
* Experienced in Normalization and Denormalization techniques for optimum performance in relational and dimensional database environments.
* Experience with Source Code Management/Version Control tools such as GitHub, Data Version Control/DVC, and Docker-Hub to coordinate work with multiple teams.

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, C++, R, Structured Query Language/SQL |
| **Libraries** | NumPy, Pandas, Matplotlib, Seaborn, SciPy, SciKit-Learn, Snscrape, BeautifulSoup, NLTK, Gensim, Spacy, Plotly, Spark MLlib. |
| **Big Data Frameworks** | PySpark, PyTorch, TensorFlow, Keras, Apache Spark, Apache Airflow, Apache Hadoop, Apache Kafka. |
| **Cloud Computing** | Amazon Web Services(S3, IAM, Lambda, API Gateway, SageMaker, Glue, Redshift, EC2, Kinesis), Snowflake |
| **Databases** | MySQL, Microsoft SQL Server, PostgreSQL, SQLite, Microsoft Access |
| **Business Intelligence** | Tableau, Microsoft Power BI, Google Analytics |
| **Integrated Development** | Visual Studio, PyCharm, Anaconda, Jupyter Notebook, RStudio |
| **Version Control** | GitHub, Data Version Control/DVC |

**Work Experience:**

**Blue Cross Blue Shield Jan 2022 – Present Chicago, IL**

**Role: Data Scientist**

**Responsibilities:**

* Worked with a cross-functional team to develop end-to-end data engineering solutions for anomaly detection product.
* Developed data pipeline using ETL method for enabling Machine learning workflow using Python, SQL.
* Experience with statistical methods like Propensity Score Matching, Causal Impact.
* Develop ETL pipeline for extracting, and pre-processing data from using Apache Airflow.
* Utilized Snowflake for data warehousing and worked to extract data from various sources for analysis and modeling.
* Data extraction from database warehouse using PostgreSQL.
* Involved in Data mapping specifying what data will be extracted from an internal data warehouse, transformed, and sent to a central data storage DVC
* Performed text analytics on claims transcript notes using NLP using Latent Dirichlet Allocation (LDA) model to perform topic modelling and enhance existing model.
* Developing experience in Amazon Cloud technologies including S3, Glue, EC2, and Kinesis.
* Big data design experience with technical stacks like Spark, Snowflake, and AWS big data technologies.
* Analyzed and found patterns and insights within Structured and Unstructured Data using Machine Learning Algorithms such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines for prediction.
* Used PyTorch framework for developing the model train and testing pipeline in a Python environment.
* Improved data distribution of the unstructured labeled images using data imputation and techniques such as Synthetic Minority Oversampling Technique (SMOTE) and trained ML models for segmentation and detection to increase precision-recall from ~30% to ~85%.
* Model hyper-parameter tuning and optimization using grid search and random search methods.
* Data Management using central data storage and tracking – Data Version Control (DVC).
* MLOps task: Deployed trained ML models to production using AWS; versioning and monitoring quality using CI/CD.
* Application of various Artificial Intelligence(AI)/machine learning algorithms and statistical modeling like decision trees, text analytics, natural language processing(NLP), supervised and unsupervised, regression models, social network analysis, neural networks, deep learning, SVM, clustering to identify Volume using scikit-learn package in python, Matlab.
* Presented data insights to senior leadership and provided actionable recommendations to improve data annotations to further improve model performance.
* Advocate the best engineering practices, including the use of design patterns, CI/CD, code review, and automated integration testing.
* Presented findings and insights to stakeholders through clear and concise reports and data visualizations generated using Tableau.

**Environment:** Python (Matplotlib, Seaborn, NumPy, SciPy, Pandas, Scikit-Learn, OpenCV), Machine Learning (Logistic Regression, Decision Trees, Random Forest, Ensemble Learning, Support Vector Machines), Computer Vision, Deep Learning, MySQL, Tableau, HDF5, Apache Airflow, ETL, AWS (Sagemaker, S3, Glue, EC2, Kinesis).

**Jenius bank June 2020 - Dec 2021**

**Role: Data Scientist**

**Responsibilities:**

* Worked with a cross-functional team to develop end-to-end data engineering solutions for anomaly detection product.
* Conducted end-to-end data science projects, from data exploration and preprocessing to model development and deployment, using tools such as Python, R, and Azure Machine Learning.
* Worked with large, complex data sets, including structured and unstructured data, to extract meaningful insights and develop predictive models.
* Developed and implemented machine learning models, such as deep learning, decision trees, and random forests, to solve real-world business problems and improve customer experiences.
* Develops predictive models using statistical techniques. Utilizes regression analysis to understand variable relationships.
* Worked on NLP machines to understand and interpret human language the way it is written or spoken using PCA, text mining.
* Designed and implemented A/B testing and experimentation frameworks to measure the impact of data science solutions on key business metrics.
* Designed and implemented recommender systems which utilized Collaborative filtering techniques to recommend courses for different customers and deployed them to AWS EMR cluster.
* Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Conducts statistical inference to draw conclusions and make predictions. Handles uncertainties and calculates confidence intervals.
* Utilized data visualization tools such as Power BI to create compelling and actionable visualizations of model outputs, presenting findings to key stakeholders.
* Stayed current with emerging technologies and trends in data science, attending conferences and training programs to expand knowledge and expertise in the field.
* Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL.
* Worked on NLP machines to understand and interpret human language.
* Identify potential problems and recommend alternative technical solutions.
* Participating in Technical Architecture Documents, Project Design, and Implementation discussions. Propose architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure
* Migration of on-premises data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake Store (ADLS) using Azure Data Factory (ADF V1/V2).
* Collaborates with domain experts to understand domain-specific statistical considerations.
* Contributed to the data science community, sharing best practices and insights with other data scientists and collaborating on open-source projects, particularly within the Microsoft ecosystem.

**Environment:** MS SQL server 2014, Teradata, ETL, SSIS, Alteryx, tableau (desktop 9. /Server 9x), python 3.x(scikit learn/SciPy/NumPy/pandas), machine learning (naive bayes, KNN, regressions, random forest, SVM, XGBoost, ensemble), AWS redshift, EMR, Spark (PySpark, MLlib, Spark SQL), Hadoop 2.x, MapReduce, HDFS, Azure Data Factory, Azure SQL Analytics, Azure Blob Storage, Azure Backup, Azure Files, Azure Data Lake Storage, Azure App Services, Azure Web Apps, Azure Logic Apps, Azure Virtual Machine (VM).

**Charter Communications Nov 2019 – May 2020**

**Denver, CO**

**Role: Data Scientist**

**Responsibilities:**

* Leveraged expertise in Data Science and Machine Learning to develop advanced models for click fraud detection in the advertising industry.
* Collaborated effectively with cross-functional teams, including data engineers, software developers, and business stakeholders, to understand requirements and implement solutions.
* Utilized big data technologies for ETL and used AWS (S3, Lambda, Glue) to handle and process large-scale click datasets efficiently and maintained a Data Warehouse with features such as IP address, operating system, device type, and time of click to predict the probability of click fraud.
* Implemented fraud detection models that resulted in a significant reduction of losses for clients by accurately identifying fraudulent clicks and helped optimize advertising revenue by ensuring genuine clicks receive the appropriate focus and investment.
* Application of various Supervised Learning Algorithms for Regression and Classification such as Linear Regression, Logistic Regression, Decision Trees, and Support Vector Machines for prediction using the Scikit-Learn package in Python.
* Employed Ensemble Methods such as Bagging, Random Forests, and Boosting Methods such as AdaBoost and Gradient Boosting to enhance model performance.
* Performed Hyperparameter Tuning to find the optimal model with Grid Search achieving the best test accuracy of 94%.
* Addressed over-fitting by implementing Algorithm Regularization methods like Ridge Regression and Lasso Regression.
* Used Principal Component Analysis in feature engineering to analyze high-dimensional data.
* Utilized statistical analysis such as Univariate Analysis, and Segmented Univariate Analysis and data visualization techniques using Python, Tableau, and Power BI to generate actionable insights for fraud detection and prevention strategies.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions about past and future behavior.
* Communicate the results with the operations team to taking the best decisions.

**Environment:** Machine Learning (Linear Regression, Logistic Regression, Random Forests, Ensemble Learning, Support Vector Machines), Python (Matplotlib, Seaborn, NumPy, SciPy, Pandas, Scikit-Learn, Plotly), MySQL, ETL, Tableau, PowerBI.

**DXC Technology Oct 2015 – Apr 2019**

**Hyderabad, India**

**Role: Data Scientist/Data Engineer**

**Responsibilities:**

* Conducted in-depth analysis of customer data to identify purchasing patterns, customer segmentation, and opportunities for upselling and cross-selling.
* Data extraction from database warehouse using PostgreSQL.
* Collaborated with cross-functional teams to define project goals, deliverables, and timelines, ensuring alignment with Agile methodologies.
* Worked on data cleaning, data preparation and feature engineering with Python including Numpy, Scipy, Pandas, Matplotlib, SQL Alchemy, Seaborn and Scikit-learn.
* Implemented predictive models to forecast demand and optimize inventory management, resulting in a 20% reduction in stockouts and overstock situations.
* Conducted market basket analysis to identify cross-selling opportunities and optimize product bundling strategies, leading to a 10% increase in average order value.
* Built recommendation systems using collaborative filtering algorithms to enhance personalized product recommendations, resulting in a 15% increase in customer engagement.
* Utilized AWS for data warehousing and worked with SQL to extract data from various sources for analysis and modeling.
* Identify multi-collinearity in data, and prepare forecasting models.
* Applied SDLC principles to data science projects, including requirements gathering, data collection, analysis, modeling, and deployment of solutions.
* Presented findings and insights to stakeholders through clear and concise reports and data visualizations generated using Tableau.
* Presented data insights to senior leadership and provided actionable recommendations to improve data annotations to further improve model performance.

**Environment:** Machine Learning (K-Means, Hierarchial), Python (Matplotlib, Seaborn, NumPy, SciPy, Pandas, Scikit-Learn), MySQL, Tableau, Google Analytics, AWS.

**Xcelvations Nov 2014 – Sept 2015**

**Hyderabad, India**

**Role: Data Engineer/Data Scientist**

**Responsibilities:**

* Developed and maintained financial models and dashboards to track and analyze business performance, utilizing tools such as Excel, SQL, and Tableau.
* Conducted analysis of financial data, including revenue, expenses, and profitability, to identify trends, variances, and opportunities for improvement.
* Built and maintained data pipelines to extract, transform, and load financial data from multiple sources, ensuring data quality and accuracy.
* Collaborated with cross-functional teams, including finance, accounting, and business operations, to develop financial forecasts, budgets, and business plans.
* Conducted ad-hoc financial analyses and prepared reports to support strategic decision-making and business planning.
* Developed and presented data-driven insights to senior management, including C-suite executives, to inform business strategy and performance optimization.
* Conducted data quality audits and implemented processes to improve data accuracy, completeness, and consistency.
* Built predictive models to forecast financial metrics such as sales, revenue, and expenses, using machine learning algorithms and statistical techniques.
* Identified and tracked key performance indicators (KPIs) for the business, developing reporting dashboards to monitor KPI performance.
* Developed and implemented data-driven solutions to address business challenges, such as reducing costs, improving efficiency, and optimizing pricing strategies.
* Prepared test plans to ensure parallel QA and development phases.
* Written and executed test cases, reviewing them with business and development teams.
* Implemented a defect tracking process using JIRA tool, assigning bugs to the development team.

**Environment**: R/R Studio, SAS, SSRS, SSIS, SSAS, Oracle Database 11g, Oracle Bi Tools, Aws, Athena, Tableau, Snowflake, MS Excel, Python, Naive Bayes, SVM, K- Means, ANN, Regression, MS Access, SQL Server Management Studio, SAS E-Miner.

**Academic Qualifications:**

**Master of Science in Data Science** – Indiana University Bloomington, Bloomington, IN

**Bachelor of Technology in Computer Science** – SRM University, Chennai, India