Ashvi Soni

+1(202)823-8515 | ashvi.y.soni@gmail.com | Linkedin | Github | Portfolio

Education

George Washington University Aug 2019 – May 2021
Master of Science (M.S.) in Computer Science *Washington, DC*

Gujarat Technological University Aug 2015 – May 2019
Bachelor of Engineering (B.E.) in Computer Engineering *Ahmedabad, India*

Technical Skills

Languages: Python, SQL, R, Java, Javascript
Libraries: pandas, numpy, sci-kit learn, matplotlib, seaborn, Tensorflow/Keras, Pytorch, PySpark, PyMySQL, NLTK, cv2, scrapy, spacy, dplyr, caret, ggplot
Frameworks and Others: REST API, Google Workspace, Google Data Studio, Machine Learning, Data Science, Data Visualization, Data Analysis, Statistical Analysis, Data Mapping, Data Cleaning, Data Wrangling, Data Integration
Developer Tools: Qlik, Tableau, PowerBI, Hadoop, Spark, Git, PyCharm, Anaconda, MySQL Workbench, MS Azure, MS SQL Server, Mongo DB
Relevant Coursework: Machine Learning, DBMS, Big Data and Analytics, DBMS for Data Analytics, Algorithms

Professional Experience

Data Engineer Jun 2022 – Present
Eclinical Solutions, Mansfield, MA

- Execute ETL processes to aggregate clinical trial data from EDC systems, vendor labs, and diverse sources into company software via AWS and sFTP, optimize data accessibility and analysis.
- Collaborated with cross-functional teams, enhanced data quality and completeness by implementing SQL data validation checks, resulting in a 25% increase in data accuracy and a 30% reduction in data errors.
- Program Python (pandas) and SQL scripts for anomaly identification, dataset reconciliation, trend detection, and customized report generation, optimizing operational efficiency and supporting data-driven decision-making.
- Develop 5 Python-based ML models with pandas and sci-kit learn for outlier detection in Lab Results and Vitals, and Time Series Analysis for atypical shift detection; visualize results using matplotlib and seaborn libraries.
- Implement diverse Python ML models using pandas and sci-kit learn to detect inappropriate medication-indication pairs in study data using Score Matching, enhancing data accuracy by 30%.
- Create data visualizations using Qlik to translate intricate clinical data into actionable insights; craft interactive dashboards, reports, and charts to effectively convey findings to key stakeholders.
- Lead client sessions to identify data needs, translating insights into actionable datasets and models with Python and SQL, resulting in a 40% boost in data accuracy.

Programmer and Data Analyst Aug 2021 – Jun 2022
Project: COVID-19 CRP *The Biostatistics Center, George Washington University, Washington, DC*

- Engaged in COVID-19 Community Research Partnership Project funded by CDC and the State of North Carolina, involving 10+ institutions; collaborated on exploratory, prescriptive and predictive data analysis using Python.
- Orchestrated a comprehensive data wrangling initiative utilizing Python, R, and SQL scripts, enhancing data quality and integrity by 40% and reducing analysis time by 25%.
- Implemented a data mapping solution utilizing Python, SQL, and R scripts to align medication data from EHR datasets with active ingredient attributes, resulting in a 40% reduction in data processing time.
- Created detailed data visualizations in Python and R, helping in comprehension and faster decision-making.
- Uncovered a substantial HIPAA violation through in-depth data analysis, leading to the implementation of enhanced security protocols and staff training, resulting in a decrease in data security incidents.

Junior Data Scientist Aug 2021 – Jun 2022
Project: EDIC *The Biostatistics Center, George Washington University, Washington, DC*

- Executed data analysis for the EDIC study funded by NIDDK, analyzing 10+ years of patient data to identify trends and correlations, resulting in a 15% improvement in study accuracy.
- Developed and executed six distinct predictive models- Logistic Regression, Random Forest, SVM, K-Nearest Neighbor, XGBoost, Neural Network to forecast the occurrence of Hypoglycemia in Type 1 diabetes participants.
- Enhanced predictive models by optimizing algorithms and fine-tuning hyperparameters, resulting in a 50% increase in sensitivity and specificity, significantly boosting overall performance and accuracy metrics.
- Generated visualizations, including the ROC curve, to illustrate the models' sensitivity-specificity trade-off.