SONALI PEDNEKAR

+1 (571) 473-9910 | ssp88@georgetown.edu | Portfolio Website | LinkedIn | GitHub | Washington DC

EDUCATION

Georgetown University, Graduate School of Arts and Science **Aug 2021 - May 2023**
Master of Science, Data Science and Analytics *4.00 GPA*
Merit-Based Scholarship Recipient

Narsee Monjee Institute of Management Studies **Jul 2017 - May 2021**
Bachelor of Technology, Data Science

WORK EXPERIENCE

Data Analyst Intern | Stratus Home Loan, Las Vegas NV ***Jul 2023 - Present***
- Established a centralized data repository, leveraging SQL and Python to extract and manipulate data for weekly reporting in Excel and Power BI.
- Handled ad hoc tasks by assessing and enhancing KPIs to ensure timely and accurate data analysis.

Research Assistant (Data Analyst) | CSET, Washington, DC **Jan 2022 - May 2023**
- Conducted research and annotation tasks to create a valuable dataset using Airtable to visualize AI industry trends.
- Employed data preprocessing and manipulation methods like cleaning stock data, pattern matching using regex, and utilizing SQL to retrieve values from BigQuery.
- Demonstrated insights into identifying potential harm and near-harm AI-related incidents, potentially leveraging an NLP API.

Graduate Teaching Assistant | Georgetown University, Washington DC **May 2022 - Dec 2022**
- Facilitated comprehensive understanding in Big Data and Cloud Computing, delivering instruction on AWS, Azure, parallel computing, Elastic Map Reduce, Hadoop, Hive, and Spark, enhancing the learning experience for over 120 students.
- Led the Data Science Bootcamp, imparting expertise in programming languages Python and R, along with core libraries including Pandas, NumPy, Tidyverse, and ggplot. Provided expert guidance and fostered interactive learning environments to facilitate students' mastery of course material and assignments.

Data Science Intern | Nielsen, India **Dec 2020 - May 2021**
- Applied ML models for Population Forecasting, utilizing current and past census data to predict future trends.
- Streamlined the tracking of weekly changes by automating data retrieval and wrangling processes in MS Excel, utilizing Power Query and Data Wrangling tools to update population forecasts.

Data Science Intern | Konsultera Solutions, India **May 2020 - Nov 2020**
- Developed automated solutions for extracting 15+ fields from legal documents, utilizing Prodigy software to develop Text Classification and Named-Entity Recognition models.
- Collaborated with stakeholders to identify trends and patterns, ensuring the retrieval of fields for efficient optimization.
- Cleaned and annotated over 6000 Mergers and Acquisitions transactions to ensure optimal accuracy. Contributed to the Fraud Detection project by annotating images with bounding boxes using LabelMe for image classification.

PROJECTS

Reddit Analysis | Big Data and Cloud Computing
- Extracted and cleaned over 2TB of large-scale dataset from the PublicFreakout subreddit and conducted EDA.
- Implemented a pre-processing NLP pipeline on 18million+ rows for Sentiment Analysis using Spark NLP and Jon Snow labs.
- Web-scaled on EC2 instances to predict score and controversiality by incorporating additional features into ML models.

Depression Analysis | Machine Learning & App Deployment (Application Link)
- Designed a model to assess the risk of depression in individuals, based on personality types and socioeconomic factors extracted from DASS-42 and deployed the application on Streamlit by storing the model as a pickle.
- Conducted Exploratory Data Analysis before building the model, selecting Logistic Regression (0.67) as the baseline model, and employing ensemble techniques with five-fold cross-validation to improve performance.

Olympics Analysis | Data Visualization
- Conducted Visual Analysis of 124 years of Olympics data using visualization tools like ggplot, Altair, Plotly, D3, and Tableau.

Autism Prediction | Neural Network and Deep Learning
- Processed and analyzed 2k MRI scans from ABIDE dataset and configured pipelines and deployed PyTorch to optimize the processing efficiency. Integrated Auto Encoders and PCA for dimensionality reduction and improved training time.
- Implemented a majority voting-based approach via TensorFlow/Keras to identify MRI scans with up to 58.66% accuracy, leveraging pre-existing CNN architectures like ResNet50, VGG16, and InceptionV3.

NYT Article Popularity Prediction | Statistical Modeling
- Utilized NYT API to extract over 17,000 articles from the New York Times, performing Exploratory Data Analysis and implementing feature engineering techniques to improve sentiment analysis and clustering for topic modeling.
- Employed classification algorithms, including Logistic Regression (0.76), LDA (0.77), Decision Trees (0.78), and Random Forest (0.79), to predict the popularity of articles, using precision and recall metrics to evaluate each algorithm.

SKILLS

Certification: AWS Certified Cloud Practitioner
Programming Languages: Python, R, SQL, AWS, Prodigy, Bash
Data tools: Git, AWS (EMR, EC2, S3, Sagemaker, Hadoop, Hue, Hive, Spark, MapReduce), Pyspark, Excel, MongoDB
Visualization tools: GGPlot, Tableau, Matplotlib, Plotly, Seaborn, Altair, R2D3, Power BI
Machine Learning models: Clustering, ARM, Decision Trees, Random Forest, Naïve Bayes, SVM, Regression
Leadership Roles: Treasurer of Rotaract Club of Bombay Airport and coordinated a team of 300+ members