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 3.16 GPA

WORK EXPERIENCE

Graduate Teaching Assistant | Georgetown University May 2022 - Present

- TA for *Big Data and Cloud Computing*: taught topics like AWS, Azure, parallel computing, Elastic Map Reduce, Hadoop, Hive, and Spark, and led doubt-solving sessions for assignments and projects for 120+ students.
- TA for *Data Science Bootcamp*: taught programming in Python, R, and their fundamental libraries, such as Pandas, NumPy, Tidyverse, and ggplot.

Center for Security and Emerging Technology, Washington, DC

- **Data Annotator Research Assistant** Jan 2022 - Present
  - Conducted research and annotation tasks to create a valuable dataset using Airtable for visualization of AI industry activity.
  - Data preprocessing, manipulation, and annotation of the information of 500+ different companies, cleaning stock data, pattern matching using regex, and data handling using SQL to retrieve values from BigQuery.
- **Regulation Survey and AI Incident Database Research Assistant** Jun 2022 – Aug 2022
  - Utilized LinkedIn API and Python to find Legal and Ethics personnel at AI startups and companies for a survey on existing regulatory structure, guidance, risk, and uncertainties in AI.
  - Increased the number of harm/near-harm AI-related incidents by potentially querying a Natural Language Processing API and interviewing Industry experts.

Data Science Intern | Konsultera Solutions, India May 2020 - Nov 2020

- Automated the process of deriving every field on the legal document website by utilizing Prodigy software to develop different models for Text Classification and Named-Entity Recognition.
- Cleaned and annotated the raw data of 6000+ Mergers and Acquisitions transactions. In the Fraud Detection project, annotated images with bounding boxes using LabelMe for image classification.

Data Science Intern | Nielsen, India Jun 2019 - Jul 2019

- Population Forecasting for predicting future population from current/past census data using different machine learning models.
- Automated tracking of weekly changes by using Power Query and Data Wrangling in MS Excel.

PROJECTS

Reddit Analysis | Big Data and Cloud Computing

- Extracted and cleaned 2TB+ of PublicFreakout subreddit data and performed Exploratory Data Analysis.
- Applied Natural Language Processing by creating a pre-processing pipeline to perform feature transformations like tokenization, lemmatization, normalization, stemming, POS tagging, removed stopwords, and text cleaning.
- Utilized Feature Extractors to create feature vectors before applying Machine Learning models to answer business questions.

Olympics Analysis | Data Visualization

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

NYT Article Popularity Prediction | Statistical Modeling

- Utilized NYT Api to extract 17k+ NYT articles, conducted EDA, feature engineering for sentiment, clustering for topic modeling.
- Used classifier algorithms like Logistic Regression (0.76), LDA (0.77), Decision trees (0.78), and Random Forest (0.79) to predict the popularity of articles by looking at the precision and recall of the algorithms.

Social Work Analytics | End-to-End Data Science Project

- Analyzed sectors of social work worldwide from a data-driven perspective using 60000+ record and text datasets from Twitter API, News API, and World Bank API and raw datasets through different websites and surveys.
- Performed EDA, Unsupervised, and Supervised Learning algorithms: Clustering, ARM and Networking, Decision Trees, Naïve Bayes, and SVM.

Placement Data Analysis | Probabilistic Modeling and Statistical Computing

- Analyzed the campus placement data of Indian students using data analysis/ visualization with ggplot in R.
- Gained insights on placement trends and predictors for salary variable using Multilinear Regression, Correlation tests, Chi Square Test, Anova Test, and T-Test using R.

Airline On-Time Statistics and Delay Causes | Database Systems and SQL

- Analyzed the On-Time Performance Airline dataset using EDA, designed a star schema, loaded data through an EC2 instance on AWS, and utilized nested SQL queries.

SKILLS

Certification: AWS Certified Cloud Practitioner
Programming Languages: Python, R, SQL, AWS, Prodigy, Bash
Data tools: MySQL 8.25, Git, AWS (EMR, EC2, S3, Sagemaker, Hadoop, Hue, Hive, Spark, MapReduce), Pyspark, Excel
Visualization tools: GGPlot, Tableau, Matplotlib, Plotly, Seaborn, Altair, R2D3
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