Prabhjot Singh

prab6421@gmail.com | **US Citizen** | Los Angeles, CA | (703) 864-6123 | LinkedIn | GitHub | Portfolio

Born and raised East Coaster with 5+ years of experience, recently transplanted out West, passionate about unraveling intricate patterns, streamlining complex tasks into elegant solutions, and utilizing machine learning to provide predictive capabilities with business impact in a diversity of domains.

Work Experience

Data Scientist | Healthcare Risk Advisors | New York, NY | 07/2022 – Present

- Developed, deployed and productionalized classical **supervised scikit-learn regression** and **classification models** on **Azure Databricks** to aid claims teams in prioritization and resourcing high-risk malpractice cases with 75% accuracy.
- Engineered, fine-tuned, and interpreted **large language models (LLM)** and other **deep learning pipelines** using **pytorch**, **tensorflow**, **keras**, **transformers**, and **huggingface_hub** packages to predict adverse patient outcomes that was less than 1% of the population.
- Conducted **hypothesis tests** by devising complex **Spark SQL** queries to analyze electronic health records (EHR), claims, demographic and legal data, leading to the discovery of over 150 **statistically significant indicators** of various target variables.
- Designed innovative **unsupervised learning** approach to objectively target hospitals that are 90% similar to other hospitals, using **distance metrics** and **clustering** methods calculated with 20 demographic, clinical, and financial features.
- Collaborated with engineering, product development, and client facing teams by performing regular data labeling, manipulating, and exploration using **pandas** and **numpy** to enhance **Power BI** dashboards, data models, and feature stores.
- Delivered technical presentations to cross-functional stakeholders on a monthly cadence by interpreting from LLMs and creating concise and understandable **visualizations** using **shap**, **alibi**, **seaborn** and **plotly** packages to illustrate rigorous testing methods, rationale for increased model adoption, and actionable insights to inform future decision-making.
- Led and mentored 3 summer interns in data science, statistics, and visualization techniques, guiding them through impactful short-term projects that led to process improvements across 2 different departments.

Senior Data Analyst | Capital One | Mclean, VA | 06/2021 – 06/2022

- Spearheaded 5 **data analytics** initiatives in internal audit, examining over 200 risks and controls across international and domestic products by writing extensive data analytical reports that identified key patterns through intricate SQL queries within a massive **snowflake database**.
- Programmed software enhancements to **algorithms** in **python** by improving sampling methodologies that leveraged **webscraping** tools such as **selenium** and **beautifulsoup** reducing overall document compilation time by over 50%.
- Conducted code reviews of lengthy scripts over 5000 lines of **SAS**, **SQL**, and **python** used on integral systems to improve infrastructure around business retention, leading to a 5% reduction in detected false positives.

Management and Support Intern | Get the Medications Right Institute | Tysons, VA | 09/2019 – 03/2020

- Researched and prepared **analytical insights** on **Excel** of user data, healthcare institutions and companies for marketing, business, and investigative purposes, leading to a 10x growth in overall membership.

Research Assistant and Technician | UVA Clinical and Psychology Research Departments | 01/2017 – 05/2019

- Performed validation and cleaning of clinical data from EHRs, clinical labs, and MRIs of patient data to ensure integrity within Epic and OnCore data repository systems and enable **hypothesis testing** on **SPSS** and **Excel**.

Education

Master of Science, Data Science | GPA: 4.0
University of Virginia, Charlottesville, VA | 2020 – 2021

BA, Cognitive Science conc. Neuroscience and Psychology
University of Virginia, Charlottesville, VA | 2015 – 2019

Machine Learning Projects and Publications

Mutational Processes in LGL Leukemia | IEEE Publication Link

- Leveraged **association analysis**, **kernel density estimation** and other **data mining** and **statistical** methods to detect over 30 statistically significant mutational hotspots along the genome that may be linked to LGL Leukemia.

Modeling Weather Data Via Natural Language Processing (NLP), Bayesian ML and Recurrent Neural Networks (RNNs)

- Experimented with **NLP** methods such as **principal component analysis (PCA)**, **sentiment analysis**, **topic modeling**, and **word embeddings** to investigate differences across meteorological storm report narratives found on NOAA **API**.
- Utilized **Bayesian**, **dense recurrent neural network** (**RNN**) and **long-short term memory (LSTMs)** architecture to devise **time series** models forecasting future wind speeds and classification models predicting tornado EF magnitude.