ARFA KHALID

(639) 382- 3258 | Edmonton, AB | arfakhalid09@gmail.com | linkedin.com/in/arfakhalid-ms/

SUMMARY OF QUALIFICATIONS

With 5+ years of expertise in statistical modeling and machine learning, I translate complex data into actionable insights across diverse fields. I built robust models for climate change forecasting, public opinion prediction, and flight arrival accuracy. Leveraging computer vision, I assessed flood impact and optimized parking lot management. Experienced in implementing and optimizing machine learning frameworks (TensorFlow, Keras) and specialize in deep learning principles. Adept in Python, R, SQL, and cloud tools, I'm a passionate data scientist driven to unlock valuable discoveries through data-driven initiatives.

SKILLS

Technical Skills: Statistical models (i.e. A/B Testing, Clustering, Regression/Classification, and Explanatory Models), statistical computing, Python (Scikit-learn, pandas, NumPy, Statsmodels), R, SQL, databases, Excel, Business analysis (data visualizations, dashboards, and reports), Spark, Machine Learning and AI techniques (supervised and unsupervised ML, decision trees, and logistic regression), deep learning (neural networks), GIS (ArcMap, QGIS), computer vision (SAM, SAMgeo), Cloud-based platforms and services (AWS, Azure), version control (GitHub)

Interpersonal Skills: Strong analytical, problem-solving, structured thinking, and organizational skills, insight & analysis, analytical and critical thinking, excellent verbal and written communication skills, detail oriented, result focused, time management, work-independently and collaboratively

WORK EXPERIENCE

QTO House **Remote**

Data Scientist *Jan 2023 - Present*

- Develop and apply advanced quantitative models using machine learning and econometric techniques to produce highly accurate and granular cost estimates for diverse project types.
- Drive collaborative knowledge sharing by building interactive dashboards and reports that effectively communicate complex cost breakdowns and risk assessments to project stakeholders, improving project transparency and buy-in.
- Streamline cost estimation workflows by automating routine tasks, data extraction, and report generation through scripting and data manipulation tools like Python, leading to increased efficiency and time saved for high-value analysis.
- Spearhead the integration of new data sources and cost databases into the company's cost estimating infrastructure, improving the comprehensiveness and accuracy of cost forecasting for future projects.
- Partner with project teams to identify and quantify project risks using Monte Carlo simulations and sensitivity analyses, informing proactive risk mitigation strategies and enhancing project confidence.

University of Regina **Canada**

Junior Data Scientist *August 2021 – April 2023*

- Boosted efficiency by 30% and extracted 5x more relevant features from massive datasets using parallel processing and distributed computing frameworks on AWS cloud platforms.
- Leveraged big data analytics and machine learning models like Support Vector Machines (SVMs) and Random Forests to tackle complex research questions in statistics and mathematics.
- Integrated Bayesian techniques into big data models, resulting in a 15% boost in model performance, leading to enhanced predictive capabilities, improved decision-making support, and a 10% reduction in overall uncertainty.
- Provided instruction and support to undergraduate students in laboratory classes, fostering their understanding of statistical data analysis tools and techniques (e.g., R, Python) on HPC and AWS cloud platforms.

Bureau of Statistics **Pakistan**

Junior Data Analyst *June 2018 - August 2021*

- Streamlined data collection processes, reducing collection time by 20% and improving accuracy by 15%.
- Implemented rigorous data quality checks and automated cleansing programs, reducing errors by 30%, solidifying the integrity of statistical reporting.
- Developed and maintained dynamic dashboards and reports, tracking key performance indicators (KPIs) and delivering data-driven insights to guide strategic decision-making across the Bureau.
- Uncovered hidden patterns in population and agricultural data, leading to targeted interventions that improved productivity and health. Collaborated with cross-functional teams to assess the impact of programs on population and agricultural trends.
- Gained experience in data collection, writing summaries, and visualizing data.

EDUCATION

University of Regina **Canada**

Master of Science in Applied Statistics *April 2023*

Course work: Probability, Statistical Inference, Statistical Modeling of Dependence and Extremes, Design of experiments, Bootstrap methods, Machine Learning, High Performance Computing

Thesis: *"Analyzing the effectiveness of COVID-19 vaccines among different age groups using multinomial logistic regression model";* This study employed multinomial logistic regression to assess the effectiveness of COVID-19 vaccines across various age groups. The mathematical proof for the model with interaction effect was derived, demonstrating significant impacts of age group and vaccination status on COVID-19 cases, offering valuable insights for policymakers in optimizing vaccination strategies and pandemic control.

University of the Punjab
Pakistan
Bachelor of Science in Mathematics and Statistics *September 2020*

Course work: Sampling Techniques, Multivariate Techniques, Applied Econometrics, Time Series Analysis, Categorical Data analytics, Machine Learning, Database development and design, Computer Science, Artificial Intelligence

PROJECTS

Compound Flood Risk Analysis: **Jan, 2024**

- Calculated annual exceedance frequencies for combined water levels, waves and precipitation from historical observations recorded by NOAA at Washington DC tide gage. Employed data cleaning and organization within R, followed by Copula analysis to investigate co-dependence of storm tide with precipitation for evaluating compound flood risk.

Satellite Imagery Segmentation for Flood Impact Assessment in Libya using Computer Vision **Dec, 2023**

- Led geospatial flood analysis in Libya using Maxar Open Data. Employed segment-geospatial and leafmap libraries for efficient data processing. Achieved accurate flood segmentation, converting raster masks to vectors. Visualized results on an interactive map, demonstrating insights into flood-affected areas.

Climate Change Analysis: Time Series Forecasting in ML **Nov, 2023**

- Led Climate Change Analysis project, utilizing Matplotlib, NumPy, Seaborn, and scikit-learn for data analysis. Built linear regression models to NASA's dataset (1951-1980) and extended insights to a broader range (1882-2014). Quantified temperature change patterns, revealing potential acceleration in climate change. Ensured model accuracy through rigorous assessment of actual vs. predicted temperature values. Identifying a 40% increase in global average temperature. Predicted further warming with 95% confidence, informing climate mitigation strategies.

Predicting Public Opinion with Deep Learning and Natural Language Processing **Oct, 2023**

- Executed sentiment analysis on the IMDB dataset (50,000 movie reviews) using Keras for neural network modeling. Achieved an 86.97% accuracy on the test set through meticulous data preprocessing, neural network design, and optimization. Developed a versatile sentiment analysis function for custom text inputs, demonstrating the model's adaptability and generalizability.

Car detection in Parking Lots: **Jun, 2022**

- Transformed raw drone-captured imagery and associated meta-data into a clean, structured dataset suitable using computer vision techniques. Leveraged AWS Mechanical Turk to obtain accurate and scalable annotations of real-world parking lot data, enriching the dataset for model training. Developed and implemented a deep learning model specifically designed for counting and pinpointing car locations within parking lots.

Enhancing Air Travel Efficiency with Classification Model **Jun, 2019**

- This project tackled the challenge of predicting on-time flight arrivals using a classification model. Harnessing a major U.S. airline dataset, I built a robust pipeline that cleans, manipulates, and prepares data for optimal model training. Employing a Random Forest classifier from scikit-learn, I achieved an impressive 86.43% accuracy in predicting on-time arrivals.

Population Growth Analysis **Dec, 2018**

- Investigated the demographic shifts and air quality trends in Lahore, Pakistan from 1972 to 2017. Examination of census data uncovered a concerning trend of exponential population growth, prompting a deeper investigation. Utilizing GIS, I intricately mapped population density alongside historical air quality index (AQI) data. Spatial visualization, coupled with statistical analysis, revealed a troubling correlation: areas experiencing the highest population growth consistently exhibited lower AQI values.

CERTIFICATIONS

Google Cloud Professional Machine Learning Engineer In-progress
AWS Cloud Practitioner In-progress
Udemy Data Science Bootcamp March 2022

AWARDS

Graduate Research Fellowship award and entrance scholarship (CAD 13,000) August 2021
Academic Excellence Award – Undergraduate October 2016