Shubhkirti Prasad

Ph. no: +1 (812) - 671-5196
Email: shubpras@iu.edu

SUMMARY

I am a result-driven Data Scientist with a strong foundation in software solutions and product design. Proven ability in implementing innovative technologies and driving customer-focused projects! My combined technical expertise with exceptional communication skills help to foster client relationships and deliver impactful solutions.

EDUCATION

MS, Data Science, Indiana University

Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, USA **(AUG '22 - Present)**

B. Tech. , Mechanical Engineering (Minor in Digital Marketing)

Manipal Institute of Technology, Manipal, India **(2017-2021)**

Work Experience

Paul O'Neill School, Luddy School - Graduate Teaching Assistant **(August '23 - Present)**

- ➢ *Course : **V506 Statistical Analysis for Effective Decision Making*** (Class average grade : A)
- ➢ *Skills : R, Statistical Modeling, Data distribution, Teaching, Hypothesis testing, A / B testing, MLS*
- Facilitated statistical modeling in R for over 40 graduate students, contributing to a class average grade of A. Innovatively integrated real-world data sets (OpenIntro Statistics) into weekly lab sessions, enhancing practical understanding and application of statistical theories.

Four Colors Technology (MA,USA)- Data Science Intern (Research) **(MAY '23 - JULY'23)**

- ➢ *Skills : NLP, LLM, SpaCy, Python, Named Entity recognition, Vertex AI, Google Bard*
- Spearheaded the customization of Google's Bison LLM, to be used as a backbone for employee support chatbot on Vertex AI, developing a context-based model that enhanced the accuracy of an chatbot by over 40% during A/B testing.

DataOrc (India) - Data Science Intern **(AUG '21 - NOV'21)**

- ➢ *Skills : Python, Convolutional NN, Jupyter Notebook, Sklearn, NumPy, Pandas, TensorFlow, Keras*
- Played a key role in a team that developed an OCR system, translating user's handwritten domain-specific text into English with 72% accuracy using CNN.
- Optimized SQL queries with Relational Algebra (decreasing the overall query runtimes by 32%) to aid the dry run and deployment of a highly scalable, data ready analytics platform.

IBM India - Data Science Intern **(APR '21 - JUL '21)**

- ➢ *Skills : Python, Statistical Analysis, Time Series, Jupyter Notebook, Sklearn, NumPy, Pandas, data mining, RegEx*
- Identified and researched a business problem to evaluate the health of of BSE-500 companies Pre/Post Covid.
- Created a Data pipeline to convert the PDF Annual Reports of BSE-500 companies for FY-18,19,20,21 into textual data and extract relevant financial / textual information.
- Collaborated with the Business team to present the firm's health data to the stakeholders.
- Lead a project *'Distribution in IT Sector with Investment Impact Analysis'* which was a comparative analysis of firms in the manufacturing sector trading in BSE500 for linkage between IT investment and the firm's financial performance.

TECHNICAL SKILLS

Languages: SQL, Python(NumPy, Pandas, Sklearn, SpaCy, NLTK, PyTorch, etc.), R, C/C++

Machine learning algorithms: Linear Models (Basis Functions, MLE, OLS, Regularization), Logistic Regression, Fisher's Linear Discriminant analysis, K-means, Principal Component Analysis, Naive Bayes, Support Vector Machines, Kernel methods, Decision Trees, Ensemble methods, Convolutional Neural Nets

Data visualization tools : Tableau, Kepler.gl, Matplotlib, Plotly, Seaborn

Software: Jupyter Notebook, PostgreSQL, Microsoft Office (Excel, Word, Power Point), Autodesk SolidWorks, Rstudio

---

ACADEMIC PROJECTS

Tropical Storm intensity estimation from Raw and Infrared Satellite images (CNN)

- Created a novel CNN model (Classifier + Regressor) based on VGG-16 to estimate the wind speed of 600 individually documented storms with their cloud formation progression images. (RAW + Infrared Images)
- The train set consisted of 70257 images and 44377 images for testing.
- Created and assessed multiple models such as: Simple Regression, Inference based (VGG-16, ResNet)
- Optimized the model using Data Augmentation (Image Generator), Max Pooling, Dropouts, Batch-Normalization, Convolutional layers, activation functions, fully connected layers, image filter, edge case hypothesis.
- Link to the paper: https://github.com/Shubhkirti24/SpicyCloud/blob/main/paper.pdf

Job-Skill Dashboard

- Created an interactive job skill mapping dashboard using Tableau based on the ONET database to visualize the current employment characteristics in the USA, based on parameters such as Age, Location, Skills, Knowledge, Education, Experience, etc.
- Implemented advanced SQL optimization techniques (Relational Algebra, set operations and SQL triggers) to reduce query executing time from 12 minutes to sub 2 minutes on a dataset containing 1000+ occupation categories over 1,000,000+ rows.
- Link to the project: https://github.com/Shubhkirti24/JobDashboard

Hand Drawn Image Classification

- From a hand drawn dataset of 500,000 images with 100 classes, classified 100,000 images using a custom CNN model with an accuracy of 72%
- Scored 2nd out of 100 models.
- Applied techniques of Data Augmentation, Max Pooling, Dropouts, Batch-Normalization

Other Projects

- Music store analysis (*Created an In-depth guide for on online project to traverse through the given relational database and answer various business questions (bestselling genre and artist, genre for a new artist, selection of songs, most featured artist, encrypted music) using advanced SQL techniques)*
- Most Sold Car Brands on German eBay (*Found the most sold cars brands and their models out of 50k postings on German eBay)*
- NYC SAT Social impact analysis (*Used the dataset provided by the NYC public schools to find the correlation between SATs and demographics to analyze the important social issues like safety scores, racial differences, gender disparity, and advanced placement)*
- Most profitable App category for iOS and Google Play Store (*cleaned and curated a dataset of 17,000 apps using Python and its libraries*)
- Analyzing Hacker News Posts (*used Date Time and OOPs concepts, consolidated the data about 20k posts from Hacker News*)
- Most Sold Car Brands on German eBay (*Found the most sold cars brands and their models out of 50k postings on German eBay)*
- Gender Gap in College Degrees (*Used matplotlib to create custom plots to represent the gender gap in College Majors in STEM fields from 1970-2012*.)

- Prediction of Car prices using Regression *(Pre-processing and wrangling•Exploratory Data Analysis (Descriptive statistical analysis and outlier detection)• Model development and refinement (Linear Regression, Multivariate polynomial regression) • Model evaluation (Cross validation, ridge regression, error correctio*n)

Mechanical Projects

- **Team Manipal Racing**
  Research, Development and Manufacturing of an off-road car (BAJA buggy) - Research, manufactured and optimized custom components of a CVT transmission system and a custom transmission assembly.
