Naveen Vatsal
Email: naveen.numpy@gmail.com **Linkedin:** https://www.linkedin.com/in/naveen-vatsal/
Phone no: (619) 693-7021 **Github:** https://github.com/ysnvatsal/

PROFESSIONAL SUMMARY

- **10 Years** overall experience in IT and **6+ years** experience in implementing data mining and statistical machine learning solutions as a data scientist to various business problems such as sales lead scoring, demand forecasting, employee churn probabilities. Possessing an extensive analytic skills, strong attention to detail and a significant ability to work in team and independently across various domains and industries.
- Developed intricate algorithms based on deep-dive statistical techniques **like Whitney test, Bayesian test, Hypothesis Testing, ANOVA, Dickey fullers test,Mc Leod and Li test, Covariate analysis, Cluster Analysis, Multivariate analysis, Discriminant Analysis.**
- Quick comprehension skills with an ability to create lean proof of concepts based on the ongoing research.
- Experience with broad range of supervised and unsupervised machine learning techniques and proficient in Machine Learning and deep learning algorithms including **Linear Regression,Logistic, K-NN, K-means, Support Vector Machines, tree based, ensembles methods and advanced techniques like NLP, RNN, CNN, ResNets, Transfer learning, GaN, BERT** etc
- Ability to data Mine of structured, semi-structured, and unstructured data. Also worked with different format of data like **JSON**, **XML, CSV, Parquet etc**
- Proficient in data visualization tools such as **looker**, **Tableau 9.x, Python** to create visually powerful and actionable interactive reports and dashboards.
- Deep understanding of **Software Development Life Cycle (SDLC)** as well as experience in implementation of Analytics as per **CRISP-DM** framework
- Worked with a team of developers to create ETL scripts in Python and SQL for data transformation, cleaning, and loading
- Training pipeline orchestration using **Kubeflow** and **MLflow** via the AI Platform engine.
- Ability to adopt best coding practices for version control, containerization, CI/CD, and MLOps.
- Expertise in providing go-live support using Continous Integration/Continous delivery(CI/CD pipelines) in cloud services like **Amazon Web Services (AWS)** and **GCP.**
- Experience in handling Big Data Tools like **Apache Spark, MapReduce, Hadoop, HBase, HDFS, Hive, and Mongo DB**

TECHNICAL SKILLS

Databases	: Oracle 11g,MySql, DB2, Hbase, MongoDB, Redshift.
Programming	: Python 3, R programming, OpenCV, SQL, Spark, Hadoop, PL/SQL,SAS, C, JavaScript, HTML, PHP
Operating System	: UNIX, Linux, Windows 10
Big Data and Business Intelligence	: Hadoop HDFS, Hive, Tableau 9.1,Looker
Statistics	: Law of large numbers, clt, Hypothesis Testing (p-values) and Test for significance (z- test, t-test and ANOVA), Machine Learning, Linear/Logistic, SVM, Ensemble Trees, Random Forests, Clustering, Gradient ,Boosted Trees, Neural Networks, Support Vector Machines, Clustering Algorithms and PCA, K-NN, Neural Networks (Tensor Flow, LSTM, GAN's, CNN, Transfer Learning),Decision Trees, Ensemble methods,
Analytics Frameworks and packages	: SKlearn 0.2, Matplotlib, Numpy, Seaborn, scipy, XGBclassifier, Keras, Tensorflow 2.0, OpenCV, Theano, Caffe, Pytorch, NumPy, SciPy, Plot.ly, Pandas, Stats Models,Pytorch Pyspark, SQLite, NLP, Conversational AI, BERT, Loss Functions
Cloud	: AWS (S3, EMR, Sagemaker), GCP(GSC, Bigquery)
Others	: MATLAB, Unix, Git, Power Scripting, TCL, Wireshark, Tableau, Excel

CERTIFICATIONS

- Certified in machine learning program an course by prof Andrew NG in C**ourse Era.**
- Successfully completed **Machine learning** handson using R and python course in **Datacamp.**
- Certifiied in '**The Analytics Edge**' an **Edx** course for Machine learning in R.
- Certified in '**Applied statistics for Data Sciences'** an Edx course for Data sciences.

PROFESSIONAL EXPERIENCE

Verizon, San Diego
Role: Data Scientist **Jun 21 – Current**

Project: Verizon -EN-FPP-CLOUD-AWS-UNIFIED-RT

Verizon has domain specific data which includes Jira, Confluence and AWS Support services internal data on which a contextual search engine should be built in order to contextualize the user queries for relevant responses to answer their questions and drive them to the specific domain related hyperlinks targets that relate to the Conersational AI engine's responses. This include usage of multiple AWS services such as AWS glue for Data Ingestion, Sagemaker Notebook, Lambda for triggering the new jobs etc, Opensearch

Responsibilities:

- Developed python scripts for ingesting data from AWS support cases using **AWS Glue, SecretManager** for API keys and **Cloudwatch** for monitoring.
- Create NLP workflows with DistilBert model using AWS **Sagemaker** to create contextual search engine, produced embedding vectors and store it on AWS **Opensearch** to store the converted embeddings.
- Experimented on different bert models using such as **MiniLM**,**MiniLM Uncased** etc and compared their performances using **cosine similarity score**.
- Involved in adding new parameters for configuring the Glue jobs and Sagemaker instance using Cloud Formation templates.
- Model evaluated based on historical data, Model re-training with new data to improve model accuracy. Designed and implemented machine learning pipeline end to end using Sagemaker notebooks.
- Used **Sagemaker studio**, **Notebooks** for model training to create the embedded vectors and save in **S3/Opensearch** to be compared with user queries for contextual search.
- Create model training scripts using NLP python in Sagemaker, create pipelines using Amazon EC2 instance to run the training job and see if the vectors are getting saved in **Opensearch**.
- Design and finetune the contextual search engine based on user's responses and add the corpuses to the embedded vectors and retrain to improve the search accuracy based with the help of loss funtions(**CosineSimilarity Loss, Triplet Loss**)
- Designed and developed dashboards using tableau and been responsible for refreshing the extracts to prevent data from becoming stale and automated using tableau server.
- Create the API calls and Fetch for sending and receiving the feedback responses from the AI search engine to take and send responses from the **training API**
- Created Fraud dictionaries for numeric patterns, name patterns, time patterns and geographic patterns with the help of **NLP text mining** techniques and found out risk-level markers and threat indicators
- Provided production support for tableau and looker on ad-hoc basis, published workboks creating user filters so that only appropriate teams can use it.
- Worked on ETL and report testing using different datawarehouse tools like AWSRDS, Databricks & Tableau

Environment: AWS Sagemaker,NLP, Conversational AI, OpenSearch, AWS Lambda, AWS Glue, Python 3 (Scikit-Learn/SciPy/NumPy/Pandas), Tableau 9.4, Oracle PL/SQL, lookml, DataBricks,Statistical Analysis,Tableau(desktop/admin/server), NLP, Bert, Machine learning,Deep learning

Regeneron Sciences, Tarrytown, NY **Oct 19 - June 21**
Role: Data Scientist

Project: Gene sequencing, Histopathologic cancer detection and Opthalmic imaging for glaucoma detection
The aim of the project is to create an algorithm that will identify the metastatic tissues in histopathologic scans of the lymph node sections using one of the deep learning techniques – Convolutional Neural Network. We aim to classify cancer tissues based on the labels - Malignant or Benign.The project is implemented using Python using CNN

Responsibilities:

- Identified the key aspects that govern our **CNN** model. Performed proper weights initialization, data augmentation and make the data ready for model training.
- Performed complex querying using **Analytical SQL** to extract the data from datawarehouse and inspect the key features using **BigQuery** under **GCP.**
- Performed Image scaling and extraction from the raw data by identifying the particulates of the images
- Trained the model on **TPU** using Googe colab under **google cloud platform** using **Vertex AI**
- Implemented convolution neural nets with required number of layers and essentials filters for model training.
- Model evaluated based on historical data, Model re-training with new data to improve model accuracy
- Designed and implemented machine learning pipeline end to end using Dataiku automation and held serving API's and enabled end user exports if needed.
- Performed feature importance techniques such as **Recursive feature elimination** RFE CV and Random forests to find the significant features.
- Used **Rshiny** to create web interface in RStudio to display the model results to the business platform.
- Implemented hyperparametric optimization to fine tune and find the best parameters for uplifting precision and recall scores
- Worked on microscopic Image data in fine tuning, resizing and cropping to train on **ConvNets**.
- Built and hosted web based dashboard on **Dataiku** using Bokeh library in python for data visualization, interaction and presentation through third party BI tools like **Tableau**.
- The model is finally designed of convultional layer, Relu and softmax layer and ran successfully by achieving an accuracy of 99%.
- Used **D-seq2, Logistic, RFECV** algorithms for gene sequencing and identified the gene that's causing tumors and percentages of becoming tumorous.
- Built and hosted web based dashboard on **AWS Sagemaker** for data visualization, interaction and presentation through **Python Flask** and third party BI tools like **Tableau**.
- Experience in managing **IAM** roles and creating roles using policies to create AWS services and intances for model training and storing purposes
- Created PoC's regarding the text analytics using natural language processing(**NLP**) with the algorithms **BERT, LSTM** and text summarization techniques
- Experience in writing complex queries, hands-on in writing scripts/programs for data analysis using Python in a Unix/Linux environment.
- Working experience on creating pipelines using bash to enable the python files using CLI under **GCP** to train using TPU.
- Model re-training plan and frequency has been finalized by thouroughly checking unbiased A/B testing results which led model accuracy scores reach the threshold levels of acceptance.

Environment: Dataiku, Python 3 (Scikit-Learn/SciPy/NumPy/Pandas), Tableau 9.4, CNN, S3,KNN, AWS, Shell scripting/Linux, Bigquery in GCP, XGBoost,K-means Random, Forests, SQLite,KNN, AWS, ConvNets, Bert, GCP, Vertex Workbench, AI Platform Engine, R, Rshiny

BFMS, State of Maine, Augusta, ME **Dec'17– Oct'19**
Role:Data Analytics/Data Scientist

***Project**: BFMS Reporting, Anomaly Detection & Time Series Analysis*
Anomaly detection in financial data has widely been ignored despite many organizations store, process and disseminate financial market data for interested customers to assist them to make informed decision and create competitive advantages. Considering the presence of anomalies in voluminous data from myriad data sources may generate

catastrophic decision through misunderstandings of market behavior. Therefore, in this project, we applied a standard set of anomaly detection techniques, based on nearest-neighbors, clustering, Time Series and statistical approaches, to detect rare anomalies present.

Responsibilities:

- Collaborated with database engineers to implement ETL process, wrote and optimized queries using **Oracle PL/SQL** to perform data extract
- Involved in Code Development, Unit Testing, Business Testing and Reconciliation of reports. Developed and executed test cases. Involved in Functional, Technical, Performance and Regression testing.
- Possess very good expertise in developing programs using Toad, SQL Developer and PL/SQL procedures to transfer data from legacy systems with knowledge in Unix Shell Scripting.
- Ensured data quality by identifying data anomalies using **Anomaly detection**.
- The LOF **(Local Outlier Factor**) model was used to detect the anamolies from the financial data.
- Generated various capacity planning reports (graphical) using Python packages like **NumPy**, **matplotlib**, **Seaborn and** used tableau for creating dashboards for forecasting annual and biennal reports over various departments.
- Experience in validating the data processed from the electronic systems using python automated scripts.
- Developed and maintained financial reports to clearly communicate actual results, forecasted performance and variances to plan forecast and budget using **Time Series Analysis** (**ARIMA/SARIMAX**).
- Measured the performance of many anomaly detection techniques using a number of metrics to highlight the best performing algorithm.
- Assisted and shadowed senior data scientists in creating POC's using deep learning algorithms for achieving better predictive power using **RNN's** and **LSTM**.

Environment: Python, ETL Data Stage, SQL views, Tableau, Statsmodel, PyoD outlier detection, Tableau.

ADP,New Jersey, US **Jan'17 – Dec'17**

Role: Data Analytics

Project: Driving the Decision Analytics Team in development of vital analytical tools for predicting the chance an employee staying in the company or leaving. Company and industry key performance indicators, trends, and patterns were analyzed and predicted using machine learning algorithms. The prediction methods are compared based on their efficiency and scalability in terms of estimation complexity and in terms of memory requirements for real-time predication.

Responsibilities:

- Created high-performance data processing pipelines in **Apache Spark** for data transformation, aggregation, and model training, loaded the data into **Hive** tables using **PySpark.**
- Integrated Hadoop into traditiona ETL, accelerating the extraction and loading massive structured and unstructured data using MongoDB
- Involved in all the phases of project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical modeling (decision trees, regression models, clustering), dimensionality reduction using Principal Component Analysis, testing and validation using **ROC** plot, **K - fold cross-validation** and data visualization
- Applied Data mining to analyze procurement of processes resulting an increase a savings about 40% of the corresponding financial year.
- Applied **Principal Component Analysis** method in feature engineering to analyze high dimensional data.
- Started baseline model by using statistical algorithms like **logistic regression**
- Improved baseline model by using advanced parameter optimization of machine learning algorithms like **XGBoost and Gradient Boosting(GBM)** methods.
- Built repeatable processes in support of implementation of new features and other initiatives
- Created various type of **data visualization** using **Tableau.**
- Used structured data lake in Amazon S3 to hold the raw, modeled, enhanced, and transformed data.
- Communicated and presented the results with product development UI team for benchmarking and presenting the results.
- Applied Data mining to analyze procurement of processes resulting an increase a savings about 40% of the corresponding financial year.

Environment: Python, PyMongo, Matplotlib, Pyspark, Sqlite, PCA, Random Forests, Tableau, Matplotlib, Seaborn

Broadcom, US, India
Role: Network Data Engineer **May'14 – Dec'15**

Project: The business objective was to solve the problem of Wi-Fi based indoor localization of IoT devices consists in determining the position of client devices with respect to access points has become a pertinent problem. Our team came up with an the automation process using **standard linear regression** tools to do real-time localization in future dynamic wireless indoor environments thus avoiding the time-consuming manual fitting and complex fingerprinting.

Responsibilities:

- Setting up the environment to make connectivity between various wireless devices and access points for a fading less transmission.
- Configuring the access point, client for wireless data transfer.
- Calibration of emulated of network traffic to begin the communication.
- Broadcasting a finite number packets to all other nodes and using trilateration algorithm to determine the relative position of the target device using **MATLAB** simulated environment.
- Plotting the graph bettween **RSSI-log(distance)** pairs using **rSSI package** in **R.**
- Use statistical modeling and analysis to understand relationships between diverse forms of network data and drawed insights for finding out key performance indicators,
- Collecting the data after packet generation in a log file using the radiaperf software.
- Conducted exploratory data analysis to check for outlier and replacing them with median.
- Generation of **Scatter plot** of all the reported averaged **RSSI** readings as a function of the distance (on a logarithmic scale) using **GGPLOT** in **R**.
- Calibrating the test bed initially by determining the intercept point and slope with the **linear regression** technique.
- Checking the correlation from the datapoints(RSSI-Log distance) using **cor.test** using **pKendall** and **Spearman** in R.
- Fitting the datapoints using a regression line and interpreting from the **pearson correlation Coefficient** and finding the perfectly linear fit line.
- Predicting the ranging bounds after modeling the data with the tested data set using **Linear Regression** and **extrapolating** to find out the optimum lossless packet transmission.
- Repeating the process with different node points and finding out the optimum access point.
- Used **R** to create various data visualization plots between RSSI and logarithmic distance among various nodes**.**
- Communicating and presenting the insights drawn by comparing various firmwares.
- Working knowledge on Feature testing, Regression testing & Performance testing.

Environment: R Studio, Rshiny,Radioperf, IoT testing, Microsoft Office,GGplot2, MATLAB FileZilla, WireShark, JIRA, ChariotIx.

Achievements:

1. Implemented few JIRA tracking tips which made my team to refer past test results in an efficient manner.
2. Promoted as a wlan-team leader to review the results before submitting the report to the senior IC engineers.

Infotech power ltd, Visakhapatnam, India **Oct'10– Mar'12**
Role: Forecast Analyst

Project: Grid Failure Prediction/Load Flow Studies.
NTPC is a power plant generation unit which also tracks station usage, real and reactive power flows, transaction losses, and deviation, develops and processes accurate and timely bills in accordance with contract provisions. It also conducts the fact analysis of energy, transmission, capacity and ancillary services. The analyst job is to maintain historical records of energy transactions, billings and generations/inter-tie meter data maintainence and associated applications.

Responsibilities

- Handle day to day decisions for the purchase and sale of electrical power and ancillary services for the site.
- Support Forecasting/Planning function for the generation and Ancillary Services Market to the business by simulation in **MATLAB**.
- Perform and/or direct activities including simulation, modeling, and field demonstration implementations of integrated energy applications including (but not limited to) microgrids, mobile energy systems, data centers, building networks, and distribution networks
- Calculated annual, monthly, and daily tie reports of power losses using **MS Excel**.
- Offer expert advice on the development of forecasts regarding the power market.
- Calculated the active and reactive power using excel to find out power factor.
- Calculate the penalty factor for the substations that consume more reactive power.
- Preparation of contingency table from all the historical data stored in the database.
- Used **MYSQL** to push as well as to extract the data.
- Developed the charts using pivot tables to find out the timely So2 and Co2 emissions from the factory.

Environment: Matlab,Mysql, MS word, Excel, Unix, Visio, Pivot tables

Bootcamp Projects:

 Traffic signs recognition using convolutional neural networks for self- driving cars
 Text Analysis of Donors choose data set and predicting the selection rate for a tutor's proposal screening the applications.
 Loan Prediction using tree-based approach like XGBoost ensembles by creating data pipelines and evaluating the performance of various classification metrics
 Face detection using OpenCV and computer vision

Tools, Libraries and packages used:

Python, Descriptive and inferential statistics, Linear and Logistic regression, K-means clustering, Naive Bayes and support vector machines, Tree based and ensemble methods, Neural networks- Tensor flow, keras and convolutional neural networks, OpenCV

EDUCATION

- **Computer Information Systems and Analytics (Masters)***, University of central Missouri*, CAP: 3.55/4 - Jan 2016 – Dec 2016
- **High Voltage Engineering(Masters)**, Jawaharlal Nehru Technological University, Kakinada, Aggregate 78% - Nov 2012 – Nov 2014
- **Electrical and Electronics Engineering(Bachelors)**, Andhra University.GPA: 3.25 - 2006-2010