**BHARGAVI KATTA  
PH NO: 571 721 148/571 520 2277   
EMAIL ID:** [bhargavi.dev2@gmail.com](mailto:bhargavi.dev2@gmail.com)

[linkedin.com/in/bhargavi4](https://www.linkedin.com/in/bhargavi4)

**PROFESSIONAL SUMMARY:**

* Over **9+** years of industry experience in as a Data Scientist with strong technical expertise, business experience, and communication skills to drive high - impact business outcomes through data-driven innovations and decisions.
* Proficient in managing entire Data Science project life cycle and actively involved in all the phases of project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical/machine learning modeling, analysis, testing, validation, evaluating the model and Data Visualization.
* Extensive experience in developing Data Analytics, Statistics, Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, Python, and Tableau.
* Worked with various large-scale data sets for data manipulation, data cleaning, quality control, data management and data analytics using Python, R and SQL.
* Cloud computing implementation experience using HDInsight, Azure Data Lake (COSMOS), Azure Data Factory, Azure Machine Learning &amp; PowerShell scripting.
* Experience in using various packages in Python like pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, Beautiful Soup and various packages in R like Rpy2, ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, C50, twitteR, NLP, Reshape2, rjson, plyr.
* Utilized Spark, Scala, MLLib, R, a broad variety of machine learning methods including classifications, regressions, dimensionality reduction etc.
* Data Scientist proficient in Python scripting with libraries such as Pandas, Scikit-learn, NumPy, and others plays a crucial role in extracting valuable insights from data.
* Proficient in statistical Modeling and Machine Learning techniques (Linear, Logistics, Decision Trees, Random Forest, SVM, Clustering, Neural Networks K-Nearest Neighbors) in Forecasting/Predictive Analytics, Segmentation methodologies, Regression-based models, Hypothesis testing, Factor analysis/ Principal Component Analysis (PCA) and good knowledge on Recommender Systems.
* Extensive knowledge of working with Artificial Neural Networks and Deep Learning models using TensorFlow, PyTorch packages in Python.
* Performing R&D in computer vision and machine learning.
* Led and managed research and development efforts for the computer vision team.
* Patented a computer vision and deep learning framework in C++.
* Experience on DevOps essential tools like Chef, Puppet, Ansible, Docker, Kubernetes, Subversion (SVN), GIT, Hudson, Jenkins, Ant, Maven and migrated VMWAREVMs to AWS and Managed Services like EC2, S3, Route53, ELB, EBS.
* Worked with various datasets, analyzed data using python/R, and implemented various supervised/unsupervised predictive models.
* Write scope scripts to handle multi-million rows and for daily loading data from COSMOS Server and scheduling the data load using Data Grid
* Within the development squad, facilitate technical implementations and assure data collection quality (testing) on a project by project basis and for each A/B test conducted
* Created private cloud using Kubernetes that supports DEV, TEST, and PROD environments.
* Skilled in Advanced Regression Modeling, Time Series Analysis, Statistical Testing, Correlation, Multivariate Analysis, Forecasting, Model Building, Business Intelligence tools and application of Statistical Concepts.
* Containerized all the Ticketing related applications- Spring Boot Java and Node.Js applications using Docker.
* Partner with other business and functional units to ensure that best practices in metric collection, decision making and A/B testing are being exposed to senior management and core website decision makers
* Skilled in performing Data Parsing, Data Manipulation and Data Preparation with methods including describe Data contents, compute descriptive statistics of Data, regex, split and combine, remap, merge, subset, reindex, melt and reshape.
* Experience in large datasets of structured and unstructured data, data visualization, data acquisition, predictive modeling, NLP / NLU / NLG / AI / Machine Learning / Deep Learning / Computer vision/ Probabilistic Graphical Models / Inferential statistics / Graph / Apache Spark / Data Validation.
* Data Scientist with PowerPoint skills can effectively communicate complex data analyses to non-technical stakeholders
* Strong SQL programming skills, with experience in working with functions, packages and triggers.
* Automated recurring reports using SQL and Python and visualized them on BI platform like Tableau.
* Excellent communication skills. Successfully working in fast-paced multitasking environment both independently and in collaborative team, a self-motivated enthusiastic learner.

# **TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Languages/Lib | Python, SQL, R, Go, TensorFlow, PyTorch, R, NumPy, NLP, UNIX |
| Database/OS | MySQL, Teradata, Oracle, Windows, Linux/Unix,COSMOS |
| BI Tools | Tableau, Power BI, Alteryx. |
| Statistics | Hypothetical Testing, ANOVA, Confidence Intervals, Bayes Law, MLE, Fish Information, Principal Component Analysis (PCA), Cross-Validation, correlation. |
| Python Libraries | Beautiful Soup, NumPy, SciPy, Pandas, Matplotlib, Python-Twitter, urllib2, TensorFlow, Keras, PyTorch, PySpark. |
| Technologies: | Git, NoSQL, SQL, Docker, Pandas, CI/CD, Azure, Kafka, Scala, Statistics, Data Mining, Big Data |
| Big Data Ecosystems | Hadoop, Hive, Pyspark |
| Others | Microsoft Excel (Advanced), PowerPoint, Visio, D3.js, Universe Designer, IDT, PHP, Jupyter Notebook, GitHub. |

# **PROFESSIONAL SUMMARY:**

**CAPITAL ONE BANK, HERNDON VIRGINIA, USA**  **MAY 2021 – TILL**

**DATA ANALYST/ DATA SCIENTIST**

**RESPONSIBILITIES**:

* Designed algorithms to identify and extract incident alerts from a daily pool of incidents.
* Reduced redundancy among incoming incidents by proposing rules to recognize patterns.
* Performed exploratory data analysis like calculation of descriptive statistics, detection of outliers, assumptions testing, factor analysis, etc., in Python and R,
* Worked on NGP Onboarding to Implement Delete Processor in Cosmos/Azure SQL
* Utilized Spark, Python, R, a broad variety of machine learning methods including classifications, regressions, dimensionality reduction based on domain knowledge and customer business objectives.
* Innovated and leveraged machine learning, data mining and statistical techniques to create new, scalable solutions for business problems.
* Collaborate with external partners such as agencies to assist with data collection, reporting and A/B test development
* Use Pandas for data manipulation and cleaning.
* Effective software development processes to customize and extend the computer vision and image processing techniques to solve new problems
* Worked with Machine learning algorithms like Regressions (linear, logistic etc...), Clustering and classification, SVMs and Decision trees.
* Build analytic models using a variety of techniques such as logistic regression, and pattern recognition technologies.
* Use PowerPoint to structure and present the data story in a clear and engaging manner.
* Involved in automating end-2-end model refresh process using COSMOS ,Azure ML AND Powershell scripting
* Created A/B testing and MVT (Multivariate) testing roadmaps for IR 500 clients of Confidential to optimize websites for Confidential elements.
* Managed Kubernetes charts using Helm. Created reproducible builds of the Kubernetes applications, managed Kubernetes manifest files and Managed releases of Helm packages.
* Container management using Docker by writing Docker files and set up the automated build on Docker HUB and installed and configured Kubernetes.
* Write Scope Scripts to handle multi-million rows and for daily loading data from COSMOS Server and scheduling the data load using DataGrid.
* Cleanse and preprocess data to handle missing values, outliers, and inconsistencies.
* Building/Maintaining Docker container clusters managed by Kubernetes, Linux, Bash, GIT, Docker, on GCP . Utilized Kubernetes and Docker for the runtime environment of the CI/CD system to build, test deploy.
* Automated applications and MySQL container deployment in Docker using Python and monitor them using Nagios.
* Virtualized servers in Docker as per test environments and Dev-environments requirements and configured automation using Docker containers.
* Performed Exploratory Data Analysis using R. Also involved in generating various graphs and charts for analyzing the data using Python Libraries.
* Work on data that was a combination of unstructured and structured data from multiple sources and automate the cleaning using Python scripts.
* Work on different data formats such as JSON, XML and applied machine learning algorithms in Python.
* Performed Data Profiling to assess data quality using SQL through complex internal database
* Extracted data from the database using Excel/Access, SQL procedures and created Python and R data frames for statistical analysis, validation, visualization, and documentation.
* Use Scikit-learn for building machine learning models.
* Developed and designed SQL procedures and Linux shell scripts for data export/import and for converting data.
* Create A/B test plans and manage implementation
* Integrate charts and graphs directly into PowerPoint slides to illustrate trends and patterns.
* Designed data profiles for processing, including running SQL, and using Python and R for Data Acquisition and Data Integrity which consists of Datasets Comparing and Dataset schema checks.

**FANNIE MAE VA REMOTE MAY 2019 – MAY 2021**

**DATA SCIENTIST**

**RESPONSIBILITIES:**

* Worked in IT team to create ETL pipelines for a centralized Datawarehouse for reporting extracting data from multiple sources.
* Collaborated with data engineers and operation team to implement the ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Worked on data cleaning and ensured data quality, consistency, integrity using Pandas, NumPy.
* Explored and analyzed the customer specific features by using Seaborn in Python and dashboards in Tableau.
* Performed data imputation using Scikit-learn package in Python.
* Participated in features engineering such as feature generating, PCA, feature normalization and label encoding with Scikit-learn pre-processing.
* Calculate and interpret descriptive statistics, including measures of central tendency, dispersion, and skewness.
* Used Python 2.x/3.X (NumPy, SciPy, Pandas, Scikit-learn, Seaborn to develop a variety of models and algorithms for analytic purposes.
* Developed ADF pipelines to move data from on - premise source systems to COSMOS, from COSMOS (with data transformation) to Azure Warehouse (staging), from Azure warehouse to Azure ML (for scoring) and appending scores back to the data in Azure warehouse
* Developed and Scheduled Data Load jobs in Data Studio Tool/Cosmos UI using Scope Scripts from both Structured and Non-Structured streams.
* Worked closely with diverse teams of Strategic Growth Manager, Technical teams and account managers to overcome objections from clients using A/B testing, MVT (Multivariate testing) as tools to overcome objections.
* Drive core insights from available data to suggest A/B tests that drive improvements in site execution and experience
* Experimented and built predictive models including ensemble methods such as Gradient Boosting trees and Neural Network using Keras to predict Sales amount.
* Use tools like Dask for parallel computing with Pandas.
* Conducted analysis and patterns on customers' shopping habits in a different location, different categories and different months by using time series modeling techniques.
* Leverage PowerPoint's features to design informative and aesthetically pleasing presentations.
* Used RMSE/MSE, R Squared Values to evaluate different models' performance.
* Designed rich data visualizations to model data into human-readable form with Tableau and Matplotlib.
* Predicted user preference based on segmentation using General Additive Models, combined with feature clustering, to understand non-linear patterns between user segmentation and related monthly platform usage features (time series data).
* Effective communication and collaboration with stakeholders are essential for translating statistical findings into actionable recommendations.
* Partner with squad members to identify, implement and analyze A/B tests
* Perform data manipulation, data preparation, normalization, and predictive modeling.
* Improve efficiency and accuracy by evaluating model in Python and R.
* Application of various machine learning algorithms and statistical modeling like Decision Trees, Random Forest, Regression Models, neural networks, SVM, clustering to identify Volume using scikit-learn package.
* Customize PowerPoint presentations to align with the organization's branding and style guidelines.
* Performed Bootstrapping techniques to the improve the model performance.

**APPLE CUPERTINO CA JULY 2018 – MAY 2019**

**DATA ANALYST/DATA SCIENTIST**

**RESPONSIBILITIES:**

* Building, publishing customized interactive reports and dashboards, report scheduling using Tableau server.
* Perform Data Profiling to learn about behavior with various features of turnover before the hiring decision, when one has no on-the-job behavioral data.
* Performed preliminary data analysis using descriptive statistics and handled anomalies such as removing duplicates and imputing missing values.
* Create visualizations that highlight key trends, anomalies, and patterns in the data.
* Time Series Analysis
* Applied various machine learning algorithms and statistical Modeling like decision trees, text analytics, natural language processing (NLP), supervised and unsupervised, regression models, social network analysis, neural networks, deep learning, SVM, clustering to identify Volume using Scikit-learn package in Python.
* Involved in creating Hive tables, loading the data and writing Hive queries which will run internally in map reduce way.
* Designed and managed API system deployment using fast HTTP server and Amazon AWS architecture
* Setup database in AWS using RDS and configuring backups for S3 bucket.
* Performed data cleaning and feature selection using MLLib package in PySpark.
* Conducted a hybrid of Hierarchical and K-means Cluster Analysis and identified meaningful segments of through a discovery approach.
* Built Artificial Neural Network using TensorFlow in Python to identify the customer's probability of canceling the connections. (Churn rate prediction)
* Understanding the business problems and analyzing the data by using appropriate Statistical models to generate insights.
* data scientist with strong Python scripting skills is well-equipped to tackle a variety of tasks in the end-to-end data science process,
* Use PowerPoint's interactive features to engage the audience during presentations.
* Having Knowledge on AWS Lambda, Auto scaling, Cloud Front, RDS, Route53, AWS SNS, SQS, SES.
* Gained Knowledge on Deploying apps using AWS Cloud Formation.
* Developed NLP algorithms coupled with Deep Learning for Topic Extraction, Sentiment Analysis.
* Identify and assess available machine learning and statistical analysis libraries (including regressors, classifiers, statistical tests, and clustering algorithms).
* Work with NLTK library to NLP data processing and finding the patterns.
* Categorize comments into positive and negative clusters from different social networking sites using Sentiment Analysis and Text Analytics.
* Partnered with account executives to define campaign strategy (including cross-channel treatments, control groups &amp; A/B multivariate testing)
* Ensure that the model has low False Positive Rate for text classification and sentiment analysis on unstructured and semi-structured data.
* Data Scientist specializing in data analysis, exploratory data analysis (EDA), data cleansing, and statistical analysis is responsible for extracting meaningful insights from data.
* Addressed overfitting by implementing the algorithm regularization methods like L2 and L1.
* Use Principal Component Analysis in feature engineering to analyze high dimensional data.
* Create and design reports that will use gathered metrics to infer and draw logical conclusions from past and future behavior.
* Perform Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python.
* Environment: Python, Linux, PySpark, Tableau, SQL.

**SIGNIFY HEALTH TX JAN 2016 - JUL 2018**

**DATA SCIENTIST**

**RESPONSIBILITIES:**

* Stored and manipulated data using SQLite 3, NumPy and Pandas.
* Visualized the data using the pie charts, chloropleth maps, line charts with a user-interface which includes a slider (for years) and dynamic suggestion (for the country name) in python.
* Identified the age category where the mortality rate is high (i.e. neonatal, infant, under 5 years or age 5-14 years), in order to understand where the country has to focus more next to reduce the child mortality.
* Applied logistic regression and various classification techniques to predict credit scores and validated the model.
* Researched various machine learning algorithms such as Decision trees, Naïve Bays, SVM, KNN and applied to predict loan acceptance to improve the previous model and increase the prediction accuracy.
* Validated the results using the cost matrix which is calculated based on cost incurred due to false positives and false negatives.
* Conduct thorough EDA to understand the distribution, relationships, and patterns within the data.
* Used Python and R Markdown to present the reports with the ROC plots, graphs and results of the models.
* Developed, recommended and assisted with the implementation of business strategies to manage credit risks, increase revenue and reduce exposure to credit losses.
* Built linear regression model to predict the interest rate for a loan application.
* Worked on a huge transactional data set of 40 million rows and performed exploratory data analysis.
* Converted continuous variables to dummy variables for performance improvement.

**ACCENTURE TX JULY 2014-DEC-2015**

**SOFTWARE ENGINEER**

**RESPONSIBILITIES:**

* Streamlined and restructured the backup jobs to optimize the use of resources.
* Used R to support the SQL analysis by implementing exploratory data analysis.
* Used Python and R Markdown to present the reports with the ROC plots, graphs and results of the models.
* Developed various reusable scripts to validate the batch files which are generated on a daily, weekly, and monthly basis, in both system integration phase and testing phase.
* Performed data conversion, consolidation, and validations, as per the business requirements.
* Developed service for sending an email, push and in-app notifications, for recommending latest offers, based on customer interests. Involved in features such as delivery time optimization, tracking, queuing, and A/B testing. Built an internal app to run batch processing for software delivery etc. (Python, Flask, MS SQL, Recommendation System).
* Simplified bulk data processing and injection service from global exchanges to CTRM and provides preprocessed data for application users. (Python, Machine Learning, NLP).
* Utilize statistical summaries, visualizations, and descriptive statistics to gain insights.
* Collaborated with cross-functional teams for collecting, understanding the business requirements, well documented them, and translated requirements into technical solutions.
* Analyzed and performed POC over Airtel Payment Bank System, using libraries (NumPy, Pandas).
* Enhanced productivity by developing a web application as MDM system, in an Agile Software development environment. (Java, Spring MVC, Hibernate, JQuery, PostgreSQL).
* Used tools in the Hadoop ecosystem (e.g., Hive, Apache Spark) to prepare complex data sets for use in building predictive models.
* Build SQL queries and procedures to retrieve data from the PostgreSQL database, by performing CRUD operations.