**SURESH** KUMAR MADDALA

[www.linkedin.com/in/suresh-datascientist](http://www.linkedin.com/in/suresh-datascientist)

rajasekhar.a@meridiansoft.com

3803883682

# Professional Summary:

Highly efficient Data Scientist with over 10 years of experience **in Machine Learning, Deep Learning and Big Data** and different **cloud technologies** with large amounts of Structured and Unstructured data**.** Adept in statistical programming languages like **Python & R**, Big Data technologies like **Spark.**

* Analyzed configuration requirements for nodes, networks, and systems. Prepared detailed plans for implementing configuration changes and integrations. Executed configuration changes according to established procedures.
* 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, feature scaling, feature engineering, machine learning modeling using a variety of algorithms like **Regression Models, Time Series(ARIMA & Holt Winters), Clustering, APRIORI, Decision Trees, KNN, Neural Networks, SVM, Ensemble(Random Forest, Boosting),** dimensionality reduction using **PCA and SVD,** testing and validation using regular performance metrics and **AUC**-**ROC Plot, K- fold cross validation** and **data visualization**.
* Experienced in verifying the successful implementation of configuration changes and integrations. Provided ongoing support for systems and applications post-rollout. Assisted different teams in the deployment and rollout of new systems. Maintain and manage system configurations, ensuring long-term performance and stability and migrating systems to new hardware or cloud environments as needed.
* Experienced in creating and utilizing **EC2 (AWS) and Compute Engine (GCP) instances.**
* Experienced in **Text Analytics and NLP**. Developed applications using **LDA**, **LSA** and advanced **text vectorizing** techniques like **TF-IDF, Word2Vec, Doc2Vec.**
* Adept and deep understanding of Statistical modeling, Multivariate Analysis, model testing, problem analysis, model comparison and validation.
* 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 using various packages in **R** and **Python** like **ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, C50, twitteR, NLP, Reshape2, rjson, plyr, pandas, numpy, seaborn, SciPy, Matplotlib, scikit-learn, tensorflow, keras, nltk.**
* Extensive experience in generating data visualizations using **R, Python** and creating dashboards using tools like **RShiny.** Proficient in **Oracle and MySQL** databases.
* Hands on experience in implementing **Naive Bayes** and skilled in **Random Forests, Decision Trees, Linear and Logistic Regression, SVM, Clustering, neural networks, Principal Component Analysis**.
* Experience in using google **big query** and **cloud storage** for implementing different projects in the retail industry.
* Extensive experience in Data Visualization including producing tables, graphs, listings using various procedures and tools such as **Tableau**.

**Technical Skills: Operating systems** Ubuntu, Windows

**Languages** R, Python, PySpark

**AI Platforms** DataRobot, GCP Cloud AI, MLFlow, AWS Sagemaker

**Databases** MySQL, Oracle

**Cloud Technologies** GCP, AWS. Azure

**Packages** ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, C50, twitteR, NLP, Reshape2, rjson, plyr, pandas, NumPy, seaborn, SciPy, Matplotlib, statmodels, Scikit-learn, tensorflow, keras, PyTorch, nltk, genism

**Reporting Tools** PowerBI, Tableau, Looker

**Version controls** GitHub

# Professional Experience: Walmart Associates May 2022 – Current

Customer Experience Score Senior Data Scientist **Responsibilities:**

* Executing full life cycle of projects through project planning, data collection, model prototyping and deployment, with responsibilities encompassing stakeholder management and communication to cross- functional partners.
* Providing leadership guideline for adherence to **MLOps** framework to create, build, and deploy data science algorithms to continuously test and monitor drifts.
* Partnering closely with Data Science Management, Business Leaders, and ML engineering leaders to inform, drive and accelerate innovations in discovery experiences via Insights, frameworks, **causal inference solutions** and **machine learning prototypes**.
* Engage with variety of business and product stakeholders to bring vision on applying broad knowledge of ML/AI capabilities to identify growth areas.

**Environment:** Big Query, GCS, Vertex AI, Scikit Learn, Causal Inference

# Niagara Bottling Apr 2021 – Apr 2022

Spot market Price Prediction Data Scientist **Responsibilities:**

* Developed a regression model using advanced boosting techniques to predict the spot market price resulted in saving **$12M** in transportation cost per.
* Performed Exploratory Data Analysis and statistical analysis on large datasets using **SQL**, **Python** and

# PySpark.

* Built Machine Learning models using Python and **PySpark** and developed Machine Learning workflows using MLOps principles.
* Developed **CI/CD** pipelines using **Azure DevOps** and used Git for source control management.
* Built MLOps infrastructure for model deployment and model monitoring using cloud technologies.

**Environment:** Azure Databricks, PySpark, Azure ML Studio, Azure Devops.

# United States Steel Nov 2020 – Mar 2021

Forecasting solution using Deep Learning Data Scientist

# Responsibilities:

* Developed a time series forecasting solution using a **Long Short Term Memory** network avoiding a **$5M production loss** for a steel manufacturer.
* Performed **Exploratory Data Analysis** and **data visualizations** using **tableau** to understand patterns in the data.
* Developed a streaming pipeline to push all the sensor data to the cloud **S3** bucket using **Kafka. Environment:** Python 3.x, AWS Sagemaker, S3, Kafka, Lambda functions

# American Electric Power Mar 2020 – Oct 2020

Dynamic Stability Signal Classification Data Scientist

# Responsibilities:

* Collaborating with the engineering team and domain experts to understand the business problem and to come up with a solution.
* Driving the project completely from data gathering, pre-processing, feature engineering and model building.
* Creating **generalized frameworks** for data **pipelines** which can read signal data from multiple files, concatenate all of them by matching the column names and merging it with metadata file to get the target.
* Created an exhaustive list of features for the machine learning model from the only one column of signal data provided by the client.
* Implemented bagging and boosting algorithms using **Random Forest, AdaBoost, XGBoost and LightGBM** to classify the signals.
* Automating the training pipeline, inference pipeline using shell script.

# Environment: Python 3.x, AWS EC2, scikit learn, tableau

**Medical Mutual Sep 2019 – Feb 2020**

Identifying Insurance Duplicate Claims Data Scientist

# Responsibilities:

* Identifying, extracting and integrating large data from **MySQL**, **Hive** Databases.
* Handling importing data from various data sources to perform transformations using **PySpark** and load data into **HDFS.**
* Performed extensive feature engineering at claim level, location level based on the business understanding.
* Extracted the optimal number of features required to be trained by performing various dimensionality reduction methods like Backward Feature Elimination, Forward Feature Construction and **PCA.**
* Developed **machine learning pipeline** using **PySpark** and integrated with mlflow to ease deployment process. Performed **SHAP** analysis to explain the model predictions.

# Environment: Python 3.x, PySpark, scikit learn, mlflow, HDFS, Hive

**HDFC Bank Mar 2019 – Aug 2019**

Project: Fraud Detection in Banking Sector Data Scientist

# Responsibilities:

* Performed **Data preparation, Data Wrangling – normalizing, transforming** all the values to a common scale, and identifying outliers from the data collected.
* Created an exhaustive list of window-based aggregates using **PySpark UDFs** at different levels of the transaction.
* Implemented various data sampling techniques like **Random Up-sampling & Down-sampling**, and **Cluster based Over-sampling** using **imblearn** library in **Python** to adjust the quantity of imbalance between classes.
* Implemented various advanced ML algorithms like Naive Bayes, Support Vector Machines, XGBoost, Random Forest, resulting in an improvement of 30% in fraud identification rate and a reduction of 45% in false positive rate across all channels.

# Environment: Python 3.x, PySpark, Hive, HDFS, scikit learn

**Edelweiss Group Sep 2018 – Feb 2019**

Scanalityx – Financial Statements Data Scientist

# Responsibilities:

* Converting all the pages of pdf documents into separate images using **pdf2image** python package.
* Used **Libelling** for creating XML annotations of different tabular structures available in the images.
* Assisted other team members in using the **LabelImg** annotation tool to create more labelled data for the modelling activity.
* Implemented various advanced **machine learning** algorithms to classify the financial document as ITR-V, Balance Sheet, Profit & Loss Statement.
* Implemented a **VGG-16** based **deep learning** approach to detect the regions of tables in the financial statements.

# Environment: Python 3.x, Tensorflow, Keras, scikit learn

**Camara de Comercio de Bogota Feb 2018 – Aug 2018**

Smart Affiliate Management – Churn Analytics Data Scientist

# Responsibilities:

* Actively participated in the **sprint planning**, creating **user stories** for the business analysis team who are at the client location.
* Creating a unified view of data for the affiliates by joining their data from the past transactions, previous renewal information.
* Creating a target column from the business logic of the renewal process.
* Implemented different machine learning models to come up with probabilities to categorize the affiliates into green, orange, and red buckets.

# Environment: Python 3.x, MySQL, scikit learn, jira

**Hinduja Global Solutions Jun 2017 – Jan 2018**

Propensity to pay – Health Insurance Data Scientist

# Responsibilities:

* Interacting with different stakeholders and domain experts in the organization to understand the business process and coming up with a solution.
* Gathering all the data that is required from multiple data sources and creating datasets that will be used in analysis.
* Performed **Exploratory Data Analysis** and **data visualizations** using **ggplot2** to understand patterns in the data.
* In Preprocessing phase, used **pandas** and **numpy** to remove or impute missing values, detect outliers, scale features, and applied feature selection (filtering) to eliminate irrelevant features.
* Performed feature extraction from various data sources and created an extensive list of features for the machine learning model.
* Divided the business problem into 5 different sub problems, developed separate models for the sub problems and created a **scoring model** to prioritize the importance of claims.
* Implemented **Ridge, Lasso regression and Elastic Net Regression** to reduce the initial model complexity and handled over fitting of the data. Performed **K-Fold Cross Validation** sampling technique to train various models.
* Built multiple machine learning algorithms using **scikit-learn,** assessed, and tested. **Decision trees** were finalized based on the feasibility, ease of explanation and accuracy of results.

# Environment: Python, R, Tableau, Oracle, scikit learn

**Axis Bank Jul 2016 – May 2017**

Smart Routing of Emails Data Scientist **Responsibilities:**

* Built frameworks which can execute various **text pre-processing & NLP** techniques like converting to lowercase, removing punctuations and stop-words, converting number to words, removing white spaces and special characters, stemming and lemmatization using **NLTK** library.
* Implemented various data sampling techniques like **Random Up-sampling & Down-sampling**, and **Cluster based Over-sampling** using **imblearn** library in Python to adjust the quantity of imbalance between classes.
* Created feature vectors by converting each text record to a vector embedding using **advanced text vectorization techniques** like **TF-IDF** and **Word2Vec** using **gensim library.**
* Extracted the optimal number of features required to be trained by performing various data dimensionality reduction methods like **Backward Feature Elimination, Forward Feature Construction** and **PCA.**
* Implemented various advanced ML algorithms like **Naive Bayes, Support Vector Machines, XGBoost, Random Forest**. Performed **K-Fold Cross Validation** sampling technique to train the data with these algorithms. The model had the capability to classify with 93% accuracy.

# Environment: R, Python, tableau, scikit learn, SQL Server

**Audifarma Droguerías Nov 2015 – Jun 2016**

Pharmacovigilance – Drug Interactions Data Scientist

# Responsibilities:

* Extracted hidden symptoms and adverse events reported by the patients from large amount of unstructured data.
* Created a data model for the client to do cohort analysis in an efficient way.
* Implemented **Association Rule Mining** algorithm called **FPGrowth** using **PySpark. Environment**: Python, Scikit-Learn, PySpark, SQL

# Axis Bank Jun 2015 – Oct 2015

Big Data Lake POC Data Scientist **Responsibilities:**

* Engage with the product owners and SMEs to understand the business problem.
* Understanding the Informatica workflows, converted the transformations written in **PL-SQL** stored procedures into **Hive** and **PIG** functions.
* Written a **shell script** by embedding the **SQOOP** functionality to fetch all the required tables from oracle database to **HDFS.**
* Created an **Oozie** workflow by incorporating all the components and written a job scheduler.

**Environment:** Hadoop, Hive, PIG, Sqoop, Oozie, Oracle

# Tata Steel Aug 2014 – May 2015

Dealer Segmentation Data Scientist **Responsibilities:**

* Collated dealer information available in different files by matching the dealer names using **FuzzyWuzzy**, **Levenshtien** distance.
* Data cleansing, designing schema for pushing data into **RDBMS** and extensive feature engineering for model building.
* Created a unified view of the data by combining data from all different sources.
* Segmented the dealers into different groups using different **clustering** algorithms.
* Implemented various advanced ML algorithms like **Naive Bayes, Support Vector Machines, XGBoost, Random Forest**. Performed K-Fold Cross Validation sampling technique to train the data with these algorithms.
* Built a **RESTful** Python API using **Flask** library which was capable of ranking the dealers and displaying the best and least performing dealers in that area.

**Environment:** R, MS Excel, Oracle

# Coromandel Fertilizers Jan 2014 – Jul 2014

Fertilizer Demand Forecasting Data Scientist **Responsibilities:**

* Extracted and explored data from different data files present in **MySQL** database using **RMySQL** library.
* Implemented R functions which can read data from different databases, performed various data cleaning and processing techniques, multiple file integrating techniques.
* Used **R** for **Exploratory Data Analysis** and identified influential data points, identified missing values and replaced them using various Imputation techniques.
* Extracted **ACF-PACF** plots from the data to analyse Trend, Seasonality and Randomness.
* Implemented various time series techniques ranging from as simple as Smoothing to complex algorithms like **ARIMA** and **Holt Winters.** Created various functions which can give our predictions for the next 7, 14, 21 days.

**Environment:** R, Shiny and MS SQL Server

**Education:**

* Master of Technology (Artificial Intelligence) 2011 – 2013, University of Hyderabad, India