### Name: Naveen kumar

**Contact: +1 (646)-722- 7112**

**Email :** [**naveenkumar.88982@gmail.com**](mailto:naveenkumar.88982@gmail.com)

**Machine Learning Researcher /**

**Data Scientist**

**SUMMARY:**

* Data/Machine Learning (ML) Scientist with 9+ years of experience in ML. Data Analysis/mining with large data sets of Structured and Unstructured data. Data Acquisition. Data Validation, Predictive modeling. Data Visualization, Web Scraping. Natural Language Processing (NLP)
* Expert in statistical programming languages like Python and familiar with Big Data technologies like Hadoop. Hive. spark. Pig PySpark, Spark SQL
* Worked with different modules of recommendation system such as content based and collaborative filtering including matrix factorization techniques.
* Experience with a variety of NLP methods for information extraction, topic modeling, sentimental analysis, parsing, and relationship extraction with developing, deploying, and maintaining production NLP models with scalability using NLTK Textblob, spacy, gensim.
* Experience in working with relational databases (Teradata. Oracle) with advanced SQL programming skills.
* Experience in using various packages in R and python-like ggplot2. NLP. Reshape2. pandas. NumPy, Seaborn, Scipy. Matplotlib, sci-kit-learn. Beautiful Soup. Keras, PyTorch, and TensorFlow.
* Hands on experience in implementing LDA, Naive Bayes and skilled in Random Forests. Decision Trees. Linear and Logistic Regression, SVM, Clustering, neural networks. Principal Component Analysis, Boosting.
* Expertise in implementing Time Series Models using RNN's, LSTM'S, ARIMA, SARIMA.
* Experience of implementing deep learning algorithms such as Artificial Neural network (ANN), Convolution Neural network (CNN) and Recurrent Neural Network (RNN), tuned hyper-parameter and improved models with Python packages TensorFlow.
* implemented, tuned, and tested the model on AWS EC2 with the best algorithm and parameters.
* Created a machine learning model for optimizing thin films parameter for best working solar cell devices using deep learning during my research.
* Machine Learning (ML) was used for providing new powerful tools to extract essential information from large amounts of data, either from experiments or simulations.
* Extracted data from HDFS and prepared data for exploratory analysis using data munging.
* Experience with NLP for sentiment analysis, topic modeling, voice recognition, language understanding, language generation, and entity extraction.
* Employing various SDLC methodologies such as Waterfall. Agile, Kanban and SCRUM methodologies.
* Engineering Machine Learning Algorithms, Validation and Visualization.
* Proficient in Python and its libraries such as NumPy, Pandas, Scikit-learn, Matplotlib and Seaborn.
* Deep knowledge of SQL languages for writing Queries. Stored Procedures. User-Defined Functions, Views, Triggers. Indexes etc.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Languages | C, C++, Java, XML. R/R Studio, Python 2.x/3.x, SQL Maven, spark 2. 2.3, Spark SQL. Spark Streaming, Shell Scripting, Scala. Hadoop, MapReduce (Packages: Stats, Zoo. Matrix, data table, OpenSSL), HDFS, Eclipse. Anaconda, Jupyter notebook) |
| Statistics | Hypothetical Testing, Confidence Intervals, Bayes Law. MLE, Fish Information, Principal Component Analysis (PCA), Cross Validation, correlation.NO SQL Databases: Mongo DB. Cassandra, Hbase |
| Algorithms | Logistic Regression, Lasso Regression. Linear Regression, Random Forest, XG Boost. KNN. SVM. Teradata. Generalized Linear Models, Al. K-Means, SVN, Clustering, PuTTY, WinSCP. Redmine (Bug Tracking, Documentation, Scrum), Neural Networks. Al. Tableau, GitHub. |
| BI Tools | Splunk. Tableau, server. Tableau Reader. SAP Business Objects, OBIEE. SAP Business Intelligence. Amazon Redshift. QlikView, Azure Data Warehouse, Azure Data Factory. SSIS |
| Data Analysis and Data Science | Deep neural networks. Logistic regression. Decision Tress, Random Forests, KNN. XG Boost. |
| Ensembles (Bagging, Boosting) | Support Vector Machines. Neural Networks, graph/network analysis, and time series analysis (ARIMA model), NLP, CNN. RNN. Recommended models Reporting Tools: Tableau, Power BI. SSRS Big Data: Hadoop, HIVE, HDFS, PuTTY, Spark Scala Shoop. |
| Spark, MongoDB. Hbase | AWS Database Design Tools and Data Modeling: MS Visio, ERWIN 4.5/4.0. Star Schema/Snowflake Schema modeling. Fact & Dimensions tables, physical & logical data modeling, Normalization and De-normalization techniques, Kimball & Inmon Methodologies |

**EDUCATION:**

* Bachelor of Computer science and Engineering (CSE) Hindustan Institute of Technology & Science, Chennai, India

**EXPERIENCE:**

**DTCC New Jersey, US April 2022 – Present**

**Machine Learning Scientist / Data Scientist**

**Responsibilities:**

* Participated in all phases of project life including data collection, in the development, validation and delivery of algorithms. creating statistical models and reports. Solved complex analytical problems.
* Created recommendation system using content based and collaborative filtering such as memory based, and model based collaborative filtering including NMF (Non-negative matrix factorization), and SVD (Singular Value Decomposition). Siamese ANN deep learning, and LightFM. Also compared and used different KNN and APRIORI recommendation models.
* Used both supervised and unsupervised anomaly detection techniques such as DBSCAN, Isolation Forests. Local Outlier Factor, One-Class Support Vector Machines, and deep learning (Autoencoder) etc.
* Worked with time series analysis to forecast the sale of product using moving average, stationarity, autocorrelation, SARIMA VAR. LSTM, and SARIMAX
* Utilize Spark SQL PySpark. Data Lake, TensorFlow. Kafka. Spark Streaming. MLLib. Python was used in a wide variety of machine learning methods including classifications, regressions, recommendation, and dimensional reduction etc.
* Apply various machine learning algorithms and statistical modeling like decision trees, text analytics, natural language processing (NLP), supervised and unsupervised, regression models, neural networks, deep learning. SVM clustering to identify Volume of product using scikit-learn package in python and pySpark
* Various data science process such as data mining, data collection, data cleansing, dataset preparation was used to create machine learning models such as trend analysis, predictive modeling, machine learning and statistics, and other data analysis techniques was used to collect, explore, and identify the data to explain customer behavior and segmentation, text analytics and big data analytics, product level analysis and customer experience analysis.
* Built and tested different Ensemble Models such as Bootstrap aggregating. Bagged Decision Trees and Random Forest. gradient boosting, involved in agile environment to implement agile management ideals such as sprint planning daily standups, managing project timelines, and communicating with clients to ensure project progress satisfactorily.
* Developed reports, charts, tables, and other visual aids using powerBI and tableau in support of findings to recommend business direction or outcomes.
* Designed and built ETL pipelines to automate ingestion of structured and unstructured data.
* Various clustering techniques such as K-means. Gaussian mixture, DBSCAN was compared and used for customer segmentation for appropriate product pricing. Develop customized marketing campaigns.
* Worked with Computer vision using Fast and Mask R-CNN in Keras for Object Detection in Photographs and find the defected components, extensively involved in writing SQL queries (Sub queries. nested queries, views. Join conditions, removal of duplicates) RDBMS, and Spark SQInvolved in team meetings, discussions with business teams to understand the business use cases.
* Performed EDA to take care of missing values, handling Categorical features. Handling imbalance data set and feature selection by using Python.
* Found the model Performance, based on prediction probabilities, Recall and Precision tradeoff and F1 score.
* Design built and deployed a set of python modeling APIs for customer analytics, which integrate multiple machine learning techniques for various user behavior pattern prediction and support multiple marketing segmentation programs.
* Segmented the customers based on demographics, geographic, behavioral, and psychographic data using K-means Clustering. Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using Python and Tableau.
* Azure Machine Learning was used for building, training, and deploying machine learning models faster using drag-and-drop designer and automated machine learning.
* Rapidly build and deploy Azure machine learning models with no code designer and visual machine learning studio. Built-in collaborative Jupyter notebooks to accelerate model creation with the automated machine learning, and access built-in feature engineering, algorithm selection, and hyper parameter sweeping to develop highly accurate models.
* Extensively used Power Bl. Pivot Table and Tableau to manipulate large data and develop visualization dashboard.
* Apache Spark was used for bigdata processing, streaming SQL Machine Learning (ML). Exploratory Data Analysis (EDA) and feature extraction Pandas Data frame, NumPy, Jupyter Notebook, SciPy, scikit-learn, TensorFlow. Keras, and Theano was used as a tool for Machine Learning and Deep Learning.
* Wrote complex SQL statements to interact with the RDBMS database to filter the data and data analytics. Design, development, and implementation of performant ETL pipelines using python API (pySpark) of Apache Spark on Azure Databricks.
* Connected to files, relational and Big Data Source using Tableau to Visually analyze, and process data. Tableau was also used to create and distribute an interactive and shareable dashboard to see the trends, density, and variations, of the data in the form of graphs and charts.
* Worked with cloud computing to store, retrieve, and share large quantities of data in Azure is the Azure Data Lake Read and wrote to Data Lake from Apache Hadoop. Apache Spark and Apache Hive PCA was used for dimensional reduction and created the K-means clustering.
* Collaborated with product management and other departments to gather the requirements. Performance of the model was Improved using K-fold cross Validation technique and the data was tested to enhance the model on the sample data before finalizing the model. Confusion Matrix and ROC Chart were used to evaluate the classification model.
* Application of 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.
* Built and tested different Ensemble Models such as Bootstrap aggregating, Bagged Decision Trees, and Random Forest. gradient boosting. XGBoost, and AdaBoost to improve accuracy, reduce variance and bias, and improve stability of a model.
* Performed MapReduce jobs and Spark analysis using Python and R for machine learning and predictive analytics models on big data in Hadoop ecosystem on AWS cloud platform as well as some data from on-premises SQL.
* Developed Spark/Scala. Python. R for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources. Used clustering technique K-Means to identify outliers and to classify unlabeled data.
* Worked in the agile environment to implement agile management ideals such as sprint planning, daily standups, managing project timelines, and communicate with clients to ensure project progress satisfactorily.

**Franklin Templeton Hyderabad, India Jan -2017- Dec-2021**

**Sr Python Developer**

**Responsibilities:**

* Integrated data from multiple data sources or functional areas, ensures data accuracy and integrity and updates data as needed using SQL and Python.
* Expertise leveraging SQL, Excel, and Tableau to manipulate, analyze and present data.
* Performs analysis of structured and unstructured data to solve multiple and/or complex business problems utilizing advanced statistical techniques and mathematical analyses.
* Developed advanced models using multivariate regression, Logistic regression. Random forests, decision trees and clustering.
* Used Pandas. NumPy, Seaborn. Scikit-learn in Python for developing various machine learning algorithms.
* Build and improve models using natural language processing (NLP) and machine learning to extract insights from unstructured data.
* Applied predictive analysis and statistical modeling techniques to analyze customer behavior and offer customized products, reduce delinquency rate, and default rate. Lead to fall in default rates from 5% to 2%.
* Applied machine learning techniques to tap into new markets and new customers and put forth my recommendations to the top management which resulted in an increase in the customer base by 5% and customer portfolio by 99%.
* Analyzed customer master data for the identification of prospective business, to understand their business needs, built client relationships and explored opportunities for cross-selling of financial products. 60% (Increased from 40%) of customers availed more than 6 products.
* Experienced in implementing Time Senes Models using RNN's, LSTM'S, ARIMA, VARIMA, SARIMA
* Collaborated with business partners to understand their problems and goals, develop predictive modeling, statistical analysis, data reports and performance metrics.
* Participate in the on-going design and development of a consolidated data warehouse supporting key business metrics across the organization.
* Designed, developed, and implemented data quality validation rules to inspect and monitor the health of the data.
* Dashboard and report development experience using Tableau.

**Teamlease Services Limited Hyderabad, India May 2014- Dec 2016**

**Python Developer**

**Responsibilities:**

* Developed pipelines using SparkML that drive data for the automation of training and testing the models.
* Supervised model types including Generalized Linear Models. Random Forests. gradient boosting machines. Support Vector Machines, Deep Learning Neural Nets, and Ensemble Learning/Stacking. Unsupervised model types like Principal Component Analysis, K-means clustering. Hierarchical Clustering, AutoEncoders
* Built models for highly unbalanced data sets. Bias/Variance tradeoff. Model quality metrics like R Squared, AUC Outlier detection and removal.
* Worked in Big Data Hadoop Hortonworks, HDFS architecture, R, Python, Jupyter, Pandas. NumPy Scikit, Matplotlib. PyHive. Keras. Hive. NoSQL-HBASE, Sqoop. Pig. MapReduce, Oozie. Spark MLlib.
* Worked in Big Data Hadoop Hortonworks, HDFS architecture, R, Python, Jupyter, Pandas. NumPy. Scikit, Matplotlib, PyHive. Keras, Hive, NoSQL-HBASE, Sqoop, Pig. MapReduce, Oozie, Spark MLlib.
* Used Cloudera Hadoop YARN to perform analytics on data in Hive, build models with big data frameworks like Cloudera Manager and Hadoop.
* Work with different data science models for Machine Learning Algorithms such as Linear, Logistic, Decision Tree. Random Forests. Support Vector Machines. Neural Networks. KNN, Deep learning.
* Causal modeling in both experimental and observational data sets. Bayesian networks. Bayesian regression.
* Advanced programming in Python using Scikit-Learn and NumPy libraries.
* Predicted the Remaining Useful Life (RUL), or Time to Failure (TTF) using Regression.
* Predicted if an asset will fail within a certain time frame (e.g. days) with Binary classification.
* Used LSTM to predict probability of failure at different time intervals compensating for independent variables reflecting states of wear.
* Worked in the agile environment to implement agile management ideals such as sprint planning daily standups, managing project timelines, and communicate with clients to ensure project progress satisfactorily.
* Detected Anomaly using data science concept of Support Vector Machine (SVM), k-means Clustering, K-nearest neighbor, Pandas, NumPy, keras. sklearn, Scikit-learn, SciPy, tensor Flows was utilized to predict categories dependent on location. time and some different highlights of Linear regression. Logistic Regression, Decision Tree. Random Forests, relapse. Deep neural network, KNN. XGBoost, k-means clustering Support Vector Machines, time arrangement analysis (ARIMA Model). Ensembles (Bagging, Boosting) Neural Networks, diagram/arrange examination.
* Trained huge rundown of models, assessed, and looked at and chose the best models for predicting and forecasting. Set up and keep up viable procedures for confusion matrix. K-fold validating and approved distinctive models. Utilized the TensorFlow deep learning library to address the image and recognition issue developing convolution neural network in python.
* Feature selection process using Random Forest. Select K best. RFE was used. Enhanced precision lessens fluctuation and predisposition, and enhances solidness of a model, using Ensemble Models in machine adapting, for example. Boosting (Gradient boosting, XGBoost, and AdaBoost), and Bootstrap accumulating (packing-Bagged Decision Trees and Random Forest), verse and convey machine-learning work processes and models into generation being familiar with Azure Machine Learning Model Management.
* Knowledge of AR (Autoregressive), MA (Moving Average), and ARIMA (Autoregressive Integrated Moving Average) time arrangement examination models. Statistical Model with Grid search and ARIMA time series was used to predict CVS market demand.
* Worked in the Agile procedure with self-sorting out, cross-useful groups run towards results in quick, iterative, gradual, and adaptive steps.