**Clearnce Teman**

**Data Scientist**

**Email: clearnce.teman@outlook.com**

Professional Profile

• Experience with a variety of NLP methods for information extraction, topic modeling, parsing, and relationship extraction

• Familiarity with developing, deploying, and maintaining production NLP models with scalability in mind

• Experience with knowledge databases and language ontologies

• Quantitative training in probability, statistics and machine learning

• Experience in the application of Neural Network, Support Vector Machines (SVM), and Random Forest.

• Creative thinking and propose innovative ways to look at problems by using data mining approaches on the set of information available.

• Identifies/creates the appropriate algorithm to discover patterns, validate their findings using an experimental and iterative approach.

• Applies advanced statistical and predictive modeling techniques to build, maintain, and improve on multiple, real-time decision systems. Closely works with product managers, Service development managers, and product development team in productizing the algorithms developed.

• Experience in designing star schema, Snow flake schema for Data Warehouse, ODS architecture.

• Experience in designing stunning visualizations using Tableau software and publishing and presenting dashboards, Storyline on web and desktop platforms.

• Experience in working with relational databases (Teradata, Oracle) with advanced SQL programming skills.

• In-depth knowledge of statistical procedures that are applied in Supervised / Unsupervised problems

• Basic-Intermediate level proficiency in SAS (Base SAS, Enterprise Guide, Enterprise Miner) & in UNIX

• Track record of applying machine learning techniques to marketing and merchandizing ideas.

• Experience in Big Data platforms like Hadoop platforms (Map-R, Hortonworks & others) , Aster and Graph Databases.

• Experience in operations research / optimization will be good to have. Experienced in working with advanced analytical teams to design, build, validate and refresh data models that enable the next generation of sophisticated solutions for global clients.

• Excellent communication skills (verbal and written) to communicate with clients and team, prepare + deliver effective presentations.

• Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design

Professional Skills

Data and Quantitative Analysis Decision Analytics

Big Data Queries and Interpretation

Data Mining and Visualization Tools (e.g. Tableau)

Machine Learning Algorithms Predictive Modeling

Research, Reports and Forecasts Probability and Statistics

Programming Languages: Java, SQL, Python (Packages- NumPy, SciPy, Pandas, Spark, SciKit-Learn, Matplotlib, seaborn). IDEs: Net Beans, Eclipse, Jupyter Notebook

Data Analytics Software: SPSS, R TensorFlow, CNN

Database Management Systems: MySQL, Oracle, PostgreSQL, Microsoft Access

Big Data Technologies: Hadoop, MapReduce, Spark, Pig, Hive, Impala, Casandra

IoT, NLP

Data science models Machine Learning Algorithms: Linear Regression, Logistic Regression, Decision Tree, Random Forests, Support Vector Machines, Neural Networks, KNN, K-means

spaCy, PyTorch, Keras clustering, Confusion Matrix Spark,

Spark Streaming

Software Applications:

Microsoft Office including Outlook, Excel, Word, PowerPoint, and other computer applications

**Experience**

**NETBASE DATA SCIENTIST**

**Mountain View, CA February 2017 – Present**

Worked on the Perspective Project to determine when online conversations turn toxic through language analysis methods using machine learning. Built a multi-headed model that’s capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate better than current models. The objective is to come up with a strong model that can be deployed in online forums, comments, reviews, social media and any online conversational arena to encourage participation by keeping participation safe.

• Designed a light-weight convolutional neural network (cNN) on TensorFlow to recognize keywords, phrases, tone and inflection.

• Achieved real-time language processing, saved 75% inference time, comparing to current solutions.

• Re-trained semantic segmentation models and applied state-of-the-art Machine Learning (ML) / Deep Learning (DL) technologies to evaluate the sentiment score.

• Used classifiers built in SciKit-learn, on top of hand-engineered linguistic features, such as sentiment polarity lexicon features extracted by spaCy, an industrial-strength NLP library.

• Developed a DL module with Keras, using GloVe (Wikipedia-trained word vectors) as the embedding layer.

• Our linguistic model got 88.2% accuracy on IMDB (marked as our benchmark performance), DL models reached the benchmark after 2 epochs of training, and the ensemble DL model achieved 90.6% after 4 epochs.

• Reported that cNN architecture in our application was leading to fast convergence and quick overfitting; Bi-direction LSTM architecture cost 10x training time, and it tended to be underfitting the training data initially.

• Migrated eDiscovery search engine from monolith to microservice architecture using NLTK, SciKit-Learn, and DynamoDB.

• Achieved a 16x performance increase compared to the industry standard, i.e. systems based on LSA embeddings.

• Worked on Cloud Services such as Amazon Web Services (AWS), S3, Redshift to assist with big data tools, solve storage issue.

• Performing Map Reduce jobs in Hadoop and implemented Spark analysis using Python for performing machine learning & predictive analytics on AWS platform.

• Engineered pipelines for tokenization and word vector encoding with CoreNLP and Python and deployed them on premise.

• Created pipeline for server-side rendering word highlighting visualizations using Flask, HTML and CSS.

• Use of data mining and machine learning algorithms, theories, principles and practices.

• Documented logical data models, semantic data models and physical data models.

• Implemented use of real-time data using Storm and Spark as well as Spark ML.

• Implemented model on batch data using Spark SQL.

• Worked on Cloud Services such as Amazon Web Services (AWS), S3, Redshift to assist with big data tools, solve storage issue.

• Performed MapReduce jobs and Spark analysis using Python for machine learning and predictive analytics models on big data in Hadoop ecosystem on AWS cloud platform as well as some data from on-premise SQL

**HONEYWELL AEROSPACE DATA SCIENTIST**

**Atlanta, GA September 2015 – February 2017**

Aeronautical Engineering – Turbofan Predictive Maintenance

This predictive maintenance project focuses on the techniques used to predict when an in-service machine will fail, so that maintenance can be planned in advance. In particular, this project illustrates the process of predicting future failure events in the scenario of aircraft engine failures.

This predictive maintenance project focuses on the techniques used to predict when an in-service machine will fail, so that maintenance can be planned in advance. In particular, this project illustrates the process of predicting future failure events in the scenario of aircraft engine failures.

The objective of this project was to learn the potential failure of in-service turbofan engines to prevent loss. The goal was to use machine learning to plan predictive maintenance based on data tracking the performance of the engine as it comes available.

This project uses simulated aircraft engine run-to-failure events in predictive maintenance modeling. The implicit assumption of modeling data is that the asset of interest has a progressing degradation pattern, which is reflected in the asset's sensor measurements. By examining the asset's sensor values over time, the machine learning algorithm can learn the relationship between the sensor values and changes in sensor values to the historical failures in order to predict failures in the future.

• Creation of real-time data pipelines using Spark Streaming, Spark

• 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, Ensemble Learning/Stacking. Unsupervised model types like Principal Component Analysis, K-means clustering, Hierarchical Clustering, AutoEnconders.

• Built models for highly imbalanced data sets. Bias/Variance tradeoff. Model quality metrics like RSquared, AUC. Outlier detection and removal.

• Advanced Statistics/Math: ANOVA/ANCOVA. Bootstrapping, confidence intervals.

• 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

• Causal modeling in both experimental and observational data sets. Bayesian networks. Bayesian regression.

• Markov Chain Monte Carlo, Gibbs sampling. Gaussian processes. Variational Inference.

• Advanced programming in Python using SciKit-Learn and NumPy libraries.

• Supervised model types including Generalized Linear Models, Random Forests, Gradient Boosting Machines, Support Vector Machines, Deep Learning Neural Nets, Ensemble Learning/Stacking. Unsupervised model types like Principal Component Analysis, K-means clustering, Hierarchical Clustering, AutoEnconders.

• Built models for highly imbalanced data sets. Bias/Variance tradeoff. Model quality metrics like RSquared, AUC. Outlier detection and removal.

• Regression: Predicted the Remaining Useful Life (RUL), or Time to Failure (TTF).

• Binary classification: Predicted if an asset will fail within certain time frame (e.g. days).

• Advanced Statistics/Math: ANOVA/ANCOVA. Bootstrapping, confidence intervals.

• Used LSTM to predict probability of failure at different time intervals compensating for independent variables reflecting states of wear.

• Causal modeling in both experimental and observational data sets. Bayesian networks. Bayesian regression.

• Markov Chain Monte Carlo, Gibbs sampling. Gaussian processes.

• Variational Inference.

• Advanced programming in Python using SciKit-Learn and NumPy libraries.

• 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

**CITIBANK DATA SCIENTIST**

**Township of Warren, NJ March 2014 – September 2015**

Using big data provides banks and finance companies the opportunity to prevent fraud to a large extent. However, the usage of big data for this purpose is till at its early stages and a lot needs to be done in this regard. Using big data requires a huge mindset change and that the companies need to “Learn to be data-centric and data-driven and then solve problems that call for bigger data, such culture change has to happen for the big data approaches to become pervasive across the industry. And yes, this includes willingness to explore, follow leads, and occasionally arrive at dead ends and algorithms that don't work." This project looks into using machine learning model with NLP and neural networks to develop models that can identify fraud attempts in real-time.

? Design and development of Statistical and ML models for the end clients. Coordinated with end users for designing and implementation of analytics solutions as per project proposals.

? Developed and deployed or trained Machine Learning Models using Azure ML Studio and Spark ML to analyze and identify Fraudulent Transactions for Risk Analysis Team.

? Responsibilities included in creating enterprise level Data Modeling and Project experience in Data mining, Segmentation analysis, Business forecasting using R with Databases like SQL Server, MongoDB, MySQL and HBase. Deployment of Production Level R models to MS Azure.

? Created Natural Language Processing model in R using OpenNLP, Stanford NLP, ANN, MS Cognitive Services to classify Citi Credit and Legal Risk Documents attitude by their contents.

? Detect and classify images using deep learning algorithms (NN, ANN, CNN with Theano in Python.

? Prepared graphs using ggPlot library and Tableau for an overview of the analytical models and results.

? Developed predictive models using Vector Machines, Decision Tree, Random Forest and Naïve Bayes, and collaborating with marketing and dev-ops teams for production deployment.

? Worked on generating reports using Neo4J graph database and Classification of key words into categories using a Neo4J graph database and make a recommendation.

? Generated AdHoc Reports using Python and Apache Spark from Large Datasets in Hadoop Clusters and created dashboards, reducing effort needed to manually review documents.

**JOHN DEERE DATA SCIENTIST**

**Moline, IL August 2012 – March 2014**

Farms are heavily reliant on small improvements in operational efficiencies and processes in order to increase crop yields, manage risk, and create greater profit. This is particularly true for large-scale agribusiness where commodity crops are involved, and small process adjustments have large impacts in terms of production.

As the global population increases, weather volatility grows, and petroleum dependent agriculture is increasingly sensitive to fossil fuel pricing, there will be more incentives to leverage new technology to increase crop yields and manage risk.

Opportunities for big data applications in agriculture include benchmarking, sensor deployment and analytics, and using better models to manage crop failure risk.

This project focused on using machine vision to reduce waste due to spoilage. Currently over a third (30%) of all food produced is either lost or wasted through the entire production process. The wasted food has about a $940 billion impact on the global economy. This was implemented using big data collected from machine vision steering system mounted on combines.

? Analyzed data from Hadoop big data system, ingested from machine vision in field; data analysis used to create custom algorithms to improve crop turnover.

? Constructed machine learning models using NumPy, SciPy, NLTK, SciKitLearn, MLPy, OpenCV

? Prototyped Convolutional and Recurrent Neural Networks to do health claims

? analysis.

Setup development environment to use Docker, and used Docker to handle deployment on heterogeneous platforms such as Google and AWS

Migrated large database from SQL Server to MySQL.

? Use of machine vision in the development of both algorithms and software toolkits in image, signal and video processing.

? Machine-learning based object detection and pattern recognition in 3D imaging applications using OpenCV, SciKit-Image and SciKit-Learn.

? Machine learning algorithm, segmentation of images using Deep Learning under OpenCV.

? Classification of images and reporting using machine learning algorithm.

? Statistical Techniques- t-tests, ANOVA, Regression (Multiple, Stepwise, Logistic, Cox), Time Series, Principal Component Analysis

? Data Mining/Machine Learning Techniques- Decision Trees (C&RT, CHAID, C5.0), Cluster Analysis, Artificial Neural Networks, Association Rule Mining, PMML deployments of DM models