Tel: +1 (732) 640-8509 **Email:** dsrneelesh8@gmail.com

Portfolio: www.neeleshrastogi.com
LinkedIn: www.linkedin.com/in/neeleshrastogi
GitHub: www.github.com/neelrast

Neelesh Rastogi

Data Scientist, Machine Learning Engineer & Researcher

PROFESSIONAL SUMMARY:

Machine Learning Engineer and Data Science Professional, specializing in Deep Learning, Computer Vision, and Natural Language Processing/Generation. With Experience in research and development within Machine Learning, Human Robotic Interaction Systems, Exploratory Data Analysis (EDA), Predictive Modelling, Data Warehousing, Data Acquisition and Validation. I enjoy combining upcoming research and existing frameworks to design end-to-end Deep Learning pipelines with an ultimate goal of creating a more human-centric and result oriented multimodal system.

My primary interests span the area of multimodal human-robotic interaction systems, behavior cloning, natural language processing and generation, time series and quantitative analysis, deep learning, computer vision, software architecture, and agile design methods.

SKILLS:

- Proficient in applying Statistical Modelling and Machine Learning techniques like Linear Regression, Logistic Regression, Decision Trees, Random Forest, SVM, K-Nearest Neighbors, Bayesian, Boosting and Bagging) in Forecasting/ Predictive Analytics, Segmentation methodologies, Regression-based models, Factor analysis, PCA, Ensembles and good knowledge on Recommendation Systems.
- Well-equipped in Statistical thinking which include Graphical and Quantitative EDA, Bootstrap Confidence Intervals, Correlation, Hypotheses modeling, Recommender systems, Time-series, Factor analysis, Pattern Recognition, Inferential Statistics, Matrix Factorization as well as Modelling Techniques to gain valid inferences.
- Advanced Knowledge of concepts within Object-Oriented, Dynamic & Functional Programming.
- Progressive understanding of Software Development Life Cycle (SDLC) with good working knowledge of conducting Requirements Analysis, Design Specification and Testing as per Development and Production cycle in both Waterfall and Agile methodologies.
- Excellent knowledge of Data Analysis, Data Validation, Data Cleansing, Data Verification and identifying data mismatch within Preprocessing and ETL Workflows.
- Good understanding of Relational Database Design, Data Warehouse/OLAP concept and methodologies.
- Well versed in Anaconda and Miniconda environments and professional understanding of libraries like NumPy, SciPy, Pandas, Scikit-Learn, Theano, TensorFlow, Keras, NLTK & OpenCV.
- Expertise in Dimensionality Reduction techniques like PCA, LDA, Singular Value Decomposition.
- Adept using Stratified k-Fold Cross Validation, Confusion Matrix, AUC-ROC, Hyper Parameter Tuning & Grid Search techniques for Model Selection.
- Proficient in generating data visualizations using PyViz, Seaborn, Matplotlib, Vega and D3.js.
- Ability writing & executing custom Data Warehousing queries using Datagrip/Dbeaver and Snowflake.
- Skilled in implementing machine learning pipelines and scaling models for production quickly.
- Adept in relational databases concepts and writing SQL Queries for Microsoft SQL Server, Oracle SQL, MySQL, and PostgreSQL.
- Working Knowledge of Big Data Ecosystems of Apache Spark, MapReduce, HDFS Architecture, Cassandra, HBase, Hive and MLlib.

- Faculty with building custom ETL workflows using Python and Apache Aiflow to perform data cleaning & mapping.
- Hands on experience in writing User Defined Functions (UDFs) for Test Driven Development (TDD) under PEP-8 coding standards for Python 3.
- Proficient with version control systems such as Git, SVN, GitHub, bitbucket.

CERTIFICATIONS:

1. AWS Big Data Specialty Certification **(In-Process)**
2. Data Science & Engineering Certification **(Anaconda + DataCamp)**
3. TensorFlow 2.0: Data and Deployment Practices **(deeplearning.ai)**
4. PyTorch Specialization **(Facebook AI + Udacity)**
5. Deep Learning Specialization **(deeplearning.ai)**
6. Machine Learning **(Stanford University)**
7. Self-Driving Car Engineer Nano-degree Program **(Udacity)**
8. Data Engineering, Big Data and Machine Learning on GCP Specialization **(Google Dev Certified)**

TECHNICAL SKILLS:

Programming Languages	Python 3.x, R, SQL, Scala, Julia, Java EE, C++ 11
Big Data Ecosystem	Apache Spark, MLlib, GraphX, PySpark 2.x, MapReduce, Hive QL 1.x/2.x, HDFS, Sqoop, Apache Pig
ETL, DAGs, & Data Warehousing	Apache Airflow, Cron, Snowflake, Google Big Query, IBM DB2, MS SQL Server, Oracle 9i/10g, MongoDB 3.x, MySQL, PostgreSQL, SQLite
Statistical Methods	Hypothesis Testing, ANOVA/MANOVA, Principal Component Analysis (PCA), Time Series, Correlation (Chi-square Tests, Covariance Analysis), Multivariate Analysis
Computation & Data Frames	Pandas, Numpy, x-array, Dask, Scipy, GeoPandas, NLTK, SpaCy, Gensim, OpenCV
Machine & Deep Learning Frameworks	Scikit-Learn, TensorFlow, PyTorch, Keras, Apache Singa, Caffe, mlpack, H20, CNN, RNN, LSTM, Deep Belief Networks, PilotNet, LeNet-5, Autoencoders, RESNET-50, VGG-19, Inception-V3, NVidia CUDA Parallel GPU Processing
Chatbot Tools	Microsoft Cognitive Bot Framework, Dialogflow, Amazon Lex, ChatScript
Machine & Deep Learning Concepts	Linear Regression, Logistic Regression, Naive Bayes, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Means Clustering, K-Nearest Neighbors (KNN), Stratified K-Fold, Gradient Boosting, Neural Networks, XGBoost, AdaBoost, LDA, Natural Language Processing, Computer Vision
Data Visualization	Tableau, Matplotlib, Seaborn, HoloViews, GeoViews, Bokeh, Panel, Datashader, hvPlot, Param, ggplot2, Vega, D3.js
Cloud Tools	AWS SageMaker/AMI/Connect/EC2/S3/Redshift/Lambda, Google Cloud Platform (GCP), Azure ML Studio, IBM Cloud Services
Reporting & Versioning	Jira, Github, BitBucket, GitLab, Git, Subversion, Travis CI, GitLab CI/CD

EDUCATION:

St. John’s University, Queens, New York **May 2019**

- Bachelor of Science in Computer Science **(Distinction - Magna Cum Laude, GPA 3.63)**
- **Certificate (Minor):** Applied Mathematics

Relevant Coursework: DBMS, Software Architecture, Algorithms, Data Structures, Operating Systems, Cryptography, Data Mining & Predictive Analytics, Natural Language Processing, Computer Vision
Math & Electives: Linear Algebra, Probability & Statistics, Calculus, Series, Differentials

PUBLICATIONS (http://bit.ly/neel-google-scholar):

1. **Tweets Classification using BERT for Disease Ontology**
   Submitted Paper in BIOSEC 2020 and IEEE Conference on ML Standards for Communications, Analysis and Networking, San Diego, California, 2020.

2. **A Multimodal Human Robot Interaction Framework based on Cognitive Behavioral Therapy Model**
   Published in ICML Conference Proceedings: H3 '18 Proceedings of the Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era. Article 2.

3. **Real-Time Mapping of Infectious Disease Analysis over Social Network Data**
   Published Abstract in The Thirty-Second International Florida Artificial Intelligence Research Society Conference (FLAIRS-32), AAAI Conference. Page 516.

PROFESSIONAL EXPERIENCE:

First Blush AI *(NYC based startup*) **Oct 2018 – Present**
Artificial Intelligence & Natural Language Processing Engineer

Project: Emotion Recognition Toolkit (**www.firstblush.io**)**:** ML Model for Emotion-Facial Recognition.
Responsibilities:

- Developed a production-ready classifier in PyTorch for POC displayed during the company's presentation for first round of funding. This involved data augmentation of images from multiple sources cameras, iPhone, Go-Pro, & other labeled open-source emotion recognition datasets.
- Applied advanced Computer Vision techniques to get the images ready for Neural Network inputs.
- Developed an interactive Facial Recognition and Keypoint detection web application in PyTorch Framework. This is scaled for production.
- Implemented a machine learning pipeline using Microsoft FER+ dataset & Mobile Nets CNN to analyze and improve model's emotion recognition accuracy over iPhone's front camera inputs.
- Extended our conversational advice chatbot with conversation scripts corpus. Implemented PyTorch chatbot prototype. Scaling to production is underway.
- Assisted the mobile engineering team to optimize and deploy models over core ML using TensorFlow Lite Converter.

- Implemented Association rule mining using python generator & pandas to handle large datasets, to suggest frequently visited app features within our user tests.
- Developed an ensemble model of Keras pre-train CNNs, Resnet-50 & inception-v3 respectively using 5-fold cross-validation to train image classification model in TensorFlow environment.
- Fine-tuned deep learning models using feature extraction and creation by applying data augmentation and transfer learning techniques.
- Developed NLP model for topic extraction using LDA (Latent Dirichlet Allocation) techniques on collected beta reviews to gauge customer retention by addressing concerns promptly, which reduced 30% of human hours spent on review reading and categorizing.
- Implemented text analysis such as classification, tagging using NLTK library (spacy, NLTK, Stanford, genism), and a good understanding of word embedding, language models, and text pre-processing over beta reviews for better insightful reports.
- Experience with defining runtime needs and software to facilitate application, runtime, and middleware components for Machine Learning Models.
- Deployed a Databricks cluster for emotion recognition data and Implemented Data Frame API, Built-in functions, User-defined functions, RDD in PySpark for data pre-processing, and maintenance.
- Integrated data streams from different source systems such as AWS, data warehouse, and Big Data platforms to write complex SQL queries, procedures, and functions to retrieve data from RDBMS.
- Performed A/B testing to determine the effectiveness of marketing campaigns to boost customer experience, to track user behavior on the mobile app, compare it against desktop activity to make sure the delivery of the right content is made to the right customer through proper channels.
- Formulated data blending and build Tableau dashboards to provide aggregated data view to the executive board to see beta app performance on a real-time basis and research in-app purchasing trends, activity cycles, and business opportunities of specific high value/opportunity customers.
- Took the lead on the app's chatbot experience and analyzed large interaction datasets to curate user-specific intents.
- Worked with cognitive researchers and copywriters to develop persona-based conversational flows & knowledge representation models for in-app chat assistant.
- Collaborated with other UX engineers & designers to communicate results & ideas at various conventional UX/ML meetups.

Environment: Python 3.x (Scikit-Learn/ SciPy/ NumPy/ Pandas/ Matplotlib/ Seaborn), XHTML, SASS, PyTorch, Tensorboard, Mobile Nets, FaceNet, NLTK, Spark NLP, Git, JIRA, Airflow, Agile/SCRUM, AWS EC2 Instances and S3 Buckets, Apple Keynote, Excel, Tableau.

NLP Lab, Collins College of Professional Studies, New York **Dec 2015 – Sept 2018**
(https://bit.ly/nlp-lab)

Project Role #1: Data Science & Natural Language Processing Researcher
Responsibilities:

- Created a Web-Based Platform to visualize, scrape, clean, and analyze 1.5 Million collected tweets for community's sentiment on potential disease outbreaks.
- Developed and documented an Affect Interaction Module for Pepper Robot, enabling Cognitive Behavior Therapy for patients showing early signs of depression.
- Brought to Nature and Presented Research Findings from the above projects at National Conferences like ICML and FLAIRS.

- Developed a system for collecting data and generating exploratory findings into PDF reports that improved the quality of research conducted at the lab.
- Synchronized data stored in several MongoDB instances with an Elasticsearch engine for better indexing of stored textual findings, articles, research papers, and publications.
- Developed a multi-class, multi-label 2-stage classification model for Unnamed Psychological reports to identify depression-related terms and classify depression- indicative symptoms. Utilized created model to calculate the severity of depression in a patient using Python, Scikit learns, & Weka.
- Worked on a pre-processing pipeline to handle large datasets for missing values, creating dummy variables for binary values and normalization in data.
- Performed data pre-processing tasks like merging, sorting, finding outliers, missing value imputation, data normalization, making it ready for statistical analysis.
- Worked on Descriptive, Diagnostic, Predictive, and Prescriptive analytics and used results for further implementing a Character Recognition Model using Support vector machine for performance optimization.
- Managed database design & implemented a comprehensive Star-Schema with shared dimensions.
- Developed and maintained, stored procedures, implemented changes to database design, including tables and views, and documented source to Target Mappings as per project guidelines.
- Utilized various techniques like Histogram, bar plot, Pie-Chart, Scatter plot, Box plots to determine the distribution and condition of captured data and features.
- Designed and developed analytical and machine learning models, with visualizations dashboard that drove research performances and provided productivity insights at all stages of prototyping to development and publication.
- Implemented various machine learning models such as regression, classification, tree-based ensemble models, also performed Hyperparameter Tuning to raise the model accuracy.
- Validated different models applying appropriate measures such as stratified k-Fold cross-validation, AUC, ROC to identify the best performing model.

Environment: Python, PySpark, R Studio, MS SQL Server, MySQL, Hive, ETL, Tableau, NumPy, Pandas, Matplotlib, Scikit-Learn, ggplot2, Shiny, TensorFlow.

Project Role #2: Chatbot Architect, Information Technology
Responsibilities:

- Utilized selenium to extract, organize and classify organizational data for conversational encoding and topic modeling.
- Created conversation models using NLTK, SpaCy, and Keras for automating low-priority IT Tickets.
- Analyzed help-desk queries for the user's spoken intents using Seq2Seq models and routed requests to specific departments.
- Implemented text classification algorithms like multinomial Naive Bayes and topic modeling to analyze the real-time incoming unstructured service requests data and provide network engineers with instant decisions of whether to reboot or replace the hardware or not.
- Worked with text features engineering techniques like n-grams, TF-IDF, word2vec, and BERT.
- Implemented Text Analytics and Sentiment Analysis feature, creating word clouds and retrieving data from social networking platforms.
- Experience building chatbots using chatbot frameworks RASA, Amazon Lex, and Dialogflow.
- Developed weighted ensemble models for predictive maintenance of devices to optimize the periodic maintenance operation and to reduce system downtimes to enable saving 10% cost by minimizing their use of resources.

- Intensified gradient boosting and bootstrapping as a challenger model to text classification algorithms, which helped drive more accurate decision making among admins.
- Tackled a highly imbalanced dataset using sampling techniques like down-sampling, up-sampling, SMOTE (synthetic minority over-sampling technique), & ADASYN (Adaptive Synthetic) using Scikit-Learn.
- Analyzed multiple statistical models for real-time Rest API consumption and developed models using H20.ai.
- Deployed models using custom APIs in a Spark environment and consolidated the entire significant data architecture for real-time scoring in MongoDB.
- Design and develop ETL workflows to migrate data from varied data sources, including SQL Server, Netezza, Kafka in batch and real-time.
- Performed data cleaning, data manipulation procedures to optimize datasets for data analysis.
- Produced database objects such as stored procedures, functions to automate data extraction.

Environment: Python, MS SQL Server, AWS Lex, API Gateway & Lambda, MS Azure Cognitive Toolkit, Hadoop, Hive, Linux Instances, Docker, Tableau, Google Analytics.

Blue Water's e-Marine Solutions, Dehradun, India **Aug 2015 – April 2016**
Application Development and Data Science, Intern

Project: Optimum Routing **(www.bwesglobal.com/optimumrout):** Unique Algorithm to Replace Traditional Weather Routing Advisory for Cargo Ships
Responsibilities:

- Designed a forecasting algorithm to predict pitch, roll & yaw to optimize fuel savings by 8-10% against the transparent benchmark.
- Trained an SVR-RBF Kernel, using engine and boiler data from 32 ships for predicting minimal fuel and energy consumption on suggested optimal routes with an optimal R2 - score.
- Collected data from the end client, performed ETL, and defined the uniform standard format.
- Wrote queries to retrieve data from the SQL Server database to get the sample dataset containing primary fields.
- Performed string formatting on the dataset converting hours from date format to DateTime format & other numerical strings to an integer, along with creating dummy values for binary labels.
- Used Python libraries like Matplotlib and Seaborn to visualize the numerical columns of the dataset, such as day of the week, age, hour, and cargo ship dimensions.
- Worked on missing value imputation, outlier identification with statistical methodologies using Pandas, NumPy. And SciPy.
- Designed and implemented K-Fold Cross-validation to test and verify the model's significance against ground truth values.
- Developed a dashboard and story in Tableau showing the benchmarks and summary of the model's measure.

Environment: Tableau, NumPy, SciPy, Scikit-Learn, Dask, PyViz, Django & Git.