Bharath Ramling

Data Scientist

Introduction: [youtube.com/watch?v=om3VDAplM\_Q](https://www.youtube.com/watch?v=om3VDAplM_Q)   
GitHub: [github.com/bharath-r](https://github.com/bharath-r)

**SUMMARY**

Data Scientist with 7+ years of experience in e-commerce, Automobile, Insurance domain, performing **Data Extraction, Data Screening, Data Cleaning, Data Exploration**, **Data Visualization, Modeling** of structured and unstructured datasets as well as implementing large-scale **Machine Learning** and **Deep Learning Algorithms** to deliver resourceful insights, significantly impacting business revenues and user experience.

* Experienced in facilitating the entire lifecycle of a data science project: **Data Extraction, Data Pre-Processing**, **Feature Engineering, Algorithm** **Implementation & Selection, Back Testing** and **Validation**.
* Expert at working with **Statistical Tests**: two-way independent & paired t-test, one-way & two-way ANOVA along with **Non-parametric tests**: Chi-squared tests, Mann-Whitney U & Kruskal-Wallis test.
* Skilled in using **Pandas, NumPy** libraries in Python for performing **Exploratory Data Analysis**.
* Proficient in **Data Transformations** using Logarithm, Square-root and Box-Cox Transformation.
* Skilled at **Data Visualization** with Tableau, PowerBI, Seaborn, Matplotlib, ggplot2, Bokeh.
* Adept at handling **Missing Data** by exploring the causes like MAR, MCAR, MNAR and analyzing **Correlations** and similarities, introducing dummy variables and various **Imputation** methods.
* Experienced in Machine Learning techniques such as **Regression and Classification** models like Linear Regression, Polynomial Regression, Logistic Regression, Decision Trees, Support Vector Machines.
* Proficient in **Ensemble Learning** using Bagging, Boosting (AdaBoost, xGBoost) & Random Forests.
* Working Knowledge in **Unsupervised Learning Algorithms** including K-means, Hierarchical Clustering.
* Experienced in developing **Supervised Deep Learning** algorithms which include Artificial Neural Networks, Convolution Neural Networks, Recurrent Neural Networks, LSTM, GRUand **Unsupervised Deep Learning** Techniques like Self-Organizing Maps (SOM’s).
* In-depth Knowledge of **Dimensionality Reduction** (PCA, LDA)**, Hyper-parameter tuning, Model Regularization** (Ridge, Lasso, Elastic net) and **Grid Search techniques** to optimize model performance.
* Adept at **Data Structures** concepts in **Python, C++** like List, Array, Dictionary, Binary Treesand **Algorithm** techniques like Search Algorithms, Greedy Algorithms and Dynamic Programming.
* Skilled at Python, C++ **Object Oriented Programming (OOP)** concepts such as Inheritance, Polymorphism, Abstraction, Encapsulation**.**
* Proficient in **Natural Language Processing (NLP)** concepts like Tokenization, Stemming, Lemmatization, Stop Words, Phrase Matching and libraries like **SpaCy** and **NLTK**.
* Skilled in **Big Data** Technologies like **Apache Spark** (PySpark, Spark Streaming, MLlib), **Hadoop** Ecosystem (MapReduce, HDFS, HIVE, Kafka, Ambari)
* Experience in **Web Data Mining and Scraping** with Python’s Scrapy and BeautifulSoup packages.
* Working Knowledge of **Cloud services** like Amazon Web Services **(AWS)** and **Microsoft Azure** for building, training and deploying scalable models.
* Proficient in using **PostgreSQL**, **Microsoft SQL server** and **MySQL** to extract data using multiple types of **SQL Queries** including **Create**, **Join, Select, Conditionals, Drop,** **Case** etc.

**SKILLS**

|  |  |
| --- | --- |
| **Languages** | Python, R, Matlab, SQL, C/C++ |
| **Analytical Techniques** | **Regression Methods:** Linear, Polynomial, Decision Trees; **Classification:** Logistic Regression, K-NN, Naïve Bayes, Support Vector Machines (SVM); **Ensemble Learning:** Random Forests, Gradient Boosting, Bagging; **Clustering:** K-means clustering, Hierarchical clustering; **Deep Learning:** Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks; **Dimensionality Reduction:** Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA); **Time-series Forecasting:** Multiple Decomposition, Additive Decomposition, Exponential Smoothing, Winter’s Multiplicative model, ARMA, ARIMA, SARIMA |
| **Database** | SQL, PostgreSQL, MongoDB, Microsoft SQL Server, NoSQL |
| **Statistical Tests** | Hypothesis Testing, ANOVA, z-test, t-test, Chi-Squared Fit test |
| **Validation Techniques** | Monte Carlo simulations, k-fold cross validation, A/B Testing |
| **Optimization Techniques** | Gradient Descent, Stochastic Gradient Descent, Gradient Optimization – Momentum, RMSProp, Adam |
| **Big Data** | Apache Hadoop, HDFS, MapReduce, Apache Spark, HiveQL, Pig, Kafka |
| **Data Visualization** | Tableau, Microsoft PowerBI, ggplot2, Matplotlib, Seaborn, Bokeh |
| **Data Modeling** | Entity Relationship Diagrams (ERD), Snowflake Schema, SPSS Modeler |
| **Operating Systems** | Microsoft Windows, iOS, Linux Ubuntu |

**PROJECTS**

**Data Scientist** December 2018 – Till Date

**Mastercard,** O’Fallon, Missouri, USA

According to a report from Nilson, the worldwide losses in card fraud related cases reached 22.8 billion dollars in 2017. The company aimed at building a robust fraud-detection model to identify future fraudulent transactions (Supervised Learning) and identify transactions which may have been wrongly labeled (Unsupervised Learning).

**Responsibilities:**

* Processed the Raw data by applying **Data Cleaning** techniques using **pandas** and **NumPy**, dealt with large data with AWS **EMR** (Elastic MapReduce).
* Extracted raw features and used **aggregation techniques** to new features that capture the **spending patterns** of the user.
* Performed **Exploratory Data Analysis** with Python and Visualization tools like **Matplotlib**, **Seaborn** to identify the patterns and correlations between the features.
* Handled the imbalanced dataset, exploring the uses of **under-sampling**, **over-sampling** and **SMOTE**.
* Used **t-SNE**, a Dimensionality Reduction technique for visualization of high-dimensional datasets to effectively visualize the transactions.
* Implemented classification algorithms **Logistic Regression, Support Vector Machines, Decision Trees** and **Random Forests** to build a classification model and predict future fraudulent customers.
* Explored applying ensemble algorithms methods like **Bagging** and **Boosting** (**AdaBoost, Gradient Boosting, xGBoost**) to enhance the performance of the model.
* Implemented **unsupervised Deep Learning** methods, **Self-Organizing Maps and Autoencoders** to identify outliers and customers potentially labelled wrongly into the other category.
* Evaluated the classification algorithms using **Precision, Recall, F1-score, Confusion Matrix,** Area under the Curve **(AUC)** for **Precision-Recall Curve and ROC Curve**.
* Assisted with the **Deployment** of the model with **Amazon SageMaker,** building the model into a **Docker Container** and storing the dataset in AWS **Elastic Block Store.**

**Data Scientist**  September 2017 – November 2018

**Kohl’s Corporation,** Menomonee Falls, Wisconsin, USA

The objective of the project was to develop a Sentiment Analysis Model to understand the customers’ sentiment towards the company’s products and assigning open-ended comments to the correct department to resolve issues quicker. The data was collected from various sources such as the company surveys, social media websites like Twitter, Reddit etc.

**Responsibilities:**

* Developed **Sentiment Analysis** and **Text Classification** algorithms by training Historical Data provided by organization and streaming data updated in **Batches**.
* Data was facilitated from various sources such as company official website using basic customer service survey questions like **NPS, CSAT** and using social media API’s such as **Twitter, Reddit**.
* Made use of the **Twitter API (Tweepy), Reddit API** and **Spark Streaming** to facilitate Data Streaming where real-time sentiments of the customers were monitored.
* Applied **Natural Language Processing** concepts like Lemmatization, Named Entity Recognition, Stop Words Removal, Tokenization, TF-IDF, Phrase Matching, Skip-grams to improve the quality of the data.
* Performed **Data Visualization** of the text data by creating interactive **Word Cloud** **Dashboards** on **Tableau**.
* Trained various Machine Learning classification algorithms such as **Logistic Regression**, **Decision Trees**, **Naïve Bayes**, Support Vector Machine (**SVM**), K-Nearest Neighbor (**KNN**).
* Enhanced the performance of Decision Trees using various **Bagging** and **Boosting** Techniques like **Random Forests, AdaBoost, xGBoost** etc.
* Evaluated the model performance by using **Confusion Matrix**, **Precision-Recall Curve, ROC Curve** additionally performed **A/B testing** to validate the model.
* Explored the possibility of applying Deep Learning (**RNN with LSTM**) supervised model in **Python** to predict the **Sentiments** on **Word2Vec** Word Embeddings of the reviews.
* Assisted with the **Deployment** ofthe model by using **Flask** **RestAPI**.

**Data Scientist** November2016 – August 2017

**Mercury Insurance,** Brea, California, USA

Mercury Insurance is a multiple-line insurance organization offering personal automobile, homeowners, renters and business insurance. My project involved predicting various levels of risk factors using machine learning models for the company approving medical insurance of the applicants based on their medical, insurance and employment history.

**Responsibilities:**

* Extracted required data from the AWS Data Lake for further exploration and visualization of the data to find insights and build prediction model.
* Performed **Data Screening,** **Data Cleaning, Data Visualization, Feature Engineering** using Python libraries such as **Pandas, Numpy, Sklearn, Matplotlib and Seaborn**.
* **Feature Engineered** raw data by implementing techniques like **imputation, normalization and scaling** while converting categorical variables to numerical values using **Label Encoder** for EDA.
* Performed **univariate, bivariate and multivariate analysis** to check how the features were related and provided regular insights about the features.
* Dealt with imbalanced dataset by using sampling methods like **SMOTE** so models train better to classify the classes and generalize the results rather than over-fitting.
* Applied Dimensionality Reduction techniques like **PCA, LDA** to reduce the correlation between features and high dimensionality of the standardized data to better use time and storage.
* Built machine learning models for classification based on **Decision Trees, Support Vector Machine and Random Forest** to predict the different risk levels of applicants and used **Grid Search**, **Random Search** for Hyperparameter Tuning to improve the performance of the model.
* Evaluated the model’s performance using various metrics like **Precision, Recall, Support and Cross Validation** to test the models with different batches of data to optimize the models.
* Prepared dashboards for **Data Visualization** with **Tableau** and generated complex reports including summaries and graphs to interpret the findings to the team.

**Robotics Machine Learning Engineer** July 2014 – September 2016

**The Future Group,** Bangalore, India

The project required to detect a red exit sign for a robot in order to depart the room with minimum travel time. Once the exit sign was identified, the algorithm required to find the distance between the robot and the sign. The project also aimed for building a localization algorithm to know the location of the robot with respect to the environment.

**Responsibilities:**

* Built a **Single Gaussian Model, Gaussian Mixture** Model for detecting a red exit sign in images.
* Parametrized Images into **RGB** space and labelled the images in the training set using roipoly function on Python.
* To find the optimal parameters, **Maximum Likelihood Estimation** was used for Gaussian Model while **Expectation Maximization** was used for Gaussian Mixture Model.
* Also build models **Linear Regressor** to predict the distance between the sign and the robot by training the regressor on the training set.
* Improved the accuracy using various methods like **Regularization, Lasso and Ridge Regression**.
* Developed a **Simultaneous Localization and Mapping (SLAM)** algorithm to build the map of the surroundings of the robot and localize the robot.
* Data was collected from **LIDAR** sensor and **RGBD** Camera mounted onto the robot.
* Developed the localization algorithm using **Particle Filter** with **Update and Predict step**, and the map was constructed using RGB data from the camera.

**Data Analyst** August 2012 – June 2014

**Ashok Leyland,** Bangalore, India

The Automobile company was in need to find the optimal operating conditions for the Diesel engine given various parameters such as Blend Ratio in the Biodiesel, Compression Ration, Fuel Injection Pressure. The project aimed at finding the optimal values in the minimal number of experiments for Performance criteria like Brake Thermal Efficiency, Emission Characteristics like NOx, CO etc.

**Responsibilities:**

* Drew statistical inferences using **t-tests, ANOVA, Chi-squared** **tests** to assess difference across levels of parameter categories, test significance of proportional differences and assess whether sample size is large enough to detect the differences.
* Employed Design of Experimental **Statistical Methods** like **Taguchi Method** in order to drastically reduce the number of experiments required to arrive at the required optimal parameter values.
* Use of **Semi-Deviation** & **Skewness-to-Kurtosis** ratio to validate the selection of levels of parameters.
* Optimum parameters levels for each of the performance criteria were calculated using **Grey Relational Analysis, Signal/Noise Ratio** and **Response Curves.**
* Built a **Polynomial Regressor Estimator** model to estimate the target variables with the given set of experimental variables, explored the possibility of **Fuzzy Logic**.
* Performed **A/B testing** in order the validate the model built.
* Created multiple custom SQL queries in **MySQL Workbench** to prepare datasets for **Tableau dashboards** and prepared dashboards using **Tableau**.

**EDUCATION**

Bachelor’s Degree in Mechanical Engineering – NIT-K, India