



PRUTHIRAJ JAYASINGH

SENIOR DATA SCIENTIST | Dell-EMC | Citigroup

Mail: data.pruthiraj@gmail.com Cell: [08884069169](tel:08884069169)

CAREER OBJECTIVE

A Data Scientist with understanding on **statistical** and **ML** to fit my solution to the problems and the ability to mine hidden gems by applying various machine learning techniques and exploratory data analysis. I am an enthusiast who learns new techniques related to data science.

EXPERIENCE

Total **5 years of experience** in data science in multiple domains like network security, storage applications, sales, operation and Banking to solve different problem located within large statements.

TECHNICAL SKILLS

- **Python**- Strong python skills related to model development and deployment. Skills in data preprocessing, data mingling and wrangling, cleaning ,function creation, loops and Popular data science packages like **Sklearn, Pandas, Numpy, Seaborn, Stata** etc. Understanding about **heat map, scatterplot, pair plot** analysis.
- **Machine Learning**- Core expertise in leveraging Machine Learning as per business requirements and a natural intuition to capture meaningful insights from data and translate them into competitive advantages for businesses. Implementation skills in multiple machine learning algorithms like **Linear Regression, Random Forest, Logistic Regression, Decision Trees, XGBoost, Clustering, K-Nearest Neighbors (K-NN)** etc. Understanding about **Ensemble Techniques** like **bagging, Boosting**. Worked with **K-fold cross validation**, and other model evaluation techniques. Worked on **Hyper Parameters** for better performance in the model. Work Experience in **Recommendation system** collaborative filtering and other techniques.
- **Statistics**: - Good Understanding about sampling, concept of **probability, mean, median, standard Deviation**. Implementation of **T-test, P-test, F-test, chi square, ANOVA** etc. Understanding about **Bias Error, Variance error, Random Error** and the characters. Controlling **underfit** and **overfit** using statistical techniques. Understanding about **PSI, CSI, Gini index, KS** etc.
- **SQL and SAS**- Strong database management and query writing skills in **SQL/Oracle/Netezza** using statements and clauses, Entity, **joins, views, when statement, ranking, row order** and store procedures etc. Strong data **manipulation skills in oracle OLAP functions**.
- **UNIX** – Deep understanding about **Linux environment**. Understanding about **shell script**. Idea in **file system**, and **Networking in Linux**. Good Skills in **vim editor, web-app** like **HTTP, FTP**, and **cron jobs**. Basic understanding about **TCP/IP, UDP, HTTP, HTTPS**, routing protocols in Data science point of view.

QUALIFICATION

B-Tech:

College: NMIET, BPUT, Odisha.

Branch: Computer Science

Year: 2010-2014

PGP:

College: Great Lakes Institute of Management

Program: Machine Learning and Artificial Intelligence.

Year: 2018-2019

ACHIEVEMENTS/CERTIFICATES:

1. Dell-EMC Information storage management (ISM) V3
2. Data Science Associate | Dell EMC
3. Certified Ethical Hacker (V9)
4. Python Programming Expert
5. Linux Admin and scripting.

PROJECTS

Project 1	Associate Data Scientist I6
Project Title	Proactive Fault Tolerance
Company	Dell-EMC
Algorithm	Random Forest, K mean ++
Techniques	Feature Engineering, K-fold Validation, Imblearn
Description	Developed a Machine Learning model to predict a Disk drive Failure prior to 12 hours to save money and time associated. This model is implemented most of the hardware to generate auto support and alerts for any drive failure as a warning.
Roles and Responsibility	<ul style="list-style-type: none">• Working with R&D team to understand business problems and finding an optimal solution.• Collecting and Sampling the Data, Preprocessing the data, treating outliers and missing values.• Finding and Developing best model for different clusters of Raid-5 Raid-6.• Perform EDA and feature engineering to develop the statically model.• Finding out the correct attribute for the model.• Computing the accuracy with K-fold cross validation.• Pickling the File for production(Pickle/joblib)• Testing the production file in Linux environment with crontab job scheduling.

Project 2	Senior Data Scientist
Project Title	Citi Recommender
Company	Citicorp Service
Algorithm	Collaborative Filtering
Techniques	User-User similarity, Product- product similarity, Cosign distance, Jaccard distance, clustering.
Description	This model was stated with a POC to challenge the existing NBO model for Citi production recommendation. Later point it becomes the priority to deploy in multiple country for Citi-product recommendation.
Roles and Responsibility	<ul style="list-style-type: none">• Collecting the data from different tables to build the recommendation engine for Citi products.• Preparing the data for model building and model validation.• Creating U-U similarity, P-P Similarity Metrix.• Deciding the model type and Building the model code using python and pyspark.• Comparing the model score using different distance algorithm like cosign, Jaccard, Manhattan etc and selecting the best out of them.• Developing the model validation procedure as the data was binary.• Model code optimization for better user experience.• Model Deployment and testing.

Project 3	Senior Data Scientist
Project Title	NNM (Net New Money)
Company	Citicorp Service
Algorithm	Regression , XGBoost
Techniques	Feature Engineering , Feature Importance, K-fold cross validation, PSI,KS
Description	Net New Money model is used to predict the percentage total relationship balance (TRB) change in the next 3 months for a customer. It helps the business in identifying the customers who might take out a lot of money from there account. This model is aimed to prevent the outflow fund from Citi and reduce the same by allocating an RM to the highlighted customers.
Roles and Responsibility	<ul style="list-style-type: none"> • Deciding the important features for the model with the help of domain expertise. • Pulling data from ADL table using Pyspark and preprocessing the data as per model requirements. • Developing the regression model using XGB and creating a benchmark model for the model performance check. • Developing the model for one county and create a model tool to deploy models for different country in one click. • Preparing production ready code and prior deployment testing. • Developing deployment ready code using spark2 and python3. • Creating library using Cython for faster processing and enabling parallel processing in the model tool using Dask.

PROFESSIONAL ENHANCEMENT

- Neural Network and Deep Learning by deeplearning.ai on Coursera
- Deep Learning for coders – Fast.ai
- Deep Learning A-Z Hands on approach – Udemy
- Machine Learning – Andrew NG
- Taming big data with Apache Spark – Udemy