  

**PARTHA PRITAM DEKA** 3026 Blackberry Avenue

Ph: 860-794-5031 email: partha.pritamdeka@gmail.com San Ramon, CA -94582

<http://www.odbms.org/2017/01/qa-with-data-scientists-partha-deka/>

GitHub: <https://github.com/ParthaPritamDeka/Machine_Learning>

https://github.com/ParthaPritamDeka/Coursera\_Deep\_Learning

<https://github.com/ParthaPritamDeka/Self_Driving_Car_Nano_degree>

Passionate Data Scientist with 9 years of experience in the electrical, electronic, manufacturing, retail & online consumer industry with a Master’s in Electrical Engineering focused in **inferential** & **descriptive** statistics and Advanced Data mining. Skilled in **statistical analysis**, **Machine learning (Python SKLEARN)**– **Supervised, Unsupervised, reinforce learning**, **Deep learning(Tensorflow)**, **Computer Vision (Image processing),** **Data Engineering**- **Python**, **Hadoop, Spark (PySpark), SQL, MPP Databases – GreenPlum, Redshift, RDMS, ETL**, Data Visualization. Strong research professional with Machine Learning / Data Science focus from **Udacity** & **Coursera**

**Education:**

**M.S. Electrical Engineering,**

**Wichita State University,** Wichita, Kansas **2007 -2009**

**B.Tech. Electrical Engineering,**

**National Institute of Technology, Silchar, Assam**, India **2002- 2006**

**Machine Learning Engineer Nanodegree, Udacity 2017**

**Deep Learning Specialization, Coursera 2017**

**Machine Learning Projects:**

* **Built a Student Intervention System** - The goal for this project is to identify students who might need early intervention before they fail to graduate. After data cleansing, transformation and feature engineering I built several supervised learning models to accurately predict where a student would fail to graduate. Then, I chose the best candidate algorithm from the preliminary analysis and further performed hyperparameter tuning to optimize the model performance. Some of the algorithms I used - Decision trees, Logistic Regression, Stochastic Gradient Descent, Multilayer Layer perceptron deep learning model, Ensembled tree based models: Random Forest, Gradient Boosting. **Gradient Boosting** was the final model with the most optimal accuracy.
* **Enron Fraud detection:** In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. The goal of this project was to use supervised learning to identify persons of interest (POIs) in the Enron corporate fraud case. The features were financial features (such as salary, bonus, and stock options) and email features (such as number of messages sent, messages sent to POIs, and number of messages received). Machine Learning allows to try to find patterns to detect POIs. I have trained an **Adaboost** (Boosted decision tree) classifier to help identify POIs on the available dataset so that on new data -- the model can then identify whether a new person may be a POI or not based on their data(features)
* **Built a stock price predictor** - This work involved research in financial data analysis as well as trying out various machine learning techniques such as Polynomial learn Regression, KNN, ARIMA for predicting stock prices. Finally, I have used Linear Regression to predict stock prices one day ,7 days, 14 days, 28 days in future with ‘**>=.85’** R-squared score on the test datasets. I also built an optimized portfolio of investment on historical data with Sharpe Ratio optimization

<https://github.com/ParthaPritamDeka/Machine_Learning/blob/master/Capstone%20Project/Udacity_Capstone_Project_Report.pdf>

* **Created Customer Segments:** In this project, I applied unsupervised learning techniques on product spending data collected for customers of a wholesale distributor in Lisbon, Portugal to identify customer segments hidden in the data. l first explored the data by selecting a small subset to sample and determine if any product categories highly correlate with one another. Afterwards, I preprocessed the data by scaling each product category and then identifying (and removing) unwanted outliers. With the good, clean customer spending data, I applied **PCA** transformations to the data and implemented K -means clustering algorithm to segment the transformed customer data. Finally, I compared the segmentation found with an additional labeling and consider ways this information could assist the wholesale distributor with future service changes. Doing so would equip the distributor with insight into how to best structure their delivery service to meet the needs of each customer by performing **A/B testing**
* **Finding Donors for CharityML –** In this project I have used several supervised learning algorithms to accurately model individuals' income using data collected from the 1994 U.S. census. Then, I chose the best candidate algorithm from preliminary results and further optimize this algorithm to best model the data. My goal with this implementation was to construct a model that accurately predicts whether an individual makes more than $50,000. This sort of task can arise in a non-profit setting, where organizations survive on donations. Understanding an individual's income can help a non-profit better understand how large of a donation to request, or whether they should reach out to begin with.
* **SmartCab** - Implemented **Reinforcement Learning based Q-Learning algorithm** to train a Smart Cab to take optimal actions in an environment. The smartcab receives positive or negative rewards based on the action it has taken. Expectedly,

the smartcab will receive a small positive reward when making a good action, and a varying amount of negative reward dependent on the severity of the traffic violation it would have committed. Based on the rewards and penalties

the smartcab receives, I implemented a **self-driving agent** which learnt an optimal policy for driving on the city roads while obeying traffic rules, avoiding accidents, and reaching passengers’ destinations in the allotted time

* **Predicting Boston Housing Prices** - In this project, I applied machine learning concepts on data collected for housing prices in the Boston, Massachusetts area to predict the selling price of a new home. I first explored the data to obtain important features and descriptive statistics about the dataset. Next, I properly split the data into testing and training subsets, and determine a suitable performance metric for this problem. I then analyzed performance graphs for a learning algorithm

with varying parameters and training set sizes. This enabled me to pick the optimal model that best generalizes for unseen data.

* **Titanic survival exploration** - In this project, I created decision functions that attempt to predict survival outcomes from the 1912 Titanic disaster based on each passenger’s features, such as sex and age. I have started with a simple

algorithm and increased its complexity until I am able to accurately predict the outcomes for at least 80% of the passengers in the provided data.

**Computer Vision and Deep Learning Projects:**

https://github.com/ParthaPritamDeka/Coursera\_Deep\_Learning

* **Image Processing - Detected Lane lines in the road on video streams using Computer Vision techniques** – Gray scaling images, Canny edge detection, Hough transform and extrapolation of the detected Lane line

<https://github.com/ParthaPritamDeka/Self_Driving_Car_Nano_degree/tree/master/CarND-LaneLines-P1-master>

* Built a **deep neural network in python** that classifies cat vs. non-cat images.
* **Built a deep neural network** using **Tensorflow** to recognize numbers from **0 to 5** and improved the accuracy **using Adam optimization.**
* Built a **Convolution Neural Network in Keras** to check whether a person is happy or not – “**Happy House Problem”**
* **Traffic Sign Classifier – Used Tensor flow to implement the LENET-5 convolutional neural networks architecture to classify traffic signs**

**Text Analytics project**:

* **Item Based collaborative filtering: Recommended** similar movies on movie-lens data using Pearson co-relation as well as cosine correlation similarity using **Spark** on **Amazon EMR**

**Environment: Python, Sklearn, Pandas, Numpy, Kaggle Datasets, , TensorFlow, OpenCV, AWS, GPU, , Spark 2.0**

**Awards, Recognitions, Patents:**

* **United States Patent Pending** 15/285,907 **“Delivery Status Diagnosis for Industrial Suppliers using Multivariate Linear Regression model”**
* Above and Beyond Silver Award **“Delivered a Boosting Algorithm for Predicting Assets for a key customer”**
* Above and Beyond Bronze Award **“Deliver Results in an Uncertain World” –** GE Digital Predix Data Science
* Above and Beyond Bronze Award **“Empower and Inspire Others” –** GE Digital Predix Data Science
* **Executive Award GE Digital–** September 2015

**Relevant Work Experience:**

**General Electric,** San Ramon California

**Staff Data Scientist**

**03/15 – Present**

* Built a **multivariate Linear Regression model** (with 95% accuracy) as well as operationalized the model in Greenplum MPP platform to predict and visualize on-time delivery metric of Suppliers. This solution processes, transforms and predicts on 10 GB of data weekly. This has improved assembly/shop inventory, supplier OTD performance & **$10M to $15M** yearly savings for business.
* Delivered 11 Statistics based Data science analytics onto Predix Analytics catalog (top contributor) This supported positioning Predix as a top Analytic IIOT Platform in Forrester report and helped to make Predix more compelling to **Toshiba, Schindler & Sun Power** (<https://www.predix.io/catalog/analytics>). Some of the key Analytics built:
* **Density Based clustering**
* **MeanShift Clustering**
* **Categorical to numerical data**
* **Max Variance feature Analysis – Principal Component Analysis**
* **Linear Discriminant Analysis**
* **Random Forrest Classifier**
* **Gradient Boosting Classifier**
* **Ensembled Classifier**
* Built a scalable Data pipeline on GE Predix platform on Cloud Foundry for processing machine learning algorithms on signals coming from various sensors on elevators- this application includes building a Gatekeeper application and configuring rabbitmq queuing engine for processing very high through data as massively parallel miniature data processes
* Built a **Gradient Boosting** ensembled model for predicting Asset Allocations for Johnson & Johnson (key customer for GE). This model deployed on GE Predix Analytic Catalog. This has made GE Predix compelling for Johnson and Johnson for building and deploying Machine Learning algorithm
* Engineered unstructured outage data as well as historical weather data and built a **Random forest based** model to predict outages. Engineered Power outage alerts on **twitter** data.
* Implemented snapshotting/archiving solution by transforming / aggregating 40GB+ data. This has enabled business to track changes in manufactured requested dates/ Promise dates/ Quantity of parts delivered by Suppliers which was never

done in legacy systems, a fundamental problem Aviation. Aviation SME appreciated the value addition this has brought to business.

* Built ETL infrastructure which purges and loads 140 GB of network & OPS data every week. Developed ETL infrastructure for structured & unstructured data for network, machine health, incidents & CI data enabling IT operations business to better support mission critical applications
* Built a Python framework for GE Healthcare to engineer Install Base config data to identify potential upgrade opportunities with a potential revenue impact of **$8MM**/yearly
* Enabled GE Digital executives to view daily Predix customer usage by designing & building dashboards on DOMO platform.

**Environment: Python, Sklearn, SQL, Talend, Greenplum, Java, Shell Scripting, Hadoop – HDFS, HIVE. Talend – ETL tool, Predix, RabbitMq, REST, Statistics**

**Hightail Inc, Campbell, CA**

**06/14 – 03/15**

**Data warehouse Developer**

Hightail is a Cloud file sharing/storage midsize company. I am responsible to provides key KPIs, sales leads to the Sales/Finance team as well as to the key Stake holders. I have designed and built a robust Datawarehouse platform to suffice the tableau dashboards as well as the adhoc reporting needs from multiple sources which includes legacy OLTP and NoSQL systems

* Work extensively on **Amazon Redshift** and **Talend** ETL to build a scalable/optimized ETL platform
* Built a Data extraction framework to pull data from multiple sources such as **MySQL, Splunk(unstructured log data), Cassandra(key value based)** and **Flurry(use REST API)**
* Work directly with Senior Sales executives to understand and deliver their reporting / KPI needs
* Developed and delivered Key User Activity, User Registration, User volume metrics etc. across the globe to the Enterprise customers
* **Optimized Redshift Clusters/Nodes** using distribution and sort keys for optimal query performance
* Designed, developed, and implemented ETL jobs to effectively process high-volume data sources and meet nightly processing windows
* Developed and Documented all ETL jobs and processes according to ETL Standards and Best Practices
* Interacted with the Product Development and Data Analytics teams to determine reporting needs and translate those needs into the Data Warehouse Data Model and ETL processing needs

**Environment: Amazon S3, Amazon Redshift, SQL, JIRA, Talend, Sqlserver, PostgreSQL, SSIS, SSRS, Java**

**Consulting at Gap Inc, Macys Inc and Pandora Media**

**06/13 – 06/14**

**Senior Data Engineer**

* Built custom python **mapreduce jobs, HIVE** ETL jobs in Hadoop clusters
* Built **PLSQL procedures/function/triggers** for Datawarehouse development
* **Performance tune PL/SQL queries using statistics, indexes, hints, partitioning , virtual private database** etc.
* Built **Unix shell scripts** for scheduling the ETL pipelines through cron
* Extensive communication with the SMEs
* Built **Data stage** ETL jobs
* Owned the code release/ deployment process following best practices

**Environment: Hadoop, Hive, Python, Unix shell script, PLSQL, Oracle 10g, JIRA, SVN, Pycharm, Perforce Version Control, Datastage 7.0 server edition, PLSQL, Oracle 10g, Espresso, JIRA, SVN**

**Cisco Systems, San Jose, CA**

**05/11 – 06/13**

**Data Engineer**

This project is based on Installed Base and Service Agreement Oracle Applications for Cisco Network Business Operations Group.

The data from IB SA oracle applications is used to build highly intuitive OBIEE Dashboards for the Customer Intelligence Center. This process therefore is carried out as an intuitive intelligent decision making step to aid in future product service renewal opportunities. Further, the IB SA data is sent through the IB Reconciliation Engine (in- house PL/SQL engine to fix IB SA data) to report error- free data in the customer-intelligence center. Further the Ib Reconciliation engine has been migrated to Hadoop environment as as a proof-of-concept for speed fast processing.

**Responsibilities**:

* Developed PLSQL **Packages/ Procedures**
* Performance tuned PLSQL jobs with **cursors** ,**bulking** techniques, **parallelism**, **Indexes** etc
* Work on analyzing and providing the correct requirements for **Dimensional hierarchies**, **Aggregate tables** for **Drill-down** and **aggregate navigations**
* Built powerful **OBIEE dashboards and reports** from scratch featuring key business matrices
* Set up a five node **Hadoop** cluster on **Cloudera** **CDH4** environmen from scratch
* Utilized **sqoop** for migration of Oracle environment to **Hadoop Hbase** environment
* Monitored the named node and the data nodes using Ganglia, Job tracker UI and Name Node UI
* Used **Sqoop** to migrate IB Reconciliations statistics from **Hbase** to **Oracle** env. and utilized **OBIEE** dashboard to publish the statistics

**Environment**: **OBIEE 10g, Oracle 10g, Teradata, PL/SQL, Toad, Linux, Informatica 8.6, Oracle Applications – Install Base and Service Agreement, Hadoop – HDFS, Hbase, Map reduce, CDH4**

**Consulting at various midsize (Leeyosoftware Inc. etc) companies via Infotech Spectrum Inc**

**02/10- 04/11**

**Data Engineer – OBIEE / ETL Developer**

Leeyosoftware Inc is the leader in revenue recognition automation. Revpro is the only “out-of-box” solution for automation of EITF 08-1 and 09-3 currently in the market. Revpro was developed to meet the needs of revenue recognition users and to eliminate the reliance on spreadsheets for revenue accounting, which is prolific amongst many companies. This Project is based on integration of OBIEE with Revpro modules for creating picture perfect reports.

* Built **OBIEE Repositories and Dashboards**.
* Built custom **PL/SQL** ETL jobs
* Built custom **ODI** ETL jobs

**Environment: Oracle BI 10.1.3.3, Oracle Data Integrator 10.1.3.4, Oracle 10g, SQL, PL/SQL, Oracle Enterprise Linux**

**Patni Computer Systems Ltd, Mumbai, India 05/06 -07/07**

**OBIEE / Siebel Analytics Developer**

**Responsibilities:**

* Built **OBIEE Repositories and Dashboards**.
* Built custom **PL/SQL** ETL jobs
* Built custom **Informatica** ETL jobs
* Performed **Unit, Integration and Regression testing** to validate **reports** and **mappings**.
* Involved in fixing invalid mappings, testing of stored procedures and functions, code review, unit and integration testing of **Informatica sessions**, batches and target data.
* Developed **PL/SQL** stored procedures for source pre-load and target pre load to verify the existence of tables

**Environment: OBIEE 10.1.3.2, Informatica 7.1.1, Oracle 10g, QL Developer, PL/SQL, Windows XP**