PRANAY KATTA

LinkedIn Profile | GitHub Profile | Tableau Public Profile | R Publications
New York City, NY | Ph. 862-208-2642 | pkkatta19@gmail.com | Personal Website

SUMMARY

- Experienced in Machine Learning, Analytics, Statistical Modeling, Business Intelligence, Data Wrangling and Visualization
- Excellent analytical, problem-solving skills in handling complex data with expertise in writing SQL queries & Data Reporting

EDUCATION

Master of Science in Data Science, Saint Peter's University **1/2016 - 1/2018**

COURSEWORK: Statistical Programming (SAS, R), Data Visualization (Tableau, D3, Plotly, leaflet), Data Mining (AWS, RapidMiner, Clustering), Machine Learning (R, Calculus, Linear Algebra, Neural Networks), Data Analysis and Decision Modeling (Statistics, Probability), Database & Data Warehousing (SQL, Postgres, RDBMS), Big Data Analytics (Hadoop, PySpark, MapReduce, Distributed Computing), Marketing Analytics (Python, Association Rule Mining, Market Basket), Financial Computing and Analytics (Sampling, Monte Carlo Simulations) , Predictive Analytics and Financial Modeling (Time Series, Tree Models)

B.Tech in Electronics and Communication Engineering, Jawaharlal Nehru Technological University, India **6/2014 - 6/2014**

Algorithms and Data Structures, Probability and Stochastic Process, Object Oriented Programming, Mathematical Methods

TECHNICAL SKILL SET

- **Tools & Languages:** R, Python, SAS, Data Structures, R, MySQL, Tableau, Power BI, MS Excel (VLOOKUP, Pivot Tables), RapidMiner, Tensorlfow, Hadoop, PySpark, D3, Google Analytics, AWS (EC2 | S3), Keras, Postgres, Hive, Git.
- **Statistical Techniques:** Decision Trees, Linear & Logistic Regression, Random Forests, Clustering, Time Series Forecasting

EXPERIENCE

Data Scientist – **NYC DEP** – New York, USA **05/2019-Present**

- Pre-processed NYC DEP's Automatic Meter Readings (AMR) Data and created visualizations using Python
- Developed a predictive model to estimate the failure of water meter using scikit-learn & reduced payback period

Data Analyst Intern – **NYC DEP** – New York, USA **05/2018-Present**

- Cleaned and pre-processed demographic data - performed exploratory data analysis and published FY 2018 Demographic Reports using Power BI
- Responsible for maintaining MySQL and Access Database by collecting and updating the data from internal sources
- Developed a predictive model using R for employee separation by using various features to reduce the employee attrition rate by visualizing the goals that should be achieved
- Writing Ad-Hoc SQL queries and generating reports based on the requests and effectively communicate results

Teaching Assistant – **Saint Peter's University** – New Jersey, USA **01/2017-05/2018**

- Taught Data Analysis and Decision Modeling, Statistical Programming, Intro to Data Science (Grad) and C++ (undergrad)

Research Assistant – **Saint Peter's University** – New Jersey, USA **05/2017-10/2017**

- Created Word2Vec Natural Language Processing models Skip-gram and CBOW using Tensorflow on New York Times data.
- Benchmarked newly setup Hadoop clusters in Big Data Lab with Hadoop Tera Sort and TestDFSIO Benchmark

Data Scientist | Programmer –**TATA Consultancy Services Ltd** – Hyderabad, India **12/2014-12/2016**

- Implemented a model which predicts Network failures by analyzing log files (unstructured data) generated by telecom systems Operational Support System and Business Support System and incorporating them into early warning system
- Worked closely with an agile team of data scientists which developed recommendation systems to optimize margins with great customer satisfaction by analyzing patterns like historical customer data, social links and purchase patterns.
- Was part of a team for creating libraries and plugins for TCS HOBS CRM across CSP's using REST API's
- Created Hybrid Automation Framework in Java for Selenium which can be integrated across different platforms. Developed an automation framework which reduced test time by 75% and costs by USD 110,000 every month. Created ETL and Cross Browser Test scenarios for deploying
- Automated CRM, SDLC & Bug life reports using R(ggplot) which helped to better understand the project bottlenecks

Software Developer Intern – **Invenio Business Solutions** – India **09/2014-11/2014**

- Analyzed business requirements for migrating from JavaCAPS to Mule ESB for a leading music company based in NY. Worked closely with senior developers by assisting in packaging, deployment and migrating to MULE ESB
- Maintaining and running scheduled jobs on Oracle DB for SAP Back End

Analyst Intern – **Frndzzy (now Edwisely)**– Hyderabad, India **01/2014-09/2014**

- Created a hybrid app-based platform to link students, universities, and the marketplace. Identified and measured the business KPIs to influence, support, and execute result-oriented business decisions.

ACADEMIC PROJECTS

Patient length of Stay |***Tool used:*** R, Python, Tableau

- Worked as a team to perform exploratory analysis and descriptive analysis on Healthcare Cost and Utilization Project (HCUP) data using regression techniques to predict the ideal length of stay of a patient in order to avoid the entire cost of readmission of patients in hospital within 30 days of discharge.
- Suggested other models like Markov Decision Process to predict the length of stay based on previous visit parameters

Forecasting Stock Returns using Neural Networks and Time Series |***Tool used:*** R

- Constructed an ARIMA Time Series Factor Model to predict log-returns in R and segregating returns in five buckets by using quarterly results of the healthcare industry. Different variables like assets, market cap, P/E ratio. 20 Quarterly results have been analyzed and portfolio has been created with five buckets. Used R statistical software for effective analysis by hypothesis testing to validate data and interpretations.
- Implemented Neural Network with deep neural networks which consistently beat the market compared to traditional time series model.

Smart Navigation | Building a Restaurant recommendation system |**Tool used:** Python, Pandas, SciPy, NumPy

- Built a recommender system which recommends user find best restaurants based on his past visits and factors like cuisine, price etc.,
- Application would help the user to navigate to nearest restaurants using his current location using various recommendation algorithms. Application could be tweaked to select recommendation based on various factors.

Pricing the Derivatives | **MCMC** | ***Tool used:*** R

- Developed an n-step Binomial Tree to compute prices of call/put options for discrete dividends using the Schroder Method
- Created Monte Carlo Simulations for random number generation and Morkov Chain Monte Carlo methods

Fake News Detection | ***Tool used***: R

- Capstone project on detecting fake news from meta data and verifying it through different sources, document similarity and algorithms
- Achieve 95% accuracy in stance detection and implemented TF-IDF based predictions using different classifier algorithms

Other Projects | ***Tool used: Python, R, Tableau***

- Predicting breast cancer using different algorithms like logistic regression, linear discriminant analysis, support vector machine, K nearest neighbors using different randomized approaches like bagging, boosting and stacking **code**
- Visualized American Airlines on-time performance metrics & performance with respect to other competitors using BTS data set containing more than 1 million rows & 147 variables for the year 2015-2016 **link**
- Implemented Expectation Maximization algorithm using R on a given dataset. The goal was to complete an expectation maximization model from the beginning which consisted of initializing the problem through K-means and then running the expectation maximization model for each K value. The dataset consisted of 71 variables, inclusive of ID and 70 vectors, and 1077 observations **code**
- Predicting the failure of the challenger based on different flight temperatures using logistic regression **code**
- Created a model which predicts mpg of a car based on different variables using linear regression **code**
- Hand Written Digit Recognition - Applied principal component analysis (PCA) to reduce the dimensionality & Cross Validation is applied on the data. Built a Neural Network with three hidden layers resulted in accuracy of 67% (without PCA) and 97% (with PCA) **code**
- Implemented L-BFGS quasi newton optimization algorithm from the scratch using R **code**

HONORS and CERTIFICATIONS

- Got shortlisted for **SSAC 2018 Hackathon** finals at **MIT (Massachusetts Institute of Technology)**
- **Rank 1** in Python Hackerrank Challenge; **Won** Saint Peter's University Data Science **Hackathon** Challenge 2018
- **Won** First Place in National Level Robotics Competition organized by **IIT Guwahati**