Pankaj Sharma- Call
Data Scientist - University of Colorado, Boulder, COAug

Avenel, NJ - Email me on Indeed: indeed.com/r/Pankaj-Sharma--Call/ba1752cd632a26fc

• Professional experience spanning 17+ years in data science, data analytics, data mining, machine learning , business analytics, business intelligence, competitive intelligence, predictive analytics, forecasting, data acquisition, data validation, predictive modeling,data visualization& project managementacross domains of biostatistics, public sector, services , retail, BFSI, econometric, academic and HR domain.
• Professional and academic work experience withhandling large data sets of structured and unstructured data.
• Hands on experience of working with APIs and doing analytics on the same
• Hands on experience in Natural Language Processing and Sentiment Analysis
• Hands on experience in undertaking unsupervised, supervised and reinforced Machine learning, analytics and visualizations using various packages of R and core analytical Python libraries such as Numpy, Scipy, Pandas and Scikit-learn and data visualization in Matplotlib,
• Extensive experience in performing the statistics test (ANOVA, Hypothesis testing and A/B test), Multivariate analysis and EFA/ PCA/CFA, etc
• Proficient in Statistical Modeling, Data Mining and Machine Learning Algorithms in Data Science/Forecasting/ Predictive Analytics such as Linear and Logistics Regression, LDA, Item and Discriminate Analysis, Apriori, Random Forest, K Means, Artificial Neural Network,Decision Tree,SVM, K-Nearest, Bayesian, Hidden Markov, etc.
• Experience on developing dashboards using Tableau, PowerBI
• Experience working with relational databases like MySQL, MS Access, NoSQL databases like MongoDB
• Exposure to AI and Deep learning platforms/methodologies like tensor flow, RNN
• Experience in statistical packages like SPSS, SAS, Crystal Ball
• Experience in dealing with structured and semi-structured data in HDFS and Hive
• Familiar with the concepts of Hadoop, MapReduce framework
• Experience in doing data visualization and report dashboard in Tableau, Qlick view, and familiar withPower BI
• Maintained version control of Python using Git, Github
• Worked in Linux as well as Windows
• Knowledge of the Software Development Life Cycle (SDLC) , Agileand Scrum
• A good team player and self-motivated by passion in data science
Willing to relocate: Anywhere

WORK EXPERIENCE

Data Scientist
University of Colorado, Boulder, COAug - Boulder, CO - 2014 to Present

About Client:- As one of 34 U.S. public research institutions belonging to the prestigious Association of American Universities (AAU) - and the only member in the Rocky Mountain region - our goal at CU Boulder is to directly affect Colorado communities through collaborative research, innovation and entrepreneurship. Our faculty, staff and students work with the broader community to establish unique connections that have lasting outcomes - both across Colorado and around the world.

Roles &Responsibilities:
• Developed applications of Machine Learning, Statistical Analysis and Data Visualizations with challenging data Processing problems in sustainability and biomedical domain.

• Compiled data from various sources public and private databases to perform complex analysis and data manipulation for actionable results.

• Gathers, analyzes, documents and translates application requirements into data models and Supports standardization of documentation and the adoption of standards and practices related to data and applications.

• Developed and implemented predictive models using Natural Language Processing Techniques and machine learning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, Random Forests, K-means clustering, KNN, PCA and regularization for data analysis.

• Designed and developed Natural Language Processing models for sentiment analysis.

• Worked on Natural Language Processing with NLTK module of python for application development for automated customer response.

• Used predictive modeling with tools in SAS, SPSS, R, Python.

• Applied concepts of probability, distribution and statistical inference on given dataset to unearth interesting findings through use of comparison, T-test, F-test, R-squared, P-value etc.

• Applied linear regression, multiple regression, ordinary least square method, mean-variance, theory of large numbers, logistic regression, dummy variable, residuals, Poisson distribution, Bayes, Naive Bayes, fitting function etc to data with help of Scikit, Scipy, Numpy and Pandas module of Python.

• Applied clustering algorithms i.e.Hierarchical, K-means with help of Scikit and Scipy.

• Developed visualizations and dashboards using ggplot, Tableau

• Worked on development of data warehouse, Data Lake and ETL systems using relational and non relational tools like SQL, No SQL.

• Built and analyzed datasets using R, SAS, Matlab and Python (in decreasing order of usage).

• Applied linear regression in Python and SAS to understand the relationship between different attributes of dataset and causal relationship between them

• Performs complex pattern recognition of financial time series data and forecast of returns through the ARMA and ARIMA models and exponential smoothening for multivariate time series data

• Pipelined (ingest/clean/munge/transform) data for feature extraction toward downstream classification.

• Used ClouderaHadoop YARN to perform analytics on data in Hive.

• Wrote Hive queries for data analysis to meet the business requirements.

• Expertise in Business Intelligence and data visualization using R and Tableau.

• Expert in Agile and Scrum Process.

• Validated the Macro-Economic data (e.g. BlackRock, Moody's etc.) and predictive analysis of world markets using key indicators in Python and machine learning concepts like regression, Boot strap Aggregation and Random Forest.

• Worked in large scale database environment like Hadoop and MapReduce, with working mechanism of Hadoop clusters, nodes and Hadoop Distributed File System (HDFS).

• Interfaced with large scale database system through an ETL server for data extraction and preparation.

• Identified patterns, data quality issues, and opportunities and leveraged insights by communicating opportunities with business partners.

• Delivered and communicated research results, recommendations, opportunities, and supporting technical designs to the managerial and executive teams, and implemented the techniques for priority projects.

Environment: Machine learning, AWS, MS Azure, Cassandra, Spark, HDFS, Hive, Pig, Linux, Python (Scikit-Learn/Scipy/Numpy/Pandas), R, SAS, SPSS, MySQL, Eclipse, PL/SQL, SQL connector, Tableau.

Data Scientist SME

Harley Davidson - Milwaukee, WI - March 2012 to March 2012

About Client: -Harley-Davidson is one of the world's strongest brands with a long history of building great motorcycles. It's more than just a business to us. We live our purpose every day- delivering dreams of personal freedom - and we do it like nobody else.

Roles &Responsibilities:

• Applied concepts of probability, distribution and statistical inference on given dataset to unearth interesting findings through use of comparison, T-test, F-test, R-squared, P-value etc.

• Applied linear regression, multiple regression, ordinary least square method, mean-variance, theory of large numbers, logistic regression, dummy variable, residuals, Poisson distribution, Bayes, Naive Bayes, fitting function etc to data with help of Scikit, Scipy, Numpy and Pandas module of Python.

• Applied clustering algorithms i.e.Hierarchical, K-means with help of Scikit and Scipy.

• Built and analyzed datasets using R, SAS, Matlab and Python (in decreasing order of usage).

• Applied linear regression in Python and SAS to understand the relationship between different attributes of dataset and causal relationship between them

• Performs complex pattern recognition of financial time series data and forecast of returns through the ARMA and ARIMA models and exponential smoothening for multivariate time series data

• Pipelined (ingest/clean/munge/transform) data for feature extraction toward downstream classification.

• NLP and sentiment analytics using social media API calls

• Developing data mining; data analytics data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.

• Performed K-means clustering, Kalman filtering, Multivariate analysis and Support Vector Machines in Python and R.

• Developed Clustering algorithms and Support Vector Machines that improved Customer segmentation and Market Expansion.

• Worked on NOSQL databases such as MongoDB and Cassandra.

• Experienced in Agile methodologies and SCRUM process.

Environment:Hadoop, Map-Reduce, HDFS, SQL, Pig,R, Python.

Team leader Data Analyst

Proctor and GambleMar - 2010 to February 2012

About Client: -P&G is made of many individual brands, each serving customers in different ways - but all with a focus on making peoples' lives a little easier. P&G Professional leverages P&G's scale, trusted brands, and strengths in market and consumer understanding to meet the specific business needs in a wide range of industries, such as foodservice, building cleaning and maintenance, education, hospitality, and care homes.

Roles &Responsibilities:

• Multivariate business and resource forecasting using machine learning algorithms.

• Applied linear regression, multiple regression, ordinary least square method, mean-variance, theory of large numbers, logistic regression, dummy variable, residuals, Poisson distribution, Bayes, Naive Bayes, fitting function etc to data with help of Scikit, Scipy, Numpy and Pandas module of Python.

• Applied clustering algorithms on market data to study the underlying data patterns. Methodologies used were PCA, Factor analysis,Hierarchial, K-means through Scikit/Scipy, R for projecting market

• Built and analyzed datasets using R, and Python.

• Applied linear regression in Python and SAS to understand the relationship between different attributes of dataset and causal relationship between them

• Performs complex pattern recognition of financial time series data and forecast of returns through the ARMA and ARIMA models and exponential smoothening for multivariate time series data

• Developed ETL based systems for data acquisition and data consumption by stakeholders.

• Developing data mining; data analytics data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.

• Performed K-means clustering, Kalman filtering, Multivariate analysis and Support Vector Machines in Python and R.

• Developed Clustering algorithms and Support Vector Machines that improved Customer segmentation and Market Expansion.
Environment:Hadoop, Map-Reduce, SQL, R, Python, In house ETL tools.

Data Analyst

SyngentaBioline Ltd - March 2009 to April 2010

About Client:- Syngenta AG is a global Swiss agribusiness that produces agrochemicals and seeds. As a biotechnology company, it conducts genomic research. It was formed in 2000 by the merger of Novartis Agribusiness and Zeneca Agrochemicals.

Roles &Responsibilities:
• Multivariate biostatical and environment data analytics.
• Applied linear regression, multiple regression, ordinary least square method, mean-variance, theory of large numbers, logistic regression, dummy variable, residuals, Poisson distribution, Bayes, Naive Bayes, fitting function etc to data with help of Scikit, Scipy, Numpy and Pandas module of Python.
• Applied decision tree and neural network based systems for forecasting and predictive analytics. Methodologies used were RNN, KnN, SMV, PCA, Factor analysis, Hierarchial, K-means through Scikit/Scipy, R, etc.
• Built and analyzed datasets using R, and Python.
• Applied linear regression in Python and SAS to understand the relationship between different attributes of dataset and causal relationship between them
• Pipelined (ingest/clean/merge/transform) data for feature extraction toward downstream classification.
• Developing data mining; data analytics data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.
• Performed K-means clustering, Kalman filtering, Multivariate analysis and Support Vector Machines in Python and R.
• Developed Clustering algorithms and Support Vector Machines that improved Customer segmentation and Market Expansion.
Environment:MS SQL, SAS, R, Vector Machines, Python.

Lead Consultant, Data base Management

Ministry of IT, Government of IndiaMar - 2001 to March 2009

Roles &Responsibilities:
• Developed the program management, program monitoring and evaluation framework of 27 mission mode project for data acquisition, data dissemination and data interoperability
• Statistical analysis and resource forecasting systems using statistical tools and packages like SPSS, SAS, etc..
• Applied linear regression, multiple regression, ordinary least square method, mean-variance, theory of large numbers, logistic regression, dummy variable, residuals, Poisson distribution, Bayes, Naive Bayes, fitting function etc to data with help of Scikit, Scipy, Numpy and Pandas module of Python.
• Applied clustering algorithms on market data to study the underlying data patterns. Methodologies used were PCA, Factor analysis, Hierarchial, K-means through Scikit/Scipy, R for projecting market
• Built and analyzed datasets using R, and Python.
• Projecting casual -effect relationship among various components of the project using open tools.
• Performing complex pattern recognition of the time series data for analytical purposes,
• Developed ETL based systems for data acquisition and data consumption by stakeholders.
• Developing data mining; data analytics data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.
Environment:Red Hat, MS, SAS, SPSS, SQL, Data Warehousing , R, Python, MS Access, In house ETL tools.

Data Technology Analyst

J.P Morgan Chase - March 1995 to February 2001

Roles &Responsibilities:
• Developing BI/CI solution for senior stakeholders
• Statistical analysis and resource forecasting systems using statistical tools and packages like SPSS, SAS, etc..
• Developing forecasting and resource projection solutions using open source and proprietary tools.
• Projecting casual -effect relationship among various components of the project using open tools.
• Performing complex pattern recognition of the time series data for analytical purposes,
• Developed ETL based systems for data acquisition and data consumption by stakeholders.
• Developing data mining; data analytics data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.
Environment: Red Hat, MS, SAS, SPSS, SQL, Data Warehousing & OLAP tools.

EDUCATION

MS in Data Analytics in Data Analytics
Clarkson University - Potsdam, NY

SKILLS

algorithms (10+ years), SAS (10+ years), Serial Attached SCSI (10+ years), SPSS (10+ years), visualization (10+ years)

ADDITIONAL INFORMATION

TECHNICAL SKILLS:
Languages/Tools/Big Data
R3.X (all major packages), Python 2.x/3.x (Numpy, Pandas, Scipy,Scikit-learn, Matplotlib),
Hadoop/HDFS/Hive/Pig
AWS
Machine Learning Algorithms
Linear and Logistics Regression, LDA, Item and Discriminate Analysis, Apriori, Random Forest, K Means, Artificial Neural Network, Decision Tree, SVM, K-Nearest, Bayesian, Hidden Markov, etc.
Visualizations
Ggplot,GoogleViz, Tableau, 3 D visualization, Power BI, Qlik , developing dashboards using Tableau, Power BI, etc
Statistical Tools
SPSS, Crystal Ball, SAS