Data Scientist/ Data Science Consultant

Data Scientist/ Data Science Consultant - Client- Macy's

San Francisco Bay Area, CA

• Data analytics professional with 6+ years of experience in delivering end to end data science projects.

• Implemented advanced analytical solutions to real business problems leveraging Machine Learning algorithms and Business intelligence tools that have impacted the business and end user experience.

• Demonstrated success in designing and executing hypothesis driven analytical projects and implementing design of experiments (DOE) methods to find cause-and-effect relationships.

• Experienced in using Python/R Studio/SQL/ SAS to perform statistical analysis and to implement machine learning algorithms utilizing different packages.

• Leveraged big data tools and supporting technologies for extracting meaningful insights from large data sets. Good knowledge on Distributed Computing, Hadoop Architecture and its ecosystem components like HDFS, Map Reduce, HIVE, IMPALA, Spark (PySpark) and Kafka.

• Experienced in using source code change management and version control tool such as Github.

• Proficient in implementing best practices for Data Visualization and adept in utilizing Tableau Desktop for creating appealing and interactive dashboards.

• Extensive exposure on analytics project life cycle CRISP-DM (Business understanding, Dataunderstanding, Data preparation, Modelling, Evaluation and Deployment).

• Capable to generate new insights, drive business decisions based on data and strong commitment to make positive impact.

Authorized to work in United States for any employer

Work Experience

Data Scientist/ Data Science Consultant

Client- Macy's - San Francisco, CA

June 2018 to Present

• Retail Analytics: Designed a predictive modelling framework in python to understand the likelihood of a customer making a purchase leveraging rule-based extraction engines and ensemble of machine learning models. The solution showed potential of 7% improvement in sales per customer with an incremental revenue of ~3M.

• Leveraged disparate data sources that provide deep customer insight including online transactional data, web data, payment and orders history and marketing campaigns exposure data.

• Performed price sensitivity and variation analysis across different marketing channels and conducted exploratory data analysis on variables such as life time value and profit score.

• Built data pipelines, implemented code modularization involving package creation and co-developed REST API's using Flask for production deployment.

• Co-designed a robust customer segmentation framework that identified behavioral groups among the customer base. Generated insights that helped marketing team to design more effective marketing campaigns and create more relevant content that improve personalization of online shoppers.

• Performed data discovery and build a stream that automatically retrieves data from multitude of sources (SQL databases, external data such as social network data, user reviews) to generate KPI's using Tableau.

• Tools: Python/ Jupyter Notebook/ Oracle SQL developer/ Unix/Tableau/HDFS/IMPALA /HIVE/Jira/Hue.

Data Scientist/ Data Science Consultant

Client- GlaxoSmithKline - San Diego, CA

December 2017 to June 2018

• Text analytics: Implemented a natural language processing and statistical modeling-based approach to find nearest-neighbor NCIs (Non-Conformance Incidents) reported for products/process that were manufactured across global manufacturing sites. Used Python NLTK package and reduced recurring incidents up to 60%.

• Performed topic modeling on incidents reported and categorized incidents to topics to tag incidents to product related or process related for further root cause analysis.
• Incident sentences were converted to tokens and compared for similarities using stop wording and word lemmatization. Computed distance match between recurring incidents using cosine similarity.
• Generated percentile scores for capturing distance between recurring incidents and integrated with complaints effectiveness metrics dashboard in Tableau to provide visual insights to business users.
• Tools: Python/ SQL Server / Microsoft Excel/Unix/Tableau/HDFS/Hive/Jira.

Data Analytics Consultant
Client- Blue Cross Blue Shield - Chicago, IL
November 2016 to December 2017
• Negative Outcomes Risk Prediction Model: Analysed Medicare resource utilization groups (RUG's) and Managed Care insurance claims data from healthcare provider and predicted residents with negative margins using Regression and CART.
• Handled class imbalance using re-sampling techniques. Utilized Logistic regression in R to identify the factors affecting margin and predict residents with negative margins. Build Gradient Boost Model utilizing H20.ai in R to analyze variable importance and evaluate model performance.
• Performed clustering analysis on historical patient level data to classify them into payment (total expense per stay) groups and identified parameters impacting expenditures and provided recommendations to drive reimbursements.
• The model showed incremental revenue increase of $1M by identifying patient groups.
• Tools: R studio/ Azure Data Studio / Microsoft Excel/ Tableau

Data Analytics Specialist
Client - Pfizer Inc - Peapack, NJ
January 2016 to November 2016
• Marketing analytics: Designed a robust customer segmentation framework based on physician prescribing potential and adoption rate of branded drugs. Predicted physician lifetime value for each segment groups leveraging APLD patient level data from Symphony, IMS Xponent, IMS Sales and Distribution data (DDD) and various internal datasets.
• Performed A/B testing by sending emails to certain physician segments in the categories while maintaining a control population to observe the incremental impact of the emails. Provided distinct segments using unsupervised techniques with key physician characteristics which helped marketing team to prioritize market segments and devise promotional messages.
• Compared conversion metrics within test and control group and identified cases that were positively correlated with segments with high prescribing potential and adoption rates. Analysed prescribing behaviour across different groups, observed segments with high physician life time values were 8 to 10 times likely to prescribe if they received email compared to other groups. Analytical model enabled marketing teams to minimize market spend and prioritize on market segments.
Tools: Python/ SQL server/ HIVE/Microsoft Excel/Power BI

Data Analytics Consultant
Client- CUNA Mutual Group - Madison, WI
November 2014 to December 2015
• Involved in the building and deployment of end to end real time fraud detection and segmentation model in Tableau and Azure ML web service to productionize claims scoring process using KNN and CART models.
• Performed text analytics on claims transcript notes using NLP using Latent Dirichlet Allocation (LDA) model to perform topic modelling and enhance existing model. Optimized and streamlined the claims model to process a claim within stipulated SLA. Implemented code modularization involving package creation, version control to push code to central repository improving code maintainability.
• Presented the model results to Claims business and helped them interpret its effects on KPI's.

• Helped in capturing required results and assess population stability over time to fine tune the model.

Sr. Data Analyst
Merck Inc
June 2013 to October 2014
• Risk analytics: Developed interactive dashboards using Tableau and made recommendations utilizing exploratory analysis that facilitated evaluation of quality and monitor the performance of BA/BE trial sites that contribute to potential risk.
• Worked closely with Business users, and interacted with ETL developers, Project Managers, and members of the QA teams for successful reporting across enterprise and ensured consistency on Key Performance Metrics (KPM's).
• Worked with DBA team for performance improvement issues. Created custom Function's (Date range, Time functions, Logical functions) for the reports. Designed, developed, tested, and maintained functional reports based on user requirements.
• Tools/Techniques: Tableau Desktop 8.0/ Tableau server / Cognos / Microsoft Excel

Data Analyst
MSN Pharmaceuticals Inc - IN
May 2012 to June 2013
Performed data pre-processing and cleaning to prepare data sets for further statistical analysis; including outlier detection and treatment, missing value treatment, variable transformation and various other data manipulation techniques using SAS programming language.
● Developed codes utilizing SAS Base/SAS SQL and prepared datasets of adverse events generated from Post Marketing Surveillance trials.
& Post-Market Surveillance for further analysis by HEOR (Health Economics & Outcomes Research) team.
● Modified existing SAS/SQL programs and created new programs using SAS macro variables to improve ease and speed of modification as well as consistency of results.

Education
Bachelor's in Information Technology
Acharya Nagarjuna University
May 2010
Skills
Apache hadoop hdfs (1 year), Apache hadoop impala (Less than 1 year), Bi (Less than 1 year), Business intelligence (5 years), clustering (1 year), Excel (3 years), Hadoop distributed file system (4 years), Hdfs (1 year), Hive (2 years), Impala (Less than 1 year), Logistic regression (1 year), Microsoft excel (3 years), Ms sql server (1 year), Power bi (Less than 1 year), python (5 years), Sas (1 year), Serial attached scsi (1 year), Sql (6 years), Sql server (1 year), version control (1 year), MS Office (Less than 1 year), Microsoft Office (Less than 1 year), Powerpoint (Less than 1 year)
Links
https://github.com/deepyion
Additional Information
TECHNICAL SKILLS
• Programming Languages: Python, Java, SAS Base, SAS Enterprise Miner, Bash Scripting, Regular Expressions and SQL (Oracle & SQL Server).
• Packages and tools: Pandas, NumPy, SciPy, Scikit-Learn, NLTK, Spacy, matplotlib, Seaborn, BeautifulSoup, Logging, PySpark, Keras and TensorFLow.
• Machine Learning: Linear Regression, Logistic Regression, Multinomial logistic regression, Regularization (Lasso & Ridge), Decision trees, Support Vector Machines, Ensembles - Random Forest, Gradient Boosting, Xtreme Gradient Boosting(xGBM), Deep Learning - Neural Networks, Deep Neural Networks(CNN, RNN & LSTM) with Keras and Tensorflow, Dimensionality Reduction-

Principal Component Analysis(PCA), Weight of Evidence (WOE) and Information Value, Hierarchical & K-means clustering, K-Nearest Neighbors.
• Data Visualization: Tableau, Google Analytics, Advanced Microsoft Excel and Power BI.
• Big Data Tools: Spark/PySpark, HIVE, IMPALA, HUE, Map Reduce, HDFS, Sqoop, Flume and Oozie
• Text Mining: Text Pre-Processing, Information Retrieval, Classification, Topic Modeling, Text Clustering, Sentiment Analysis and Word2Vec.
• Cloud Technologies: Google Cloud Platform Big Data & Machine Learning modules- Cloud Storage, Cloud DataFlow, Cloud ML, BigQuery, Cloud Dataproc, Cloud Datastore, BigTable. Familiarity on AWS - EMR, EC2, S3.
• Version Control: Git