Data Analyst/ Python Developer Data Analyst/Python Developer Data Analyst/ Python Developer - Capital Group * Highly efficient Data Scientist/Data Analyst with 6+ years of experience in Data Analysis, Machine Learning, Data mining with large data sets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Data Visualization, Web Scraping. Adept in statistical programming languages like R and Python including Big Data technologies like Hadoop, Hive. * Proficient in managing entire data science project life cycle and actively involved in all the phases of project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical modeling (decision trees, regression models,clustering), dimensionality reduction using Principal Component Analysis and Factor Analysis, testing and validation using ROC plot, K- fold cross-validation and data visualization. * Adept and deep understanding of Statistical modeling, Multivariate Analysis, model testing, problem analysis, model comparison, and validation. * Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data. * Skilled in performing data parsing, data manipulation and data preparation with methods including describe data contents, compute descriptive statistics of data, regex, split and combine, Remap, merge, subset, reindex, melt and reshape. * Experience in using various packages in R and python-like ggplot2, caret, dplyr, Rweka, gmodels, twitter, NLP, Reshape2, rjson, plyr, pandas, NumPy, Seaborn, SciPy, Matplotlib, sci-kit-learn, Beautiful Soup. * Extensive experience in Text Analytics, generating data visualizations using R, Python and creating dashboards using tools like Tableau. * Hands on experience with big data tools like Hadoop, Spark, Hive, Pig, PySpark, Spark SQL,PySpark * Hands on experience in implementing LDA, Naive Bayes and skilled in Random Forests, Decision Trees, Linear and Logistic Regression, SVM, Clustering, neural networks, Principle Component Analysis. * Good Knowledge in Proof of Concepts (PoC's), gap analysis and gathered necessary data for analysis from different sources, prepared data for data exploration using data munging. * Good industry knowledge, analytical &problem-solving skills and ability to work well within a team as well as an individual. * Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data. * Experience in designing stunning visualizations using Tableau software and publishing and presenting dashboards, Storyline on web and desktop platforms. * Experience and Technical proficiency in Designing, Data Modeling Online Applications, Solution Lead for Architecting Data Warehouse/Business Intelligence Applications. * Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales, PivotTables and OLAP reporting. * Highly skilled in using Hadoop (pig and Hive) for basic analysis and extraction of data in the infrastructure to provide data summarization. * Highly skilled in using visualization tools like Tableau, ggplot2 ,dash, flask for creating dashboards. * Worked and extracted data from various database sources like Oracle, SQL Server, DB2, regularly accessing JIRA tool and other internal issue trackers for the Project development. * Highly creative, innovative, committed, intellectually curious, business savvy with good communication and interpersonal skills. * Extensive experience in Data Visualization including producing tables, graphs, listings using various procedures and tools such as Tableau. Work Experience Data Analyst/ Python Developer Capital Group - Los Angeles, CA February 2018 to Present Capital Group is an American financial services company. Capital offers a range of products, more than 40 mutual funds through its subsidiary, American Funds, as well as separately managed accounts , private equity, investment services for high net worth investors in the U.S., and a range of other offerings for institutional clients and individual investors globally. Worked with a Fixed Income front office client to build a model for the Investment managers, Portfolio Managers, Traders to help make better investment decisions. Created an aggregated report daily for the client to make investment decisions and help analyze market trends. Built an internal visualization platform for the clients to view historic data, make comparisons between various issuers, analytics for different bonds and market. The model collects, merges daily data from market providers and applies different cleaning techniques to eliminate bad data points. The model merges the daily data with the historical data and applies various quantitative algorithms to check the best fit for the day. Captures the changes for each market to create a daily email alert to the client to help make better investment decisions. Built the model on Azure platform using Python and Spark for the model development and Dash by plotly for visualizations. Built REST APIs to easily add new analytics or issuers into the model. Automate different workflows, which are initiated manually with Python scripts and Unix shell scripting. Create, activate and program in Anaconda environment. Worked on predictive analytics use-cases using Python language. Clean data and processed third party spending data into maneuverable deliverables within specific format with Excel macros and python libraries such as NumPy, SQLAlchemy and matplotlib. Used Pandas as API to put the data as time series and tabular format for manipulation and retrieval of data. Helped with the migration from the old server to Jira database (Matching Fields) with Python scripts for transferring and verifying the information. Analyze Format data using Machine Learning algorithm by Python Scikit-Learn. Experience in python, Jupyter, Scientific computing stack (numpy, scipy, pandasand matplotlib). Perform troubleshooting, fixed and deployed many Python bug fixes of the two main applications that were a main source of data for both customers and internal customer service team. Write Python scripts to parse JSON documents and load the data in database. Generating various capacity planning reports (graphical) using Python packages like Numpy, matplotlib. Analyzing various logs that are been generating and predicting/forecasting next occurrence of event with various Python libraries. Created Autosys batch processes to fully automate the model to pick the latest as well as the best bond that fits best for that market. Created a framework using plotly, dash and flask for visualizing the trends and understanding patterns for each market using the history data. Used python APIs for extracting daily data from multiple vendors. Used Spark and SparkSQL for data integrations, manipulations.Worked on a POC for creating a docker image on azure to run the model Environment: Python, Pyspark, Spark SQL, Plotly, Dash, Flask, Post Man Microsoft Azure, Autosys, Docker Data Scientist/Data Analyst (Python) Anthem - Richmond, VA March 2017 to January 2018 Anthem Inc. is an American health insurance company founded in the 1940s, prior to 2014 known as WellPoint, Inc. It is the largest for-profit managed health care company in the Blue Cross and Blue Shield Association. Data Lake Creation Data is collected from different sources as flat files, and exported to cloud storage using AWS and Hadoop technologies. Data analyzed, reported and presented using Tableau dashboards. Worked on creation of Orchestration engine that was controlling the complete flow of data ingestion using Sqoop pipelines dumping data into Hive and DQM layers were enabled. Data processed in Hive and end results were reported using Tableau dashboards. Customer Segmentation Developed 11 customer segments using unsupervised learning techniques like KMeans and Gaussian mixture models. The clusters helped business simplify complex patterns to manageable set of 11 patterns that helped set strategic and tactical objectives pertaining to customer retention, acquisition, spend and loyalty. Responsibilities * Implemented Data Exploration to analyze patterns and to select features using Python SciPy. * Built Factor Analysis and Cluster Analysis models using Python SciPy to classify customers into different target groups. * Designed an A/B experiment for testing the business performance of the new recommendation system. * Supported MapReduce Programs running on the cluster. * Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs. * Participated in Data Acquisition with Data Engineer team to extract historical and real-time data by using Hadoop MapReduce and HDFS. * Communicated and presented default customers profiles along with reports using Python and Tableau, analytical results and strategic implications to senior management for strategic decision making * Developed scripts in Python to automate the customer query addressable system using python which decreased the time for solving the query of the customer by 45% * Collaborated with other functional teams across the Risk and Non-Risk groups to use standard methodologies and ensure a positive customer experience throughout the customer journey. * Performed Data Enrichment jobs to deal missing value, to normalize data, and to select features. * Developed multiple MapReduce jobs in java for data cleaning and pre-processing. * Analyzed the partitioned and bucketed data and compute various metrics for reporting. * Extracted data from Twitter using Java and Twitter API. Parsed JSON formatted twitter data and uploaded to database. * Developed Hive queries for analysis, and exported the result set from Hive to MySQL using Sqoop after processing the data. * Created HBase tables to store various data formats of data coming from different portfolios. * Worked on improving performance of existing Pig and Hive Queries. * Created reports and dashboards, by using D3.js and Tableau 9.x, to explain and communicate data insights, significant features, models scores and performance of new recommendation system to both technical and business teams. * Utilize SQL, Excel and several Marketing/Web Analytics tools (Google Analytics, Bing Ads, AdWords, AdSense, Criteo, Smartly, SurveyMonkey, and Mailchimp) in order to complete business & marketing analysis and assessment. * Used Git 2.x for version control with Data Engineer team and Data Scientists colleagues. * Used Agile methodology and SCRUM process for project developing. * KT with the client to understand their various Data Management systems and understanding the data. * Creating meta-data and data dictionary for the future data use/ data refresh of the same client. * Structuring the Data Marts to store and organize the customer's data. * Running SQL scripts, creating indexes, stored procedures for data analysis * Data Lineage methodology for data mapping and maintaining data quality. * Prepared Scripts in Python and Shell for Automation of administration tasks. * Maintained PL/SQL objects like packages, triggers, procedures etc. * Mapping flow of trade cycle data from source to target and documenting the same. * Performing QA on the data extracted, transformed and exported to excel. * Participated in all phases of data mining; data collection, data cleaning, developing models, validation, visualization and performed Gap analysis. * Extracted data from HDFS and prepared data for exploratory analysis using data munging * Built models using Statistical techniques like Bayesian HMM and Machine Learning classification models like XG Boost, SVM, and Random Forest. * A highly immersive Data Science program involving Data Manipulation & Visualization, Web Scraping, Machine Learning, Python programming, SQL, GIT, Unix Commands, NoSQL, MongoDB, Hadoop. * Used pandas, numpy, seaborn, scipy, matplotlib, scikit-learn, NLTK in Python for developing various machine learning algorithms. * Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python. Environment: ER Studio 9.7, MDM, GIT, Unix, Python (SciPy, NumPy, Pandas, StatsModel, Plotly), MySQL, Excel, Google Cloud Platform, Tableau 9.x, D3.js, SVM, Random Forests, Na ve Bayes Classifier, A/B experiment, Git 2.x, Agile/SCRUM., MLLib, SAS, regression, logistic regression, Hadoop, NoSQL, Teradata, OLTP, random forest, OLAP, HDFS, ODS, NLTK, SVM, JSON, XML, MapReduce Data Scientist Yum

Brands - Louisville, KY March 2016 to February 2017 Yum! Brands, Inc., or Yum! is an American fast food company. A Fortune 500 corporation, Yum! operates the brands Taco Bell, KFC, Pizza Hut, and WingStreet worldwide. Based in Louisville, Kentucky, it is one of the world's largest fast food restaurant companies in terms of system units - with 43,617 restaurants around the world in over 135 countries and territories. Project Description: The project involves data extraction and applying data integrity and analytical techniques for story telling from the data. The major key performance indicators such as In store behavior, price optimization and distribution and logistics optimization and improved order handling times etc. using supervised and unsupervised machine learning techniques. Responsibilities * Applied Lean Six Sigma process improvement in plant and developed Capacity Calculation systems using purchase order tracking system and improvement inbound efficiency by 23.56%. * Worked with Machine learning algorithms like Linear Regressions (linear, logistic etc.) SVMs, Decision trees for classification of groups and analyzing most significant variables such as FTE, Waiting times of purchase orders and Capacities available and applied process improvement techniques. * And calculated Process Cycle efficiency of 33.2% and identified value added and non-value added activities * And utilized SAS for developing Pareto Chart for identifying highly impacting categories in modules to find the work force distribution and created various data visualization charts. * Performed univariate, bivariate and multivariate analysis of approx. 4890 tuples using bar charts, box plots and histograms. * Participated in features engineering such as feature creating, feature scaling and One-Hot encoding with Scikit-learn. * Converted raw data to processed data by merging, finding outliers, errors, trends, missing values and distributions in the data. * Generated detailed report after validating the graphs using R, and adjusting the variables to fit the model. * Worked on Clustering and factor analysis for classification of data using machine learning algorithms. * Developed Descriptive statistics and inferential statistics for Logistics optimization, Average hours per job, Value throughput data to at 95% confidence interval. * Written MapReduce code to process and parsing the data from various sources and storing parsed data into HBase and Hive using HBase - Hive Integration. * Created SQL tables with referential integrity and developed advanced queries using stored procedures and functions using SQL server management studio. * Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis. * Used packages like dplyr, tidyr & ggplot2 in R Studio for data visualization and generated scatter plot and high low graph to identify relation between different variables. * Worked on Business forecasting, segmentation analysis and Data mining and prepared management reports defining the problem; documenting the analysis and recommending courses of action to determine the best outcomes. * Worked on various Statistical models like DOE, hypothesis testing, Survey testing and queuing theory. * Experience with risk analysis, root cause analysis, cluster analysis, correlation and optimization and K-means algorithm for clustering data into groups. * Coordinate with data scientists and senior technical staff to identify client's needs and document assumptions. Environment: SQL Server 2012, Jupyter, R 3.1.2, Python, MATLAB, SSRS, SSIS, SSAS, MongoDB, HBase, HDFS, Hive, Pig, Microsoft office, SQL Server Management Studio, Business Intelligence Development Studio, MS Access. Data Analyst JP Morgan - New York, NY February 2014 to December 2015 Credit Card Fraud The purpose of this project was to fight against credit card fraud. My team mainly focused on rebuilding credit card fraud detection model, monitoring the model in production, taking action if model performance degrades and working closely with business team to onboard new model. Monthly Dashboard Project is to develop monthly dashboard that will help the business to see the usage of the cards (Credit & Debit) on monthly basis with respect to their card categories like Signature, Platinum, Gold and Silver etc. This dashboard also provides the activation status of newly issued cards. With this dashboard business can easily track the history over latest one year. Responsibilities * Built scalable and deployable machine learning models. * Utilized Sqoop to ingest real-time data. Used analytics libraries Sci-Kit Learn, MLLIB and MLxtend. * Extensively used Python's multiple data science packages like Pandas, NumPy, matplotlib, Seaborn, SciPy, Scikit-learn and NLTK. * Performed Exploratory Data Analysis, trying to find trends and clusters. * Built models using techniques like Regression, Tree based ensemble methods, Time Series forecasting, KNN, Clustering and Isolation Forest methods. * Worked on data that was a combination of unstructured and structured data from multiple sources and automated the cleaning using Python scripts. * Extensively performed large data read/writes to and from csv and excel files using pandas. * Tasked with maintaining RDD's using SparkSQL. * Communicated and coordinated with other departments to collection business requirement. * Tackled highly imbalanced Fraud dataset using undersampling with ensemble methods, oversampling and cost sensitive algorithms. * Improved fraud prediction performance by using random forest and gradient boosting for feature selection with Python Scikit-learn. * Implemented machine learning model (logistic regression, XGboost) with Python Scikit- learn. * Optimized algorithm with stochastic gradient descent algorithm Fine-tuned the algorithm parameter with manual tuning and automated tuning such as Bayesian Optimization. * Developed a technical brief based on the business brief. This contains detailed steps and stages of developing and delivering the project including timelines. * After sign-off from the client on technical brief, started developing the SAS codes. * Wrote the data validation SAS codes with the help of Univariate, Frequency procedures. * Summarising the data at customer level by joining the datasets of customer transaction, dimension and from 3rd party sources. * Separately calculated the KPIs for Target and Mass campaigns at pre-promo-post periods with respective to their transactions, spend and visits. * Also measured the KPIs at MoM (Month on Month), QoQ (Quarter on Quarter) and YoY (Year on Year) with respect to pre-promo-post. * Measured the ROI based on the differences pre-promo-post KPIs. * Extensively used SAS procedures like IMPORT, EXPORT, SORT, FREQ, MEANS, FORMAT, APPEND, UNIVARIATE, DATASETS and REPORT. * Standardised the data with the help of PROC STANDARD. * Implemented cluster analysis (PROC CLUSTER and PROC FASTCLUS) iteratively. * Worked extensively with data governance team to maintain data models, Metadata and dictionaries. * Used Python to preprocess data and attempt to find insights. * Iteratively rebuild models dealing with changes in data and refining them over time. * Created and published multiple dashboards and reports using Tableau server. * Extensively used SQL queries for legacy data retrieval jobs. * Tasked with migrating the django database from MySQL to PostgreSQL. * Gained expertise in Data Visualization using matplotlib, Bokeh and Plotly. *

Responsible for maintaining and analyzing large datasets used to analyze risk by domain experts. * Developed Hive queries that compared new incoming data against historic data. Built tables in Hive to store large volumes of data. * Used big data tools Spark (Sparksql, Mllib) to conduct the real time analysis of credit card fraud based on AWS * Performed Data audit, QA of SAS code/projects and sense check of results. Accomplishments: * Accomplished 75% reduction in cycle time for automation of gathering and reporting of performance issues in Hadoop applications used by the company for ETL in Enterprise Data Management, reducing team effort resulting in a net savings of $80,000 per annum for the company's budget. * Won the "Harbinger" award for Value Engineering from the clients for lead the Standards and Compliance team in overseeing programming standards, debugging and authorized migration to production environment 15+ largescale data management software applications in Hadoop ecosystem. * Built Hadoop based data warehouse, streamlined data ingestion, distributed data storage - data lake (HDFS), standardized and provided scalable data processing to monetize data effectively for a Bank of America. * Migrated raw data from Mainframes and extracted it to HDFS and Hive using sqoop to for preprocessing and structuring. * Collaborated with team directly interacting with clients in large scale warehousing of sensitive data using Hadoop ecosystem (UNIX scripting, Map Reduce and Hive) and Extraction Transformation Loading. Environmen:t Spark, Hadoop, AWS, SAS Enterprise Guide, SAS/MACROS, SAS/ACCESS, SAS/STAT, SAS/SQL, ORACLE, MS-OFFICE, Python (scikit-learn, pandas, Numpy), Machine Learning (logistic regression, XGboost), Gradient Descent algorithm, Bayesian optimization, Tableau. Data Scientist Travelers Insurance - Pune, Maharashtra August 2012 to August 2013 The Travelers Companies is an American insurance company. It is the second largest writer of U.S. commercial property casualty insurance and the third largest writer of U.S. personal insurance through independent agents. Project: Predicting Customer Churn Project Description: To predict the attrition in personal insurance products. The Churn prediction model predicts a customer's propensity to churn by using information about the customer such as household and financial data, transactional data, and behavioral data. The inputs for the Churn prediction model are customer demographic data, insurance policies, premiums, tenure, claims, complaints, and the sentiment score from past surveys. Responsibilities * Aggregate all available information about the customer. The data that is obtained for predicting the churn is classified in the following categories. * Demographic data, such as age, gender, education, marital status, employment status, income, home ownership status, and retirement plan. * Policy-related data, such as insurance lines, number of policies in the household, household tenure, premium, disposable income, and insured cars. * Claims, such as claim settlement duration, number of claims that are filed and denied. * Complaints, such as number of open and closed complaints. * Survey sentiment data. Sentiment scores from past surveys are captured in the latest and average note attitude score fields. The note attitude score is derived from customer negative feedback only. If the note attitude is zero, the customer is more satisfied while as the number increases, satisfaction level decreases. * Responsible for building data analysis infrastructure to collect, analyze, and visualize data. * Data elements validation using exploratory data analysis (univariate, bivariate, multivariate analysis). * Missing value treatment, outlier capping and anomalies treatment using statistical methods. * Variable selection was done by making use of R-square and VIF values. * Deployed Machine Learning (Logistic Regression and PCA) to predict customer churn. Environment: Statistical tools: R - 3.3.0, Python 3.0, SQL Server, MS-Excel, MS-PowerPoint. Data Analyst Travelers Insurance - Pune, Maharashtra June 2011 to August 2012 Intern) Responsibilities * Developed and implemented predictive models using Natural Language Processing Techniques and machinelearning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, RandomForests, K-means clustering, KNN, PCA and regularization for data analysis. * Designed and developed Natural Language Processing models for sentiment analysis. * 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 relationaltools like SQL, No SQL. * Built and analyzed datasets using R, SAS, Matlab and Python (in decreasing order of usage). * Participated in all phases of datamining; datacollection, datacleaning, developingmodels, validation, visualization and performed Gapanalysis. * DataManipulation and Aggregation from different source using Nexus, Toad, BusinessObjects, PowerBI and SmartView. * Implemented Agile Methodology for building an internal application. * Good knowledge of HadoopArchitecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Secondary Name Node, and MapReduce concepts. * As Architect delivered various complex OLAPdatabases/cubes, scorecards, dashboards and reports. * Programmed a utility in Python that used multiple packages (scipy, numpy, pandas). * Implemented Classification using supervised algorithms like LogisticRegression, Decisiontrees, KNN, NaiveBayes. * Used Teradata15 utilities such as FastExport, MLOAD for handling various tasks data migration/ETL from OLTP Source Systems to OLAP Target Systems * Maintenance in the testing team for System testing/Integration/UAT. * Involved in preparation & design of technical documents like Bus Matrix Document, PPDM Model, and LDM & PDM. * Understanding the client business problems and analyzing the data by using appropriate Statistical models to generate insights.   Environment: R 3.0, Erwin 9.5, Tableau 8.0, MDM, QlikView, ML Lib, PL/SQL, HDFS, Teradata 14.1, JSON, HADOOP (HDFS), MapReduce, PIG, Spark, R Studio, MAHOUT, JAVA, HIVE, AWS. Education Bachelor of Technology in Information Technology Jawaharlal Nehru Technological University - Hyderabad, Telangana Skills APPLICATION DEVELOPMENT, Hadoop, HBase, HDFS, Hive, MapReduce, Pig, PYTHON, FLASK, GGPLOT2, NUMPY, REPORTING TOOLS, VISIO, XML, JDBC, MS ACCESS, SQL, CASSANDRA, IMPALA, MAPREDUCE Additional Information SKILLS   Languages Python, R, Java 8 Packages ggplot2, caret, dplyr, Rweka, gmodels, RCurl, C50, twitter, NLP, Reshape2, rjson, plyr, pandas, numPy, Seaborn, sciPy, matplot lib, sci-kit-learn, Beautiful Soup, Rpy2.   Web Technologies JDBC, HTML5, DHTML and XML, CSS3, Web Services, WSDL Data Modelling Tools Erwin r 9.6, 9.5, 9.1, 8.x, Rational Rose, ER/Studio, MS Visio, SAP Power designer Big Data Technologies Hadoop, Hive, HDFS, MapReduce, Pig, Databases SQL, Hive, Impala, Pig, Spark SQL, Databases SQL-Server, My SQL, MS Access, HDFS, HBase, Teradata, Netezza, MongoDB, Cassandra. Reporting Tools MS Office (Word/Excel/Power Point/ Visio), Tableau, Crystal reports XI, Business Intelligence, SSRS, Business Objects 5.x/ 6.x, Cognos7.0/6.0, Flask, Dash   ETL Tools Informatica Power Centre, SSIS. Version Control Tools SVM, GitHub Project Execution

Methodologies Ralph Kimball and Bill Inmon data warehousing methodology, Rational Unified Process (RUP), Rapid Application Development (RAD), Joint Application Development (JAD). BI Tools Tableau, Tableau Server, Tableau Reader, SAP Business Objects, OBIEE, QlikView, SAP Business Intelligence, Amazon Redshift, or Azure Data Warehouse Operating System Windows, Linux, Unix, Macintosh HD, Red Hat

Name: Justin Knox

Email: robert28@example.org

Phone: 663.954.3915x8959