Data Scientist / Big Data Data Scientist / Big Data Data Scientist / Big Data - Cigna Health Insurance Bloomfield, CT Professional qualified Data Scientist/Data Analyst with over 9 years of experience in Data Science and Analytics including Artificial Intelligence/Deep Learning/Machine Learning, Data Mining and Statistical Analysis Involved in the entire data science project life cycle and actively involved in all the phases including data extraction, data cleaning, statistical modeling and data visualization with large data sets of structured and unstructured data, created ER diagrams and schema. Experienced with machine learning algorithm such as logistic regression, random forest, XGboost, KNN, SVM, neural network, linear regression, lasso regression and k-means Implemented Bagging and Boosting to enhance the model performance. Strong skills in statistical methodologies such as A/B test, experiment design, hypothesis test, ANOVA Experience in implementing data analysis with various analytic tools, such as Anaconda 4.0 Jupiter Notebook 4.X, R 3.0 (ggplot2, Caret, dplyr) and Excel [ ] Solid ability to write and optimize diverse SQL queries, working knowledge of RDBMS like SQL Server 2008, NoSql databases like MongoDB 3.2 Developed API libraries and coded business logic using C#, XML and designed web pages using .NET framework, C#, Python, Django, HTML, AJAX Strong experience for over 5 years in Image Recognition and Big Data technologies like Spark 1.6, Sparksql, pySpark, Hadoop 2.X, HDFS, Hive 1.X Experience in visualization tools like, Tableau 9.X, 10.X for creating dashboards Excellent understanding Agile and Scrum development methodology Used the version control tools like Git 2.X and build tools like Apache Maven/Ant Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making Ability to maintain a fun, casual, professional and productive team atmosphere Experienced the full software life cycle in SDLC, Agile, Devops and Scrum methodologies including creating requirements, test plans. Skilled in Advanced Regression Modeling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools and application of Statistical Concepts. Proficient in Predictive Modeling, Data Mining Methods, Factor Analysis, ANOVA, Hypothetical testing, normal distribution and other advanced statistical and econometric techniques. Developed predictive models using Decision Tree, Random Forest, Na ve Bayes, Logistic Regression, Social

Network Analysis, Cluster Analysis, and Neural Networks. Experienced in Machine Learning and Statistical Analysis with Python Scikit-Learn. Experienced in Python to manipulate data for data loading and extraction and worked with python libraries like Matplotlib, Numpy, Scipy and Pandas for data analysis. Worked with complex applications such as R, Python, Theano, H20, SAS, Matlab and SPSS to develop neural network, cluster analysis. Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing data mining and reporting solutions that scales across massive volume of structured and unstructured data. Skilled in performing data parsing, data ingestion, data manipulation,data architecture, data modelling 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. Strong C#, SQL programming skills, with experience in working with functions, packages and triggers. Extensively worked on Python 3.5/2.7 (Numpy, Pandas, Matplotlib, NLTK and Scikit-learn) Experienced in Visual Basic for Applications and VB programming languages C#, .NET framework to work with developing applications. Worked with NoSQL Database including Hbase, Cassandra and MongoDB. Experienced in Big Data with Hadoop, HDFS, MapReduce, and Spark. Experienced in Data Integration Validation and Data Quality controls for ETL process and Data Warehousing using MS Visual Studio SSIS, SSAS, SSRS. Proficient in Tableau,Adobe Analytics and R-Shiny data visualization tools to analyze and obtain insights into large datasets, create visually powerful and actionable interactive reports and dashboards. Automated recurring reports using SQL and Python and visualized them on BI platform like Tableau. Worked in development environment like Git and VM. Excellent communication skills. Successfully working in fast-paced multitasking environment both independently and in collaborative team, a self-motivated enthusiastic learner. TOOLS AND TECHNOLOGIES: Languages Java 8, Python, R, C#, Powershell Packages ggplot2, caret, dplyr, Rweka, gmodels, Edward, RCurl, tm, C50, twitteR, NLP, Reshape2, rjson, plyr, pandas, numPy, TensorFlow, seaborn, sciPy, matplot lib, scikit-learn, Beautiful Soup, Rpy2. Web Technologies JDBC, HTML5, DHTML and XML, CSS3, Web Services, WSDL, Django, AWS 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, Presto, MapReduce, Pig, Kafka, oozie. Databases SQL, Hive, Impala, Pig, Spark SQL, HQL, VQL, Databases SQL-Server, My SQL, MS SQL,MS Access, HDFS, HBase, Teradata, Netezza, MongoDB, Cassandra. Reporting Tools MS Office (Word/Excel/Power Point/ Visio), Tableau, Spotfire, Crystal reports XI, Business Intelligence, SSRS, Business Objects 5.x/ 6.x, Cognos7.0/6.0. ETL Tools Informatica Power Centre, SSIS. Version Control Tools SVM, GitHub. Project Execution Work Experience Data Scientist / Big Data Cigna Health Insurance - Bloomfield, CT August 2017 to Present Description: The Cigna Information Management & Analytics (CIMA) unit offers solutions that provide actionable insights to internal and external business partners and customers that help reduce health costs, improve outcomes, provide financial security and measure and forecast business performance Responsibilities: Several visualizations (density plots, forest plots, leverage plots, network plots, covariant adjustment plots etc.) were made using packages such as GGPLOT2, GGMCMC. Successfully delivered multiple NLP projects like building a chatbot that assists a customer to trouble shoot claim issues and recommend actions.Further the bot could handle questions asked in natural language related to common issues with the customer e.g. when is my premium due, what is my plan deductible, what is my copay for a sick reject. Extracted data from multiple sources like Medicare, Medicaid, ACA claims. Performed data pre-processing like data cleaning, text preprocessing, noise removal, lexicon normalization and object standardization. Perform featuring engineering like Word Embedding using word2vec models. Build seq2seqmodels using structured data & word embedding. Seq2Seq model take an input and returns as desired output for e.g. it can take a question as an input and returns an answer. The benefit is it can take any arbitrary length question and returns and answers in natural language. It uses a recurrent neural network (LSTM/Memory Network) at the back-end. Performing Map Reduce jobs in Hadoop and implemented Spark analysis using Python for performing machine learning & predictive analytics on AWS platform. Analyzed administrative claims data - Medicare and ACA Marketplace - to answer health services research questions on costs, utilization or outcomes, using advanced statistical and econometric methods. Used Tensorflow packages to train machine learning models. Developed

Oozie workflows to ingest/parse the raw data, populate staging tables and store the refined data in partitioned tables in the Hive. Hand-on experience with data ingestion into Big Data platform from disparate data sources using Sqoop, Hive, Pig, Flume and Spark. Created an End-to-End data analytical solutions and models by manipulating large data sets and integrating diverse data sources. Worked with team of developers to design, develop and implement BI solutions in Tableau to measure Point of Sale KPIs at micro and macro level. Environment: Tableau, Python, PyCharm, Statistics, Machine Learning, Tensorflow, Alteryx, Hadoop, Hive, Pig, No SQL, PL/SQL, Excel, AWS Data Scientist Opera Solutions, New Jersey January 2016 to July 2017 Description: Opera Solutions, LLC is a technology and analytics company mainly focused on big data. The firm uses a combination of machine learning science, advanced predictive analytics, technology, large-scale data management, and human expertise. Opera Solutions delivers predictive analytics as a service, and offers hosted, cloud-based systems for specific business problems, e.g., predicting the behavior of individual consumers, stopping revenue leakage in hospitals, warning of threats to corporate security or brand health, etc. Responsibilities: Performed Data Profiling to learn about behavior with various features of USMLE examinations of various student patterns using Tableau, Adobe Analytics and Python Matplotlib. Evaluated models using Cross Validation, Log loss function, ROC curves and used AUC for feature selection and elastic technologies like ElasticSearch, Kibana etc Addressed overfitting by implementing of the algorithm regularization methods like L2 and L1. Implemented statistical modeling with XGBoost machine learning software package using Python to determine the predicted probabilities of each model. Created master data for modelling by combining various tables and derived fields from client data and students LORs, essays and various performance metrics. Formulated a basis for variable selection and GridSearch, KFold for optimal hyperparameters Utilized Boosting algorithms to build a model for predictive analysis of student's behaviour who took USMLE exam apply for residency. Used numpy, scipy, pandas, nltk(Natural Language Processing Toolkit),matplotlib to build the model. Formulated several graphs to show the performance of the students by demographics and their mean score in different USMLE exams. Extracted data from HDFS using Hive, Presto and performed data analysis using Spark with Scala, pySpark, Redshift and feature selection and created nonparametric models in Spark Application of various Artificial Intelligence(AI)/machine learning algorithms and statistical modeling like decision trees,text analytics, Image and Text Recognition using OCR tools like Abbyy, natural language processing(NLP), supervised and unsupervised, regression models. Used Principal Component Analysis in feature engineering to analyze high dimensional data. Performed Data Cleaning, features scaling, features engineering using pandas and numpy packages in python and build models using deep learning frameworks. Created deep learning models using Tensorflow and keras by combining all tests as a single normalized score and predict residency attainment of students. Used XGB classifier if the feature is an categorical variable and XGB regressor for continuous variables and combined it using FeatureUnion and FunctionTransfomer methods of Natural Language Processing. Used OnevsRest Classifier to fit each classifier against all other classifiers and used it on multiclass classification problems. Implemented application of various machine learning algorithms and statistical modeling like Decision Tree, Text Analytics, Sentiment Analysis, Naive Bayes, Logistic Regression and Linear Regression using Python to determine the accuracy rate of each model. Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behavior with cloud based products like Azure ML Studio and Dataiku. Generated various models by using different machine learning and deep learning frameworks and tuned the best performance model using Signal Hub and AWS Sagemaker/Azure Databricks. Created data layers as signals to Signal Hub to predict new unseen data with performance not less than the static model build using deep learning framework. Environment: Python 2.x,3.x, Hive, AWS, Linux, Tableau Desktop, Microsoft Excel, NLP, Deep learning frameworks such as TensorFLow, Keras, Boosting algorithms etc Data Scientist Mercedes Benz Financial Services, Michigan July 2014 to December 2015 Description: Mercedes-Benz Financial Services is a leading, captive financial services provider and the global financial services company of Daimler AG. Doing business as Mercedes-Benz Financial Services and Daimler Truck Financial, we provide financing for automotive and commercial vehicle dealers and their retail consumers in the United States, Canada, Mexico, Brazil and Argentina. Responsibilities: Performed Data Profiling to learn about behavior with various features such as traffic pattern, location, time, Date and Time etc using Adode Analytics. Application of various Artificial Intelligence(AI)/machine learning algorithms and statistical modeling like decision trees,text analytics, natural language processing(NLP), supervised and unsupervised, regression models, social network analysis, neural networks, deep learning, SVM, clustering to identify Volume using scikit-learn package in python, Matlab. Utilized Spark, Snowflake, Presto, Scala, Hadoop, HQL, VQL, oozie, pySpark, Data Lake, TensorFlow, HBase, Cassandra, Athena, Redshift, MongoDB, Kafka, Kinesis, Spark Streaming, Edward, CUDA, MLLib, AWS, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc. and Utilized the engine to increase user lifetime by 45% and triple user conversations for target categories. Created and connected SQL engine through C# to connect database, developed API libraries and business logic using C#, XML and Python Exploring DAG's, their dependencies and logs using AirFlow pipelines for automation Performed data cleaning and feature selection using MLlib package in PySpark and working with deep learning frameworks such as Caffe, Neon etc Developed Spark/Scala, Python,R for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources. Used clustering technique K-Means to identify outliers and to classify unlabeled data. Utilized AWS Lambda in created user-friendly interface for quick view of reports by using C#, JSP, XML and developed expandable menu that show drilldown data on graph click Evaluated models using Cross Validation, Log loss function, ROC curves and used AUC for feature selection and elastic technologies like ElasticSearch, Kibana etc Categorised comments into positive and negative clusters from different social networking sites using Sentiment Analysis and Text Analytics and have done Image Recognition Tracking operations using sensors until certain criteria is met using AirFlow technology. Responsible for different Data mapping activities from Source systems to Teradata using utilities like TPump, FEXP, MLOAD, BTEQ, FLOAD etc Analyze traffic patterns by calculating autocorrelation with different time lags. Ensured that the model has low False Positive Rate and Text classification and sentiment analysis for unstructured and semi-structured data.

Addressed overfitting by implementing of the algorithm regularization methods like L2 and L1. Used Principal Component Analysis in feature engineering to analyze high dimensional data. Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behavior. Performed Multinomial Logistic Regression, Random forest, Decision Tree, SVM to classify package is going to deliver on time for the new route. Performed data analysis by using Hive to retrieve the data from Hadoop cluster, Sql to retrieve data from Oracle database and used ETL for data transformation. Used MLlib, Spark's Machine learning library to build and evaluate different models and used AWS Rekognition for image analysis. Implemented rule based expertise system from the results of exploratory analysis and information gathered from the people from different departments. Performed Data Cleaning, features scaling, features engineering using pandas and numpy packages in python and build models using SAP Predictive Analytics. Developed MapReduce pipeline for feature extraction using Hive and Pig. Created Data Quality Scripts using SQL and Hive to validate successful data load and quality of the data. Created various types of data visualizations using Python and Tableau/Spotfire. Communicated the results with operations team for taking best decisions. Collected data needs and requirements by Interacting with the other departments. Environment: Python 2.x, CDH5, HDFS, C#, Hadoop 2.3, Hive, Impala, AWS, Linux, Spark, Tableau Desktop, SQL Server 2012, Microsoft Excel, Matlab, Spark SQL, Pyspark. Data Scientist First Data - Atlanta, GA January 2013 to June 2014 Description: First Data Corporation is a global payment processing company headquartered in Atlanta, Georgia, United States. The company's portfolio includes merchant transaction processing services; credit, debit, private-label, gift, payroll and other prepaid card offerings; fraud protection and authentication solutions. Responsibilities: Provided Configuration Management and Build support for more than 5 different applications, built and deployed to the production and lower environments. Implemented public segmentation using unsupervised machine learning algorithms by implementing k-means algorithm using Pyspark. Using AirFlow to keep track of job statuses in repositories like MySQl and Postgre databases. Explored and Extracted data from source XML in HDFS, used ETL for preparing data for exploratory analysis using data munging. Responsible for different Data mapping activities from Source systems to Teradata, Text mining and building models using topic analysis, sentiment analysis for both semi-structured and unstructured data. Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS Used R and python for Exploratory Data Analysis, A/B testing, HQL, VQL, Data Lake, AWS Redshift, oozie, pySpark, Anova test and Hypothesis test to compare and identify the effectiveness of Creative Campaigns. Computing A/B testing frameworks, clickstream and time spent databases using Airflow Created clusters to Control and test groups and conducted group campaigns using Text Analytics. Created positive and negative clusters from merchant's transaction using Sentiment Analysis to test the authenticity of transactions and resolve any chargebacks. Analyzed and calculated the lifetime cost of everyone in the welfare system using 20 years of historical data. Created and developed classes and web page elements using C# and AJAX. JSP was used for validating client side responses and connected C# to database to retrieve SQL data Developed LINUXShell scripts by using NZSQL/NZLOAD utilities to load data from flat files to Netezza database. Developed triggers, stored procedures, functions and packages using cursors and ref cursor concepts associated with the project using Pl/SQL Created various types of data visualizations using R, C#, python and Tableau/Spotfire also connected Pipeline Pilot with Spotfire to create more interactive business driven layouts. Used Python, R, SQL, Tensorflow to create Statistical algorithms involving Multivariate Regression, Linear Regression, Logistic Regression, PCA, Image Recognition, Random forest models, Decision trees, Support Vector Machine for estimating the risks of welfare dependency. Identified and targeted welfare high-risk groups with Machine learning/deep learning algorithms. Conducted campaigns and run real-time trials to determine what works fast and track the impact of different initiatives. Developed Tableau visualizations and dashboards using Tableau Desktop. Used Graphical Entity-Relationship Diagramming to create new database design via easy to use, graphical interface. Created multiple custom SQL queries in Teradata SQL Workbench to prepare the right data sets for Tableau dashboards Perform analyses such as regression analysis, logistic regression, discriminant analysis, cluster analysis using SAS programming. Environment: R 3.x, HDFS, C#,Hadoop 2.3, Pig, Hive, Linux, R-Studio, Tableau 10, SQL Server, Ms Excel, Pypark. Data Scientist TripAdvisor - New York, NY November 2011 to December 2012 Description: TripAdvisor, Inc. is an American travel website company providing reviews of travel-related content. It also includes interactive travel forums. TripAdvisor was an early adopter of user-generated content. The website services are free to users, who provide most of the content, and the website is supported by an advertising business model. Responsibilities: Involved in Design, Development and Support phases of Software Development Life Cycle (SDLC) Performed data ETL by collecting, exporting, merging and massaging data from multiple sources and platforms including SSRS/SSIS (SQL Server Integration Services) in SQL Server. Programming experience with .NET framework, C#, Visual Studio 2005/2008 to build web based, client/server architecture and to produce reports with C# and JSP. Worked with cross-functional teams (including data engineer team) to extract data and rapidly execute from MongoDB through MongDB connector for Hadoop. Performed data cleaning and feature selection using MLlib package in PySpark. Performed partitional clustering into 100 by k-means clustering using Scikit-learn package in Python where similar hotels for a search are grouped together and Image Recognition. Used Python to perform ANOVA test to analyze the differences among hotel clusters. Implemented application of various machine learning algorithms and statistical modeling like Decision Tree, Text Analytics, Sentiment Analysis, Naive Bayes, Logistic Regression and Linear Regression using Python to determine the accuracy rate of each model. Determined the most accurately prediction model based on the accuracy rate. Used text-mining process of reviews to determine customers' concentrations. Delivered analysis support to hotel recommendation and providing an online A/B test. Designed Tableau bar graphs, scattered plots, and geographical maps to create detailed level summary reports and dashboards. Developed hybrid model to improve the accuracy rate. Environment: Python, PySpark, C#, Tableau, MongoDB, Hadoop, SQL Server, SDLC, ETL, SSIS, recommendation systems, Machine Learning Algorithms, text-mining process, A/B test Data Scientist Bank of America - Wilmington, DE October 2010 to October 2011 Description: Bank of America is a multinational banking and financial services corporation. It is ranked 2nd on the list of largest banks in the United States by assets. As of 2016, Bank of America was the 26th largest company in the United States by total revenue. Responsibilities: Participated in all phases of research including data collection, data cleaning, data mining, developing models and visualizations. Collaborated with data engineers and operation team to collect data from internal system to fit the analytical requirements. Redefined many attributes and relationships and cleansed unwanted tables/columns using SQL queries. Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries and also used C# connector to perform SQL queries by creating and connecting to SQL engine. Performed data imputation using Scikit-learn package in Python. Performed data processing using Python libraries like Numpy and Pandas. Worked with data analysis using ggplot2 library in R to do data visualizations for better understanding of customers' behaviors. Implemented statistical modeling with XGBoost machine learning software package using R to determine the predicted probabilities of each model. Delivered the results with operation team for better decisions. Environment: Python, R, SQL, Tableau, Spark, Machine Learning Software Package, recommendation systems. Python Developer Cenvien Technologies - Hyderabad, Telangana June 2009 to August 2010 Description: Cenvien technologies gather the requirements by listening and understanding to the client's business requirement to deliver quality products. It is highly qualified and strongly dedicated developing team that produces unique solutions. Responsibilities: Developed entire frontend and backend modules using Python on Django Web Framework. Implemented the presentation layer with HTML, CSS and JavaScript. Involved in writing stored procedures using Oracle. Optimized the database queries to improve the performance. Designed and developed data management system using Oracle. Environment: MySQL, ORACLE, HTML5, CSS3, JavaScript, Shell, Linux & Windows, Django, Python Programmer Analyst Pennar Industries Limited - Hyderabad, Telangana April 2008 to May 2009 Description: As a backend developer of web applications and data science infrastructure. The main area of focus is to come up with comprehensive solutions that need massive capacity and throughput. Responsibilities: Effectively communicated with the stakeholders to gather requirements for different projects Used MySQL db package and Python-MySQL connector for writing and executing several MYSQL database queries from Python.

Implemented Client/Server applications using C++, C#, JSP and SQL Created functions, triggers, views and stored procedures using My SQL. Worked closely with back-end developer to find ways to push the limits of existing Web technology. Involved in the code review meetings. Environment: Python, MySQL, C#. Education Bachelor of Computer Science in Computer Science JNTU Anantapur - Anantapur, Andhra Pradesh Skills APPLICATION DEVELOPMENT, TABLEAU (8 years), LINUX (5 years), SAP (1 year), BI (1 year), Python, R, Hadoop, Machine Learning, Spark Additional Information 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,Android.

Name: Kenneth Wilson

Email: frodgers@example.org

Phone: 368.360.0709x4002