Data Scientist Data Scientist Data Scientist - Zoetis Inc Kalamazoo, MI Efficient Data Scientist with around 6 years of experience in, Statistical Modeling, Machine Learning, Data Mining with Large Data Sets of Structured and Unstructured Data and Performed Data Acquisition, Data Validation, Predictive Modeling and Data Visualization. Significant industry experience and domain knowledge in Healthcare, Retail, Banking, Energy and got some domain knowledge in Telecom industries. Experience in feature extraction, creating Regression models, Classification, Predictive data modeling, and Cluster analysis. Expertise in Python (2.x/3.x) programming with multiple packages including NumPy, Pandas, Matplotlib, SciPy, Seaborn and Scikit-learn. hands-on experience with all Python libraries for Data Acquisition, Data Cleaning, Data Validation, Predictive modeling, and Data Visualization tools. Experience in designing stunning visualizations using Tableau software and publishing and presenting dashboards, Storyline on web and desktop platforms. Worked on technologies like slack, Git, SVN and Openpyxl for reading and writing. Strong business judgment and ability to take ambiguous problems and solve them in a structured, hypothesis-driven, and data-supported way. Hands on experience in implementing LDA, Na ve Bayes and skilled in Random Forests, Decision Trees, Linear and Logistic Regression, SVM, Clustering, Neural Networks, Principle Component Analysis and good knowledge on Recommender Systems. Implementation experiences in Machine Learning and deep learning, including Regression, Classification, Neural network, object tracking, Natural Language Processing (NLP) using packages like Tensor Flow, Keras, NLTK, Spacy. Highly skilled in advanced Regression Modelling, Time Series Analysis, Correlation and Multivariate Analysis. Experienced in Machine Learning Classification Algorithms like Logistic Regression, K-NN, SVM, Kernel SVM, Naive Bayes, and Decision Tree. Experience in tuning algorithms using methods such as Grid Search, Randomized Search, K-Fold Cross Validation and Error Analysis. Worked with outlier analysis with various methods like Z-Score value analysis, Liner regression, Dbscan (Density Based Spatial Clustering of Applications with Noise) and Isolation forest Worked on Gradient Boosting decision trees with XGBoostto improve performance and accuracy in solving problems. Also worked with several boosting methodologies like ADA Boost, Gradient Boosting and XGBoost. Implemented

various statistical tests like ANOVA, A/B testing, Z-Test, T-Test for various business cases. Validated the machine learning classifiers using Accuracy, AUC, ROCCurves and Lift Charts. Worked on Artificial Neural Networks and Deep Learning models using Theano and Keras packages using Python. Implemented and analyzed RNN based approaches for automatically predicting implicit relations in text. The disclosure relation has potential applications in NLP tasks like Text Parsing, Text Analytics, Text Summarization, Conversational systems. Worked with various text analytics or Word Embedding libraries like Word2Vec, Count Vectorizer, GloVe, LDA etc. Solid knowledge and experience in Deep Learning techniques including Feed forward Neural Network, Convolutional Neural Network (CNN), Recursive Neural Network (RNN) Worked with numerous data visualization tools in python like Matplotlib, Seaborn, ggplot, pygal. Worked and extracted data from various database sources like Oracle, SQL Server, DB2, MongoDB and Teradata. Highly skilled in using Hadoop, HBase, Spark, and Hive for basic analysis and extraction of data in the infrastructure to provide data summarization. Good knowledge of Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Secondary Name Node, MapReduce concepts, and ecosystems including Hive and Pig. Handled importing data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS. Experience working with MS Word, MS Excel, MS PowerPoint, MS SharePoint, and MS Project. Work Experience Data Scientist Zoetis Inc - Kalamazoo, MI September 2018 to Present Zoetis Inc. is the world's largest producer of medicine and vaccinations for pets and livestock. Zoetis delivers quality medicines, vaccines and diagnostic products, which are complemented by genetic tests, bio devices and a range of services. The project is to collect data from different sources and create a master data set. And we do predictions on sales and profits. Measures to be taken for improving the sales by applying machine learning strategies and statistical analyses to support animal health projects and products. Responsibilities: Developing data analytical databases from different sources and create a master data set. Responsible for data identification, collection, exploration, cleaning for modeling Data entry, data auditing, creating data reports and monitoring all data for accuracy We do predictions on sales and profits using machine

learning and deep learning strategies. Performed Time Series analysis on sales data to consider what measures to be taken for improve the Sales. Manage large data sets from a wide variety of sources and apply analytics and statistical analyses to support animal health projects and products. Analysis of biological and spatial data to develop insights into precision animal management and precision medicine Implementing analytics algorithms in Python, Reprogramming languages Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, to visualization of the data after removing missing and outliers to fit in the model Applied isolation forest, local outlier factor from Sklearn, where local filters are used unsupervised outlier detection and score each sample. Applied deep learning libraries (Tensor Flow, Theano, Torch, etc.) and scalable event stream processing architectures (e.g. Lambda, CEP, etc.) Performed training Natural Language models and reinforcement learning engines to optimize intelligent agents that automate task execution. Worked with dimensionality reduction techniques like PCA, LDA and ICA Performed k-Means clustering in order to understand customer itemized bought products and segment the customers based on the customer products for animal medicine and vaccines behavior information for customized product offering, customized and priority service, to improve existing profitable relationships and to avoid customer churn, etc using Python. Applying Clustering algorithms to group the data on their similar behavior patterns. Performed animal medicines and vaccine's sales Predictive Modelling by using Decision Trees and Regressions in order to get the risk involved by giving individual scores to the customers Work with data analytics team to develop time series and optimization. Performed Time Series Analysis on animal medicine and vaccine product sales datain order to extract meaningful statistics and other characteristics of the data to predict future values based on previously observed values. Used Expert level understanding of different databases in combinations for Data extraction and loading, joiningdata extracted from different databases and loading to a specific database in SQL Performed Advanced SQL queries for script executions like Update, Insert, and Delete. Worked on Hadoop ecosystem components like HadoopMapReduce, HDFS, HBase, Hive, Sqoop, Pig including their installation and configuration. Used Hive to store the data and perform data cleaning steps for huge datasets. Used self- service

environment Cloudera Data Science Workbench (CDSW) to manage the data analytics pipelines, including built-in scheduling, monitoring, and email alerting Created various Proof of Concepts (PoC) and gap analysis and gathered necessary data for analysis from different sources, prepared data for data exploration using data munging. Implemented Agile Methodology for building an internal application Used Tableau to generate reports with internal records, secondary sources of data, JSON, CSV and more.Which helped the support team for better marketing. Data Scientist Mars Solutions Group, WI March 2017 to August 2018 The Client is the largest Healthcare Company and offers health care products, insurance services, Data Analytics, Payment Integrity, and The project was to build predictive models for customer value analysis by applying machine learning methods, principal component analysis, and regression on large dataset. Responsibilities:

Creating statistical machine learning models for implementing Customer Churn, Ticket routing techniques, invoice premium predictions and claim classification. Collaborated with other departments to collect and understand client business requirements. Collaborated with Data Engineers to gathered business requirements and filtered the data according to project requirements. Worked in importing and cleansing of data from various sources like Teradata, Oracle, flat files, SQLServer 2005 with high volume data Performed feature engineering including feature intersection generating, feature normalize and labelencoding with Scikit-learn preprocessing.

Congregated data from multiple sources and performed resampling to handle the issue of imbalanced data. Treated missing values and outliers with several techniques Boxplots, Z-Score and DB Scan Explored and visualized the data to check the pattern, distribution, descriptive statistic, and correlation using Python, Matplotlib, and Seaborn. Performed NLP tasks with NLP library CoreNLP, NLTK and Gensim. Performed text representation techniques (such as n-grams, bag of words, sentiment analysis etc.), Installed HDFS storage and data analysis tools in Amazon Web Services (AWS) cloud environment computing infrastructure. Developed ETL processes for data conversions and construction of data warehouse using INFORMATICA. Updated Python scripts to match training data with our database stored in AWS Cloud Search, so that we would be able to assign each document a response label for further classification. Import/export data from

Teradata database to HDFS using Sqoop. Performed data analysis by using Hive to retrieve the data from Hadoop cluster, Sql to retrieve data from Oracle database Creating data pipelines using big data tools like Hadoop, spark etc. Good knowledge on Hadoop components such as HDFS, Job Tracker, TaskTracker, Name Node, Data Node, and Map Reduce concepts. Responsible for managing and reviewing Hadoop Log files. Created bucketing and partitions in HIVE to handle the data. Applied different dimensionality reduction techniques like principle component analysis (PCA) and t-stochastic neighborhood embedding (t-SNE) on feature matrix Worked with various customer analytics such as segmenting the customers, Product Recommendations and NLP Tasks Worked with Clustering algorithms like K-Means, K-Means++, DBSCAN and Agglomerative Hierarchical Clustering to target specific group of customers to generate profitable revenue. Using NLP to sorting the email to automatically updating the records in Customer Relationship management (CRM) We can run natural language processing algorithms against the data and automatically extract the features or risk factors from the notes in the medical record. Performed Multinomial Logistic Regression, Random forest, Decision Tree, SVM and more machine learning algorithms Using graphical packages produced ROC Curve to visually represent True Positive Rate versus FalsePositive Rate. Equally produced visualization of Precision Recall Curve for Area under the Curve. Used Market Basket Analysis, association rules analysis to identified patterns, data quality issues and leveraged insights Addressed over fitting by implementing of the algorithm regularization methods like L2 and L1 Improved model's accuracy by using Gradient Boosting technique like Light GBM and gained around 82% accuracy with Random Forest and 77% with Logistic Regression. Used K-fold cross validation technique to increase the model performance and worked with hyper parameter tuning methods like Grid Search. Worked with visualization tools like Tableau, Cognos and Micro Strategy to create business reports for higher management and used Python visualization libraries like Seaborn, Matplotlib and ggplot depending on business requirements. Provided schedules, status reports, and issue resolutions to the Project team, Business Users, and Project Managers Data Analyst / Data Scientist CMS Energy - Jackson, MI January 2016 to February 2017 CMS Energy is an energy company that is focused principally on

utility operations. I was responsible for building a new data science department with the help of other departments and I was able to learn how the business is operated and helped the company to grow and stay ahead of the competition. By using machine learning we improvised the predictive algorithm for pricing strategy. And we creating alerts that would notify customers of potential issues that their system has solely based on the data available to me. Responsibilities: Worked on Data Manipulation & Visualization, Machine Learning, Python, SQL, NoSQL, MongoDB, Hadoop Performed Advanced SQL queries for script executions like Update, Insert, and Delete. Used Expert level understanding of different databases in combinations for Data extraction and loading, joining data extracted from different databases and loading to a specific database in SQL Programmed utilities in Python that uses packages like SciPy, NumPy, pandas, stats model, scikit learn, XG boost, matplotlib, plotly, NLTK, seaborn, bokeh. Transformed the 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. Have done Normalization& Denormalization techniques for optimum performance in relational and dimensional database environments. Worked on customer segmentation using an unsupervised learning technique - clustering. Implemented Classification using supervised learninglike Logistic Regression, Decision trees, KNN, Naive Bayes. Built models using Statistical techniques and Machine Learning classification models like XG Boost, SVM, and Random Forest. Created various Proof of Concepts (PoC) and gap analysis and gathered necessary data for analysis from different sources, prepared data for data exploration using data munging. Used jupyter notebook for spark to make data manipulations Developed ETL processes for data conversions and construction of data warehouse using INFORMATICA. Worked on Hadoop ecosystem components like Hadoop MapReduce, HDFS, HBase, Hive, Sqoop and Pig including their installation and configuration. Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS Used Hive to store the data and perform data cleaning steps for huge datasets. Extracted data from source XML in HDFS, preparing data for exploratory analysis using data munging Interacted with the other departments to understand and identify data

needs and requirements and work with other members of the IT organization to deliver data visualization and reporting solutions to address those needs Used visualization tools like Tableau for the interactive graphs Used python libraries Matplotlib and Seaborn for creating dashboards Data Analyst Karvy Financial Services Limited November 2014 to December 2015 Karvy Financial Services Limited is a company which has been playing a very proactive role in the economic growth of India by providing loans to Micro & Small Business segments and individuals like credit for the requirements of different sectors of economy. Industries, exports, trading, agriculture, infrastructure and the individual segments. We worked on various projects which handle customer analytics; Credit Risk analysis and assessing risks associated with loans like identify and prevent fraudulent loans, identify and prevent fraud detection for transactions. Responsibilities: Compiled data from various sources public and private databases to perform complex analysis and data manipulation for actionable results Applied concepts of probability, distribution, and statistical inference on the given dataset to unearth interesting findings using comparison, T-test, F-test, R-squared, P-value etc Applied linear regression, multiple regressions, ordinary least square method, mean-variance, the theory of large numbers, logistic regression, dummy variable, residuals, Poisson distribution, Naive Bayes, fitting function etc to data with help of Scikit, SciPy, NumPy and Pandas module of Python. Applied Principal Component Analysis (PCA) based unsupervised technique to determine unusual VPN log-on time. Performed Clustering with historical, demographic and behavioral data as features to implement the personalized marketing to the customers Also created classification model using Logistic Regression, Random Forests to classify dependent variable into two classes which are risky and okay Used F-Score, Precision, recall evaluating model performance Built user behavior models for finding activity patterns and evaluating risk scores for every transaction using historic data to train the supervised learning models such as Decision trees, Random Forests and SVM Real time analysis of customer's financial profile and providing recommendation for financial products best suited. Collected historical data and third-party data from different data sources and performed data integration using Alteryx. Forecasted demand for loans and interest rates using Time Series analysis like ARIMAX,

VARMAX and Holt-Winters. Obtained better predictive performance of 81% accuracy using ensemble methods like Bootstrap aggregation (Bagging) and Boosting (Light GBM, Gradient) Tested complex ETL mappings and sessions based on business user requirements and business rules to load data from source flat files and RDBMS tables to target tables. Developed visualizations and dashboards using ggplot, Tableau Prepared and presented data quality report to stakeholders to give understanding of data. Python Developer / Data Analyst Symbiosys Technologies - Visakhapatnam, Andhra Pradesh January 2014 to October 2014 Genius Brands International is our client and we performed exploratory data analysis on corporate purchase orders, contracts and projects data using sampling and statistical methods. Identified strata, improved precision and accuracy. Works with other team members, including DBA's, Other ETL developers, Technical Architects, QA, and Business Analysts & Project Managers Responsibilities: The work will involve the development of workflows triggered by events from other systems. Develop easy to use documentation for the frameworks and tools developed for adaption by other teams. Applied k-means and hierarchical clustering on the data. Identified and analyzed business insights. Developed Hive UDFs and Pig UDFs using Python in Microsoft HDInsight environment Implemented end-to-end systems for Data Analytics, Data Automation and customized visualization tools using Python, R, Hadoop and MongoDB. Used pandas, NumPy, seaborn, SciPy, matplotlib, scikit-learn in Python for developing various machine learning algorithms. Performed data profiling to merge the data from multiple data sources Worked on csv, json, excel different types of files for the data cleaning and data analysis. Used Python for statistical operations on the data and ggplot2 for the visualizing the data. Participated in feature engineering such as feature intersection generating for adding potential powerful features, plotting feature correlation matrix for feature selection and reducing, feature normalization for ease to implement machine algorithms, Principal Component Analysis (PCA) for dimensionality reduction and label encoding with Scikit-learn preprocessing. Worked with several use cases like campaign sales analysis, forecasting sales, KPI analysis and NLP models Worked with Clustering algorithms to target specific group of customers to generate profitable revenue Worked with word embedding techniques like

Word2Vec, GloVe for sentiment analysis and text classifications Worked with text to vector representation methods including Counter Vectorizer, Tf-idf and Latent Dirichlet Allocation (LDA) for topic modeling Performed time series analysis using Tableau Developed and executed Ad hoc reporting's according to the business needs. Managed offshore projects and coordinated work for 24 hour productivity cycle. ETL Developer Sutherland Global Services - Hyderabad, Telangana February 2013 to December 2013 Sutherland builds processes for the digital age by combining the speed and insight of design thinking with the scale and accuracy of data analytics. Sutherland has customers across industries like financial services to Healthcare. My role is to assist Analytics department for the data extraction and cleaning as a data preprocessing steps to build models. Responsibilities: Involved with Business Analysts team in requirements gathering and in preparing functional specifications and changing them into technical specifications Involved in Data mapping specifications to create and execute detailed system test plans. The data mapping specifies what data will be extracted from an internal data warehouse, transformed and sent to an external entity. Managed full SDLC processes involving requirements management, workflow analysis, source data analysis, data mapping, metadata management, data quality, testing strategy and maintenance of the model Involved in extensive DATA validation by writing several complex SQL queries and Involved in back-end testing and worked with data quality issues. Designed SSIS packages to extract, transform and load existing data into SQL Server, used lots of components of SSIS, such as Pivot Transformation, Fuzzy Lookup, Merge, Merge Join, Data Conversion, Row Count, Sort, Derived Columns, Conditional Split, Execute SQL Task, Data Flow Task and Execute Package Task. Created SSIS Packages that involved dealing with different source formats (flat files, Excel, XML, OLE DB) and different destination formats Debugged and troubleshot the ETL packages by using a breakpoint, analyzing the process, catching error information by SQL command in SSIS Developed SQL queries in SQL Server management studio, Toad and generated complex reports forth end users. Automated and scheduled recurring reporting processes using UNIX shell scripting and Teradata utilities such as MLOAD, BTEQ, and Fast Load Experience with Perl. Performed data analysis and data profiling using complex SQL on various sources systems

including Oracle and Teradata Education Master of Science in Information Technology Management Campbellsville University Bachelor of Technology in Electronics and Communication Engineering Jawaharlal Nehru Technological University - Kakinada, Andhra Pradesh

Name: Scott Gillespie

Email: robinporter@example.org

Phone: 001-868-483-1090x93366