

Data Scientist Data Scientist Data Scientist - VF Corporation Reading, PA Data Scientist with 7+ years of professional experience in the E-commerce, Retail and Music Streaming domain, performing Statistical Modelling, Data Extraction, Data screening, Data cleaning, Data Exploration and Data Visualization of structured and unstructured datasets as well as implementing large scale Machine Learning and Deep Learning algorithms to deliver resourceful insights, inferences and significantly impacted business revenues and user experience. Experienced in facilitating the entire life-cycle of a data science project: Data Extraction, Data Pre-Processing, FeatureEngineering, Dimensionality Reduction, Algorithm implementation and Validation. Proficient in Data transformations using log, square-root, reciprocal, differencing and complete box-cox transformation depending upon the dataset. Knowledge of normality tests like Shapiro-Wilk, Anderson-Darling. Adept at analysis of Missing data by exploring correlations and similarities, introducing dummy variables for missingness, and choosing from imputation methods such iterative imputer on Python. Experienced in Machine Learning techniques such as regression and classification models like Linear, Polynomial, Support Vector Machines, Decision Trees, Logistic Regression. Experienced in Ensemble learning using Bagging, Boosting & Random Forests; clustering like K-means. In-depth Knowledge of Dimensionality Reduction (PCA, LDA), Hyper-parameter tuning, Model Regularization (Ridge, Lasso, Elastic Net) and Grid Search techniques to optimize model performance. Adept with Python and OOP concepts such as Inheritance, Polymorphism, Abstraction, Association, etc. Hands on experience in Artificial Neural Networks, Deep Learning, Convolution Neural Networks. Expertise in creating executive Tableau Dashboards for Data visualization and deploying it to the servers; Skilled in using Tidyverse framework in R and Pandas in Python for performing exploratory data analysis. Proficient in Data Visualization tools such as Tableau and PowerBI, Big Data tools such as Hadoop HDFS, Spark and MapReduce and Microsoft Excel (VLOOKUP, Pivot tables). Experience in Web Data Mining with Python's ScraPy and BeautifulSoup packages along with working knowledge of Natural Language Processing (NLP) to analyze text patterns. Experience with Python libraries including NumPy, Pandas, SciPy, Scikit-Learn&statsmodels, Matplotlib, Seaborn, NLTK and R libraries like ggplot2,

dplyr. Work Experience Data Scientist VF Corporation - Greensboro, NC June 2018 to Present

Description: VF Corporation is an American worldwide apparel and footwear company. Worked on the recommender system by implementing Sentiment Analysis on other people reviews and extract the best product for the user to buy. Responsibilities: Reviewed business requirements to analyze the data sources and worked closely with the business analysts to understand business objectives. Extracted data by web-scraping through the reviews using Beautiful Soup. Involved in various pre-processing phases of text-data like Tokenization, Stemming, Lemmatization and converting the raw text data to structured data. Performed data collection, data cleaning, feature scaling, feature engineering, validation, visualization, report findings, develop strategic uses of data by Python libraries like NumPy, Pandas, Scipy, Matplotlib, Scikit-Learn. Used Tableau for visualizing and analyzing the data to facilitate the understanding of the team about the data. Implemented various statistical techniques to manipulate the data like missing data imputation, Principal Component Analysis for dimension-reduction. Constructed new vocabulary to convert the data into numbers to be processed by the machine by using the approaches like Bag of Words model, tf-idf, Word2Vec. Employed statistical methodologies such as A/B test, experiment design and hypothesis testing and deployed models on Docker. Performed Na ve Bayes, KNN, Logistic Regression, Random Forest, SVM and KMeans to categorize customers into certain groups. Employed various metrics such as Cross-Validation, LogLoss function, Confusion Matrix, ROCAUC to evaluate the performance of each model. Using NLP developed deep learning algorithms for analyzing text, over the existing dictionary-based approaches. Environment: Python(NumPy, Pandas, Matplotlib), TensorFlow, NLP Data Scientist SiriusXM - Washington, DC April 2017 to May 2018 Description: SiriusXM is a broadcasting company that provides satellite radio and online radio. The project required to build a recommender system by extracting and analyzing different features from raw audio files which took new songs into account. Responsibilities: Performed Data Collection, Data Cleaning, Data Visualization using Python, Deep Feature Synthesis and extracted key statistical findings to develop business strategies. Since sound is represented in the form of audio signals, parameters like frequency, decibel, timbre, pitch were usedfor analysis. Used

Librosa (Python library) to analyze the audio signals and plot wave plot and create a spectrogram to analyze the behavior for the sound. Cleaned the audio files to remove the audio with no noise by setting a threshold and retrieve the audio above the set threshold. Created 2-D Convolution Neural Networks using Keras on GPU's by extracting different number of MFCC features using Librosa. Used global-temporal pooling layer to effectively compute statistics of learned features across time. Implemented regularization methods like Dropout, Lasso Regression and Ridge Regression to prevent the model from overfitting. Final model was selected by evaluating them using various metrics like Accuracy, ConfusionMatrix, Precision, Recall. Environment: Python (Pandas, Scikit, Numpy), TensorFlow, Keras, Librosa.

Data Scientist Swiggy - IN November 2014 to January 2017

Description: Swiggy is online food delivery company in India. The project was predicting deals and coupons for frequent customers of the company.

Responsibilities:

Participated in all phases of project life cycle including data collection, data mining, data cleaning, developing models, validation and creating reports. Performed data cleaning on a huge dataset which had missing data and extreme outliers from Hadoop workbooks and explored data to draw relationships and correlations between variables. Performed data-preprocessing on messy data including imputation, normalization, scaling, and feature engineering using Scikit-Learn. Conducted exploratory data analysis using Python Matplotlib and Seaborn to identify underlying patterns and correlations between features. Build classification models based on Logistic Regression, Decision Trees, Support Vector Machine to predict the probability of a customer using the application. Employed Ensemble Learning techniques such as Random Forests and Ada Gradient Boosting to improve the model performance by 10%. Used various metrics such as F-Score, ROC and AUC to evaluate the performance of each model and 5-Fold Cross Validation to test the models with different batches of data to optimize the models. Implemented and tested the model on AWS EC2 and collaborated with development team to get the best algorithms and parameters. Prepared data-visualization designed dashboards with Tableau, and generated complex reports including summaries and graphs to interpret the findings to the team.

Environment: Python(NumPy, Pandas, Matplotlib), Amazon Web Services, Jupyter Notebook, Tableau Data

Analyst - Python Developer BigBasket - IN July 2012 to October 2014 Description: Bigbasket is the Indian online grocery delivery service. My responsibilities included working on RestFul Web Services on Python Flask and working on a team building a predictive model to enhance the online shopping for the users. Responsibilities: Worked on both legacy data and new data mostly built around the user experience and grocery inventory available. Performed Data Analysis on target data after transfer to Data Warehouse. Created ETL solution using MS SQL Server and worked with Agile and Test-Driven development within SDLC. Worked on RESTful Web Services on Python Flask and built primary functions for classification. Conducted data preparation and outlier detection using Python and implemented Logistic Regression, Random Forest, Na ve Bayes Classifier for classification for recommendation. Employed K-Fold Cross-validation to test and verify the model accuracy. Worked with the team to host data and certain web interfaces on Amazon Web Services EC2 and store data on S3 bucket. Worked with Team manager to develop a lucrative system of classifying auditions and vendors best fitting for the company in the long run. Presented executive dashboards and scorecards to visualize and present trends in the data using Excel and Python (Matplotlib). Environment: Python(NumPy, Pandas, Matplotlib), Amazon Web Services, Python Flask, REST APIs, Linux Education Bachelor's Skills Amazon web services, Hadoop, Hdfs, Mapreduce, Python, Ggplot2, Matplotlib, Anova, Mapreduce, Kafka, Data visualization, Hadoop, Mongodb, Database, Microsoft sql server, Sql server, Mysql, Oracle, Postgresql, Sql Additional Information SKILLS Languages Python, R, Matlab, SQL Database MySQL, PostgreSQL, Oracle, MongoDB, Microsoft SQL Server Statistical Tests Hypothesis Testing, ANOVA tests, t-tests, Chi-Square Fit test. Validation Techniques k-fold cross validation, Out of the Box Estimates, A/B Tests. Optimization Techniques Gradient Descent, Stochastic Gradient Descent, Mini-Batch Gradient Descent, Gradient Optimization - Adam, Momentum, RMSProp Data Visualization Tableau, Microsoft PowerBI, ggplot2, MatplotLib, Seaborn Big Data Apache Hadoop, HDFS, Kafka, MapReduce, Spark Cloud Technologies Amazon Web Services Tools and Software PyCharm, XCode, Jupyter Notebook, Microsoft SQL, Linux, Unix, Microsoft Office

Name: Mary Gonzalez

Email: gillespiehannah@example.net

Phone: (504)876-1935x103