**JACK ZHANG**

DATA SICENTIST

Email id: [jz134068@gmail.com](mailto:jz134068@gmail.com)

Contact number: 661-247-1236

**Summary:**

* A Passionate, team-oriented Data Scientist with over 6 years of experience in **Data Extraction, Data Modelling, Statistical Modeling**, **Data Mining, Machine Learning** and **Data Visualization**.
* Expertise in transforming business resources and tasks into **regularized data** and **analytical models**, **designing algorithms**, **developing data mining** and **reporting solutions** across a massive volume of **structured and unstructured data**.
* Involved in entire data science project life cycle, including **Data Acquisition**, **Data Cleansing**, **Data Manipulation**, **Feature Engineering, Modelling**, **Evaluation**, **Optimization**, **Testing** and **Deployment**.
* Experienced in **Natural Language Processing (NLP), Text Mining, Topic Modelling, Sentiment Analysis, Association Rules Analysis** and **Market Basket Analysis.**
* Proficient at **Machine Learning algorithms** and **Predictive Modeling** including **Linear Regression**, **Logistic Regression**, **Naive Bayes**, **Decision Tree**, **Neural Networks**, **Random Forest**, **Ensemble Models**, **SVM**, **KNN** and **K-means clustering**.
* Solid knowledge and experience in Deep Learning techniques including **Feedforward Neural Network,** **Convolutional Neural Network (CNN), Recursive Neural Network (RNN)**, **pooling**, **regularization.**
* Excellent proficiency in model validation and optimization with **Model selection**, **Parameter/Hyper-Parameter tuning**, **K-fold cross validation**, **Hypothesis Testing**, **Principle Component Analysis (PCA)**.
* Proficient with Python 3.x including **Numpy**, **Scikit-learn, NLP, Pandas**, **Matplotlib** and **Seaborn**.
* Extensive experience in **RDBMS** such as **SQL server 2012**, **Oracle 9i/10g** and non-relational database such as **MongoDB 3.x**.
* Hand on experience on **Hadoop 2.x ecosystem** and **Apache Spark 2.x framework** such as **Hive**, **Pig**, **and PySpark**.
* Proficient at data visualization tools such as **Tableau, R ggplot, Python Matplotlib** and **Seaborn**.
* Experienced in **Amazon Web Services (AWS)**, such as AWS EC2, EMR, S3, RD3, and Redshift.
* Experienced designing and developing T-SQL queries, ETL packages and business reports using **SQL Server Management Studio (SSMS)** and **BI Suite (SSIS/SSRS)**.
* Adept in developing and debugging **Stored Procedures**, **User-defined Functions (UDFs)**, **Triggers**, **Indexes**, **Constraints**, Transactions and Queries using **Transact-SQL (T-SQL)**.
* Knowledge and experience working in **Waterfall** as well as **Agile** environments including the **Scrum process** and using **Project Management** tools like **ProjectLibre, Jira/Confluence** and version control tools such as **Github.**

**TECHNICIAL SKILLSETS:**

|  |
| --- |
| * Statistical Methods: Hypothetical Testing, Exploratory Data Analysis (EDA), Confidence Intervals, ANOVA, Bayesian, Principal Component Analysis (PCA), Chi-square test, Correlation Analysis |
| * Machine Learning: Linear/Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine, K-Means Clustering, Adaptive Boosting, Gradient Boosting, Random Forests, Deep Learning |
| * Hadoop Ecosystem: Hadoop 2.x, Spark 2.x, MapReduce, Hive, Pig, HDFS |
| * Cloud Services: Amazon Web Services (AWS) EC2/S3/Redshift |
| * Deep Learning: Keras, Tensor Flow, Torch, Theano for Feedforward/Convolutional (CNN)/Recurrent (RNN) Neural Network |
| * Databases: MS SQL Server 2008/2008R2/2012/2014, Oracle, Amazon Redshift, MS SQL Server 2016/2014/2012/2008 R2/2008, MongoDB 3.x, Teradata |
| * Reporting Tools: Tableau Suite of Tools 10.x, Power BI, SQL Server Reporting Services (SSRS) |
| * Data Visualization: Tableau, MatPlotLib, Seaborn, ggplot2 |
| * Languages: Python (2.x/3.x), R, T-SQL, XML, C#, Java, PL/SQL |
| * Operating Systems: Microsoft Windows, Linux (Ubuntu), Mac OS |

**Professional Experience:**

**Inspira Health Network, Woodbury, NJ 01/2017 – Present**

**Data Scientist**

Inspira Health Network is the region’s leading network of health care providers, delivering the full continuum of primary, acute and advanced care services. The project was to analyze data from Inspira Health database and historical data from their legacy database in order to build solutions for increasing patient flow as well as improve efficiency of patient appointments. The project also involved designing different reports for business users and maintaining business intelligence solutions for different purposes.

**Responsibilities:**

* Gathered business requirements, definition and design of the data sourcing, worked in conjunction with the data warehouse architect on the development of logical data models.
* Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Conducted reverse engineering based on demo reports to understand the data without documentation and redefined the proper requirements and negotiated with our client.
* Generated different Data Marts for gathering the tables which were needed (Member info, Claim info, Transaction info, Appointment info, Diagnose info) from SQL Server Database and Teradata.
* Processed data using Python pandas to examine transaction data, identify outliers and inconsistencies and conducted exploratory data analysis using python NumPy and Seaborn to see the insights of data and validate each feature.
* Performed univariate and multivariate analysis on the data to identify any underlying pattern in the data and associations between the variables.
* Performed feature engineering including feature intersection generating, feature normalize and label encoding with Scikit-learn preprocessing.
* Used Python 3.X (numpy, SciPy, pandas, scikit-learn, seaborn) and Spark 2.0 (PySpark, MLlib) to develop variety of models and algorithms for analytic purposes.
* Developed and implemented predictive models using machine learning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, Random Forests, K-means clustering, KNN.
* Used Grid Search to evaluate each model and to find best hyper-parameters for each model.
* Designed and implemented a recommendation system which utilized Collaborative filtering techniques to recommend course for different customers and deployed to AWS EMR cluster.
* Designed rich data visualizations to model data into human-readable form with Tableau and Matplotlib.

**Environments:**

Python 3.3, Tableau, TensorFlow, Keras, AWS RedShift, EC2, EMR, Hadoop Framework, S3, HDFS, Spark (Pyspark, MLlib, Spark SQL), Python 3.x (Scikit-Learn/SciPy/Numpy/Pandas/Matplotlib/Seaborn, Agile/SCRUM, SQL Server 2012, Teradata 15.0, SQL Server Data Tools 2010, SQL Server Integration Services,

**Better Homes and Gardens Real Estate, Madison, NJ 11/2014– 12/2016**

**Data Scientist**

Better Homes and Gardens Real Estate is a dynamic real estate brand that offers a full range of services to brokers, sales associates and home buyers and sellers. Using innovative technology, sophisticated business systems and the broad appeal of a lifestyle brand, Better Homes and Gardens Real Estate embodies the future of the real estate industry while remaining grounded in the tradition of home. The project involved preprocessing, cleaning and analyzing the grand customer feedback data and finding key factors toward customer’s likelihood to recommend level.

**Responsibilities:**

* Gathered, analyzed, documented and translated business requirements into data models and applications.
* Acquired image dataset of products from different data sources and aggregated into one dataset on Amazon Redshift.
* Analyzed Net Promoter and Detractor customer group based on customer feedback dataset.
* Used market basket analysis, association rules analysis to identified patterns, data quality issues, and leveraged insights.
* Converted unstructured pure text consumer comments data to structured dataset using NLP techniques and feature engineering.
* Explored and visualized the data to get descriptive statistics and inferential statistics for better understanding the dataset.
* Built predictive models including support Vector Machine, Decision tree, Naive Bayes Classifier, Neural Network plus ensemble methods of the models to evaluate how the likelihood to recommend of customer groups would change in different set of service by using python scikit-learn.
* Implemented training process using cross-validation and test sets, evaluated the result based on different performance matrices and collected feedback and retrained the model to improve the performance.
* Conducted different tests like A/B Testing and conversion rate to evaluate our result.

**Environment:**

Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn), Tensorflow, Keras, AWS RedShift, EC2, EMR, Hadoop Framework, S3, HDFS, Spark (Pyspark, MLlib, Spark SQL), Agile/SCRUM

**Investors Bank, Iselin, NJ 07/2013- 11/2014**

**Data Analyst**

Investors Bank is a public traded, full-service bank that is based in Short Hills, New Jersey. The bank operates over 150 branches across New Jersey and New York. The Project was to analyze the data patterns and implement machine learning techniques to identify and recommend financial service and product to customers as well as to maintain operations of transactional databases and applications to support business intelligence platform and build reports.

**Responsibilities:**

* Collaborated with database engineers to implement ETL process, wrote and optimized SQL queries to perform data extraction and merging from SQL server database.
* Stored data from SQL Server database into Hadoop clusters which were set up in AWS EMR.
* Performed data integrity checks, data cleansing, exploratory data analysis, and feature engineer using Python and data visualization packages such as Matplotlib, Seaborn.
* Tackled highly imbalanced dataset using oversampling with SMOTE (Synthetic Minority Over-Sampling Technique) and cost sensitive algorithms with Python Scikit-learn.
* Implemented different machine learning algorithms in Python, including Logistic Regression, Support Vector Machine, and Random Forest classification.
* Evaluated parameters with K-Fold Cross Validation and sample free transaction data for test sets optimized the performance of models.
* Deployed the model on AWS Lambda, collaborated with develop team to build the business solutions.
* Conducted K-means clustering and Hierarchical clustering to segment customers into different groups and ranked the result based on customer lifetime value.
* Performed data visualization and Designed dashboards with Tableau, and generated complex reports, including charts, summaries, and graphs to interpret the findings to the team and stakeholders.
* Researched and designed large scale visual recommendation engine to give better recommendations for financial products.

**Environments:**

Python (Scikit-Learn/Scipy/Numpy/Pandas), Hadoop HDFS, AWS EC2, EMR, Lambda, Tableau, MS SQL Server 2012, Windows 8/XP, JIRA

**Lime Energy, Newark, NJ 04/2012 – 06/2013**

**BI/ETL Developer**

As a leading national provider of energy efficiency for small business customers, Lime Energy designs and implements direct install programs for our utility clients which consistently exceed program savings goals. The project was to implement data warehouse development and maintenance to support existing business processes and the establishment of customer data platform.

**Responsibilities:**

* Maintained and developed complex SQL queries, stored procedures, triggers, clustered index & non-clustered index, Views, and functions that meet user requirements using Microsoft SQL Server 2008.
* Created Views and Table-valued Functions, Common Table Expression (CTE), joins, and complex subqueries to provide the reporting solutions.
* Optimized the performance of queries with modification in T-SQL queries, removed the unnecessary columns and redundant data, normalized tables, established joins and created index.
* Designed report using T-SQL, SSMS and SSRS, illustrated by interactive dashboards, linked reports, sub-reports, parameterized reports and drill down, drill through reports.
* Refactored the database schema for margin matrix to display the report into an organized structure.
* Built SSIS ETL packages for processing fact and dimension tables with transforms and slowly changing dimensions to ensure the database platforms are configured, hardened and maintained to capture record changes.
* Used various SSIS transformations such as conditional split, derived column for data scrubbing, including data validation, checks during staging before loading the data into the data warehouse.
* Communicated with all lines of business and management, proactively identified cross functional or technical issues to improve application stability.

**Environment:**

SQL Server 2008 R2, SQL Server Management Studio (SSMS), MS BI Suite (SSIS/SSRS), SQL Server Data Tools (SSDT), Team Foundation Server (TFS), SharePoint, Visual Studio

**EDUCATION:**  
MS in Information Science and Studies, Syracuse University, NY