**Sam Young sam.young65@outlook.com**

**DATA SCIENTIST**

**SUMMARY**

* A pioneering Data Scientist with deep expertise and experience with statistical data analysis such as linear models, multivariate analysis, Statistical Analysis, Data Mining and Machine Learning Skills.
* High Proficiency in Advanced Excel and Access including complex data analysis and manipulation.
* Data Driven and highly analytical with working knowledge and statistical model approaches and methodologies (clustering, Segmentation, Variable reduction, Regression analysis, Hypothesis testing, Decision trees, Machine learning), rules and ever evolving regulatory environment.
* Experience with analyzing online user behavior, Conversion Data(A/Testing) and customer journeys, funnel analysis.
* Strong Data Analysis skills using business intelligence, SQL and / or MS Office Tools.
* Experience in Agile Methodologies.
* Experience with applied statistical techniques and machine learning, including Bayesian methods, time-series modeling, classification, regression, mixture models, clustering, dimensionality reduction, model selection, feature extraction, experimental design, and choice modeling.
* Strong understanding of how analytics supports a large organization including being able to successfully articulate the linkage between business objectives, analytical approaches &findings and business decisions.
* Professional working experience in Machine Learning algorithms such as LDA, linear regression, logistic regression, Naive Bayes, Decision Trees, Clustering, and Principle Component Analysis.
* Hands on experience building regression, classification, and recommender systems with large datasets in distributed systems and constrained environments
* Hands on experience in writing queries in SQL and R to extract, transform and load (ETL) data from large datasets using Data Staging.
* Highly skilled in using Hadoop (Pig and Hive) for basic analysis and extraction of data in the infrastructure to provide data summarization.
* Experience with traditional analytics tools (Excel, Tableau and QlikView).
* Developed and deployed dashboards in Tableau and RShiny to identify trends and opportunities, surface actionable insights, and help teams set goals, forecasts and prioritization of initiatives
* Working experienced of statistical analysis using R, SPSS, Matlab and Excel.
* Ability to work with large transactional databases across multiple platforms (Teradata, Oracle, HDFS, SAS).
* Deep understanding of Software Development Life Cycle (SDLC) as well as Agile/Scrum methodology to accelerate Software Development iteration.
* Good oral and written communication skills.
* Strong interpersonal skills to successfully build long-term relationships with colleagues and business partners.
* A results-driven individual with a passion for data/analytics who can work collaboratively with others to solve business problems that drive business growth.
* Demonstrated leadership and self-direction. Demonstrated willingness to both teach others and learn new techniques.
* Ability to work with managers and executives to understand the business objectives and deliver as per the business needs and a firm believer in team work.

***TECHNICAL SKILLS:***

|  |  |
| --- | --- |
| Programming & Scripting Languages | C, R, HTML, Python, Hadoop,XML,Hive SQL |
| Database | SQL, MySQL, MS Access, Oracle |
| Statistical Software | SPSS, R, SAS |
| Web Packages | Google Analytics, Adobe Test & Target, Web Trends |
| Development Tools | R Studio, Notepad++ |
| Visualization Tools | Tableau, Matplotlib |
| Packages | Dplyr, rjson, GGPLOT2 |
| Techniques | Machine learning, Regression, Clustering, Data mining. |
| Machine Learning | LDA, Naive Bayes, Decision trees, Regression models, Random forests, Time-series, k-means |
| Business Analysis | Requirements Engineering, Business Process Modeling & Improvement, Financial Modeling |
| Operating Systems | Microsoft windows 7/8/8.1/10/Vista/XP, Linux (Ubuntu) |

***PROFESSIONAL EXPERIENCE:***

***DATA SCIENTIST***

***SYNOVUS BANK-COLUMBUS, GA MAY 2018 TO PRESENT***

**Description:**

The aim of the project is to analyze individual customer behaviors and spending activities. Applying unsupervised and supervised machine learning techniques the team have been able to predict customer behaviors.

**Responsibilities:**

Analyzed individual customer behavior.  
● Segmented customers based on spending activities.  
● Categorized risky customers based on the days past due parameter.  
● Categorized active and inactive customers based on their utilization.  
● Designed, developed and deployed statistical data models R.  
● Utilized machine learning techniques for predictions & forecasting based on the training data.  
● Developed data mining algorithms using Machine learning (Random Forest, Regression, Clustering) for decision making using R, Mahout on Hadoop.  
● Partnered with ETL team to extract data from Hadoop environment.  
● Prepared Dashboards using calculations, parameters in QlikView.  
● Created and assisted users in Tableau dashboard development.

**Environment:** Machine Learning, R, Hadoop, Mahout, QlikView, Excel.

***DATA SCIENTIST***

***AMERICAN FIRST NATIONAL BANK- HOUSTON, TX DEC 2017 – MAY 2018***

**Description:**

The intent of this project was to improve on the state-of-the-art in credit scoring by predicting probability that whether a customer will experience default in the next two years, and automated customer segmentation system using machine learning algorithms to support marketing department.

**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 are 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 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.

**Environment:** Python (Scikit-Learn/Scipy/Numpy/Pandas), Machine Learning (Logistic Regression, Support Vector Machine, Random Forest, K-means Clustering), Hadoop HDFS, AWS EC2, EMR, Lambda, Tableau, MS SQL Server 2012, MS SSIS, Windows 8/XP, JIRA

***DATA SCIENTIST***

**PNC BANK*– PITTSBURG, PA NOV 2015 TO NOV 2017***

**Description:** The project was to build an algorithm that accurately classifies credit card holders among multiple classes based on the historical data available on multiple variables. The objective of the project was to improve bank's efficiency by reducing default rate while offering new products.

**Responsibilities:**

 Responsible for predictive analysis of credit scoring to predict whether or not credit extended to a new or an existing applicant will likely result in profit or losses.  
• Primarily used Python packages for the data mining tasks.  
• Participated in all phases of data mining; data collection, data cleaning, developing models, validation, and visualization.  
• Data for modeling was collected using SQL by querying several tables. The extracted tables were further appended or merged to create tables for modeling using Python.  
• Computed Credit Risk Parameters such as Probability of Default and Loss Given Default and Exposure at Default.  
• Used logistic regression, clustering and multivariate modeling to provide valuable analytical insights.  
• Used Python for generating various graphs and charts for analyzing the different features.  
• Used k-fold cross-validation to avoid overfitting.  
• Used Kolmogorov-Smirnov test to measure the quality of the models.

**Environment:** Python, MS SQL, Hadoop, Hive, Pig, MS SQL Server 2012.