**Becky Zhu**

**Email id: [beckyzhu35@gmail.com](mailto:beckyzhu35@gmail.com)**

**Contact number: 661-247-1236**

**Summary**

* Around 6 years of experience as Data Analyst/Data Scientist working in multiple industries. Highly experienced in **Data Extraction, Data Manipulation, Data Visualization, Statistical Modeling, Feature Engineering,** **Machine Learning,** and **Big Data**.
* Significant industry experience and domain knowledge in **Healthcare, Retail, Telecom,** and **Banking** industries.
* Familiar with the entire life cycle of data science project including **data modeling**, **exploration**, **model selection**, **evaluation**, and **deployment**.
* Proficient in **Machine Learning Algorithms** including **Linear Regression, Logistic Regression, Decision Tree, Random Forests, K-Nearest Neighbors, K-means, Naïve Bayes, Support Vector Machines,** and **Ensemble Models** etc.,.
* Experienced with **Natural Language Processing** in sentiment analysis.
* Strong business judgment and ability to take ambiguous problems and solve them in a structured, hypothesis-driven, and data-supported way.
* Experienced in statistical software (e.g., **Python, pandas**) and database languages (e.g., **SQL**).
* Expertise in **Python (2.x/3.x)** programming with multiple packages including **Numpy, Pandas, Scipy,** and **Scikit-learn**.
* Experienced in working with big data solutions such as **Hadoop ecosystem** and **Apache Spark framework** including **HDFS, MapReduce, Spark SQL**, and **PySpark**.
* Proficient in **data visualization** tools such as **Tableau (9.x/10.x), Python Matplotlib, Plotly, Seaborn,** and **ggplot** to create interactive reports and dashboards.
* Hands-on experience in using **Microsoft SQL Server, SQL Server Integration Services (SSIS), Business Intelligence Development Studio,** and **Excel (pivot tables)**.
* Experienced in advanced SQL and relational databases including queries, database definition, and schema design.
* Highly skilled in **T-SQL** scripting for developing complex **stored procedures, triggers, views**, and **user functions** using various RDBMS such as **SQL Server, MySQL,** and **Oracle**.
* Experienced in designing, implementing, and enhancing **Extract, Transform,** and **Load (ETL)** processes.
* Familiar with **Cloud Service servers** such as **AWS, Google Cloud**, and **Azure**.
* Strong understanding of **SDLC** in **Agile** methodology and **Scrum** process.
* Experienced in **ticketing systems** such as **Jira** and **Azure DevOps** and version control tools such as **Git**.
* Interpreted business needs from requests and rapidly implemented effective technical solutions.

**TECHNICIAL SKILLSETS**

|  |  |
| --- | --- |
| **Databases** | MySQL, Oracle, Microsoft SQL Server 2016/2014/2012/2008, and Amazon Redshift. |
| **Statistical Methods** | Hypothetical Testing, ANOVA, Time Series, Confidence Intervals, Bayes Law, Principal Component Analysis (PCA), Dimensionality Reduction, and Cross-Validation. |
| **Machine Learning** | Regression analysis, Naïve Bayes, Decision Tree, Random Forests, Support Vector Machine, Neural Network, Sentiment Analysis, K-Means Clustering, KNN, and Ensemble Methods, and Sentiment Analysis. |
| **Hadoop Ecosystem** | MapReduce, Spark 2.x (PySpark, SparkSQL, SparkMLLib), MapReduce, Hive, HDFS, and Pig. |
| **BI Tools** | Tableau Suite of Tools 10.x, 9.x, 8.x which includes Desktop, Server and Online, and Server Integration Services (SSIS). |
| **Data Visualization** | Tableau, Matplotlib, Seaborn, Plotly, and ggplot. |
| **Languages** | Python (2.x/3.x), R, SQL, T-SQL, and LaTex. |
|  |  |

**EXPERIENCE**

**Boxed New York, NY**

**Data Scientist/Business Intelligence Engineer 09/2017 –** **Present**

Boxed.com is an online and mobile wholesale retailer that offers direct delivery via the Boxed app or the website. The project applies machine learning techniques and develops statistical models to conduct experiments on demand anomaly detection. The project has forecasting science team that predicts product demand and anomaly detection for downstream teams by generating dashboard, which helps in digital strategy and implementation of online retail operations for owned and marketplace sites.

**Responsibilities:**

* Wrote efficient customized SQL query based on user’s inputs.
* Acquired aggregated data at different levels from data warehouse.
* Conducted experiments on demand anomaly detection at multiple levels using statistical process control methods, which employed statistical methods to monitor and control the process.
* Introduced statistical method to identify irregularities in data by calculating rolling average across the data points.
* Applied density-based anomaly detection, clustering-based detection, and support vector machine-based anomaly detection.
* Implemented ARIMA model and clustering methods for smoother operation of anomaly detection.
* Promoted a novel model evaluation and selection process for anomaly detection using kappa score and loss ratio.
* Created an interactive dashboard which contains the modeling analysis, backend database, and frontend UI interface by using Plotly.

**Environment:** Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas), Machine Learning, AWS Redshift, and Jupyter Notebook.

**DV Trading, LLC New York, NY**

**Data Analyst 06/2016 – 09/2017**

DV Trading, LLC is a North American based proprietary trading firm with a significant presence on derivatives and securities exchanges worldwide. The project provides technical support to clients by generating reports with the information clients need such as account information and the historical trading data. The reports are for internal use and the clients are the traders in the company.

**Responsibilities:**

* Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the user’s requirements.
* Processed new incoming data and deleted noise of data by detecting syntax errors, transforming data type, converting values to conform to certain range, eliminating duplicates, and saving datasets in excel or csv formats.
* Used SSIS to create ETL packages to Validate, Extract, Transform, and Load processed data into Data Warehouse and Data Mart.
* Maintained and developed complex SQL queries, stored procedures, views, functions, and reports that meet customer requirements using Microsoft SQL Server 2008.
* Generated reports with information clients need using Python visualization packages and Tableau.

**Environments:** AWS EC2, EMR, S3, Tableau Desktop (9.x/10.x), Tableau Server (9.x/10.x) and Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn).

**Wolfram Research Champaign, Illinois**

**Data Analyst 05/2015 – 05/2016**

Wolfram Research is one of the world’s most respected computer, web, and cloud software companies, with specialization on developing mathematical software for engineers, financial analysts, researchers, and students. The project works on analyzing customer usage, user retention and providing feasible advice for improving retention rate with statistical and mathematical models.

**Responsibilities:**

* Extracted, cleaned and reconsolidated data from multiple sources via Python API and SQL for further analysis.
* Computed customer lifetime value and used customer lifetime cycle to improve CLV.
* Increased retention rate via email campaigns to target users based on lifetime pattern.
* Applied A/B test on main product webpage and improved subscriptions through further optimization of subscription flow.
* Implemented ARIMA model of time series analysis to new subscriptions and revenue prediction.
* Created and maintained monthly reports for main products including the key metrics, trends, and insights of product behavior.
* Implemented market analysis, financial analysis and competitor analysis to detect the potential opportunities.

**Environment:** Python 2.x/3.x(Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn), Jupyter Notebook, SQL, and Tableau.

**Pavlov Media, Inc. Champaign, Illinois**

**Data Analyst 05/2014 – 05/2015**

Pavlov Media, Inc. is one of the nation’s largest private provider of Internet and video services to off-campus student housing. The project builds predictive pipeline for customer churn by developing machine learning algorithms and providing reports for departments to interact with clients.

**Responsibilities:**

* Queried historical labeled data which is stored in the database by connecting Jupyter notebook to SQL database.
* Performed exploratory analysis including checking the missing values, data type of each column, and the distribution of data to get a better understanding of data using built-in functions in Python Pandas library.
* Processed data set by filling in missing values, parsing dates, and numbers in incorrect formats, and converting categorical values to numeric values.
* Plotted charts to figure out useful variables/features and got a visual indication of correlation between features.
* Done feature selection by analyzing feature importance, identified top factors that may affect the result, and visualized feature importance using plot tree library in Python.
* Split data sets into training set and testing set by randomly sampled split.
* Initialized and trained a variety of different machine learning models including decision tree, logistic regression, KNN, and random forest, and selected the random forest which had the best performance to refine.
* Evaluated the model by getting the score of accuracy and compared it with the actual probabilities of a customer churning calculated by using predic\_proba ().
* Constructed confusion matrix and ROC curve to dig further into the quality of the results.
* Applied methods to avoid overfitting such as regularization and k-fold cross validation.
* Visualized the results and important factors and generated the report.

**Environment:** Python 2.x/3.x(Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn), Machine Learning, AWS Redshift, Jupyter Notebook, and SQL.

**Christie Clinic Champaign, IL**

**Business Intelligence/Database Developer 01/2013 – 05/2014**

Christie Clinic is one of the largest physician-owned multi-specialty group medical practices in Illinois. It has been providing the people of east central Illinois with the highest quality health care. The project creates SSIS packages to integrate the data contained in various external file systems and SQL server database to generate the reports using SSRS to fulfill data solution requirements from business departments.

**Responsibilities:**

* Used SSIS to create ETL packages to Validate, Extract, Transform, and Load data into Data Warehouse and Data Mart.
* Maintained and developed complex T-SQL queries in Microsoft SQL Server such as stored procedures, views, temp tables, and user-defined functions to meet clients’ requirements.
* Designed SSIS packages to build ETL pipeline and employed various functions such as aggregations, merge joins, count, and data flow task, etc.
* Created SSIS packages to extract, migrate, and integrate healthcare data from multiple data sources such as database source, text files, and excel files.
* Supported and maintained reporting platform – Tableau Server administration.
* Worked closely with internal and external stakeholders to discuss organizational needs and gather dashboard requirements.
* Supported data visualization with Tableau.

**Environment:** SQL Server 2008, SQL Server Management Studio (SSMS), MS BI Suite (SSIS/SSRS), SQL Server Data Tools (SSDT), Visual Studio, and Tableau.

**EDUCATION**

Master of Science in Data Science, New York University, New York, NY.