**Ricky Ward**Data Scientist

Email: Ricky.Ward84@outlook.com |

**SUMMARY**

Data Scientist with 8+ years of professional experience in the **Banking, E-commerce, Transportation** and **Supply Chain** 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** 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**, **Feature Engineering, Dimensionality Reduction**, **Algorithm implementation, Back Testing** and **Validation**.
* Expert at working with statistical tests: two-way **independent** & **paired t-test, one-way & two-way ANOVA** along with non-parametric tests: **chi-sq. tests, Mann-Whitney U, Wilcoxon rank tests, Shapiro-Wilk & Kruskal-Wallis** **test** using **RStudio**.
* Proficient in Data transformations using **log, square-root, reciprocal, differencing** and **complete box-cox transformation** depending upon the dataset.
* Adept at Analysis of **Missing data** by exploring **correlations** and similarities, introducing **dummy variables** for **missingness**, and choosing from imputation methods such as **MICE** in R and **iterative imputer** on Python.
* Experienced in Machine Learning techniques such as regression and classification models like **Linear, Polynomial, Support Vector, Decision Trees, Logistic Regression, Support Vector Machines**.
* Experienced in Ensemble learning using **Bagging, Boosting & Random Forests**; clustering like **K-means, DBSCAN**; Association Rule learning with **Apriori, Eclat**.
* 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.
* Proficient at Data Cleaning process of outlier detection and removal using **Grubb’s test** for univariate analysis, **Leverage test, Mahalanobis and Cook’s distance** for multivariate analysis;
* Adept with R, Python and OOP concepts such as **Inheritance, Polymorphism, Abstraction, Association**, etc.
* Experienced in developing algorithms to create **Artificial Neural Networks** to implement AI solutions.
* Expertise in creating **executive Tableau Dashboards** for Data visualization and deploying it to the servers; Skilled in using **tidyverse** 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**, **MySQL**, **Oracle SQL** and **Redshift SQL** and Microsoft Excel (**VLOOKUP, Pivot tables)**
* Skilled in Big Data Technologies like **Spark, Spark SQL, PySpark, HDFS (Hadoop), MapReduce** & **Kafka.**
* 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.
* Excellent exposure to Data Visualization with **Tableau, PowerBI, Seaborn, Matplotlib and ggplot2**.
* Experience with Python libraries including NumPy, Pandas, SciPy, SkLearn & statsmodels, MatplotLib, Seaborn, Theano, Tensorflow, Keras, nltk and R libraries **ggplot2, dplyr, Esquisse, CRAN**.
* Working knowledge of Database Creation and maintenance of Physical data models with **Oracle, DB2 and SQL server** databases as well as normalizing databases up to third form using **SQL** functions.

**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, Regression. |
| **Validation Techniques** | Monte Carlo simulations, 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 |
| **Validation Techniques** | Gradient, K-fold cross Validation, Monte-Carlo simulations, Out of bag sample estimate |
| **Data Visualization** | Tableau, Microsoft PowerBI, ggplot2, MatplotLib, Seaborn and Bokeh |
| **Data modeling** | Entity relationship Diagrams (ERD), Snowflake Schema |
| **Big Data** | Apache Hadoop, HDFS, Kafka, MapReduce, Spark |

**PROJECTS**

**Kohl’s Corp Menomonee Falls, Wisconsin**

**Data Scientist: Jan 2018 - Present**

**Description:**

Kohl’s Corporate is a chain of American department stores that offers in-store as well as e-shopping options. The first half of the project involved predicting Customer Life-Time Values based on the historical data gathered and the second half of the project employed clustering of products in different segments and classes.

**Responsibilities:**

* Performed **Data Collection, Data Cleaning** and **Data Visualization** using RStudio**, Deep Feature Synthesis** and extracted key statistical findings to develop business strategies.
* Initiated various pre-processing phases of text like **Tokenizing, Stemming & Lemmatization** and converting the raw text to structured data.
* Constructed new vocabulary to encode the variables in a machine-readable format using **Bag of words** and **TF-IDF**.
* Executed processes in parallel using distributed environment of **Tensorflow** across multiple devices (CPUs & GPUs).
* Implemented **sampling, PCA and LDA** for high dimensional data and drew visual statistical conclusions as well as statistical inferences.
* Employed **NLP** to classify text within the dataset. Categorization involved **labeling** natural language texts with relevant categories from a predefined set.
* Analyzed and grouped products into different clusters based on product description, purchase and historic data using techniques such as **k-means clustering**.
* Employed **auto-classification** of products based on customer database by drawing inferences from products ordered together. This assisted in creation of cohorts.
* A gradient boosted **Decision Tree Classifier** was trained using **Extreme Gradient Boosting** to identify whether a cohort was a promoter or detractor.
* Optimized the performance of the neural network using the regularization and choosing the right number of hidden layers and neurons per layer.
* The **NLP** text analysis monitored, tracked and classified user discussion about product and service in online discussion. (ScraPy and BeautifulSoup)

**Environments:** R, Tableau,Python – NLTK, SpaCy, Sci-Kit learn.

**Amway Ada Township, MI**

**Data Scientist: July 2016 - December 2017**

**Description:**

Amway is an American multi-level marketing company that offers health, beauty and home care products on their website. The project required me to build and facilitate a robust model capable of predicting total purchase amount based on customer demographics and product categories followed by maximizing the revenue of the company.

**Responsibilities:**

* Performed Data cleaning in a huge dataset which had many missing data & extreme outliers from **Hadoop workbooks** and explored data to draw relationships and correlations between variables.
* Used **MICE** in **R** to impute missing observations based on the existing observations & tracked outliers using **Mahalanobis** distance & **leverage** statistics: **chi square** **cut off** to remove extreme outliers.
* Used **cook’s distance** to detect distinct observational influence on the dataset and removed the outliers.
* Used **two sample independent t-tests** to assess the differences in mean purchases across dichotomous variables such as gender and marital status, used **one-way ANOVA** and **tukey** parameter to assess difference between mean purchases across polychotomous variables such as occupation and age.
* Used **Multiple Linear Regression**, **Decision Tree Regression**, **Support Vector Regression** & ensemble learning like **Bagging**, **Random Forests** & **Gradient Boosting Machine** to train 70% of the model & the models were optimized using **Grid Search** & the predictions were made on the test set using each trained model.
* Implemented an **artificial neural network** to predict the total purchase amount of customer using 64 neurons close to the **input** and 15 neurons in the **outer layer** close to the output.
* Computed **Absolute** and **Relative** return based on the simulation and plotted **histograms** for the selections to find the best ad and strategy based on a **Reinforcement Learning** algorithm (**Thompson Sampling**).
* The final model was selected using **Gain Plot Curve Relative Gini Score**, **Root Mean Squared Error**, and **Mean Absolute Error** and validated using **ten-fold cross validation** technique.
* The results were summarized as a dashboard in **tableau** and presented to the client.

**Environment:** Tableau, Excel, Python (Pandas, Scikit, Numpy), TensorFlow, Keras, R, MySQL, HDFS.

**Deutsche Bank New York City, NY**

**Data Scientist: Sep 2015 – June 2016**

**Description:**

Deutsche Bank is a leading global investment bank with strong and profitable private client’s franchise. The project was to implement Machine Learning techniques and develop statistical models to improve trade surveillance spoofing detection.

**Responsibilities:**

* Developed and enhanced spoofing surveillance by building trader profiles, utilizing historical trade activity to detect potential spoofing patterns and behaviors related to disruptive trading practices.
* Tackled highly imbalanced Fraud Dataset using **under-sampling, over-sampling with SMOTE** and cost sensitive algorithms with Python.
* Conducted **Data blending, Data Preparation** using **Alteryx** and **Python** for Tableau consumption and published data sources to Tableau server.
* Developed **PySpark** modules for predictive analysis and machine learning.
* Worked on Data Cleaning and ensured Data Quality, consistency and integrity using Pandas and Numpy.
* Participated in feature engineering such as **feature intersection generating, feature normalize and label encoding** with **Scikit-Learn**.
* Improved fraud prediction performance by using **Random Forest** and **LightGBM** for feature selection.
* Performed **Naïve Bayes, KNN, Logistic Regression, Random Forest, SVM** and **LightGBM** to identify spoofing pattern and collusion pattern.
* Employed various metrics such as **RMSE, MAE, Confusion Matrix, ROC and AUC** to evaluate the performance of each model.
* Employed statistical methodologies such as A/B test, experiment design and hypothesis testing.
* Implemented **Bagging** and **Boosting**, using **AdaBoost**, **Gradient Boosting** and **Extreme Gradient Boosting** to enhance the model performance.
* Performed data analysis by using **Spectrum** to run **Redshift SQL** queries against **S3** to directly retrieve data.
* Created multiple custom SQL queries in **MySQL Workbench** to prepare datasets for **Tableau dashboards** and retrieved data from multiple tables using join conditions to efficiently extract data for **Tableau workbooks**.

**Environment:** Teradata, Alteryx, Tableau, AWS RedShift, Spark (PySpark), LightGBM.

**Wells Fargo Charlotte/NC**

**Data Analyst: June 2013 – Aug 2015**

**Description:**

My responsibilities included developing a classification model to segregate customers and direct them to subscription through App Behavioral Analysis.

**Responsibilities:**

* Used Python to develop different models & algorithms to predict the probability of customer subscribing for premium using different variables.
* Deployed advanced techniques such as **text mining, statistical analysis** and successfully formulated the problem.
* Used RStudio libraries like **ggplot2 and dplyr** to visualize the data and draw inferences from different features such as age, survey, minigame, etc.
* Built a classification model to classify customers for promotional deals to increase likelihood of subscription using **Logistic Regression** and **Decision Tree Classifier**.
* Developed and implemented predictive models like Decision Tree, **Support Vector Machine** and Logistic Regression to predict the probability of enrollment.
* Employed Ensemble Learning techniques such as **Random Forests** and **Ada Gradient Boosting** to improve the model performance by 15%.
* Picked the final model based on **ROC & AUC** and fine-tuned the hyper parameters of the above models using **Grid Search** to find the optimum model.
* Employed **K-Fold cross-validation** to test and verify the model accuracy.
* Prepared a dashboard and story in **Tableau** showing the benchmarks and summary of model’s measure.

**Environment:** R, Python, Tableau, SQL.

**Centene St. Louis/MO**

**Data Analyst: April2011- May 2013**

**Description:**

My responsibilities included drawing statistical inferences related to Economic Order Quantity. Analyzed the model to optimize raw materials and vendor list to reduce re-order cost and time.

**Responsibilities:**

* Drew statistical inferences using t-tests, ANOVA, chi-sq tests and performed Post-Hoc Analysis using Tukey’s HSD and Bonferroni correction to assess difference across levels of raw material categories, test significance of proportional differences and assess whether sample size is large enough to detect the differences.
* Provided statistical insights into **semi-deviation** & **skewness-to-kurtosis** ratio to guide vendor decisions and inferences into optimum pricing for raw material order quantities.
* Performed **Data Analysis** on target data after transfer to Data Warehouse.
* Developed interactive executive dashboards using **PowerBI** and **Excel VBA** to provide a reporting tool that facilitates organizational metrics and data
* Worked with Data Architects on **Dimensional Model** with both Star and Snowflake Schemas utilized
* Created ETL solution using **Informatica** tool to read Product & Order data from files on shared network into **SQL** **Server** database.
* Created Database designs through **data-mapping** using ER diagrams and normalization upto the 3rd normal form and extracted relevant data whenever required using joins in Postgre SQL and Microsoft SQL Server.
* Conducted **data preparation** and outlier detection using Python.
* 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 **VBA-Macros**.

**Environment:** Microsoft Office, Microsoft PowerBI, ARIBA, SQL, Informatica.