**Emily Donatien**

[**emily.donatien@outlook.com**](mailto:emily.donatien@outlook.com)

**Professional Summary:**

* 6+ years hands-on experience in data analysis and machine learning projects.
* Strong understanding of data science project life cycle and experience in developing variety of statistical models, machine learning and data mining solutions for business requirements and data visualizations using R, Python and Tableau.
* Adept in statistical techniques such as descriptive statistics, correlation, hypothesis modeling, inferential statistics, multivariate analysis, model comparison and validation.
* Proficient in data wrangling and data mining of structured data using SQL, PL/SQL, Talend, R programming and python.
* Familiar with building supervised and unsupervised machine learning experiments in Microsoft Azure, Python and R programming to perform detailed predictive analytics and building Web Services models for different data types.
* Experienced in implementing linear & logistic regression, classification modeling, decision-trees, cluster and segmentation analysis, Time Series analysis, Principle Component Analysis using Python and R programming.
* Strong understanding of all aspects of data warehousing and experienced in ETL techniques using SQL, Toad and Informatica (PowerCenter, IDQ).
* Strong knowledge of RDBMS concepts and familiar with various relational database platforms such as Oracle, DB2, NoSQL, etc.
* Experienced in visualizing and reporting real-time insights using Tableau, ggplot, matplotlib to increase project visibility and ensue better business decision.
* Substantial experience in Big Data processing in Hadoop ecosystem using Hive, Spark, Hive, Pig, Impala, MapReduce.
* Experienced in using various packages in Python and R like ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, C50, twitteR, NLP, Reshape2, rjson, plyr, pandas, numpy, seaborn, scipy, matplotlib, scikit-learn, Beautiful Soup, Rpy2.
* Proficient in research of current process and emerging technologies which need analytic models, data inputs and output, analytic metrics and user interface needs.
* Good Knowledge in Proof of Concepts and gap analysis.
* Excellent understanding of Systmes Development Life Cycle (SDLC), Agile, Scrum and waterfall.
* Sound business intelligence and analytical skills with ability to extract insights and identify risk factors through careful analysis of statistical data.
* Effective team player with strong communication and interpersonal skills, possessing strong ability to adapt and learn new technologies and new business lines promptly.

**Technical Skills:**

**Statistical Techniques:** Descriptive statistics, hypothesis modelling, t- test, ANOVA and its variants, correlation modelling, principle component analysis, inferential statistics, time series analysis.

**Machine Learning Techniques:** feature scaling, regularization, model selction, Linear and Logistic Regressions, Boosted Decision Tree, K-Nearest Neighbhors, Random Tree, SVM, neural network.

**Programming languages and Tools:** SQL,Python, Scikit\_learn, Pandas, Numpy, R Programming, RStudio, Talend, Tableau, Hadoop, Gretl, Matlab,, Informatica (PowerCenter and IDQ), Toad, Jupyter Notebook**,** Microsoft Azure, PL/SQL, Microsoft Office

**Databases:** Oracle**,** MS Access, DB2, NoSQL

**Big Data Ecosystems:** Hadoop, MapReduce, HDFS, HBase, Hive, Spark

**Operating Systems:** Windows XP/7/8/10, Mac OS, Linux, Unix

**Professional Experience:**

**Staples Inc. Framingham, MA Mar 2017 – Till Date**

**Data Scientist**

**Responsibilities:**

* Modeled statistical algorithms against data sets and deployed predictive models using R-Studio and Python to provide solutions for decision making.
* Applied machine learning algorithms/methods on data sets to predict credit risk, fraud detection, customer churn, and target marketing.
* Built ARIAM models with high accuracy for SKU demand prediction.
* Characterized false positives and false negatives (confusion matrix) to improve a model for predicting customer churn rate.
* Performed customer segmentation with clustering algorithms like k-means clustering in Python to optimize marketing strategy and promote stability of customer base.
* Worked on products recommendation using collaborative filtering and Boosted Tree to better meet the needs of existing customers and acquire new customers.
* Used Boosting method on predictive models to optimize model performances.
* Used Hadoop and its ecosystem such as Hive, Pig for processing big data, Sqoop for importing big data.
* Performed data transformation method for rescaling and normalizing variables.
* Closely worked with ETL team in performing data profiling, mapping and analysis on large datasets using Informatica IDQ and analysis tools.
* Optimized extraction and summarization routines by restructuring using PL/SQL.
* Delivered Interactive visualizations/dashboards using ggplot and Tableau to present analysis outcomes in terms of patterns, anomalies and predictions.
* Worked closely with marketing team to deliver actionable insights from huge volume of data, coming from different marketing campaigns and customer interaction matrices such as web portal usage, email campaign responses, public site interaction, and other customer specific parameters.

**Cummins Columbus, IN June 2015 – Feb 2017**

**Data Scientist**

**Responsibilities:**

* Worked with team members to create and enhance models and data pipeline written in the R and Python by parallelizing code and/or rewriting portions of it to make sure the algorithm can be executed in a production such as Azure.
* Leveraged predictive models created in conjunction with team members on datastream from the process line to improve equipment uptime and performance optimization.
* Designed machine learning algorithms to enhance customer’s equipment performance using data obtained through loT getway.
* Assisted in deployment process of validated and tested models to production as API.
* Designed a static pipeline in MS Azure for data ingestion and dashboarding.
* Created scripts and integrated several systems like Python, SQL, R and MapRduce to datamine and efficiently analyze large amounts of current and historical data.
* Closely worked with the ETL team to determine data needed to support data flow to deployed algorithms.
* Used multiple visualization tools to translate analysis solutions and present insights.
* Participated in the Agile process and met deadlines for time-sensitive projects.
* Collaborated with SME and team members to understand the business requirements.
* Participated in maintenance and upgrades of machine learning models as necessary.

**CVS Pharmacy** **Irving, TX Feb 2012 – Mar 2015**

**Data Scientist**

**Responsibilities:**

* Conducted descriptive ad-hoc and in-depth analysis of clinical and operational data using statistical modelling approaches to reduce cost and improve efficiency.
* Translated analytical findings into predictive modellings and evaluation results.
* Implemented machine learning algorithms such as regression and clustering models to effectively target the population based on project requirements to improve the HEDIS ratings to meet annual targets.
* Performed exploratory data analysis to understand delays in the insurance approval process and developed a modeled that led to 30% increase in number of cases.
* Performed data manipulation, data modeling and feature engineering using ICD codes in Python and SQL
* Collaborated with business partners and performed data profiling using Informatica IDQ and Oracle SQL.
* Identified data source using data lineage (metadata) and assessed and improved data quality.
* Developed predictive models using regression, GLM, GAM and machine learning technique such as cluster analysis, Random Forest, SVM in Python programing language.
* Designed dashboards using Tableau for efficient reporting.
* Handled importing data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS.