Bradley James

Data Scientist

Bradley.James79@outlook.comSummary

* Data Scientist with 7 years of experience and expertise in **Machine Learning, Predictive Analytics, Data Analytics, Dashboards** to deliver insights and action-oriented solutions to complex business problems.
* Experience working in various domains like **“Telecom”, “Energy”, “Finance” and “eCommerce”** domains
* Analyzed and processed complex data sets using advanced querying, visualization, analytics tools and worked with several database technologies like **Oracle**, **SQL server, MongoDB, and MySQL**
* Developed **Contextual Semantic Search** including **Sentiment Analysis** and **Opinion Mining** in both python and spark environments and utilized deep learning resources like **Recurrent Neural Networks** and **LSTM**
* Performed **Text Data Analytics** and **Text Classification** using various **Natural Language Processing** techniques like **tokenization, lemmatization, stemming, parsing** and used **Deep learning** applications
* Proficient in managing entire data science project life cycle and actively involved in all the phases of project life cycle including **Data acquisition, Data cleaning**, **Feature scaling, Dimension reduction techniques**, **Feature engineering, Statistical modeling and** **Ensemble learning**
* Implemented several statistical methodologies like **Classification** (K nearest neighbors, support vector machines, decision trees, Naïve-Bayes classifier) and **Regression models** (multiple, logistic regression and regression trees, SVR, and k-means clustering in **Python, R** and SAS JmpPro
* Expertise in developing time series forecasting models like **ARIMA,** and relevant models, Exponential and Seasonal Exponential Smoothing, and **Volatility Modeling** using **GARCH** in R programming
* Strong skills of optimized sampling methodologies like **Synthetic Minority Oversampling Technique** to deal with **oversampling** or **Under-sampling** issues which are helpful in churn modeling
* Used various metrics such as F-Score, ROC, and AUC to evaluate the performance of each model and K -fold cross-validation to test the models with different batches of data to optimize the models
* Performed **Data visualization**, designed **Dashboards** with **Tableau** and **Power BI** and generated complex reports including charts, summaries, and graphs to interpret the findings to the team and stakeholders
* Expertise in **statistical methodologies** like parametric and non-parametric tests like **t-test, chi-squared goodness of fit test, and ANOVA, hypothesis testing**
* Worked with Big data tools like **Pig, Hive, Spark**, and worked with data engineers on deployment tools like **Flask, Kubernetes and Docker** for validating and maintaining model performance
* Experience with Python libraries including **NumPy, Pandas, SciPy, SkLearn, MatplotLib, Seaborn, Theano, Tensorflow, NLTK** and R libraries including **ggplot2, dplyr**
* Creative in finding solutions to problems and determining modifications for optimal use of organizational data and expert at providing realistic projections and establishing various scenarios to determine viable process strategies

Skills

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Scripting Languages and platforms** | Python (Numpy, pandas, Scikit-learn, tensorflow, re, pickle, lstm, tkinter, etc.), R (xts, zoo, quantmod), Hive, SQL, C, Google Colab, Jupyter Lab | | **Statistical Tests** | Hypothesis testing, ANOVA, MANOVA and ANCOVA tests, t-tests, Chi-Square Goodness of Fit test, Linear and Logistic Regression, Discriminant Analysis | | **Machine Learning Algorithms** | **Regression Models:** Linear, Polynomial, Support Vector, Decision Trees  **Classification Models:** Logistic Regression, k-Nearest Neighbors, Decision Trees, Naïve-Bayes, Random Forest, Support Vector Machines  **Clustering**: K-means, Hierarchical, Expectation maximization  **Association Rule Learning**: Apriori, Eclat  **Ensemble Learning** Random Forest, Bagging Trees, Gradient Boosting Machine  **Time Series Forecasting** AR, MA, ARIMA, ARCH, GARCH, MSGARCH, eGARCH  **Dimensionality Reduction**: Principal component Analysis (PCA), Linear discriminant Analysis (LDA), Autoencoders  **Text Data Analytics:** Natural Language Processing, NLTK, Spacy, LSTM, RNN | | **Validation Tests** | Monte Carlo methods, k-fold cross validation, Out of the Box Estimate | | **Analytical tools** | Google Analytics, R Studio, SAS | | **Data Visualization** | Tableau, Microsoft Power BI, R ggplot2, plotly, Python matplotlib, seaborn, bokeh | | **Database Systems** | MS SQL, Oracle, MYSQL, PostgreSQL, MongoDB, Teradata, DB2, Amazon Dynamo DB | | **Big Data Tools** | Apache Ambari, Pig, Hive, Hadoop, Spark, Kafka, Hive, Flask, Kubernetes, Docker | |

# **Experience**

## ***Kohl’s,*** **Data Scientist** Menomonee Falls, WI Mar 2018 - Present

**Objective:** To build an opinion mining and sentiment analysis of user reviews for various products and develop a search algorithm to boost sales and reduce the bounce rate.

**Responsibilities**:

* Performed **opinion mining** and **sentiment analysis** of user reviews at a document level, sentence level and aspect level to optimize the sentiment of the users about the products and improve the **contextual semantic search** which improved the Search Rank algorithm accuracy by 5% and reduced bounce rate by 13% in some cases
* Used common NLP pre-processing techniques, such as (**tokenization, lemmatization/stemming, POS tagging, parsing**) on text data for analytics over products and user reviews which in turn were used in search criteria
* Utilized several natural language processing techniques like POS tagging, bag of words model, **word2vec, count vectorizer** and modelled using **PySpark MLLib** and python.
* Performed Latent Semantic Analysis to understand the contextual usage of words by statistical computations
* Initiated the application of deep learning into existing use cases and implemented Machine Learning/Deep learning models to build **Text Classification, Topic Modelling**, utilized **tf-idf, Random Forests** and **Naive Bayes** to perform topics classification and sentiment analysis
* Trained models including **Logistic Regression, Random Forest** and **K-Nearest Neighbors**, and **Support Vector Machine** and applied **regularization** with optimal parameters to overcome overfitting
* Performed Product Matching, created sequential model for product matching across various retail websites
* Evaluated the performance of 4 classifiers using **k-fold cross-validation** technique and generated **ROC curves** and **PR curves** for comparison, analyzed **feature importance** to identify top factors that influenced prediction results
* Used PySpark machine learning library MLlib to build and evaluate different machine learning models
* Worked with **chi-squared analysis** for **feature engineering** that involves converting the arbitrary data to well-behaved data such as dealing with text features
* Performed data mining using very complex **SQL** queries and discovered pattern, used extensive SQL for **data profiling/analysis** to provide guidance in building the **data model**

**Environment:**  Python 3.6, R 3.3.1, PySpark 2.4.1, Tableau 10, Linux, Hive, SQL Server.

## ***Ally Financial***, **Data Scientist/Quantitative Analyst** Charlotte, NC Sept 2017 - Mar 2018

**Objective:** Perform time series analysis for trading team interested in options markets for regular and trading around earnings, volatility modelling using **MSGARCH** and other techniques

**Responsibilities**:

* Developed **time series forecasting** models such as **Autoregressive Integrated Moving Average model, Exponential Smoothing model, Seasonal Exponential Smoothing model, and Holt-Winters model** in R programming and python
* Developed volatility models such as **ARCH, GARCH and Markov Regime switching GARCH models** on required stocks and simulated using **maximum likelihood estimation** and **Bayesian MCMC methods**
* Developed a **statistical arbitrage strategy** using multiple characteristics for each stock including **size, value, momentum**, and their interactions with **macroeconomic** variables by applying **machine learning algorithms** including **generalized linear models, boosted regression tress and support vector machines**, tuned the **hyperparameters** and **back-tested** the models using validation and out-of-sample data
* Implemented and maintained scalable **python** code for daily automated data update, technical indicators generations which are used to build **Ensemble models** to predict the expected revenue of targeted companies
* Implemented **Markowitz Mean-Variance Model** and **Risk parity models** in Python which are 15% better performing compared to S&P index and maximize drawdown
* Created and presented models for potential holdings to fund managers, achieved 10% better returns against historical performance.
* Created **Machine Learning tools** that computes **adjusted P/E values** and few other **custom** **visualizations** to internally used application required for various teams based on **tkinter module in python**
* Extracted news titles using news feed trade logs for past 10 years and back-tested keyword trading strategy with historical prices
* Selected 30 features by performing stocks ranking, portfolios grouping and back-testing on different types of factors such as **Trailing P/E, Debt Ratio, Sentiment** on a 10-year daily data
* Automate and run monthly monitoring reports to check that the model inputs and outputs are stable and behave reliably and assess whether model performance is deteriorating over time
* Developed data analytical databases from complex financial data source and Performed daily system checks, data auditing, created reports & monitored data for accuracy

**Environment:**  Python 3.5, R 3.1.1, Tableau, MSSQL, SQL Server.

## ***Wells Fargo***, Jr. **Data Scientist** Charlotte, NC Mar 2015 - July 2017

**Objective:** Develop machine learning models, churn modeling, customer segmentation and identify key business performance indicators

**Responsibilities**:

* Built **customer churn models** for which addressed unbalanced classification problem using the **Synthetic Minority Over-sampling Technique**, built the model based on **CAR, CHAID** and other machine learning algorithms and achieved an overall accuracy of 83%
* Evaluated the performance of using **k-fold cross-validation** and generated **ROC curves and PR curves** for comparison, analyzed feature importance to identify top factors that influenced prediction results
* Performed Segmentation: Business requirement was to be able to effectively customize the marketing campaigns to perform clustering which effectively segments the customers
* Utilized classification models like **logistic regression, decision and boosted trees**, **random forest** and performed **cross validation** based on **grid search and k-fold cross validation**
* Worked with **AWS EC2** instances to create GPU heavy models and worked with data engineers to deploy models
* Involved in defining the source to target data mappings, business rules, data definitions
* Deployed various machine learning models and regularly updated them with quarterly development with new improvements
* Moved the data science eco system into **Git version control** to track changes across teams
* Created Data Quality Scripts using **SQL** and **Hive** to validate successful data load and quality of the data, various types of **data visualizations using Python and Tableau**.
* Documented the complete process flow to describe program development, testing, logic, and implementation, application integration, coding.
* Involved in defining the business/transformation rules applied for sales and service data.
* Worked with internal architects and assisted in the development of current and target state data architectures.

**Environment:**  Python, R, R, Linux, Power BI, MSSQL, SQL Server, Hive, Git, AWS.

## ***BCBS,*** **Data Analyst** Chicago,IL May 2012 - Feb 2015

**Responsibilities**:

* Developed Machine learning models that predict and optimize product performance
* Co-supported the pilot **data science** group with **data architecture** and **data cleaning** whose projects generated a savings of $130,000 in company operating costs
* Redefined many attributes and relationships in the reverse engineered model and cleansed unwanted tables/ columns as part of data analysis responsibilities
* Collaborated with key business executives to understand organizational needs and appropriate use cases for nascent and existing machine
* Involved in **Normalization** (up to 3rd normal form), **De-normalization** (**Star Schema** for Data Warehousing.) of **databases** and setup the **pipelines** for various reports generation
* Prepared weekly and quarterly required **Data Visualization** reports for management using **R programming**
* Responsible for **quantitative analysis** of **structured and semi-structured data**, working in small teams to develop, test, and harden advanced analytical models as required.
* Performed extensive **requirement analysis** and developed use cases and **Workflow Diagrams**