**Jay Kelley**

**Data Scientist**

**Jay.Kelly92@outlook.com**

**Professional Summary:**

* 6+ years of IT experience which includes **Machine Learning**, Datamining with large datasets of Structured and Unstructured data, Data Acquisition, Data Validation, **Predictive modeling**, **Data** **Visualization**.
* A deep understanding of **Statistical Modelling**, **Multivariate Analysis**, **Bigdata analytics** and Standard Procedures Highly efficient in Dimensionality Reduction methods such as PCA (Principal component Analysis), **Factor Analysis** etc. Implemented bootstrapping methods such as **Random Forests** (Classification), **K-Means** **Clustering, KNN, Naive Bayes, SVM, Decision Tree, BFS, Linear** and **Logistic Regression** Methods.
* The experience of working in **text understanding, classification, pattern recognition, recommendation systems, targeting systems** and **ranking systems** using **Python**.
* Experience with **Natural Language Processing** (NLP).
* Proficiency in application of **statistical prediction modeling, machine learning classification techniques** and **econometric forecasting techniques**.
* In-depth knowledge of statistical procedures that are applied in **Supervised / Unsupervised problems**.
* Experience in the application of **Neural Network, Support Vector Machines** **(SVM)**, and **Random Forest**.
* Experienced in working with advanced analytical teams to design, build, validate and refresh data models that enable the next generation of sophisticated solutions for global clients.
* Extensively worked on **Python** **3.5/2.7** (NumPy, Pandas, Matplotlib, NLTK and Sci-kit learn), PHP.
* Extensive experience in **Text Analytics**, developing different **Statistical Machine Learning, Data Mining** solutions to various business problems and generating data visualizations using **R, Python** and **Tableau**.
* Experience in implementing data analysis with various analytic tools, such as **Anaconda 4.0 Jupyter Notebook 4.X, R 3.0**(ggplot2) and **Excel**.
* Experience in designing **star schema, Snowflake schema** for **Data Warehouse, ODS architecture**.
* Hands on experience in business understanding, data understanding, and preparation of large databases.
* Experience in working with **relational databases (MySQL, Oracle)** with advanced **SQL programming** skills
* Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing **datamining and reporting solutions** that scales across massive volume of **structured and unstructured data**.
* Creative thinking and propose innovative ways to look at problems by using data mining approaches on the set of information available.
* Experience in using one or more cloud computing frameworks, such as **AWS, Azure, Google Cloud**, etc.
* Mapping and tracing data from system to system in order to establish **data hierarchy** and **lineage**.
* Used Data Lineage and reverse engineering as a way to track back errors in data till the data source.
* Documenting new data to help source to target mapping. Also updating the documentation for existing data assisting with **data profiling** to maintain **data sanitation, validation**.
* Identifies what data is available and relevant, including internal and external data sources, leveraging new data collection processes such **geo-location information**.
* Good understanding on **data preparation techniques**.
* Experience with distributed data/computing tools, **Map/Reduce, Hadoop, Hive, Spark, MySQL.**
* Worked on **Tableau** to create dashboards and visualizations.
* Proficiency in various type of **optimization, Market Mix modeling, Segmentation, Time Series, Price Promo models** etc.
* Experience analyzing data from 3rd party providers such as **Google Analytics**, Facebook Insights, etc.
* Identifies/creates the appropriate algorithm to discover patterns, validate their findings using an experimental and iterative approach.
* Strong skills in statistical methodologies such as **A/B test, Experiment design, Hypothesis test, ANOVA, Crosstabs, T tests** and **Correlation Techniques.**
* Applies advanced statistical and predictive modeling techniques to build, maintain, and improve on multiple real-time decision systems.
* Experience in designing stunning **visualizations** using **Tableau** software and publishing and presenting dashboards, Storyline on web and desktop platforms.
* Proficiency in **SAS** **(Base SAS, Enterprise Guide, Enterprise Miner)**
* Excellent communication skills (verbal and written) to communicate with clients and team, prepare deliver effective presentations.
* Strong experience in **Software Development Life Cycle (SDLC)** including Requirements Analysis, Design Specification and Testing as per Cycle in both **Waterfall** and **Agile methodologies**.
* Strong experience in interacting with stakeholders/customers, gathering requirements through interviews, workshops, and existing system documentation or procedures, defining business processes, identifying and analyzing risks using appropriate templates and analysis tools.
* Closely works with product managers, Service development managers, and product development team in productizing the algorithms developed.

**Technical Skills**

|  |  |
| --- | --- |
| **Predictive Modelling Technique** | Supervised Learning- Decision trees, Naive Bayes classification, Natural Language Processing (NLP), Ordinary Least Squares regression, Logistic regression, Neural networks, Support vector machines, Unsupervised Learning- Clustering Algorithms and Reinforcement Learning |
| **Languages** | Python, Scala, SQL, R, MATLAB, HTML, CSS, Angular JS, PHP |
| **Visualization** | Matplotlib, seaborn, sci-kit image, Big Data (HDFS, Pig, Hive, HBase, Sqoop, Spark), Excel, TensorFlow, Tableau, Power BI |
| **Bigdata Technologies** | Spark, Hadoop, Hive, HDFS, MapReduce |
| **Statistical Methods** | t-test, Chi-squared and ANOVA testing, A/B Testing, Descriptive and Inferential Statistics, Hypothesis testing |
| **Database** | Hadoop, Spark, Postgres, Access, Oracle, SQL Server, NoSQL (Mongo DB), Cassandra, HBase, MySQL, Netezza, Dynamo DB |
| **Version control** | Git, GitHub |
| **Cloud Computing** | Amazon AWS, Microsoft Azure, Google Analytics, Amazon Redshift, OpenShift |

**Professional Experience:**

**Kindred Health, Louisville, KY July 2018 – Till date**

**Data Scientist**

**Description:** Kindred Healthcare is a healthcare services company that operates hospitals, nursing centers, and contract rehabilitation services across the United States. The project is involved in predictive analysis and trend analysis using the patient historical data.

**Responsibilities:**

* Extensively used Python’s multiple data science packages like **Pandas, NumPy, Matplotlib, SciPy, Sci-kit learn** and **NLTK**. Worked on data cleaning, data preparation and feature engineering with **Python 3.X.**
* In depth understanding of **NLP** for **tokenization, lemmatization and stemming using NLTK, Spacy, Pattern libraries.**
* Experience in evaluating and proposing solutions to **NLP** problems through various approaches.
* Deep background in information retrieval (**OCR, Speech-to-Text** etc.), Natural Language Processing (**NLP**), knowledge representation or computational linguistics.
* Understanding of **NLP** techniques for text representation, semantic extraction techniques, data structures and modeling.
* Experience in Java, **Python** and **NLP/ML** frameworks and libraries.
* Worked to apply a broad array of capabilities spanning machine learning, statistics, **text-mining/NLP**, and modeling to extract insights to structured and unstructured healthcare data sources, pre-clinical, clinical trial and complementary real world information streams to apply a broad array of capabilities spanning machine learning, statistics, text-mining/**NLP**, and modeling to extract insights to structured and unstructured healthcare data sources, pre-clinical, clinical trial and complementary real world information streams.
* Developed **User Defined Functions** in Python for rapid analysis and performed data imputation using Sci-kit learn package in **Python**.
* Developed the applications, models, used appropriate **algorithms** for arriving at the required insights by analyzing business requirements.
* Worked with unsupervised (**K-means, DBSCAN**) and supervised learning techniques (**Regression, Classification**) for feature engineering and did **Principal Component Analysis** for dimensionality reduction of features.
* Used the Classification machine learning algorithms **Naïve Bayes, Linear regression, Logistic regression, SVM**, **Neural Networks** and used **Clustering Algorithm K Means**.
* Performed text classification task using **NLTK** package and implemented various natural language processing techniques.
* Worked collaboratively with senior management to develop strategy and approach to defining business challenges to be answered by **data science**. Established partnerships with product and engineering teams and work closely with other teams.
* Used **Spark-Streaming APIs** to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from **Kafka** in near real time.
* Worked on end to end pipe line in **Spark** and on **Apache spark** for analyzing the live streaming data.
* Worked on **Spark Python** modules for machine learning and predictive analytics in Spark on **AWS**.
* Explored and analyzed the customer specific features by using **Spark SQL**.
* Created **Hive scripts** to create external, internal data tables on Hive. Worked on creating datasets to load data into HIVE.
* Worked on Tableau for **data visualization** to create reports, dashboards for insights and business process improvement. Created the dashboards and reports in **tableau** for visualizing the data in required format.

**Environment:** Spark, Apache Spark, Hive, Machine learning, Python, NumPy, NLTK, Pandas, SciPy, MySQL, Tableau, Sqoop, HBase, HDFS, Tableau, DynamoDB, Mongo DB, SQL Server, and ETL.

**BNY Mellon, Boston, MA March 2017 – June 2018**

**Data Scientist**

**Description**: Bank of New York Mellon is an American worldwide banking and financial services holding company. The project involves creating analytics reports by connecting to databases and performing data analysis by maintaining and tracking as well as scheduling the payment system information, cancellation of the money transfer and send notifications to the user through email.

**Responsibilities:**

* Incident and problem management, coordinating resolution of data movement disruptions.
* Capacity analysis and planning for a sizable application server environment. Performance metering and tuning for application servers, database, monitors and alerts.
* Risk identification and evaluation, managing and improving upon internal controls which mitigate risks
* Working with development teams to identify enhancement opportunities and resolve code defects
* Code deployment and configuration management, Coordination and execution of changes within a complex testing environment
* Data collection procedures enhancement to include information that is relevant for building analytic systems processing, cleansing, and verifying the **integrity** of data used for analysis.
* Building analytic models using a variety of techniques such **as logistic regression, risk scorecards** and **pattern recognition technologies**.
* Analyze and understand large amounts of data to determine suitability for use in models and then work to segment the **data, create variables, build models** and test those models.
* Excellent understanding of machine learning techniques and algorithms, such as **Logistic Regression, SVM**, **Random Forests**, Deep Learning etc.
* In depth understanding and experience in **NLP** and **Deep learning.**
* Expertise in **Sentiment Analysis**, Entity Extraction, Document Classification, Topic Modeling, **Natural Language Understanding** (**NLU**) and **Natural Language Generation** (**NLG**).
* Leads all **NLP/NLG** driven solutions for the project
* Design **NLP** models for searching structured/unstructured data in real/near-real time
* Interacts regularly with product team members ensuring successful integration of **NLP** solutions in the product architecture
* Keeps the solutions updated with the recent developments in **NLP/NLG**.
* Perform a proper **EDA**, Univariate and bi-variate analysis to understand the intrinsic effect/combined.
* Performing **ad-hoc** analysis and presenting results in a clear manner and constant tracking of model performance.
* Worked with Data governance, Data quality, **Data lineage, Data architect** to design various models.
* Designed data models and data flow diagrams using **Erwin** and **MS Visio**. Independently coded new programs and designed Tables to load and test the program effectively for the given **POC’s**.
* Developed Implemented &amp; maintained the Conceptual, Logical &amp; **Physical Data Models** using **Erwin** for **Forward/Reverse Engineered Databases**.
* Experience with common data science toolkits, such as **R, MATLAB, Python, Spark**, etc.
* Applied statistics skills, such as statistical **sampling, testing, regression**, etc. Work with technical and development teams to deploy models. **Build Model Performance Reports** and
* Designing Technical Documentation to support each of the models for the product line.
* Exploratory Data Analysis and **Data Visualizations** using **R**, and **Tableau**. Established Data architecture strategy, best practices, standards, and **roadmaps**.
* Lead the development and presentation of a data analytics **data-hub prototype** with the help of the other members of the emerging solutions team.
* Involved in analysis of Business requirement, Design and Development of High level and Low-level designs, **Unit** and **Integration** testing. Interacted with the other departments to understand and identify data needs and requirements.
* Worked with several R packages including **knitr, dplyr, SparkR, CausalInfer, spacetime**.

**Environment:** UNIX, Python 3.5.2, MLLib, SAS, regression, logistic regression, Hadoop, NoSQL, MySQL, OLTP, Random forest, OLAP, HDFS, ODS

**UnitedHealth Group, Minneapolis, MN Oct 2016 – Feb 2017**

**Data Analyst/ Data Scientist**

**Description:** UnitedHealth Group is an American non-profit managed fortune top 10 health care company which offers health care products and insurance services. The project involves creating a generalizable workflow with custom Python functions to detect insurance roadblocks and scale entire analysis to large pool of medical conditions.

**Responsibilities:**

* Experience working in project with machine learning, big data, data visualization, R and Python development, UNIX, SQL.
* Extraction, transformation and loading data from given source for analysis.
* Experience in Spark SQL and Data frames for faster execution of Hive queries using **Spark SQL** Context.
* Explored different regression and ensemble models in machine learning to perform forecasting and presented Dashboards to Higher Management for more Insights using **Power BI**.
* Used Random Forest and Logistic Regression classification techniques to quantify the likelihood of each user referring.
* Extensive experience in **NLP** and used **K-means** Clustering for customer segmentation based on demographics.
* Experience in feature engineering and **NLP.**
* Analyze and decompose the written/spoken word, the story, the author and conduct research on problems specific to the **Natural Language Processing (NLP)** field, develop prediction models that can go directly into production, and build working prototypes.
* Knowledge of evolving data science concepts and best practices including big data, **NLP**, graph analysis, non-regressive methods, and unstructured data synthesis.
* Improved efficiency of the model by performing **Boosting** method on predicted model.
* Worked on Spark, Hadoop and Kafka to load data in to **HDFS** and move data into **NoSQL** databases.
* Extensive knowledge in **Map Reduce** using **Python**, **Sqoop queries,** **Pig scripts** and using Hive queries.
* Applied Backward and forward filling methods on datasets for handling missing values for performing Data cleaning process.
* Integrated multiple machine learning techniques for various user behavior prediction and support multiple marketing segmentation programs to design, build and deploy a set of python modelling **APIs** for customer analytics
* Good hands on experience on **Amazon Redshift** platform.
* Development of Tableau visualizations and stories using tableau desktop. Documenting Business requirements and plans for creating dashboards. Implementation of **R, MATLAB, Python, Hadoop** and **Tableau** to extract and import data.
* Generated graphs and reports using **matplotlib, Seaborn** and **Pandas** packages in python for analytical models.

**Environment:** Python, Regression, logistic regression, Hadoop, NoSQL, Random forest, Informatica, SQL/PLSQL, Oracle 10g, MS-Office, and Tableau.

**Western Union, Chicago, IL Nov 2015 – Sep 2016**

**Data Analyst**

**Description:** Western Union is the American financial services and communications company which is best known for exchanging business telegrams. The project is related to person-to-person money transfer, business payments and commercial services.

**Responsibilities:**

* Implemented and updated analytical methods such as regression modelling, classification tree, statistical tests and data visualization techniques with **Python**.
* Performed exploratory Data Analysis, Data Wrangling and development of algorithms in **R** and **Python** for data mining and analysis.
* Analysis of customer data and other operational data in **MySQL** and **MS Access** to provide insights that enable improvements to customer experience.
* Utilized p**andas** and **NumPy** packages in Python to improve data collection and distribution processes as well as to enhance reporting capabilities to provide clear line of sight into key performance trends and metrics.
* Performed data analysis, data manipulation, data transformation and data mapping of source data from the **MySQL** **server**.
* Understanding and adherence to the principles of data quality management including metadata, lineage, and business definitions.
* Analyzed historical demand, filter out outliers/exceptions, identify the most appropriate statistical forecasting algorithm, develop base plan, understand variance, propose improvement opportunities, and incorporate demand signal into forecast and executed data visualization by using **plotly** package in Python.
* Participated in all phases of research including data collection, data cleaning, data mining, development models and visualizations.
* Examine customer feedback and activity for use in detecting or confirming fraud, using a combination of text analytics, statistical modeling, and classification.
* Deep understanding of Software Development Life Cycle (**SDLC**) as well as Agile/Scrum methodology to accelerate Software Development iteration.
* Improvisations and maintenance to existing automated solutions. Used **MS Visio**, **MS Project** to assist the team in project planning, quality plan, risk management, requirements management, change management, defect management, change management and release management.

**Environment:** MySQL, Statistical modelling, Python libraries, Pandas, NumPy packages, R, MS Visio, MS Project, MS Access.

**United Health Group, NY Feb 2013 – Sep 2015**

**Data Analyst/ Python Developer**

**Responsibilities:**

* Designed and developed transformation logic for **BI** tools (Informatica) for data transformation into various layers of **Data warehouse**.
* Performed code development with the help of internal **Python** library that speeds up database querying and allow users to write more resilient ETL jobs.
* Implemented **PostgreSQL, SQL** servers to develop stored procedures, views to create result sets to meet varying reporting requirements.
* Ensured data integrity using advanced excel formulas (lookup functions, pivot table, If Statements etc.) for analyzing data.
* Hands on experience in writing queries in SQL and R to extract, transform and load (**ETL**) data from large datasets using Data Staging.
* Participated in the **A/B** testing conducted by **BI Analytics** team for data extraction and exploratory analysis.
* Generated dashboards and presented the analysis to researchers explaining insights on the data.
* Provided analytical insights and decision support tools for executives for accurate decision making by identifying, measuring and recommending improvement strategies for **KPIs** across all business areas.
* Submitted summaries, **charts**, and **graphs** to team and stakeholders that help to interpret findings based on complex **excel** reports.
* Responsible for Credit data related warehouse creation that could help with **Risk Assessment** for Commercial loans.
* Performed competitor and customer analysis, risk and pricing analysis and forecasted results for credit card holders on demographical basis.
* Worked in manipulating various management reports in **MS Excel** for sales metrics using **VLOOKUP** and Pivot tables.
* Involved in estimating, defining, implementing, and utilizing business metrics calculations and methodologies.

**Environment:** Excel 2010, R, Informatica Power Center 9.0, Python, PostgreSQL, MS SQL Server 200.