**Tommy Wang**

[wangtommy091@gmail.com](mailto:wangtommy091@gmail.com)

*732-889-4796*

**SUMMARY**

* 10+ years of experience in **Data Engeering, Machine Learning, Data Analytic, Data mining, Data Architecture, Data Modeling, Data Mining, Data Analysis, NLP** with large data sets of **Structured** and **Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Data Visualization, Web Crawling, Web Scraping**, Adept in statistical programming languages like **R** and **Python** including **Big Data technologies** like **Hadoop**, **Hive, HDFS, MapReduce** and **NoSQL** **Based Databases**.
* **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, data engineering, features scaling, features engineering, statistical modeling (decision trees, regression models, neural networks, SVM, clustering), dimensionality reduction** using **Principal Component Analysis** and **Factor Analysis, testing** and validation using **ROC plot, K- fold cross validation** and **data visualization.**
* Professional experience manipulating the **Google Cloud platform (GCP,Google Analytics, Tag Manager, BigQuery).**
* Experience in **Deep Learning frameworks like MXNet, Caffe 2, Tensorflow, Theano, CNTK, and Keras** to help our customers build DL models.
* Good knowledge of **Hadoop Architecture** and various components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Secondary Name Node, MapReduce concepts, and ecosystems including Hive and Pig.**
* Strong Experience operating in **Big Data Pipelines (PySpark, Hive, Presto, SQL engines)** batch and streaming.
* Implement **statistical and machine learning models, large-scale, cloud-based data processing pipelines** and off the shelf solutions for test and evaluation; interpret data to assess algorithm performance.
* Expert working within enterprise data warehouse environments platforms (**Teradata, Netezza, Oracle,** etc.) and working within distributed computing platforms such as **Hadoop** and associated technologies such as **SQL, HQL, MapReduce, PySpark, Storm, Yarn, datameer,Kafka, Sqoop and Hive.**
* Very good experience and knowledge in provisioning virtual clusters under **AWS** cloud which includes services like **EC2, S3,** and **EMR.**
* Experience in Deep Learning frameworks like **MXNet, Caffe 2, Tensorflow, Theano, CNTK, and Keras to help our customers build DL models.**
* Experience in using **GIT Version Control System**. Implemented **datameer&Kafka** for building **data pipeline and analytic modules.**
* Experience in using **Statistical procedures and Machine Learning algorithms** such as **ANOVA, Clustering, Regression and Time Series Analysis to analyze data for further Model Building.**
* Extensive hands-on experience and high proficiency in **writing complex SQL queries** like **stored procedures, triggers, joins and subqueries along with that used MongoDB for extraction data.**
* Proficient knowledge on **Mathematical Matrix Operations, Statistics, Linear Algebra, Probability, Differentiation, Integration and Geometry.**
* Skilled in **System Analysis, E-R/Dimensional Data Modeling, Database Design** and implementing **RDBMS** specific features.
* Excellent knowledge of **Machine Learning, Mathematical Modeling** and **Operations Research**. Comfortable with **R, Python, SAS and Weka, MATLAB, MS Access,Relational databases.**
* Extensive experience in Text Analytics, generating data visualizations using **R, Python** and creating dashboards using tools like **Tableau, PowerBi,ggplot2 and d3.JS.**
* Leverage a **broad stack of technologies** — **Python, Docker, AWS, GCP,Airflow,salesforce and Spark** to reveal the insights hidden within **huge volumes** of **numeric and textual data**.
* Experience using **SparkML and Amazon Machine Learning (AML)** to build **ML models**.
* Experience with medical terminologies such as **UMLS, SNOMED CT, ICD-9, ICD-10.**
* Extensive hands-on experience in navigating complex relational datasets in both structured and semi-unstructured formats.
* Expertise in **Excel Macros, Pivot Tables, vlookups** and other advanced functions and experience with working in **Agile/SCRUM** software environments.
* Expert level proficiency with statistical probabilistic modeling techniques such as **regression, tree-based methods (Random Forest, GBM), neural networks, support vector machines, supervised/unsupervised clustering techniques (k-means, DBSCAN, Expectation Maximization), principal component and factor analysis.**
* In-depth knowledge of **databases, data modeling, Hadoop, and distributed computing frameworks.**
* Hands on experience in implementing **LDA**, **Naive Bayes** and skilled in **Random Forests, Decision Trees, Linear** and **Logistic Regression, SVM, Clustering, neural networks, Principle Component Analysis.**
* Worked and extracted data from various database sources like **Oracle, SQL Server,My SQL, DB2, Regularly accessing JIRA tool and other internal issue trackers for the Project development**.

**SKILLS**

|  |  |
| --- | --- |
| **Languages** | Java 8, Python, R,JavaScript |
| **Packages** | ggplot2, caret, dplyr, Rweka, gmodels, RCurl, C50, twitter, NLP, BERT,Reshape2, rjson, plyr, pandas, numPy, Seaborn, sciPy, matplot lib, sci-kit-learn, Beautiful Soup, Rpy2. |
| **Web Technologies** | HTML, CSS, Javascript, JQuery, Bootstrap, AngularJS |
| **Machine Learning** | Decision Tree, SVM, KNN, K-Means, EM, Apriori, PageRank, AdaBoost, Deep-Learning |
| **Big Data Technologies** | Hadoop, Hive, HDFS |
| **Databases** | SQL, Hive, Impala, Pig, Spark SQL, Databases SQL-Server, My SQL, MS Access, HDFS, HBase, Teradata, Netezza, MongoDB, Cassandra,Aurora |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio), Tableau,PowerBi |
| **Version Control Tools** | SVM, GitHub,Git |
| **Operating System** | Windows, Linux, Unix, Macintosh HD, Red Hat |
| **Analysis Tools** | Python (Pandas, Numpy, scikit-learn, matplotlib), R, SAS, Tableau, Advanced MS Excel, A/B testing,PowerBi |
| **Machine Learning** | Decision Tree, SVM, KNN, K-Means, EM, Apriori, PageRank, AdaBoost, Deep-Learning |
| **Databases** | Oracle, MySQL, Microsoft SQL Server, MS Access |
| **Big Data Technologies** | Hadoop, HDFS, MapReduce, Sqoop, Hive, Spark,Data Lake |
| **Cloud** | Azure, GCP,Amazon Web Services (AWS) |
| **Application & System** | Linux |
| **Web Development** | HTML, CSS, Javascript, JQuery, Bootstrap, AngularJS |

**EDUCATION**

Shanghai University - Shanghai, China

*Bachelor of Computer Science and Technology (Graduated 2005)*

Shanghai University - Shanghai, China

*Master of Finance (Graduated 2009)*

**EXPERIENCE**

|  |
| --- |
| ***McKinesey & Company -*** *Dallas, TX Jun 2020 – Present* ***Data Scientist /Data Engineer***  McKinsey & Company is a management consulting firm that advises on strategic management to corporations, governments, and other organizations. McKinsey is the oldest and largest of the "Big Three" management consultancies (MBB), the world's three largest strategy consulting firms by revenue. It has consistently been recognized by Vault as the most prestigious consulting firm in the world.  ***Responsibilities:***   * Write Scope script to extract the data of user subscription and usage from Cosmos and Analysis different features.Created the connection between **Cosmos** and **Azure data lake storage gen1** and transferred the data to Data Lake.Used **DataBricks,python** to process data, Analysis the relationship and relevance about the features and subscription.**Used Logistic Regression, Random Forest** to build the Machine learning model to identify and target ‘at risk’ users. **Deploy the Machine learning model on Azure**, help the Microsoft Xbox marketing team to optimize and personalize campaigns to move members from high churn risk to low-risk segment. * Write shell script to extract data from **Kafka** and transfer to **GCP buckets,** create different **schemas** **in Big Query** and use **ETL** **tool(Automic)** to write pipeline and land data. Write query to get data and create dashboard. * Build machine learning based regression models using **scikit-learn python frameworks** to estimate the customer propensity to purchase based on attributes such as customer verticals they operate in, **revenue, historic purchases, frequency** and **regency behaviours**. These predictions helped estimate propensities with higher accuracy improving the overall productivity of sales teams by **accurately targeting** the **prospective clients**. * Created **SQL** tables with referential integrity and **developed** queries using **MYSQL,SQLSERVER, SQL\*PLUS** and **PL/SQL**. * Design, coding, unit testing of **ETL** package source marts and subject marts using **Informatica ETL** processes for **Oracle database.** * Collaborated with **data engineers** and operation team to implement the **ETL process**, wrote and optimized **SQL queries to perform data extraction** to fit the analytical requirements, using **Airflow,SSIS** framework,including **salesforce and snowflake platform.** * **Use NLP(NLTK,TF-IDF) to deal with the text data and use machine learning model to do the sentiment analysis.** * **Develop algorithms** for **telecommunications** service vendors to predict customer churn probability using **python**. * Use machine learning algorithms to do the **risk prediction and the Fraud detection in life insurance.** * Train supervised machine learning models including **logistic regression, random forest and k-nearest neighbors (KNN)** and applied **regularization** with optimal parameters to overcome overfitting. * Use Python Libraries including **NumPy, Pandas, Scipy, Sklearn, Matplotlib, Keras and Tensor flows**. * Used various machine learning algorithms such as **Linear Regression, Ridge Regression, Lasso Regression, Elastic net regression, KNN, Decision TreeRegressor, SVM, Bagging Decision Trees, Random Forest, AdaBoost, and XGBoost.** * Leverage a broad stack of technologies — **Python, Docker, AWS, Airflow,Salesforce,Data lake,SSIS,DataBricks and Spark to reveal the insights hidden within huge volumes of numeric and textual data.** * Use a combination of **business focus, strong analytical** and problem solving skills and programming knowledge to be able to quickly cycle hypotheses through the discovery phase of the project. * Participated in feature engineering such as feature intersection generating, feature normalize and label encoding with **Scikit-learn pre-processing.** * Used **PySpark's Machine learning library** to build and evaluate different models. * Loaded unstructured data into **Hadoop File System (HDFS).** * Performed multiple MapReduce jobs in **PIG and Hive** for **data cleaning and pre-processing.** * Created a text classification model using **RNN and LSTM with TensorFlow** * Evaluate model performance of classification using **K-fold cross-validation technique** and **confusion matrix**. * Apply unsupervised learning model and **clustering customers** into groups according to review’s texts by **Python**. * **Implement tokenization** and **stemming technology** to remove delimiters and stop words from each review text. * Derive formulae to **calculate cross-elasticity** between products for **retail cannibalization effect**. * **Implement new price optimization** rules in **python** to enhance price and **promotion optimization**. * Create term frequency – inverse document frequency (**TF-IDF**) matrix to weight factors and reflect the importance of a word in a corpus. * Used **Principal Component Analysis and t-SNE** in feature engineering to analyze high dimensional data. * Work with **Data Engineers** to determine how to best **source data**, including **identification of potential proxy data sources**, and design **business analytics solutions**, considering current and future needs, infrastructure and security requirements, load frequencies, etc. * Developed predictive Model using historical and current data to identify interested customers for Email Campaign * **Train K-Means** clustering model and Latent Dirichlet Allocation (**LDA**) to identify each review into different topic groups.   ***Mount Sinai Health -*** *New York, NY Jan 2018 – Jun 2020*  ***Data Scientist*** |
|  |

The Mount Sinai Health System is a hospital network in New York City. It was formed in September 2013 by merging the operations of Continuum Health Partners and the Mount Sinai Medical Center.

***Responsibilities:***

* **Developed starch** of perception model and **logistic regression model**, used **loss function** to find the best parameters.
* Applied visualization tools for the **loss function, data exploration** and **feature engineering**.
* Perform **Data Profiling** to learn about behavior with various features such as **traffic pattern, location, Date and Time** etc.
* Develop **Spark,DataBricks, Python, R** for **regular expression (regex)** project in the Hadoop environment with Linux for big data resources. Used clustering technique K-Means to identify outliers and to classify unlabeled data.
* Use Principal Component Analysis in feature engineering to analyze **high dimensional data.**
* Develop **MapReduce pipeline** for feature extraction using **Hive and Pig**
* **Preprocessed data** set by **data cleaning, categorical feature transformation** and **standardization, etc**.
* **Trained supervised machine learning model** with perception model and **logistic regression model, evaluated** the model performance (**The accuracy of perceptron model is 0.787 and logistic regression model is 0.792**)
* Built using **python** (**NumPy, Pandas**) in **Jupiter notebook**.
* Responsible for the **design, development** and **production support** of **interactive data visualizations** used across the project.
* Work in **RDBMS** on User **Defined Data Types, Indexes, Stored Procedures,** and **Views**.
* Used **Pandas, NumPy, Scikit-learn** in **Python** for developing various machine learning models such **Random forest** and **step-wise regression.**
* Perform data analysis by using Hive to retrieve the data from **Hadoop cluster, datalake,** **PL/SQL** to retrieve data from **Oracle database and used ETL for data transformation**.
* Involved in creating **Hive ORC tables**, loading the data into it and **writing Hive queries to analyze the data**.
* Interacted with **Team and Analysis**, Design and Develop database using **ER Diagram, involved in Design, Development and testing of the system.**
* Worked on **Python OpenStack APIs** and used **NumPy for Numerical analysis**.
* **Analyzed data** to identify emerging trends in **COVID-19 positivity** **among staff, patients, visitors** and **contacts** at healthcare facilities with timely communication of trend to lead and facility
* Collaborated with the health care team to strategize effective ways to improve the **data collection, evaluation** and **analysis**
* Worked on **NLTK,BERT** library in **python** for doing **sentiment analysis** on customer product reviews and other **third party websites** using **web scrapping.**

***First National Bank Omaha -*** *Omaha, Nebraska**Nov 2014 – Dec 2017*

***Data Analyst***

*Chartered and headquartered in Omaha, Nebraska, United States, First National provides corporate banking, investment banking, retail banking, wealth management and consumer lending services at locations in Nebraska, Iowa, Colorado, Texas, Kansas, South Dakota and Illinois.*

***Responsibilities:***

* Implemented **Data Exploration** to **analyze patterns** and to select features using **Python SciPy**.
* Built **Factor Analysis** and **Cluster Analysis** models using **Python SciPy** to classify customers into different target groups.
* **Built training data** by pulling in usage and **customer profile data stored** in different tables
* Tested various algorithms like **Logistic Regression, Decision Tree, Random Forest** and **SVM** to arrive at the best model and assign a propensity score for which was subsequently used in **targeted promotion activity**
* Built predictive models including **Support Vector Machine, Random Forests** and **Naïve Bayes Classifier** using **Python** **Scikit**-Learn to predict the personalized product choice for each client.
* Using **R’s** **dplyr** and **ggplot2 packages**, performed an extensive graphical visualization of **overall data, including customized graphical representation of revenue reports, specific item sales statistics** and **visualization.**

|  |
| --- |
| ***Lufax - Shanghai,*** *China**Dec 2011 – May 2014*  ***Data Analyst*** |

*Lufax, full name Shanghai Lujiazui International Financial Asset Exchange Co., Ltd., is an online Internet finance marketplace headquartered in Lujiazui, Shanghai. Founded in 2011, it is an associate of China Ping An Group. The company was founded in September 2011, and started with P2P lending as the only service.*

***Responsibilities:***

* Defined and create a distinctive financial product vision **including requirements, definitions,** and **specifications**.
* **Analyzed market** and **competitive product** to understand opportunities and coordinate with various internal partners to capture overall business needs and **strategic direction/alignment**.
* Leaded project teams and manage activities associated with **cross-departmental** or **organizational, entity-wide scope, coordinating multiple projects** with impeccable follow-up and **organization for time-sensitive deadlines**.
* Involved in **designing conceptual, logical** and **physical models** using **Erwin** and build datamarts using **hybrid** **Inmon** and **Kimball DW methodologies**.
* Worked closely with **Business team, Data Governance team, SMEs**, and **Vendors to define data requirements**.
* Used **Microsoft Excel** for **formatting data as a table, visualization** and **analyzing data** by using certain methods like **Conditional Formatting, removing Duplicates, Pivot** and **Unpivot tables**, **create Charts,** **sort** and **filter Data Set**.
* Worked with **data investigation, discovery** and **mapping tools** to scan every **single data** record from many sources.
* **Designed** the **prototype** of the **Data mart** and **documented** possible outcome from it for end-user.
* Involved in business process modeling using **UML**.
* **Developed** and **maintained data dictionary** to create metadata reports for **technical** and **business purpose**.
* Created **SQL** tables with referential integrity and **developed** queries using **SQL, SQL\*PLUS** and **PL/SQL**.
* Design, coding, unit testing of **ETL** package source marts and subject marts using **Informatica ETL** processes for **Oracle database.**
* **Developed efficient** and reliable metrics to help measure the success of store-based strategic initiatives, and act as the Subject Matter expert for the **development of product marketing** and **go-to market strategies**.