JINGSI GONG

gongjingsi@hotmail.com 201-204-8214

SUMMARY

- Professionalized in **Data Mining, Data Warehouse, ETL, Data Modelling, Statistical Modeling, Machine Learning,** and **Data Visualization over 3 years' experience**.
- Proficient in SQL with **Queries, Stored Procedures, triggers, views, constraints, indexes, and User-Defined Functions .**
- Expertise in directing entire data science project workflow process including **Target Framework, Data Acquisition, Data Preprocessing, Data Visualization,** and **Analysis, Data Modelling** and **Testing, Data Optimization** and **Deployment**.
- Efficient in transforming business requirements into meaningful and effective **Statistical** and **Analytical Models** as well as **Algorithms Designing** and **Solution Reporting.**
- Expertise in applying **Tableau** to publish and deliver **reports and dashboards.**
- Experienced in utilizing cloud services Amazon Web Services (AWS) including Aurora, S3, and EC2.
- Experienced in designing Data Warehouse, creating ETL packages, migrating data, and delivering daily reports by **Informatica Job** and **Autosys** Job.
- Experienced in **Hadoop ecosystem**, Spark with Big Data tools such as **PySpark, Apache Hive, Apache Pig, Databricks** and **Airflow**.
- Excellent in Statistical Methods that consists of **Hypothesis Testing, Likelihood Ratio Test, Student's t-test, Chi-Squared Test, z-test, f-test,** and **ANOVA**.
- Experienced in building various **unsupervised, supervised** machine learning algorithms and **deep learning** such as Linear Regression, Logistic Regression, Decision Tree, Random Forest, K-means, DBSCAN, PCA, Natural Language Processing (NLP), Ensembles(bagging, boosting, stacking) and more.
- Experienced in **Time Series Analysis** using **AR, MA, ARIMA, GARCH, and ARCH Models**.
- Strong experience in Python (3.x) and its libraries including **NumPy, Pandas, Scipy, Scikit-learn, Keras, NLTK, Statsmodels, Beautiful Soup, Selenium, Matplotlib,** and **Seaborn**.
- Experienced in **Agile**, Waterfall methodology, and **SCRUM** process.
- Domain knowledge in **Finance** industries.

TECHNICAL SKILLS

- **Languages:** Python(3.x), R, T-SQL, MATLAB, HTML, MySQL, PostgreSQL, C
- **Hadoop Ecosystem:** MapReduce, Spark, Pig, Hive, Airflow
- **Cloud Services:** Amazon Web Services (AWS) Aurora/EC2/S3/Redshift
- **Machine Learning:** Linear Regression, Logistic Regression, Stepwise Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Naive Bayes, KNN, Support Vector Machine(SVM), Gradient Boosting, XGBoost, Bagging, Boosting, Stacking, K-means Clustering, Hierarchical Clustering, DBSCAN Clustering, Probabilistic Clustering, Neural Network; Dimensionality Reduction; NLP.
- **Data Visualization and Dashboards:** Tableau, Snowflake, Qlik Sense

Education

Stevens Institute of Technology, Hoboken, NJ **Jan.2019**

Master of Science in Applied Mathematics

Merit Award: Graduate Teaching Assistantship

GPA: 3.7/4.0

University of Waterloo, Waterloo, ON, Canada **May.2016**

Bachelor of Science in Physics

Industry Projects

Itlize Global LLC **Feb.2019 - Present**

Client – T. Rowe Price

Product Master Enhancement and Dashboard Development Project

- Assisted project manager in preparing project management plan and roadmap while creating the work breakdown structure
- Apprehended transformation logics for individual data sources to design a data repository capable of capturing insights of user behaviors and interactions for investment assets
- Prepared the data modeling by creating **conceptual** and **logical diagram**, identifying the **entity relationship,** and designing database schema to build the data warehouse architecture
- Build **low fidelity and high-fidelity prototypes** like **mockups** and **wireframes** to analyze the user interface and present to primary stakeholders
- Utilized **Postman** to perform **API testing** to validate the request and response for **JSON based RESTful** methods like GET/POST to capture the response error codes and data issue
- Facilitated **UAT sessions** like alpha and beta testing ensuring the product quality with a thorough understanding of testing concepts pertaining to both manual and automation
- Designed the Intake Request Form Workflow to facilitate business process on ad-hoc requests
- Implemented visualization tools like Qlik Sense to build ad-hoc dashboard report

Trust Modernization Project

- Gathered, analyzed, documented, and transformed business requirements into technical requirement
- Participated in data modelling design for new application initiation
- Provided guide implementation by running insights analysis programmatically support new projects and initiatives
- Communicated and coordinated with coworkers for collecting data and performed ETL packages to define the standard process on data migration
- Retrieved, cleansed, and analyzed business' financial data by writing complex SQL queries to identify abnormality and assess the risks of misstatement of the financial statement
- Provided data quality validations for each pipeline resources by Python Script
- Assisted in testing of the application by running smoke testing, sanity testing, API testing, performance testing, created testing artefacts such as test cases and test scripts wherever necessary
- Implemented AI models like GAM (Generalized Additive Model), Hybrid GAM, Random Forest, XG boosting to improve the efficiency and accuracy of the application

Sentiment Analysis and Recommendation System Project

- Implemented ETL process and optimized SQL queries to perform data migration and data integration from AWS Redshift to Spark.
- Explored and analyzed customer behavior features by applying Spark SQL.
- Worked on data cleaning and ensured data quality, consistency, integrity using Python.
- Participated in features engineering such as feature generating, PCA, feature normalization and label encoding.
- Used MLlib to develop variety of models and algorithms for analytic purposes to identify a shopper's preferences to develop a descent shopping experience.
- Conducted analysis and patterns on customers' shopping habits in different location, different categories and different months by using clustering algorithms and classification models.
- Designed rich data visualizations to model data into human-readable form with Tableau

Technology Stack: Spark (MLlib), SSMS, Python 3.x, Machine Learning, Tableau

Ampla Health Data Warehouse Project

- Gathered, analyzed, documented and transformed requirements into data models.
- Designed Data Warehouse using snowflake schema and determined the type of dimensions with SQL Server Management Studio (SSMS).
- Developed SSIS ETL packages for processing fact and dimension tables and slowly changing dimensions to ensure the database platforms are configured and maintained to update record changes.
- Built up high performance ETL packages to migrate and manipulate data from various data sources such Flat, Excel and SQL Server by using SSIS.
- Applied various components of SSIS to ensure data quality and data consistency such as Pivot Transformation, Fuzzy Lookup, Derived Columns, Condition Split, Aggregate, Execute SQL Task and Data Flow Task.
- Generated various Data Marts which collected the interested fact tables and dimension table (Patients Info, Member Info, Claim Info and Transaction Info) from SQL Server.
- Optimized the performance of queries with modification in T-SQL queries, removed unnecessary columns and redundant data, normalized tables, established joins and created indexes.
- Explored and visualized claim, transaction and payment data to check the pattern, distribution, descriptive statistics and correlation using Python Matplotlib, Seaborn and Tableau.
- Built regression model including linear regression, lassso to predict the payment delay duration and applied feature engineering to measure which factors impact the payment delay duration.
- Implemented, tuned and validated model with the best performing algorithm and parameters.
- Integrated the geographic maps with the feature engineering results to delivery dashboards and analysis to provide solutions on long payment delay issues for clients.

Technology Stack: SQL Server 2017, SSMS, SSIS, Python 3.x (Numpy, Pandas, Scipy, Scikit-Learn, Matplotlib, Seaborn, etc), Machine Learning, Tableau

Opentrial System Project

- Collected requirements from business clients, and designed report models to meet demands.
- Identified and analyzed various data sources such as csv files, SQL database and other platforms to enrich reporting.
- Used SSIS to create ETL packages to Validate, Extract, Transform, and Load data into Data Warehouse and Data Mart.
- Worked on data cleaning and preprocessing to ensure the consistency and quality of datasets.
- Experimented with predictive models including Decision Trees, Random Forest, Logistic Regression, XGBoost to determine the patterns and trends in datasets.
- Designed and implemented Cross-validation and statistical tests including k-fold, stratified k-fold, hold-out schema to test and verify the model's significance.
- Designed, developed summary, trending and benchmark reports in Tableau.

Technology Stack: SQL Server 2017, SSMS, SSIS, Python 3.x (Numpy, Pandas, Scikit-Learn, Potly, Matplotlib, Seaborn, etc), Tableau

Stevens Institute of Technology, Hoboken, NJ **Fall.2018**

Undergraduate Teaching Assistant

- Directed the recitations and workshops of MA124 Introduction to Multivariable Calculus
- Conducted exams, grade papers and assisted with students in solving problems.

Academic Projects

Stevens Institute of Technology

Topic Trends Prediction in Biomaterial Research **Fall.2018**

Sponsor: John Wiley & Sons, Inc

- The datasets were extracted from Web of Science, which contains 43480 papers from 1972 to 2018.
- Exploratory data analysis and data visualization with Python plot packages and VOSViewer.
- Extracted target topics with natural language processing (NLP) method of TF_IDF filter and latent dirichlet allocation (LDA) model.
- Confirmed the impact of social media on indicating topic trends by using Multi Linear Regression Model.
- Predicted the topic trends based on the social media datasets using Time Series Model
- Systematized the objective and coordinated effectively with teammates

Amazon's Choice Analysis and Modeling

- The datasets were crawled and preprocessed from Amazon website by Python using Selenium package, which contains 20000 records of products.
- Applied Smote method to imbalanced datasets and feature engineering to multiple features.

- Modeled datasets with various machine learning classification algorithms including Linear Regression, KNN, Decision Tree, Random Forest and Grid Search via Scikit-Learn and successfully predicted the target tag in high accuracy.

University of Waterloo

Biosensor Development **Oct.2015 - Apr.2016**

- Ascertained and attributed a biosensor based on stabilized lipid bilayers in fluid situation with Localized Surface Plasmon Resonance device
- Quantified and recognized a high-quality lipid-based biosensor with several parameters
- Illustrated the topology of gold-lipid interaction by using Atomic Force Microscopy
- Preprocessed and analyzed data into mathematical models created by Origin
- Structured the experimental process and resolved measurements with partners