联系方式

www.linkedin.com/in/siyu-xiao-aa322611b (LinkedIn)

热门技能

Data Analysis
Machine Learning
Mathematics

Languages

English (Full Professional)
Chinese (Native or Bilingual)
Japanese (Elementary)

Siyu Xiao

Data Scientist at Edwards Lifesciences
Irvine, California

个人简介

I am currently working at Edwards Lifesciences as a data scientist. I have 6 years experience including academy and industry. I also did projects for GE Aviation, Viacom and so on. I am good at programming and mathematical theories about data science especially machine learning.
Software Tools: Python, R, SQL, MATLAB, HTML, CSS, C, Lisp, Tableau, Excel, SPSS, Spark, RapidMiner, Linux, Neo4j, Orange
Framework and Platform: PyTorch, TensorFlow, Keras, Scikit-learn, AWS, Azure, Cloudera, Databricks

---

工作经历

Edwards Lifesciences
Data Scientist
2019 年 9 月 - Present (7 个月)
Irvine, CA

• Scraped and cleaned FAQs from multiple sources in Python using Beautifulsoup and inserted into a SQL database
• Used Dash framework to develop a dashboard UI application for heart disease patients in Jupyter Notebook, Group by the count of different variables, write Callbacks functions to make interactive tables and graphs for users.
• Developed an Alexa skill to integrate the chatbot with Alexa and created lambda function using AWS toolkit to process the JSON data.
• Implemented Deep Learning LSTM using TensorFlow in python to build a chat bot system for patients.
• Calculated sentence similarity scores using word2vec embeddings and similarity measures from sklearn and NLTK to handle semantic and syntactic differences.
• Used different NLP similarity score functions (including word2vec, ngram, TFIDF, topic modelling) to match an input question with our target answers.

• Used GRU to predict sentiment with the accuracy about 90%. Input analyzed results with cURL or used py2neo as a RESTful API to input data into Neo4j database. Made an app about sentiment with R Shiny.
• Conducted A/B testing for the Edwards webpage and delivered simplified reports to senior managers weekly.
• Ensured high quality data collection and maintaining the integrity of the data. Designed and developed the UI of the website using HTML, CSS and Bootstrap.
• Designed and developed the data management system using MySQL.
• Used KNN and Decision Tree to detect breast cancer more efficiently, used z-score standardization or normalization to find nearest neighbors. With different k values, the highest diagnosing predictive Recall and Precision were around 95% and 92%.
• Predicted the power consumption based on hospitals real-time condition. Mined the data like ranking power-consuming activities. Used RNN in Keras from TensorFlow to predict the energy usage and applied Dropout to regularize the model with accuracy of 83%.

GE Aviation
Data Scientist
2019 年 4 月 - 2019 年 8 月 (5 个月)
Greater Boston Area

• Merged all the datasets into one by Primary Key, which made it easier to be analyzed. Processed outlier, missing and useless data.
• Drew the Entity-Relationship Diagrams (ERD) like Snowflake Schema with Normal Form in MySQL Workbench, made Exploration of Data Analysis (EDA) for the merged dataset, visualized the rank of indicators that trigger security alerts with Matplotlib.
• KL-divergence was used to check the weight among explanatory variables.
• Embedded and normalized the data, used Supervised Learning in sklearn including XGBoost, SVM and Logistic Regression to classify the dataset with 96% Recall and 95% Precision which helped GE to find the employees who had malicious intent.
• Performed preliminary risk analysis and implemented methods like Monte Carlo and Bootstrapping to estimate the Value of Risk for a malicious behavior.
• Applied Maximum Likelihood Estimation (MLE) to fit a function which can show the real data distribution.
• Analyzed and worked with all aspects of regression models (Least Squares, etc) and time series analysis.

• Worked with managers and directors to design solutions and strategies enhancing security monitor platform.
• Leverage information design concepts and principles to create compelling and effective charts, tables, presentations and other visuals using Python and Excel that convey analytical results clearly and effectively.
• Applied Microsoft Azure ML Studio to train all the 50 million rows of data with Logistic Regression plus Random Forest, drew ROC figure and calculated AUC with the result of 0.9.
• Coached and mentored new trainees and consulted struggling advisors to help them meet monthly target goals.

VitaData.io
Data Scientist / Machine Learning Engineer
2018 年 10 月 - 2019 年 3 月 (6 个月)
Cambridge, MA

• Performed data migration and developed Python / Django based web application, PostgreSQL Database and integrations with 3rd party email, messaging and storage services.
• Python Object Oriented Design code for manufacturing quality, monitoring, logging and debugging code optimization.
• Validated huge data and worked on python backend scripting.
• Automated the developed web application/portal and developed Python Automation Scripts using Selenium IDE.
• Analyzed CoinMarketCap website and customer data to identify market, product trends and profitable revenue growth opportunities using Python.
• Predicted the price of cryptocurrency with Regression Tree and Linear Regression in MLlib of Spark with accuracy around 75%.
• Redesigned market risk model originally implemented in R to use map reduce in Cloudera’s Hadoop cluster using unsupervised learning /principal components analysis.
• Used SQL to get tables from Teradata warehouse, randomly sampled data using RAND function.
• Performed logistic regression on newly selected variables to minimized information cross entropy and then delete the redundant ones.
• Visualized the moving averages of the cryptocurrency over the years to obtain the trends and estimate the growth of the companies using Seaborn in Python.
• Created LIFT probability table and GAIN chart.
• Created shared Object repository, Selenium Library Function, saved all components functions in Library Functions in Selenium library.

• Developed entire frontend and backend modules using Python on Django Web Framework.
• Used AWS for application deployment and configuration.
• Designed and developed the UI of the website using Python, HTML and CSS.
• Performed debugging and troubleshooting the web applications using Subversion version control tool to coordinate team-development.
• Created Python scripts to validate based on the keyword-driven testing and test cases.
• Developed for fully automated continuous integration system using Python and Bash scripting.

Citizens Bank
Data Scientist / Data Engineer
2017 年 9 月 - 2018 年 7 月 (11 个月)
Greater Boston Area

• Applied the Entity Embedding method to automatically learn the representation of structured data in multi-dimensional space, which is better than one-hot encoding and manually designed features based on experience.
• Applied RNN layer in PyTorch to process time-series data, applied dense layer to process fixed-length data.
• Merged dense layer and LSTM hidden layer after feature extraction, then used two dense layers to get output layer, calculated loss with Binary-Cross-Entropy. Learned parameters of all networks with Stochastic Gradient Descent.
• Quantitative analysis and software development using data sets forecasting Economic Capital models and Regulatory Capital models for managing risk-based capital for Citizens bank.
• Developed a web scraper to collect historical financial data of technology giants from Yahoo Finance using DataReader in Python.
• Worked with credit-risk models (PD, LGD, EAD) in use for loan applicants' credit risk.
• Involved in validating the Web Services related to Customer, Account and Transaction Management using the RESTful UI.
• Actively involved in Agile Methodologies and SCRUM Process and worked closely with different stakeholders to understand their system needs.
• Gathered loan data, made cross-validation and designed new credit evaluation policies, created statistical data models using Python, Excel and SQL which lowered bad debts by 5% for the personal loan segment.

• Collaborated with cross-functional stakeholders and senior managers to design credit check procedures that eliminated 15% of monthly customers at source who did not meet full criteria prior to loan underwriting process.
• Developed scripts in Python to automate the customer query addressable system using python which decreased the time for solving the query of the customer by 45%.
• Visualized results with Plotly, provided technical or analytical guidance as needed for issue management, project assessments and reporting.

Information Shield
Data Analyst
2016 年 9 月 - 2017 年 8 月 (1 年)
Beijing City, China

• Extracted and adjusted data frame from database with MySQL.
• Applied one-hot encoding and manually designed features to embed the original data into the data structure that can be input into our training models.
• Used K-means algorithm to cluster the market segments of Viacom social channels data, targeted ads to users who had the same interests, which improved 11% of ads effect. Visualized clustering data and ads effect with ggplot2.
• Calculated the Correlation Coefficient of each variable in Excel.
• Transformed original text into the words and visualized them with wordcloud package. Applied algorithms like Naïve Bayes in AWS Sagemaker to filter the ads with the Precision about 92% and Sensitivity about 93%.
• Communicated and presented default customers profiles along with reports using Python and Tableau, analytical results and strategic implications to senior managers for strategic decision making.
• Visualized analyzed results with Bokeh and Tableau and made presentation to managers and customers.

Harbin Bank
Business Analyst
2014 年 9 月 - 2016 年 6 月 (1 年 10 个月)
Heilongjiang, China

• Conducted detailed industry analysis, research, drafted reports and developed analytics insights on Small and Medium Sized (SME) enterprises.
• Served as a key strategic partner to uncover underlying business sector needs and information gaps.
• Coordinated with internal and external stakeholders to gather key compounding industry insights and proactively communicated industry news.

• Implemented approaches like Process Capability Analysis and Root Cause Analysis to determine the reasons for problems in food, mining and textiles.
• Maintained traceability among business requirements, technical requirements, design and testing.
• Assisted to build analytic tools to manage data and streamline data analysis using R and SQL Server.
• Created reporting documentation that identified metrics and data required for display as well as identification of filtering criteria and input.

教育经历

Northeastern University

Master's degree, Analytics · (2017 - 2019)

Harbin University of Science and Technology

Bachelor's degree, Electrical, Electronics and Communications Engineering · (2012 - 2016)