

Mark Ye

Data Scientist

Profile Summary

- Professional Data Scientist/Data Analyst with extensive work experience in Machine Learning/Deep Learning, Statistical modeling/Analysis, Data Mining, Data Validation and Data Visualization.
- Hands on experience in NLP tasks such as OCR, Sentiment analysis, Named entity recognition (NER) and Natural language generation. Experienced in NLP techniques like Bag of Words (BOW), TFIDF, Word2Vec, Hidden Markov Model (HMM), LDA, CNN, RNN (LSTM, GRU), 1D Convolutional Layer (Conv1D) for text analysis/ mining.
- Working knowledge in big data ecosystem, including Hadoop, (HDFS), MapReduce, Hive and AWS (EC2, S3, Amazon DynamoDB).
- Expertise in Data Visualization tools like Tableau 10.5, MS Power BI. Excellent Knowledge in using TensorFlow 8.0/12.0 for Deep learning (Neural Networks, Reinforcement Learning).

Domain Expertise

- E-commerce
- Healthcare
- Financial Services
- Technology
- Transportation

A Brief Summary of Experience

Client: TIAA-CREF

Responsibilities:

- Dealt with massive imbalanced datasets by different methods such as: oversample minority class, undersample majority class, generate synthetic samples.
- Implemented different statistical modeling including Random Forest, AdaBoost, CatBoost.
- Worked with large, complex data sets and developed machine learning models for financial program chatbot, like building digital assistants by using Natural Language Processing.
- Extracted information from the user-customer service text interaction (like sentiment analysis and topics) and add things like response time and user feedback right after the ticket.
- Experienced in spaCy, NLTK, chatterBot python library for Tokenization, text classification, rule-based matching and the Chatbot development.
- Participated in all phases of data visualization, delivered weekly dashboards to multiple levels of stakeholders with strong data transformation skills (Python/R/Tableau/MySQL).

Client: WizSolution Research LLC

Responsibilities:

- Applied Cross Validation for hyperparameter tuning on different classification models and compare the performance among different models. Validated the machine learning classifiers using ROC.
- Topic identification, Multi-label classification of printed media articles to topics, Automatic text summarization using NLP.
- Developing Virtual conversational assistant framework based on Python backend with MongoDB/NoSQL database.
- Engaged in developing a virtual AI chatbot for Chinese customers which will understand and reply in Mandarin and English language, with TextBlob python package.
- Experienced in dealing with Time Series data mining with RNN, KNN, LSTM, Conv1D using TensorFlow for forecasting.
- Used Git, TFVC for version control and Jira for bug tracking.

Certification and Knowledge

- Nanodegree Program, Natural Language Processing | Udacity
- A/B Testing by Google | Udacity
- Applied Data Science with Python by University of Michigan | Coursera
- Python, R, MySQL, Java, HTML5, CSS, Javascript
- Pandas, NumPy, Scikit-learn, Seaborn, Caret
- Matplotlib, Keras, TensorFlow, GGPLOT2
- forecast, tseries, MASS, rjson, plyr, glmnet
- www.yezhengkun.com