Zidong Xu

Washington, DC 20007 | (202) 644-2783 | zidong2020@gmail.com | LinkedIn | GitHub

EDUCATION

Georgetown University Washington, DC
Master of Data Science and Analytics (STEM field), GPA: 3.9 / 4.0 Expected Graduation: May. 2023
• Relevant coursework: Statistical Learning for ANLY, Neural Nets and Deep Learning, Big Data and Cloud Computing, Natural Language Processing, Time Series, Advanced Data Visualization, Digital Storytelling

Central China Normal University Wuhan, China
Bachelor of Information management, GPA: 3.61/ 4.0 (top 5%) Sept. 2016 – Jun. 2020
• Relevant coursework: Business Environment and Data Analysis, Principles of Database System, Website Design and Management, Analysis and Design of E-Commerce System, Basic Accounting, Marketing

SKILLS

Programming: **R, Python** (Pandas, Numpy, Scikit-Learn), Pytorch, Keras, tensorflow, spaCy, nltk, **SQL**, AWS (S3, EMR, EC2, Lambda, Cloud9, Docker), HDFS, Azure, Pysaprk, Hadoop, Linux, Google Colab, HTML
Machine Learning: Regressions & Classification Methods (Tree-based models, SVM, KNN, K-means, PCA)
NLP: Bag-of-words (BOW), TF-IDF, skip-gram & CBOW models, CNN, RNNs, Transformers (BERT, GPT)
Visualization: Tableau, Plotly, Seaborn, Matplotlib, ggplot2, d3.js

EXPERIENCE

Georgetown University Washington, DC
Graduate Teaching Assistant Jan. 2023 - Present
• Instructed a class of **50+** graduate students in a **Statistical Learning for ANLY** course for the Georgetown University Graduate School of Arts & Sciences (GSAS).
• Mentored students to use statistical techniques to find structure or patterns in given data sets (**unsupervised learning**) or use given data instances to predict outcomes in new cases (**supervised learning**).

Trip.com Group Limited Shenzhen, China
Air Ticket R&D Department, data analyst Intern July. 2019 – Aug 2019
• Extracted customer data from the database using **SQL**, preprocessing the data, and Built **machine learning models** (Logistic Regression, Tree-based models, Naive Bayes) and **Time series methods** to find important features and predict the customers' purchase decisions, ended up receiving an **81%** precision score.
• Visualized the results through **Tableau** and **PowerPoint** and showed them to the team and clients.

PROJECTS

Classification: Analyze Reddit data with Azure Sept. 2022 – Nov 2022
• Analyzed Reddit data **(1TB)** which focused on teenager-related subreddits, created new variables for analysis,
• Performed **exploratory data analysis (EDA)**, such as Word Cloud, histogram, box plot, correlation plot, and time series chart, to gain crucial general information. Also conducted **under-sampling** and **outliers detection**.
• Used **Regular expressions** (python's re library) and **NLTK** to preprocess text data. Also, used johnsnowlabs **sparkNLP** to build Pipelines, which include several stages, such as documentAssembler, tokenizer, normalizer, stop_words, stemmer and finisher. These pipelines convert text data into readable tokens and can be easily reused.
• Created a dummy variable as the label and used **Machine Learning Pipelines (ML Spark)** to build models and find the most important predictors. The prediction accuracy of the **Random Forest** model on the test data is **75.1%** better than the result of the **Logistic Regression** model, in which prediction accuracy is 70%. Used Feature Importance Plot and confusion matrix (heatmap) to show the results.

NLP: Text Classification on the DBpedia14 dataset Sept. 2022 – Nov 2022
• Conducted exploratory data analysis and Data preprocessing, solving **uneven distribution labels** of the data set.
• Build linear text classification models (the **perceptron, linear SVM, multinomial naive Bayes**). The best test accuracy was in the 92-93% range and then trained a **Feed-Forward Neural Network** with CBOW features, the accuracy of it achieved ~97%.
• Created batched inputs using Huggingface's DistilBERT tokenizer, Loaded the **pre-trained Distilbert** model, and implemented a training loop. This model achieved **>99% test accuracy** with minimal hyperparameter tuning and performed much better than the above models (Use Google Colab+, GPU).