Ming (Jerry) Gao

Tel: (201)-377-8408 | Email: minggao077@hotmail.com | Linkedln: http://www.linkedin.com/in/ming-jerry-gao

EDUCATION

New York University, ***Center for Data Science*** New York City, NY
Master of Science in Data Science Expected Graduation 05/2024
Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing with Representation Learning

Yunnan University, ***School of Mathematics and Statistics*** Kunming, China
Bachelor of Engineering in Data Science and Big Data Technology | GPA: 3.83/4 09/2018-06/2022
Relevant Coursework: SQL, Data Mining, Data Visualization, Big Data Exploratory Analysis, Big Data Preprocessing

SKILLS

- **Programming**: Python (Scikit-Learn, Pandas, Numpy, Matplotlib, Seaborn, etc.), R (dplyr, caret, ggplot2), SQL
- **Data Science:** Data Analysis, Machine Learning, NLP, Data Visualization, Statistics, Web Scraping, A/B Testing
- **Other tools**: PowerPoint, Excel, Jupyter Notebook, Pycharm, Rstudio

PROFESSIONAL EXPERIENCES

Henan Junyou Digital Technology Co., Ltd. Zhengzhou, China
Data Scientist Intern 02/2022-04/2022

- Compared the similarity between Jingdong and Taobao category names and matched the Jingdong category to the Taobao category with the highest similarity layer-by-layer using the **Alibaba Cloud word vector API interface** and **cosine similarity**
- Applied the **SMOTE algorithm** to process the imbalanced **4M+** data of Jingdong in a certain month
- Constructed the **logistic regression**, **Naive Bayes**, **AdaBoost**, and **XGBoost classifiers** to predict price reduction for products, and found the optimal model with high interpretability and high accuracy of 87.6%
- Provided the company with a **project example** as it was the company's **first project** to use machine learning for data analysis during the transformation process

RELEVENT PROJECTS

Bank Credit Card Customer Churn Warning Based on Multi-class Logistic Regression Model 02/2022-05/2022

- Applied the **K-prototype clustering** algorithm to cluster **10k+** credit card customers data, dividing customers of a bank into three categories: loyal, churn and potential churn
- Conducted **One-Hot encoding**, performed **feature engineering** based on the Extra-Trees model
- Applied One-vs-Rest method to construct a **three-classification logistic regression** model, achieving 99.67% accuracy
- Explored the factors affecting the churn status of different credit card customer groups and delivered **recommendations** on customer churning prevention and improvement of customers' loyalty

Research on the Accurate Discriminative Model for Heart Disease Patients 06/2021-07/2021

- Preprocessed **3k+** data, applied a C4.5 **decision tree** model, a **back propagation neural network** model, and a **logistic regression** model to construct different discriminant models for heart disease patients using **Python**
- Compared the 3 models and reached 73.33% accuracy via the logistic regression model, explored the factors affecting the likelihood of heart disease, which could improve the efficiency of predicting heart disease (**by 30%**)

Natural Language Processing (NLP) Text Classification 12/2020-01/2021

- Led a team of 3 to scraped **1k+** pieces Covid-19 news in Chinese using **Python**, and labeled the news texts manually
- Segmented the texts using **jieba** library, removed non-text characters and stop words
- Performed a **TF-IDF** model and a **PCA** model for text feature extraction
- Constructed **AdaBoost** models and **XGBoost** models and optimized models with fine-tuning hyperparameters
- Compared the models, found that a certain AdaBoost model was the optimal model for classifying the scraped news with accuracy of 80%

Predictive Analysis of Car Dealer A's Customer Churn Warning and Car Return to Factory 05/2020-06/2020

- Preprocessed **430k+** data, used **Python** to connect the **Microsoft SQL Server** to store the data, and constructed **discriminant analysis**, **logistic regression** and **k-nearest neighbor** models to predict customer churn
- Constructed **OLS Regression**, **Ridge Regression**, **Lasso Regression** and **Principal Component Regression** models to predict the car dealer A's car return situations to the factory
- Compared the models and applied the optimal ones, founding the churn rate of the car dealer A's customers as 32.81%
- Explored the specific factors affecting the customer churn and delivered **recommendations** on reducing customer churn rate