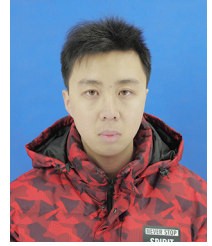


Zhong junyan

Guangzhou,Guangdong province (city)
510080 (postcode)

17665458796 (phone)
pzsx005@gmail.com (mail)
github.com/W55699 (GitHub)



EDUCATION

- **M.U.S.T** · *Applied Mathematics and Data Science* 2019.09 – 2021.08 M.S.
 - Numerical Linear Algebra, Applied Statistics, Machine Learning, Time Series Analysis, Data Mining, Database Systems
- **ECJTU** · *Vechile engnieering* 2015.09 – 2019.07 B.S.C
 - Mathematical modeling, Numerical calculation, Microcomputer principle, C++, Automatic control principle, Networke control system of the train

Skills and Research area

- **Computer skills:** Python, C/C++, MATLAB, \LaTeX , Java
- **Database management tools:**MySQL,Hive
- **Reasearch area:**Graph neural network, recommendation system application and security

Work experience

- **HSBC Technology** *Data science* 2022.07- 2023.03
 - I participate in the money laundering prevention prediction project.
This project mainly uses the user's transfer and consumption characteristics,buils a deep learning pipeline through tensorflow to predict the user's risk level offline and output the dashboard of risky user characteristics.
- **iPinyou.inc** *Machine learning engineer* 2021.02- 2022.06
 - I bulid the pipeline for data prepatation using the hive and sqoop tools.
 - I use python to build the machine learning pipeline for the the rebuy problem of car,including feature engnieering and model train and optimize.
 - I work with the aispeech inc to build the ASR(audio speech recognition) pipeline(use python and linux shell) to recognize the audio and analyze the result file.

Intern experience

- **TCL Industrial research lab** *Data mining intern* 2020.07-2020.09
 - I participated in the formulation of the business indicators of the recommendation system CTR through discussions with members of the group, and conducted statistical analysis based on the above indicators.
 - I use spark to complete data cleaning and verification of abnormal data.
 - I participate in the recommendation strategy discussion within the group, and participate in the classification of large-scale feature data (Random Fourier features SVM), clustering (minitach kmeans), and participate in feature selection and feature crossover work.
 - I participate in the daily maintenance of the crawler code in the group to enrich my own data mining experience.
- **AIATSS** *Data analysis intern* 2020.04-2020.6
 - I write SQL and python scripts to check the company's data.

- I use jira to monitor the progress of the workflow in real time, and draw a report on the progress of the test in the group through the Excel pivot table.
- I have a deeper understanding of the testing process and the ETL development process

Project experience

- 2017 MCM/ICM** *Captain* 2017.02
 - We use the analytic hierarchy process to carry out statistical analysis on the factors affecting the urban living environment, and at the same time analyze the factors affecting the urban living environment and the weight of the factors.
 - We use fuzzy comprehensive evaluation and gray prediction methods to predict the changes of various indicators.
 - We study urban sustainable development models, and after joint discussions within the group, we write the sustainable development papers to describe changes in various indicators.
- TaiDi competition** *Captain* 2018.03 – 2018.04
 - Firstly, we cleaned the original data set and used regular expressions to extract the names of TV programs in the original data and categorized the programs. We also implicitly score TV programs based on the user's viewing time and frequency.
 - Secondly, we constructed a user label system table and a product label system table, and performed user portraits based on the time characteristics of user viewing information and we matched programs and program categories to categorize the TV programs.
 - To better implement the recommendation, we use user-base collaborative filtering, item-base collaborative filtering, SVD and hybrid Recommendation algorithm. The SVD algorithm has the best effect.
 - We use K-means method to package users and products, and recommend users without any historical behavior, effectively solving the problems of user cold start and product cold start.
 - After the competition, we optimized the original method and tried to use Text-cnn to recommend the original data, which improved the accuracy of the recommendation.
- Douban web crawler data analysis** *Personal project* 2019.09 – 2019.12
 - Project website: <https://github.com/W55699/doubanbook-web-crawler>
 - I use regular expressions, and by creating a thread pool, multi-threaded to crawl Douban books information
 - I converted the information obtained by the crawler into csv file and stored it in the mysql database.
 - I use pandas to read csv, and do data visualization analysis and statistical analysis.
 - I use pca to reduce the dimensionality of the data, extract key information, and then use the k-means algorithm to mine similar types of book information.
 - I divided the data into training set and test set based on the information after reducing dimension, combined with the label of the data, and divided the data into two categories, and compared the pros and cons of various classification methods such as SVM, LR, decision tree, and random forest algorithm.
- IMDB sentiment analysis** *Personal project* 2020.10 – 2020.12
 - I use stop-words to clean the data set and use wordcloud to visualize the keyword.
 - I use word2vec to vectorize the text.
 - I used the bi-lstm model to classify comments with different emotions, and trained the model, and the accuracy of the model reached 85% in the test set.
 - I use docker, Tensorflow-serving, and streamlit to deploy the model on the web and complete the visualization of the model.

- IMDB sentiment analysis** *Personal project* 2020.10 –2020.12
 - I use stop-words to clean the data set and use wordcloud to visualize the keyword.
 - I use word2vec to vectorize the text.
 - I used the bi-lstm model to classify comments with different emotions, and trained the model, and the accuracy of the model reached 85% in the test set.
 - I use docker, Tensorflow-serving, and streamlit to deploy the model on the web and complete the visualization of the model.

- IMDB sentiment analysis** *Personal project* 2020.10 –2020.12
 - I use stop-words to clean the data set and use wordcloud to visualize the keyword.
 - I use word2vec to vectorize the text.
 - I used the bi-lstm model to classify comments with different emotions, and trained the model, and the accuracy of the model reached 85% in the test set.
 - I use docker, Tensorflow-serving, and streamlit to deploy the model on the web and complete the visualization of the model.

- GHAC car repurchase machine learning project** *Ipinyou.inc* 2021.02 –2022.06
 - The project mainly uses the characteristics of vehicle repairs and the characteristics of car owners on the official Weibo and WeChat to recommend car owners to repurchase.
 - Data preprocessing: Import data from database to Hive using sqoop. Then use hive SQL to preprocess the data (data cleaning and defining positive and negative samples)
 - Machine learning model training and prediction: use pandas to import data and do feature engineering. Then we use catboost, xgboost, logistic regression model to train the model and predict. At the same time, we adjust the output weights of each model based on the feedback we get.
 - Data output: Combined with business logic filtering, the top 0.5% prediction results of the model are output to the output database, and finally the ETL team sends the data to each 4s store. The ETL team will transfer these data to the 4S store to call customers for redemption promotion.

- Credit card financial risk analysis** *HSBC* 2022.07 –2023.02
 - This project mainly uses user characteristics and consumption characteristics to conduct risk ratings for HSBC credit cards.
 - Data preprocessing: use python, hive, and shell scripts to clean financial data, and use jenkins and git to manage CI/CD codes.
 - Data modeling: Use python to perform feature engineering on the data on the gcp platform, use tensorflow to build a tabnet model for prediction, and compare it with LR and Xgboost models.
 - Model output: output 1% of the highest risk users, and use seaborn to visualize their characteristics.

Awards

- **2017 ICM/MCM** *Honorable prize* 2017.02
- **The 6th TaiDi Data mining contest** *Third prize* 2018.04

Publications and patents

- Application code:201721495952.0 Bogie capable of monorail travel 2017.11
- Application code:201721495896.0 Networked miniaturized multifunctional clothes drying machine 2017.11
- Application code:201721495318.7 Shared bicycle area escort system based on laser detection and independent supply 2017.11
- JUNYAN ZHONG,HUIBIN WANG,KIN TAK U. Matryoshka Attack: Research on an Attack Method of Recommender System Based on Adversarial Learning and Optimization Solution,ICWAPR, 2020.
- JUNYAN ZHONG,HUIBIN WANG, Robust graph embedding recommendation against data posioning attack Big Data Intelligence and Computing: International Conference, DataCom 2022.