PENGFEI(ALLEN) XIAO

University of Melbourne, Parkville, Victoria, Australia
+61-(0)450868869 | pengfei123xiao@gmail.com
LinkedIn: https://www.linkedin.com/in/pengfei-allen-xiao-37449b104/

PROFILE

- A self-motivated executor with hands-on experience in data analytics who is passionate to derive insights from data to solve complex problems
- Solid programming skills: **Python**, **Java**, **MySQL**, and **R**
- Experienced in manipulating large data with **PySpark**
- In-depth knowledge in machine learning algorithms, intermediate experience in implementing machine learning models with **Python Pandas**, **NumPy**, **Scikit-Learn** and **TensorFlow**
- Specialized in data visualizations using **Python matplotlib**, **Seaborn** and **Tableau**
- A music lover and keen on piano

EDUCATION

University of Melbourne (UoM), Melbourne, Australia 02/2018-12/2019(expected)
Master of Information Technology (Distributed Computing)

University of Electronic Science and Technology of China (UESTC), Chengdu, China 09/2013-06/2017
Bachelor of Science in Electronic and Information Engineering

University of Glasgow (UoG) (joint program with UESTC) 09/2013-06/2017
Bachelor of Engineering in Electrical and Electronic Engineering (Second Upper-Class Honor Degree)

PROFESSIONAL EXPERIENCE

Singulariti, Melbourne, Australia 08/2019-11/2019
Data Engineer Intern

Responsibility

- Customise docker images and write YAML files for bitbucket CI/CD pipeline
- Assist in building data preprocessing pipeline via PySpark
- Assist in reforming a prediction model based on Convolutional Neural Networks to predict mining machine failure probability

Royal Melbourne Institute of Technology (RMIT University), Melbourne, Australia 11/2018-03/2019
Data Science Research Assistant Intern - Center of Information Discovery and Data Analytics

Responsibility

- Deployed an efficient and reliable data pipeline on Melbourne On-Street Parking Data for further analysis and visualization
- Proposed and realized a hybrid model that combines the time-series clustering model and Recurrent Neural Network model (LSTM) in predicting Melbourne parking area occupancy rate in the next 5 minutes, 15 minutes, 30 minutes
- Extracted and analyzed useful features, including nearby office/restaurant number, nearby restaurant business hour, public transportation condition, parking rules, etc., that influence the occupancy rate by analyzing the cluster distribution and the occupancy rate of different clusters
- Visualized experimental results via Python and Tableau

Achievement

- The hybrid model reduces the error rate by around 20% compared with traditional (non-cluster) regression model
- Summarized the experimental and analytical results and prepared an international research paper for IEEE Transactions on Intelligent Transportation Systems

PROJECT EXPERIENCE

University of Melbourne, Melbourne, Australia

Australia Politician support rate analysis via Twitter 03/2019-06/2019

- Applied Twitter's restful and streaming API to crawl political-related tweets in Australia from Twitter and stored the tweets in MongoDB
- Applied simple natural language processing techniques on tweets contents to analysis sentiments towards 280 politicians in Australia
  - In specific, applied TF-IDF, BoW, N-Gram methods to extract the features from Twitter contents, then trained Naïve Bayesian, SVM models with IMDB dataset, and finally predicted the sentiment of Twitter contents with trained models
- Visualized overall support rates and daily changes in support rates; visualized politicians' word cloud and daily popular hashtags
- Attempted to predict the 2019 Australian Federal Election results based on the above analysis and the accuracy is around 72% among 139 electoral districts

Kaggle Competition

Elo Merchant Category Recommendation - Bronze Medal (Top 7%) 02/2019

- Predicted user loyalty based on credit card records and therefore customised special offers and discounts for different users
- Motivated by the RFM (Recency - Frequency - Monetary) model, created some additional features.
  - The features are roughly divided into basic information features, time features, and amount features. They are filtered based on variance, Pearson correlation coefficients, etc.
- Built prediction models based on XGBoost, LightGBM, Multilayer perceptron and blended the prediction results of these three models

HONOURS

Community Service Scholarship, UESTC 2014-2015

Excellent Student Cadre, Student Union, UESTC 2014-2015

Second Class People's Scholarship in successive three years, UESTC 2013-2016

EXTRACURRICULAR ACTIVITIES

Student Representative of INFO90002, UoM, Melbourne, Australia 02/2018-06/2018

Class Representative

- Collected students' feedbacks about the teaching mode and other course-related questions by doing survey, interviews and focus groups and then reflected feedbacks

International Exchange Office, UESTC, Chengdu, China 10/2014-06/2015

Cultural Exchange Ambassador

- Facilitated the visits of foreign campus guests, including sharing university's history and helping them with necessary arrangements
- Arranged the receptions for the professors from The Institution of Engineering and Technology (IET) and organized academic seminars with the professors from UESTC

Volunteer of the 2011 Summer Universiade, Shenzhen, China 08/2011

Volunteer

- Offered volunteer services to pedestrians and foreign visitors in a U-Station