Tong Liu

Los Angeles, CA | 213.322.7621 | liu553@usc.edu | https://www.linkedin.com/in/tong-liu-797497140/

EDUCATION

University of Southern California (USC) *Los Angeles, CA* Jan 2018 - Dec 2019

- Master of Science in Data Informatics, GPA 4.0
- Coursework: Database Management, Machine Learning, Data Mining, Natural Language Processing

Beijing University of Posts and Telecommunications (BUPT) *Beijing, China* Aug 2013 - June 2017

- Bachelor of Engineering in Computer Science, Graduated with Honors (5%)

TECHNICAL SKILLS

Programming Languages: Python (sklearn, NumPy, pandas, seaborn, SciPy), C++, Shell

Tools: Hadoop, Hive, MySQL, NoSQL (Firebase), Jupyter Notebook, QT, MongoDB, AWS, Spark, Tableau

Analysis Techniques: Supervised Learning (SVM, Linear Regression, XGBoost, Random Forest, FM), Unsupervised Learning (K-means, KNN, hierarchical clustering, LSTM), Penalized Regression Methods (Lasso, Ridge), PCA (Principal Component Analysis), Regularization, Model Evaluation

EXPERIENCES

ByteDance - Data Science Intern May 2018 - Aug 2018

- Worked for *toutiao.com* car channel and *toutiao* APP. Modified on-line and off-line recommendation models for advertisements, improved advertisement push, banners, and SMS performance.
- Wrote HiveQL and SQL to gather original data, preprocessed data set by data cleaning, categorical feature transformation and normalization.
- Enriched user-profile features and optimized user-profile update strategies to supply reliable features for machine learning models.
- Trained XGBoost and Logistic Regression model, increased AUC from 0.67 to 0.8. Did data resampling and undersampling for imbalanced data, imputed missing values, evaluated model by recall and precision.
- Built MapReduce ETL pipeline on Hadoop, which could automatically fetch original data in HDFS, generate instances, calculate feature and data coverage rate, trained machine learning models, calculated model evaluation indexes, recorded debug log.
- Designed A/B test to evaluate model performance. Increased push service CTR from 3% to 5.5%, ROI for SMS advertising from 8 to 12.

PROJECTS

Kaggle Challenge: Toxic Comment Classification Feb 2018 - May 2018

- Build a multi-headed model detecting different types of toxicity. Used published Wikipedia comments data which was labeled by human raters.
- Implemented TF-IDF and Glove to convert words to vectors for data preprocessing.
- Fitted several supervised and unsupervised models including Random Forest, LSTM(Long Short Term Memory), Logistic Regression, XGBoost, LDA (Latent Dirichlet Allocation), and CNN (Convolutional Neural Network).
- Evaluated model performance of classification via k-fold cross-validation technique and analyzed feature importance to identify top factors that influenced the results, achieved 98% accuracy rate.

San Francisco Crime Analysis in Apache Spark May 2018 - Aug 2018

- Studied real crime data to offer trip advices for citizens and give patrol guidance for police officers.
- Implemented Spark SQL to analyze data. Performed spatial and time series analysis for a 15 years dataset for reported incidents from SFPD.
- Trained and fine-tuned an ARIMA model with Spark MLlib to forecast the number of theft incidents per month.
- Visualized variation of the spatial distribution of the incidents over time.