**­­**

**Wei Liu**

https://github.com/wliu40 | 315-560-8306 | wliu0828@gmail.com

**Summary of Skills**

• Proficient in Python/pyspark, Familiar with R, C++, Linux

• Complete knowledge of statistics, data structures, algorithms and object-oriented programming.

• Proficient with data cleansing and engineering technique using python libraries and Spark.

• Complete knowledge of Big Data technologies (Spark, Hive, HDFS, MapReduce)

• Proficient with computer vision/deep learning: object detection/semantic segmentation/image classifications.

• Proficient with SQL/NoSQL (MySQL, MongoDB).

• Hands on experience on AWS machine learning.

• Passion of learning new technologies in ML/DS/AI area, MOOC enthusiast.

**Work Experience**

**Data Scientist @ Norfork Southern Railways Corp.** Jan.2018-present

**Projects in Big-data analytics with Spark and python:**

• Anomaly detection of VSWR/Losing cooling water/Losing Lubricant oil of locomotives.

• Rail-wear prediction.

• Rail-defects matching and report.

Brief introduction of my work and contributions:

Retrieving data from HDFS using Hive/Spark, applied/design different ML algorithms, include SVM, Linear Regression,

K-means, Time Series decomposition/forecasting, e.g.,

Anomaly detection with one-class SVMs/RNN (LSTM)/autoencoders/twitter anomaly detection (Locomotive Anomaly Detection projects).

Feature dimension reductions, multicollinearity analyze and high dimension visualizations with t-SNE.

Big data geographical matching with KD tree design (improved the matching efficiency by order of magnitude.

Change-point Detection and outliers cleansing (Railwear/Defects Predictions projects) and etc.

**Projects in Computer vision and deep learning:**

• Intermodal container OCR recognition, Railway tie-plate and spikes recognition

• Rail defect/jointbar/frog/heel-block/weld/diamond detection

• Car-safety inspection (missing cotter-key detection/missing bolts detection)

Brief introduction of my work and contributions:

Apply and developed deep-learning algorithms for image semantic segmentation/object detection/image classification using FCN, Unet, EAST, YOLO-v3, VGG16/Mobilenets-v2/Resnet/etc. models with Tensorflow, keras, cv2, e.g.,

Handle the extreme imbalance data by various augmentation methods.

Increase the development efficiency with auto-labeling methods.

Enhance the model performance with boosting sampling.

**Tools and libraries used:**

•Hive, pyspark, numpy, pandas, stats models, scikit-learn, tensorflow, keras, opencv, folium, matplotlib and etc.

**Other job responsibilities:**

•Project feasibility research, Data exploration/cleansing, Ad-hoc analysis with big data, peer code-review, QA analysis,

Apache NIFI flow design and code version control(git)/documentation/productionization.

**EDUCATION**

**Degree:**

Syracuse University, College of Engineering & Computer Science 2017

***Master of Science*** *in Computer Science*

***PhD*** *in Mechanical & Aerospace Engineering*

Chongqing University 2009

***Bachelor*** *in Thermo-physics* Engineering

**Certifications:**

Nano Degree in Deep Learning (Udacity)

Deep Learning Specialization (Coursera)

**AcaDEMIC aND OTHER Projects**

**Humpback Whale Identifications**

•With the stacking (SVM) on eight different DL classifiers (in keras) to boost the overall performance.

**Fraud Detections in Financial Payment Services**

•Handle the extreme skewed data, build/compared different models of Logistic Regression, KNN, Random Forest, XGBoost and Light GBM.

**Credit Cards Fraud Detection**

• Three different models of SVM/LightGBM/Random Forest were built and tested. LightGBM with best AUC of 0.97 was achieved.

**Bitcoin Price Prediction**

• Build LSTM/XGBoost/ARIMA models to predict the Bitcoin prices.

**Recommendation Engines**

•Built movie recommendation engine with: 1) memory based collaborative filtering algorithms, 2. Latent factor matrix factorization model with SGD optimizer & L2 regularization, 3. ALS matrix factorization method in Spark.

**Natural Language Processing**

•Text classification (20newsgroups) with sklearn(NB, SVM) and Spark (NB). The highest accuracy of 97.6% was achieved by multinominal NB model in Spark.

•Built LSTM RNN models with Tensorflow and generated Simpson TV scripts with existing scripts as training dataset.

•Built a text generator with LSTM-RNNmodel which was trained by popular Chinese novels.

**Amazon AWS machine learning and Cloud computing**

•Trained models with default & customized recipe. Performed ML model quality evaluation with confusion matrix, interactive cutoff threshold adjustment and other hyper-parameters.