

WENKANG WEI

Current Address: Central, South Carolina, 29630 || **Phone:** 864-324-4885 || **Email (Preferred):** wenkanw@g.clemson.edu
Tech Blog: wenkangwei.github.io || **LinkedIn:** www.linkedin.com/in/wenkang-wei-588811167/ || **Academic Profile:** meritpages.com/wenkan

SUMMARY

I'm a second-year master student with 5 years of experience in programming and 3 years of experience in Machine Learning. I'm actively looking for a full-time job or summer internship of **Data Scientist**.

EDUCATION

Clemson University, SC

- Master of Computer Engineering, Minor of Computer Science Expected in May 2021 *GPA 3.79/4.0*
- Bachelor of Electrical Engineering May 2019 *GPA 3.84/4.0*

TECHNICAL SKILLS

- **Programming:** Python (PyTorch, Tensorflow, Numpy, Pandas, Sklearn, etc), PostgreSQL, MatLab, C/C++, Markdown
- **Software Toolkits:** Linux, Git/Github, Docker, GCP (Google Colab), AWS (EC2, RDS)
- **Distributed Computing and Data ETL:** Apache PySpark, Hadoop MapReduce, Databrick Distributed Cluster
- **Data Analysis and Visualization:** Seaborn, Matplotlib
- **Feature Engineering:** Image processing (Mixup), Natural Language Processing (Word Embedding, TF-IDF), PCA, One-Hot Encoder, etc
- **Machine Learning:** Recommendation System (Matrix Factorization, etc), Deep Learning (CNN, RNN), Classification, Regression, Clustering
- **Model Evaluation and Improvement:** Cross-Validation, ROC, AUC, Feature Importance, Ensemble Learning, etc.
- **Statistic:** Hypothesis Testing (T-test, Chi-Square, etc), A/B testing, Bayesian Theorem

PROFESSIONAL EXPERIENCE

Machine Learning Research Assistant

Clemson University, Summer 2020-Current

- Proved convergence and convergence rate of Multiple Update Algorithm (MUA) in Non-Negative Matrix Factorization Problem
- Formulated Matrix Factorization Problem into Constraint **Optimization** Problem, simplified problem by Linear Algebra, Lagrange multiplier
- Utilized Lipschitz gradient, convex optimization to prove the convergence and convergence rate of MUA algorithm
- Implemented MUA and **ALS (alternative least square)** algorithm in Google Colab and Matlab to verify convergence results
- Wrote a paper in AAAI format using **Latex** (unpublished due to copyright)

Team Leader in Kaggle Competition: Cassava Leaf Disease Image Classification

Kaggle, Fall 2020- Current

- Lead 2-person team to build a multi-task image classification system to classify cassava leaf diseases via a noisy cassava dataset from real world
- Construct data pipeline in **PyTorch** to extract and load Cassava leaf disease image dataset (5.76GB compressed data)
- Leverage data reduction methods to analyze data distribution and applied Image augmentation (cutout, mixup, etc) to transform images
- Apply and tune **efficient-net** and **visual transformer** models in multi-task classification task with label smoothing, early stopping, weight decay and improve model accuracy in public score by 3% using ensemble learning
- Achieve **0.905** accuracy in public score and rank **top 3% out of 4000 teams** in kaggle leaderboard currently

Leader in Human Activity Time Series Data Collection (20GB) and Analysis

Clemson University, Fall 2020-Current

- Write tutorial documents in Markdown in Github and coach **10 students** to collect wrist motion data from daily life to analyze eating behaviors
- Mentor and assist each student to collect, clean and label 2GB individual data in 2 weeks and transform time series data for data wrangling
- Construct robust data pipeline to solve buffer overflow problem to extract, load and transform **large-scale time series dataset (20GB) in 1 min**
- Visualize and analyze imbalanced data with seaborn and smooth data using moving average for data augmentation using Pandas, Numpy
- Build Convolution Neural Network to classify and segment eating period and achieve the **best weighed accuracy 96%** in cross validation

SELECTED RELATED PROJECTS

Youtube Comments Analysis and Pet Owners Classification (PySpark, SQL, Databrick Cluster)

Fall 2020

- Utilized **PySpark** and **PostgreSQL** to load, query and explore Youtube comment text data to classify if user is owner of dog or cat (about 1GB)
- Built data pipeline and applied Term-Frequency-Inverse Document-Frequency(**TF-IDF**) to transform text data into numerical data
- Applied Logistic Regression, Random Forest, Gradient Boosting machine in PySpark to classify cat or dog owners from comments
- Achieved **92%** prediction accuracy on test set by using grid search and cross validation to select the best model

Bank Customer Churn Prediction on Kaggle Bank Customer Dataset (Python, Sklearn, Git)

Fall 2020

- Visualized and analyzed bank customer dataset by using visualization toolkits: **seaborn, matplotlib**
- Preprocessed and transformed categorical data for machine learning model training using **pandas** toolkit and normalization techniques
- Established data pipeline and ML models like Random Forest, Logistic Regression, SVM, and evaluated models using ROC,AUC
- Improved Models Accuracy from **80% to 86%** by model selection, cross validation and feature selection, L1 Regularization techniques

California Housing Analysis and Prediction (Python, Sklearn, Git)

Summer 2020

- Explored California housing dataset from Statlib CMU repository to predict housing price using **pandas, seaborn** toolkits
- Utilized Inter-Quantile-Range(IQR) method to remove outliers and transformed data by normalization and one hot encoding
- Leveraged **sklearn** to build data pipeline to preprocess data and apply regression models: Linear Regression, Polynomial Regression, KNN
- Applied **Hypothesis Testing, F-test statistic** and **R2-statistic** test to measure and analyze contribution of features for feature selection in linear regression and achieved **9% improvement** (from 0.55 to 0.62) on R2-score on test set

IMDB Movie Rating Positive/Negative Sentiment Classification (NLP, Tensorflow)

Summer 2020

- Extracted IMDB movie rating text dataset (1.4GB) using BeautifulSoup and cleaned data by stemming, removing stop words
- Applied **Word Embedding, Bag of Word** model, **TF-IDF** Techniques to transform text data into different representations for model training
- Designed Convolution Neural Network in **Tensorflow** and applied ML models (SVM, Random Forest, etc) for classification
- Evaluated model performance and achieved model test accuracy **88%**

HONOR

Eta Kappa Nu (HKN) (Spring 2018 - present); Golden Key Honor Society (Fall 2017- present); President's List (Summer 2019);
Dean's List (Fall 2016 – May 2019); Best overall hack award of CUhackit Competition (Spring 2018);