Meixian Wu

Cell: (858) 866-6591 | mxwu86@gmail.com | Los Angeles, CA 90017 LinkedIn

SUMMARY

Master student who aims at finding data analytics opportunities, with sufficient project experience in **visualizations**, **machine learning and data analytics**. Strong knowledge of Statistics and solid programming in **Python and SQL**.

EDUCATION

University of Southern California | Viterbi School of Engineering *Los Angeles CA* 2021- Expected May 2023
Master of Science in Applied Data Science GPA*: 3.65/4.0*
Coursework: Cloud Database, Deep Learning, Relational Modeling, Storage System, Spark RDD, Hadoop, Machine Learning

University of California, San Diego | Jacob School of Engineering *La Jolla, CA* 2017- 2021
Bachelor of Science in Bioengineering: Bioinformatics (with honor) | Minor in Statistics GPA*: 3.85/4.0*
Coursework: Machine Learning, Statistical Analysis, Data Science in Practice, Data Structure, Algorithm Design

SKILLS

Programming: Python (sklearn, pandas, numpy), SQL/MySQL, Spark, AWS, Java, MongoDB, advanced Excel
Machine Learning: Exploratory Data Analysis, Visualization, Classical & Penalized Regression Methods (Lasso, Ridge), Decision Tree, Regularization, Clustering, K Nearest Neighbors, K-means, Principal Component Analysis (PCA)
Statistical Analysis: Hypothesis Testing, Time Analysis, Association Rule, Bayesian Classification, Metrics, Cross Validation

WORK EXPERIENCE

J. Craig Venter Institute La Jolla, California

Research Intern July 2020 - June 2021

- Studied the techniques to represent high dimensional data through transformation to work with machine/deep learning pipeline.
- Designed experiment to investigate the possibility to make reasonable diagnostic prediction of Leukemia by using CNN to study the imagified high-dimensional patient sample data.
- Built model for CNN learning pipeline with GridSearch parameter tuning; and concluded that CNN is possible to make acceptable prediction studying the sample's distribution.
- Leveraged the performance and computational resources needed to reach final recall of 88%.
- Overcame the limitation of the small-size dataset by data augmentation through the templates of dimension reduction UMAP.

PROJECTS

Customer Churn Prediction in Telecommunication Industry

- Developed algorithms for telecommunications service vendors to predict customer churn probability based on scaled data via Python programming and Apache Spark.
- Preprocessed dataset by data cleaning, categorical feature transformation and standardization, etc.
- Trained supervised machine learning models including Logistics Regression, Random Forest, and K-Nearest Neighbors, and applied regularization with optimal parameters to overcome overfitting.
- Evaluated model performance (F1 scores 0.94) of classification via K-Fold cross validation techniques and identified top factors that influenced the churn probability using Random Forest, including age, estimated salary, and credit scores that scored 0.238, 0.148 and 0.143 respectively.

IEEE-CIS Credit Fraud Detection

- Cleaned and preprocessed 590,000+ transactions with 40+ features through exploratory analysis; spotted severe disproportion between normal and fraud transactions data, and handled by under sampling true transactions.
- Trained with three boosting algorithms: LightGBM, XGBoost and CatBoost; Tuned parameters with Bayesian Optimization and evaluated models using AUC metric
- Tuned final prediction by resembling the lowly correlated models, improving accuracy to 93%, ranked 17% in Kaggle competition

Topic modeling with Natural Language Processing on Product Review Dataset

- Clustered customer reviews into groups and discovered that latent semantic structures suing Python.
- Preprocessed review text by tokenization, stemming, removing, stop words and extracted features by Term Frequency -Inverse Document Frequency (TFIDF).
- Trained unsupervised learning model of K-Means clustering and Latent Dirichlet Analysis.
- Identified latent topics and keywords of each review for clustering. Keywords that dominate the top clusters are words about price, quality, purpose, appearance respectively.
- Visualized model training results by dimension reduction using Principal Component Analysis (PCA).