Zehao Li

zehao.li.1997@gmail.com || (510) 598-9373 || linkedin.com/in/leo-zehao-li

SUMMARY

Highly-motivated Master student looking for data science related job opportunities, with sufficient background knowledge and project experience in **Statistics, Machine Learning, and Data Analytics**. Solid programming skills in **Python and SQL**.

EDUCATION

University of Southern California Los Angeles, CA Expected May 2023
Master of Science in Applied Data Science
Coursework: Principles of Programming, Machine Learning

Johns Hopkins University Baltimore, MD September 2020 - July 2021
Master of Science in Business Analytics and Risk Management GPA: 3.2/4.00
Coursework: Data Analytics, Big Data Machine Learning, Data Visualization

University of California, Santa Barbara Santa Barbara, CA September 2016 - June 2020
Bachelor of Science in Statistics and Data Science GPA: 3.33/4.00
Coursework: Probability and Statistics, Design of Experiments, Regression Analysis, Stochastic Process, Time Series Analysis

SKILLS

Programming: Python, SQL, R, SAS, C++
Technology: MySQL, Spark, PyTorch, Hadoop, Tensorflow, Keras, Tableau, AWS, Github
Machine Learning Techniques:

- Linear Regression, Logistic Regression, Decision Tree, Random Forest, KNN, Gaussian Mixture Model (GMM), FastMap, SVM, Gradient Boosting Decision Tree, CNN, RNN, LSTM, GRU, Natural Language Processing, Transformer
- PCA, K-means Clustering, Model Evaluation, Latent Dirichlet Allocation (LDA), Regularization, Cross Validation
- Exploratory Data Analysis, A/B Testing, Hypothesis Testing, Experiment Design

PROJECTS

Customer Churn Prediction in Banking

- Developed algorithms and built machine learning models to predict bank customer churn probability based on labeled user data by using Python programming.
- Understood the dataset by performing data exploration analysis, and also got insights into possible features of customer churning.
- Preprocessed the raw data set by data cleaning, transforming categorical features, and normalization.
- Trained supervised learning models such as Logistic Regression, K-Nearest Neighbors, and Random Forest with 5-fold Cross Validation and Regularization applied to find optimal parameters and prevent overfitting.
- Evaluated models performance (Accuracy: 0.86, AUC: 0.85) and explored Feature importance ranking to find out leading factors that affects the results such as Age and Estimated Salary.

Natural Language Processing and Semantic Analysis on E-commerce Product Reviews Dataset

- Clustered unlabeled customer reviews into different groups and explored their latent semantic topics by using machine learning models in Python programming.
- Used tokenization and stemming with Natural Language Toolkit (NLTK) in Python to preprocess review texts.
- Extracted features and transformed them into numerical data by using Term Frequency - Inverse Document Frequency (TF-IDF).
- Identified latent topics of each customer review by training unsupervised learning models of 5-cluster K-Means Clustering and Latent Dirichlet Allocation (LDA).

Car Images Classification using Deep Learning

- Trained and fine-tuned a Convolutional Neural Network (CNN) model in PyTorch using a pre-trained model with transfer learning.
- Augmented images data with random transformations in PyTorch transform library to reduce model overfitting.
- Accelerated training process by running on GPU.
- Deployed the model to predict a new image and obtained an accuracy of 0.99.

Stock Prices Prediction based on Deep Learning

- Performed a time series prediction to predict future stock price using a Recurrent Neural Network (RNN) regressor in PyTorch.
- Cut the time series into sequences and treat the time series prediction problem as a regression problem in order to apply RNN.
- Deployed RNN model using the network architecture of Long Term Short Memory (LSTM)
- Trained LSTM model using Adam Optimizer and MSELoss function via PyTorch on GPU.
- Deployed the built model (RMSE: 6.57) to predict the variation of future stock price.