FRANK XU

frankxu0124@gmail.com franknb.medium.com 9193083187 Dallas, TX frank-xu-huaze/ franknb

SUMMARY

Data scientist and Machine Learning Engineer with 4 years of experience in data analytics and machine learning, optimized billions of dollars of advertising spend for one of the clients with cloud-based data engineering & modeling pipelines, built models to predict millions of driving dynamics using real-time vehicle data for major automotive manufacturer. Data Science content creator on TowardsDataScience.com with 100K+ views.

EXPERIENCE

Data Scientist & Machine Learning Engineer, Klaviyo, Boston, MA — Oct. 2022 - Current

- Developed transformer-based text classification models to automate the reviewing process of transactional messages sent, providing interpretations using Shap, achieving a 99% accuracy and saving 30% labor for security team.
- Developed an identity-matching algorithm for incoming Salesforce leads and reduced approximately 25% of duplication leads, alleviating hundreds of hours of manual work from the previous process.

Data Scientist, Credera, Addison, TX — June 2020 - Oct. 2022

- Built models to predict vehicle driving dynamics for one of the world's largest automotive manufacturer. Leveraged real-time data from vehicle head units and constructed time-series classification models using LightGBM/PySpark ML with AWS EMR.
- Built PySpark based ETL pipelines with AWS, enabling a streaming process from data cleaning, data modeling to data reporting using SQL, Python, R & Power BI, optimizing billions of dollars of advertising spend annually.
- Trained a transformer-based deep learning pipeline with PyTorch, and built a text summarization application which utilizes both Extractive & Abstractive algorithm that generates text summarizations from Internal Microsoft Stream Videos.

Machine Learning Engineer, Trade Pending, Carrboro, NC — Aug. 2019 - Apr. 2020

- Handled the 72-million rows dataset and deployed dashboards used Google Cloud Services, built Neural Networks and Random Forest, provided 92% accuracy on predicting vehicle transaction prices.
- Enabled the product to make viable predictions when selected data points are scarce, bringing at least 5% potential profit increase for the product.

Data Science Intern, Wells Fargo, San Francisco, CA — June 2019 - Aug. 2019

- Analyzed customer data using Random Forests and Gradient Boosting Machines to recommend banking products.
- Researched on 5 different methods of interpreting ensemble tree models including Shap and LIME, bringing more accurate decisions and potential profits for the team.

EDUCATION

Duke University — 2018 - 2020
Master in Interdisciplinary Data Science

Fudan University — 2014 - 2018
Bachelor in Environmental Science

PROJECTS

Realistic Face Images generated from Sketches — Jan. 2021 - Feb. 2021

- Implemented the DeepFaceDrawing-Jittor model that's consists of Component Embedding, Feature Mapping and Image Synthesis (GAN) modules.
- Created a docker image in DockerHub for a faster and portable deployment of Ubuntu based Jittor environment.

Comparing Self-attention GAN with DCGAN — May 2020 - June 2020

- Built Deep Convolutional Generative Adversarial Networks with and without self-attention layers.
- Researched and applied the non-local self attention modules.

Exploring methods of Feature Interpretation — Feb. 2020 - Mar. 2020

- An article exploring new methods of feature interpretation for ensemble tree models available on Python and R.

Kaggle Competition: Identifying Solar PV in Aerial Imagery — Jan. 2019 - Feb. 2019

- Used a Convolutional Neural Network built based on MobileNet to achieve a 97.8% accuracy on identifying solar panels in aerial images and ranked 10% tier.

Analyzing Text Data for Real Talk — Sept. 2018 - Feb. 2019

- Real Talk is a mobile app available on iOS facing teenager users.
- Ran text & regression analysis with R on a dataset of around 1400 stories, unveiled the connection between teenager psychometric status with demographic information.

SKILLS

Python, R, PyTorch, TensorFlow, PySpark, Scikit-learn, SQL, AWS, Google Cloud Platform, Docker, Azure, React