Qin Yang

+1 (201) 565-5998 | New York, NY | qy692@nyu.edu

New grad with broad-based experience in building data-intensive applications, overcoming complex architectural, and scalability issues in diverse industries. Proficient in predictive modeling, data processing, and machine learning algorithms, as well as scripting languages and version control tool, including python, SQL, and Git. Capable of creating, developing, testing, and deploying highly adaptive diverse services to translate business and functional qualifications into substantial deliverables.

EDUCATION

New York University, Center for Data Science **New York, NY**
Master in Data Science Expected May 2023
GPA: 3.6/4.0
Relevant Coursework: Applied ML in Finance, NLP and Computational Semantics, Deep Learning, Machine Learning, Computer Vision, Time Series Analysis, Probability and Statistics, Big data

Shandong University, School of Management Science and Engineering **China**
Bachelor in Management Science and Engineering/Master in Econ Sep 2014 – Jul 2021
GPA: 3.8/4.0
Relevant Coursework: Advanced Micro and Macro Economics, Econometrics, Finance, Calculus, Linear Algebra, Data Principles and Applications, Python and Data Analysis, Database, C++

EXPERIENCE

Data Science Department, Campana Schott Inc. May – Sep 2022
Position: Data Science Intern

- Bayer: Product Sales Forecasting *(with NLP + Time Series Models)*

- Identify key drivers for a downshift in product sales by novel analysis of both consumer market dynamics and mobility data from internal and external sources, determined if sales can be expected to normalize over time.
- Analyzed >2m mobility data points over 3 years as a proxy for consumers' social interactions, scraped >5m Tweets to measure and plot the sentiment of products and analyze statistical relationship to company sales, and implemented NLP and feature importance techniques to understand the explanatory value of large-scale product reviews and opinions of target consumers.
- Ran the Facebook Prophet time series forecasting model to capture the general growth trend, seasonalities, and holiday effects of our products as benchmark results; adopted historical sales data to train deep learning model N-HiTS for capturing the variation of sales and making forecasts at different horizons and achieved 84% explanatory power of product sales when forecasting at a 24-week horizon.
- Present solutions to core clients, leadership teams, and forum audiences.

- Radius Health: Extraction of Customer Satisfaction Topics Regarding Product Effectiveness *(with NLP)*

- Utilized Zero-shot learning and Few-shot learning with Hugging Face pre-trained models in extracting consumers' post-purchase reactions from unstructured texts, assessed opportunities for service or product upgrade.
- Developed a real-world dataset by crawling 360,000+ users' reviews from internal and external websites, analyzed and annotated a set of reviews on a sentence level with different labels (e.g., beneficial effects vs. adverse effects).
- Fine-tuned pre-trained sentence classification models from Hugging Face (e.g., BART-large) on the manually labeled dataset, achieved 90% accuracy and delivered our research outcome in the form of a 20-page report and presentation to core stakeholders.

R&D Department, IFLYTEK Jun – Sep 2019
Position: Data Analyst

- Developed a nationwide retention program with Python, SQL, and Excel, saved 1000+ hours of labor per year.
- Identified procedural areas of improvement through 10+ million customer data, queried and analyzed data from department database system using SQL, and created 20 dashboards and presented findings to 10 stakeholders.
- Trained Linear Regression model, predicted repair costs for vehicles on the market, and increased the profitability of a nationwide retention program by 8% (~100 million).

PROJECT

Improve Speech Recognition Performance with Unpaired Audio and Text Data *(with Hugging Face Transformer and Pytorch, ongoing)*

- Fine-tune Hugging Face pre-trained ASR models on 360+ hours of audio dataset to boost speech recognition performance on highly accented and disfluent data.
- Analyze large audio datasets, measure, and benchmark model performance, as well as present model accuracy with visualization tools (e.g., Tensorboard, W&B).

Adversarial Learning on Neural Network Models for Text Classification *(with Hugging Face Transformer)*
Link to the research paper:https://drive.google.com/file/d/1fIH3axJJcmxTTtmSDQtVNDMrm7_Hrbmi/view?usp=sharing

- Extended the idea of self-supervised learning and combined it with the fine-tuning approach for boosting text classification performance in an adversarial setting.
- Pre-trained a generative model to predict the representative of adversarial input and fine-tune it with one additional output layer for downstream NLP tasks (e.g., sentiment analysis and textual entailment).

- Constructed 10,000+ strong adversarial texts dynamically by attacking STOA Transformer models (e.g., BERT and RoBERT) on three datasets (IMDB, AG's News, and SNLI), generated another 200,000+ weakly augmented texts by following the implementation of SSMBA and NLPAug methods as a comparison with baseline results.
- Improved the accuracy of classification and entailment tasks on adversarial texts by 30% on average, compared to Hugging Face pre-trained STOA models.

Bitcoin Price Forecast with Time Series Models *(with ARIMA & GPR)*
- Predicted future value of Bitcoin using Time Series techniques (ARIMA and Gaussian Process Model) in a 4-year period.
- Applied stationary check (ADF & KPSS tests) and transformation techniques (e.g., power transformation and difference) to preprocess 10,000+ price records of Bitcoin; determined the values of autocorrelation and partial autocorrelation of bitcoin price series by plotting ACF and PACF.
- Made predictions and evaluated model performance by tracking diverse metrics (RMSE or MAE); demonstrated findings in the form of a presentation to 40+ audiences and a 20-page written report.

Recommendation System for MovieLens *(with Apache Spark and Hadoop)*
- Implemented Alternating Least Squares (ALS) and popularity baseline model for collaborative filtering in making movie recommendations.
- Partitioned the rating data (58,000 movies and 280,000 users) into stratified training, validation, and testing samples based on user ID; made prediction and evaluation based on predictions for the top 100 items for each user with ranking metrics (e.g., precision at k, NDCG at k, RMSE, and MAP).
- Executed hyperparameter tuning with a number of latent factors, regularization, and iteration epochs to produce observable differences in evaluation score; make comparison to single-machine implementations (e.g., lightfm and lenskit) and achieved a 20% accuracy boost.

Sentiment Analysis of Comment Texts Based on Bi-LSTM *(with Keras and Tensorflow)*
- Encoded 50,000 sentences with a word-level Bi-LSTM sentence encoder and trained a neural network classifier for sentiment analysis.
- Performed NLP-based tokenization, lemmatization, and vectorization in machine understandable language and applied pre-trained Glove embedding to initialize embedding layer of word representations.
- Added dropout layer, early stopping, and max-norm constraints as regularization methods, and max/avg pooling layers to reduce variance and computation complexity, trained deep neural network Bi-LSTM, Bag-of-Words Classifier, and LSTM, and achieved the highest accuracy (97%) with Bi-LSTM.

SKILL

Key Skills: Data Visualization, Predictive Analysis, Statistical Modeling, Clustering & Classification
Technical Skills:
- **Tools:** Python, Spark, Hadoop, Dask, SQL, Tableau, Pytorch, MapReduce, Linux, HPC
- **Packages**: Scikit-Learn, Numpy, Scipy, Pandas, NLTK, Matplotlib, Seaborn, Jupyter Notebook
- **Statistics/Machine Learning**: Statistical Analysis, Linear/Logistic Regression, Clustering, Decision Tree, GBM, Deep Learning, Natural Language Processing
- **Link to Github**: https://github.com/opal-1996.