MICHAEL JEHL
DATA SCIENTIST

CONTACT

mjehl1012@gmail.com
847-521-0197
Chicago, IL 60647
mtjehl
mjehl1012/
mjehl1012

SKILLS

LANGUAGES: Python, SQL
LIBRARIES: pandas, scikit-learn, XGBoost, NumPy, BeautifulSoup, Selenium, Keras, TensorFlow, PyTorch
MACHINE LEARNING ALGORITHMS: Linear Regression, Logistic Regression, Naive Bayes, k-Nearest Neighbors, Decision Trees, k-Means Clustering, Principal Component Analysis, Gradient Boosted Machines, Convolutional Neural Networks, Long Short-Term Memory Networks, Facebook Prophet, ARIMA, SARIMA
DATA VISUALIZATION TOOLS: Matplotlib, Seaborn, Plotly, Tableau
CLOUD & DATA STORAGE: AWS, GCP
CLOUD & DATA STORAGE: AWS, MongoDB, PostgresQL

PROJECTS

Forecasting Electricity Demand
Mar. 2021 to Current

This project uses advanced time series analysis to predict hourly electricity demand in Seattle.

- Data Acquisition: Hourly demand data from July 2015 through early March 2021 was acquired through the EIA API and hourly weather data through a NOAA data order request.
- Data Exploration and Preprocessing: Important preprocessing steps include handling seasonality, ACF and PACF interpretation, ADF/KPSS tests to help determine stationarity, differencing, and visualizations showcased in a Tableau dashboard.
- Modeling: Baseline model is an ARIMA(5,1,0) with RMSE of 28.199MWh with a forecast bias of 0.003889. Plan to use grid search to find optimal parameters of SARIMA model and also try LSTM and Prophet models.

Clean Energy Jobs Act Interpretation Using NLP
Feb. 2021 to Feb. 2021

This project uses natural language processing to aid human understanding of the Clean Energy Jobs Act, a 906-page bill proposed in the Illinois state legislature.

- Data Acquisition: I obtained a pdf of the bill through TrackBill.com and parsed it into python using pdfPlumber.
- Methods and Tools: Topic modeling with NMF, keyword highlighting, and auto-summarization with NLTK. Parts of speech tagging and named entity recognition using spaCy extract the rules and obligations set forth in the bill to identify the burdens and subjects responsible for compliance. t-SNE plot demonstrates clear separation of bill topics. Cosine distance metrics were applied to quantify the similarity between burdens across major subjects.
- Recommendation App (in development): Advanced distance calculations can be applied across legislation to enhance TrackBill's bill similarity recommendation system.
- Communicated results to class with a focus on political use case using Google Slides.

MTA Exploratory Data Analysis Project
Jan. 2021 to Jan. 2021

This group project involved extensive data cleaning, exploration, and teamwork to provide recommendations for a hypothetical charity marketing campaign.

- Data Acquisition: MTA turnstile data obtained by custom script to loop through hundreds of .csvs on the MTA website, income data downloaded through CCC New York's website, and location data obtained through the Google Maps API.
- Data Cleaning and Preprocessing: handled duplicates, outliers, null values, and joined disparate datasets by bringing in income data.
- Methods and Tools: Wrote custom plotting functions with matplotlib and used pandas groupby operations extensively to explore and make sense of data.
- Presented results to class with a focus on hypothetical charity use case using Google Slides.

COVID-19 Intubation Classifier
Jan. 2021 to Feb. 2021

This project uses a binary classification algorithm to determine whether someone needs a ventilator.

- Data Acquisition: Dataset containing 263,000 results of real-time PCR testing in Mexico, including patient-level clinical and epidemiological data, obtained through the website of the Mexican General Directorate of Epidemiology.
- Data Cleaning and Preprocessing: Thorough data cleaning and preprocessing steps including imputing missing values, scaling the data, and handling class imbalance with RandomOverSampler vs. SMOTE.
- Methods and Tools: Started with several out of the box classification models as baselines, including Logistic Regression, XGBoost, kNN, Tree based models, and Naive Bayes. Observed Confusion Matrices and ranked ROC AUC on validation set to identify candidate model, then ran GridSearchCV to determine optimal hyperparameters of best performing model: Random Forest Classifier.
- Model Results: Random Forest best balanced false positives and false negatives for my use case with a ROC AUC of 0.93. Visualizations of model separation conclude model predicts outcomes with high confidence.
- Presented results to class with a focus on healthcare workers use case using Google Slides.

EXPERIENCE

Metis
Chicago, IL
Data Scientist · Jan. 2021 to Current

- Metis is a licensed 12 week immersive data science bootcamp focused on project oriented learning.
- The core curriculum is centered around Python, statistics, supervised and unsupervised machine learning, exploratory data analysis, databases, and visualization techniques.
- Completed five self-designed data science projects from conception to presentation; including data collection, data management, exploratory data analysis, modeling, and visualizations.

bswift
Chicago, IL
Application Support Analyst · Mar. 2020 to Dec. 2020

- Triaged, investigated, and resolved time-sensitive product issues using SQL and liaised with other departments including client services, marketing, and development to assist with product questions.
- Performed rigorous quality assurance testing and analysis prior to new product releases.
- Wrote guide with common SQL queries to be shared across teams and help with training future team members.

Chicago, IL
Channel Partners Consultant · Feb. 2019 to Mar. 2020

- Managed client relationships by troubleshooting client tickets and providing next response time. Crafted and presented training materials to limit ticket volume.
- Identified client data issues and developed mitigation strategies using Excel macros to minimize errors.
- Mentored new implementation and channel partner analysts and helped lead new hire trainings.

Chicago, IL
Client Services Analyst · July 2017 to Feb. 2019

- Supported clients by updating benefit records, performing audits, and testing system configuration.

Stats Perform
Chicago, IL
Operations Analyst · July 2016 to July 2017

- Collected data during live soccer matches and wrote play-by-play commentary for Fox Sports.

ACTIVITIES

Chi Hack Night · Member
Current

Chi Hack Night is a weekly event to build, share, and learn about civic tech in Chicago where data professionals code together, network, and attend talks by local organizations and companies working at the intersection of technology and government.

VOLUNTEERING

Illinois Environmental Council · Data Management Consultant
Current · Chicago, IL

Clean and manage constituent donor database of over 30,000 records using Python. Interact with IEC leadership and IT staff of third-party CRM to push thousands of updates into production and suggest ways to improve data management and data entry practices. Data cleaning tools include building custom functions, regex, and OpenRefine.

Urban Rivers · Data Scientist
Current · Chicago, IL

Working with the leadership team at Urban Rivers to build an image classifier to automatically identify the type of benthic macro-invertebrate in a water sample collected from the Chicago River. Techniques include data cleaning, data labelling, image augmentation, and transfer learning. Model deployment in a user-friendly app planned for Summer 2021 after more data collection, labelling, and fine-tuning.

EDUCATION

Washington University in St. Louis Aug. 2008 to Dec. 2013
Studied Marketing, Psychology, Neuroscience, and Psychology
Awarded merit and need-based scholarship to attend Washington University in St. Louis