PAUL KANG

FULL STACK DATA SCIENTIST

pkang0831@gmail.com +1-416-908-5421 Toronto, ON pkang0831 pkang0831

SKILLS

PROGRAMMING LANGUAGE: Python, MATLAB, Java, MySQL
DATA ENGINEERING / BACK END: Apache Spark, Data Engineering, Version controls & Agile (Git & Jira), OLTP & OLAP, Data Standardization, Apache Kafka, SQL query optimization, Table Transformation
DATA SCIENCE & MACHINE LEARNING: Multiclass/Binary Classification, Regression & Correlation Analysis, Clustering, Dimensionality Reduction Techniques, Neural Networks , Deep learning, Time series forecasting, Anomaly Detection, Microsoft Azure / Databricks, A/B Hypothesis testing, Multivariate Design (MVT), Experimental Design, Natural Language Processing, Recommender Systems, Tensorflow / Keras / Pytorch, Python (Scikit Learn, Pandas, Numpy, Plotly, statsmodels, etc)
COMPUTER SCIENCE: RESTful API , CRUD, Git, OOP, Data Structure & Algorithms, Flask / Django - Web based frameworks, Modern CI/CD
VISUALIZATIONS / FRONT END: Power BI, Tableau, CSS3/CSS5, React.js, Seaborn/Matplotlib/Plotly

EDUCATION

University of Texas At Austin (Online, Part time) Aug. 2021 - Current
Master of Science Data Science 2023
Austin, TX, United States
Relevant coursework: Probability & Inference, Bayesian Statistics, Principles of Machine Learning, Data Structure & Algorithms

University of Waterloo Sept. 2012 - May 2018
Bachelor's of Applied Science, Honours Chemical Engineering 2018
Waterloo, ON, Canada
Cumulative GPA: 3.8/4.0
Relevant Courses: Advanced Statistics for Industrial Engineers, Applied MATLAB

EMPLOYMENT

BASF SE Group, Toronto, ON, Canada Jan. 2022 - Current
Full Stack Data Scientist

Chemical Catalyst Market Modelling (Market Intelligence)

Designed a batch ETL pipeline to ingest & process 500 GB of raw data and developed a Power BI dashboard app that presents the calculated domestic / international market demand of chemical catalysts: automated market sizing activities performed by marketing professionals.

- Utilized Azure Data Lake and Databricks as primary back-end development platform, Power BI & DAX as main front-end development. main stack include Pandas, PySpark, MS SQL, for dataframe manipulation, text processings/standardizations
- Developed an in-house python package that lemmatizes the names of chemical variations into its original form to capture relevant market data from 5 major industrial chemical dataframe sources; used Regex, fuzzy matching, python string manipulations
- Details include PDF parsings to workable data frames (data ingestion), text standardization through spaCy/NLTK lemmatization and Fuzzy matchings, table transformations, and feature engineering to correctly aggregate 118k+ plant catalyst consumption and market demand from 200+ countries and registering the transformed data frames to Hive.
- Developed Power BI dashboard to present the calculated market demand of the catalysts to guide the domain experts in investing marketing resources for favorable returns

Product Recommender (Business Competitiveness)

Developed a collaborative filtering recommendation system using frequent itemset minings on big data - association rules (FPGrowth / Apriori) for identifying cross-selling opportunities to BASF clients. 40% of recommendation to actual sales; increasing company profit by 2.8 M USD /yr

- Transformed and optimized SQL query from SAP Hana to acquire monthly transaction records - 1.9 GB / month and transformed to implicit market data in CSR format
- Developed in-house package of Apriori & FPGrowth scripts using Pyspark, Pandas and Scipy, which replaced existing XGBoost (tree-based algorithm recommender) due to improved accuracy and performance (increased accuracy from 3% to 14%, reduced runtime of 13 hours to 0.75 hours)
- Performed hypothesis testing to validate the performance of the model and the quality of recommendations to customers. Successfully proved that the 40% of recommendations are increasing sales profit by 2.8M USD/yr
- Wrote a requested business report on the recommendation engine and successfully initiated a product launch for the system.

EMPLOYMENT

Suncor Energy Inc, Canada Apr. 2019 - Jan. 2022
Data Scientist & Process Engineering

Carbon Footprint Reduction (Transformational Management)

Developed an oil leak detection model (binary classification) using a boosting ensemble (XGBoost) to minimize the environmental impact caused by inefficient oil plant processes. Projected successfully implemented detection model of abnormal plant operating conditions which trigger immediate corrective action guidance to operators and reduced oil leaks by 31% on annual basis with 98% user engagement.

- Optimized SQL query to retrieve and transform past 11 years of plant process data to workable dataset format for rapid model prototyping.
- Employed Pandas & Spark for feature engineering and performed Exploratory Data Analysis (EDA) to conduct correlation analysis.
- Utilized clustering techniques and dimensionality reduction (PCA) to represent 3 major conditions of the plant, identifying the range of plant attributes in an abnormal condition and used as a main model accuracy booster.
- Trained 23 Scikit-Learn pre-designed classification models and shortlisted 2 models XGBoost and Random Forest with a best F1 score of 0.75 on the target class with assessing Precision and Recall, minimizing type II error by 45%.
- Used React, NodeJS, HTML/CSS, to develop the front-end design of the stand-alone product, constructed a data pipeline with Kedron and python Flask to develop a web framework, and deployed the minimal viable product to Azure cloud services.
- Presented the result to the Advanced Analytics and process engineering Area Cross-Functional Team and wrote a requested executive summary to initiate the product launch.

Suncor Oilsands Baseplant Production Optimizer (Business Competitiveness)

Support for the development of real-time plant optimizer model using multilayer perceptron (MLP) Neural Network to generate hourly recommendations of 5 Key Performance Indicators (KPIs) which helped site operations to maximize the oil production (uplift by 1.2% / year, equivalent to $ 7.3M/yr)

- Defined success target & user engagement tracking metrics of the Plant Optimizer Product during MVP phase and full package deployment.
- Acted as a domain subject matter expert during sprint reviews, knowledge transferring for the feature engineering EDA process, validating observations and logic to ensure well-versed communication with the other site experts in an agile environment.
- Collaborated with McKinsey Data Scientists to develop MLP (deep learning) with test R2 of 0.95 with MAPE < 6%, grid search optimization & hyperparameter tuning of the models.
- Constructed SHAP (feature importance) plots for assessing variable impacts for the model outputs.
- Trained plant operators (clients) with the usage of the web-based application for successful launching and safe landing product implementation.
- Liaison with the oilsands mine operation, ensuring strong communication between site technicals and data scientists for understanding the ML recommendation behaviors.

PROJECTS

Design of Experiment (2^k factorial) for oil recovery improvement (Suncor) Sept. 2019 - Apr. 2020

Designed 2^k factorial experiments to test and quantify the impact of deteriorating raw material quality having on overall oil recovery. Plant operational guidelines are updated as per the recommendations from the analysis for maintaining sustainable operation while achieving 95% recovery.

Financial Market Volatility prediction - Competition (Kaggle) July 2021 - July 2021

Designed a stacked supervised regression algorithm with pre-defined 24 regression algorithms and designed deep learning Neural Network regression in CNN architecture using TensorFlow, visualized with Parallel Coordinates View. Scored R2 of 0.8 with RMSPE of < 22%

Stock Price Prediction - Dataset (Kaggle) Apr. 2021 - Apr. 2021

Developed LSTM, RNN sequential supervised models to predict stock price of 40 different companies using 10 years of 6000 stock/ETF price history data, historical location weather data to yield R2 of 0.81 and maintain positive reward return at epoch of 30

Digit Recognizer, Computer Vision, MNIST - Competition (Kaggle) May 2021 - May 2021

Employed 15 different ML models and selected ensemble classifier supervised algorithms from sklearn library to yield validation accuracy score of 0.97

PKTripp - Python GUI app development June 2021 - Current

Integrated API calls to CRUD trip related data: Flights, Hotels, Restaurants, Weather, Things-to-do
Based on Reviews & Ratings, application provides upto 3 recommendation optimized trip plans and calculates required budgets
Tech used: Python, REST API, JSON, Git