Contact

www.linkedin.com/in/jay-kim-0a0433120 (LinkedIn)

Jay Kim

Senior Data Scientist at SoCalGas (Advanced Analytics)
Orange County, California Area

Summary

Highly efficient and results-oriented data scientist with strong quantitative skills, development experience and strong education background with a MSc (Imperial College London (World Rank within Top 10 QS)). Responsible self-starter with demonstrated experience in statistical programming language (R, Python, SAS, Scala) and programming language python for API's. High ability holder on visualization with tools such as Tableau as well as good understanding of relational database such as SQL and oracle and non-relational database such as hbase, mongoDB and redis. Expert in Machine learning, Deep Learning, AI, as well as data science. Good at ML tools such as Hadoop, Spark, H2O, sparkling-water, pysparkling, SAS etc. as well as deep learning tools such as Keras, Tensorflow, Theano, MXnet, PyTorch. GPU cuda programming. Scaling data science. Expert in Predictive Modeling such as XGBoost, regression, Logit, Probit, GBM, RandomForest, Neural Network (generative model, GAN, VAE, RNN, CNN, word2vec etc.) , Naive Bays, K-nearest learn, PCA etc. (supervised learning, unsupervised learning, semi-supervised learning , reinforcement learning etc.) and also probabilistic modeling (PyMC3, Edward, Pyro) such as MCMC, HMC, NUTS, bayesian linear regression, variational models etc, Data mining skills such as parsing, nlp (natural language processing) and proficient in language modeling such as topic model, text clustering, word embedding, Word2Vec, Glove, text classification, RNN, Convolutional RNN etc. familiar with all the development environment such as Hadoop, Cloud (AWS, GCP, Azure) , GPU, Spark. Docker etc.
Strong communication and relationship-building skills with diverse parties;
contact me at jay.kim07@yahoo.com

Experience

SoCalGas

2 years 6 months

Senior Data Scientist
August 2018 - Present (1 year 8 months)
Downtown LA

Works at SoCalGas as senior data scientist. Leading projects building models as well as statistical analyses. Working on multiple projects. * Next generation neighbor comparison project.* Meter pipe image recognition modeling (Vision)*Fast Meter Predictive modeling*Gas Engineering projects, tensorflow_extended, tensorflow_probability, tensorflow_federated, etc. *building predictive models (ml & dl) and productionizing in Emergency Services and Safety Monitoring (ESSM) project.* Transfer learning, GAN model. DCGAN, Bi-directional LSTM, Ladder Network, Ladder VAE, Auxiliary VAE, Info VAE,Disentangled Sequential Autoencoder. Triplet Loss Model etc * Perceptron, MLP, Conv1, LSTM, Conv1-LSTM, LSTM-VAE, ARIMA, VAR etc.•Time series clustering modeling (Dec. 2018): k-means time series clustering, K-Shape time series clustering, global assignment k-means clustering etc.• variational auto encoder, LSTM-VAE model : applied VAE model & LSTM-VAE model to detect various gas leaks & water leaks in huge dataset. *drf model (gap model) (Aug. 2018 – Oct. 2018): distributed random forest model to predict gaps of transmissions •Text model (Aug. 2018 – Nov. 2018):predicting gas leaks & water leaks by text model that is built from nlp & adaboost learning algorithm. •Time-series modeling.(Nov. 2018 – present) : time dependent consumption data and weather data with generated pilot data, time series models are attempted to predict gas leaks & water leaks. •Bad-debt model (Logit) implementation to production. (April. 2018 – Nov. 2018) : SAS predictive model built previously by me is implemented to SoCalGas system. End to end test, data pipeline etc. • Battery model.(Sep. 2018 – present) : building predictive model that classifies the battery is defect or not. • Speaker at PyData Conference. (Oct. 2018) : presented to experienced level audience with topic of “hot water leak detection using variational autoencoder” https://pydata.org/la2018/schedule/presentation/25/

Data Scientist
October 2017 - August 2018 (11 months)
downtown LA

• Gas Leak Detection: There are 6 million customers, which means there are 6 million meters. This is a pilot to implement a machine learning algorithm to forecast what the next daily hour of usage will be. This could increase safety because, if this number is way out of bounds, it could be the result of a gas leak, or, it can be used to better market to customers to save energy. Applied

deeplearning model (segmentation analysis, Text Analytics using service order and language model. Variational Autoencoder (generative model(semi-supervised learning)).

• High Bill Predictive model: There are people who call for high bill inquiries. This model predicts the likelihood for them to call. Who is calling and when they call. This is time series problem and requires to build VAR and LSTM to predict the right timings as well as batch analysis such as XGBoost etc.
Did clustering analyses and built clusters. And built predictive models in each cluster.

• Drive a Bad Debt : The credit group wants to know what customer attribute is driving/increasing bad debt. By doing features engineering, select right features (attributes) of customers.

- Calculate correlation coefficients and important rate of features.
- By business & statistical analysis, select important features.
- Check variance loadings between variables.
- Calculate p-value, and see how significant the features are.
- built predictive models .

Packt
Technical reviewer (data science)
December 2018 - October 2019 (11 months)

Review and edit technologies and contents in data science books.

Published "what's new in tensorflow 2.0" as a technical reviewer
Published "unsupervised learning with python" as a technical reviewer

Atmospheric Data Solution & Trace3
Data Scientist
March 2016 - August 2017 (1 year 6 months)
Santa Ana & Irvine

Atmospheric Data Solution :
Worked with SDG&E’s data and weather data such as wind speed, humidity, dewpoint, temperature and etc. Creating prediction model to predict outages and customer impacts. Perform data mining and bootstraps for resampling. Various resampling techniques such as are performed. Used AWS for S3 and RedShift. Used Scrapy for data mining of some texts. With NLP,

found keywords and important contexts from the text data. Applied some deeplearning models such as RNN and ANN.
• Predicted utility's damages and the number of outages from the weather data and historical outage and customer impact (damages) data.
• Did image processing with satellite weather data.
• Did text mining and NLP from event title column and event description.
• Applied XGboost algorithm, RandomForest algorithm, GLM, SVM, decision tree and etc.

Trace3 :
Worked with Wash Laundry Multifamily System and Boeing Company (CDG) as their contractor. Dealt with sensor data and textual data. Topic modeling, Text classification & clustering. Built simulations for pricing models and quantity models. Analyzed data and feature engineering. Leading data science group of Advanced Analytics. Analysis of full potential of market. Also dealt with demographic data. Predictive Modelling, Algorithms etc. For these project, I have used python, R, SQL languages to obtain the dataset. H2O, SparklyR and MLlib in Spark. Scrapy and beautifulSoup to extract texts from XML files. NLP doing text clustering, text classification and topic modeling. SVM, GBM, Randomforest, CNN and ANN learning algorithms were tried and tested. For deep learning, Keras, MXNet, theano, PyTorch and tensorflow. . GCP for ml and dl APIs and Big-query. Working on the unstructured data in Boeing Company (CDG) project. Applied Unsupervised learning such as Topic model, Text Clustering etc. Vectorization of unstructured data such as bag of words, TF-IDF, bi-gram, tri-gram, skip-gram etc.

OspreyData, Inc.
Data Scientist
May 2015 - January 2016 (9 months)
San Juan Capastrano

Worked with data from ESP (electricity submersible pump) : ESPs are installed in production wells to optimize production rate of petroleum at the production sites. The ESP records many data automatically every second. The data is often noisy with significant data volumes. Our company leverages the data to diagnose the operational state of ESP pump. For these project, I have used python, R, SQL languages to obtain the dataset. To apply machine learning model to big data, I used pySpark. For data preprocessing, I used moving average to make data smooth since sensor data were so noisy And applied NLP doing text clustering, text classification and topic modeling. For predictive models, Decision-Tree, SVM, GBM, RandomForest and ANN

learning algorithms were tried and tested. For centralization of sensor data for IoT, we used Microsoft Azure (IoTHub) to collect data from sensors. For deeplearning, I used tensorflow. Wrote algorithms for engineering model that are to solve engineering problems in the project.

• Predicted operational and failure modes from physical and calculated sensors from ESP.
• Wrote unique classifier that predicts gaslock, watercut, impeller wearing, etc with 95% confidence. Expected to save 10 million dollars by predicting pumps failure in downhole.
• Leveraged Machine Learning to create diagnostic classfiers and clustering analysis.
• Analyzed and designed feature set to maximize model fit with R.
• Implemented the machine learning algorithm into production software utilizing Python.
• Applied SVM machine learning algorithm to non-linear data to fit and predict.
• Wrote Algorithm programming with R and python. Sometimes Matlab.
• Did Data mining using various models.
• Worked with Shapelet and Time series data Warping.

OilQuest & Seoul National University
Data Scientist
November 2013 - March 2015 (1 year 5 months)

OilQuest

Worked with data from clients (Conoco Phillips, SK). The data includes well logging data, production data, seismic data, microseismic data, and geographic data..

• Utilized various industrial geophysical and engineering data, Identified 3 new reserviors with an estimated 2 billion barrels of oil in North East Australian offshore Australia. Found where petroleum and water are located.
• Analyzed property of rocks underground based on data utilizing wavelets from well data integrated with seismic data to create synthetic map.
• Applied machine learning models to dataset and predicted.
• Designed and ran uncertainty analyses.
• Accurately predicted size of petroleum reserves underground with 90% confidence.
• Did Visualization and interpretation (2D, 3D, 4D)

Seoul National University

Worked with the Data from European Research Center (Kezin,Germany). The data is time lapse data set including geographic data, microseismic data, geological data, seismic data, well logging data and production data..

• Worked with various geological, geographical and engineering data. With various well logging data and production data, we found out how water layers have been changed depending on time going by.
• Analyzed property of rocks underground based on data.
• Ran AVO(Amplitude Versus Offset) analyses, Fuzzy Analyses, Uncertainty Analyses
• Programmed full waveform inversion models with Fortran, C, C++.
• Analyzed CO2 layer change when time flew from seismic map after Data processing and Tomography
• Modelled of CO2 expansion underground and timing when the CO2 overflow the storage of reservoir.

Education

Imperial College London
Geophysics

University of Nottingham
Process engineering, Signal Processing

Yonsei University
Engineering