Data Scientist

Dublin, CA

- Achievement-driven professional with experience of 10 years in IT and 5 years in applying Data Analytics, Machine Learning, NLP, Linear part of Data Analytics.
- Supervised Learning : Worked on multiple analytics solutions based on Linear Regression. Used multiple Model selection methods, diagnostics and cross validation techniques.
- Supervised Learning : Worked on multiple classification problems using various techniques - Logistics Regression, Random Forest, Naïve Bayes, k-NN, SVM, Neural Net, Bagging, Boosting and various other Ensemble techniques.
- Unsupervised Learning : Used K-means clustering for segmentation, PCA methods for dimension reductions, Anomaly detection model.
- NLP/Text Mining : Sentiment analytics based on unigram, bi-gram, and tri-gram bag of words model using Python NLTK. LDA based Topic model using R & Python. Social Media Analytics, Text Clustering & visualization.
- Knowledge of Deep Neural Networks (ANNs, CNNs, RNNs, Autoencoding) using TensorFlow and Keras.
- Visualization using ggplot2, MatPlotLib, Seaborn, Tableau.
- Knowledge of cloud based systems like Azure, AWS, Google Cloud.

Work Experience Data Scientist
Oppenheimer Funds - San Francisco, CA
Oppenheimer is a fund manager and wealth management company. Primary objective of this project is to predict redemption rate of funds and make performance scores for funds. Responsibilities :

- Beginning of the month, process monthly funds raw data by ingesting thru Jenkins.
- Input files are assets and flows, flowsonly, advisorsIndicative, FundPerformance, Fund indicative.
- A snapshot of each input file is saved on AWS S3.
- Aggregate fund, firm, channel level in SQL.
- Predict redemption in the next 3 months using binary classification using XGBoost.
- Predict redemption rate in the next 3 months using linear regression using XGBoost.
- Model ensemble and obtain final score, then aggregate to composite fund level..
- Process the data using XGBoost model and arrive at classification and regression and score.
- Save the output data to vertica db tables and backup output files on AWS S3.
- Retrain retail fund at Risk model every month and predict redemption rate.
- Retrain retail channel spike model every month and make scores.
- Prepare retail campaign dashboard to pick fund to defend and generate balanced campaign list.
- Every 3 months analyze the flows data.
- Used R language to develop Database connection utility, ingest advisor file, spark implementation code to aggregate daily transaction to monthly level, ingest fund indicative file, and calculate scorecards for individual advisors.
- Used SparklyR to aggregate and ingest monthly assets and flows data.
- Used R to create training data for Fund at Risk model, execute model training and prediction and push the prediction to Vertica Db.
- Used Python and sql scripts to execute model training and prediction for Advisor at Risk and push prediction to Vertica Db.

- Used R to generate monthly scorecards for funds, R Shiny dashboard..
- Used Jenkins to create spark job aggregate monthly flows data and Assets data from Bitbucket repository, Advisor at Risk model.
- RMSE is used for linear regression validation and AUC is used for classification validation.