JANIS LYNN CORONA | 951.358.9116 | Corona, CA|janiscorona1982@yahoo.com | rpubs.com | LinkedIn | GitHub

Data Scientist | Scalable Code Development | Machine Learning & Artificial Intelligence (ML/AI) | Quantitative Analysis

Data Scientist leveraging a **master's in data science** to solve complex problems by leveraging state-of-the-art machine learning models and advanced statistical analysis. Proficient at retrieving and aggregating raw data from multiple sources, reshaping it to ensure suitability for analysis, analyzing it, and compiling data into digestible, actionable formats and dashboards.

DATA SCIENCE TOOLS & TECHNIQUES: R, Python ,MySQL, SQLite3, PostgreSQL, SAS, HTML, CSS, PuTTY SSHL, Tableau, Google Analytics, Deep Learning, Big Data Analytics, Natural Language Processing (NLP), Time Series Analysis, Logistic Regression, Random Forest, Naïve Bayes, K-Nearest Neighbor, K-Means, Principal Component Analysis (PCA), Support Vector Machines (SVM), Decision Trees, General Linear Models (GLM), Gradient Boosted Trees, Exploratory Data Analysis (EDA), Classification Report, Receiver Operating Curve (ROC), statistical modelling, Image Classification with neural nets, VirtualBox, Cloudera), Impala, Microsoft Azure, Ambari, HDFS, Hive SQL (HQL, Amazon Web Services (AWS) EC2 and S3, Stata, Multivariate Testing, Regex feature extraction in text mining, automated programming, and more.

PROFESSIONAL EXPERIENCE

Data Scientist, The Massage Negotiator 08/2018 - present

Independently ran website connecting clients with mobile service providers. Developed algorithms to find and acquire legitimate customers, provide advertising services, and expand the network of service providers; managed web content with WordPress, managed best return on investment of advertising, developed new network methods to establish marketing of services. Disseminated Amazon Web Services (AWS) cloud computing knowledge to other data scientists by creating a series of YouTube video tutorials; developed text mining, Elastic Compute Cloud (EC2), and user account and permission management demos.

- Optimized ad photos displayed on the Yelp page by leveraging mastery of neural network and deep learning mechanisms.
- Identified high-value clients and service providers by utilizing Python and R text mining and NLP using TextBlob, nltk, scipy, and sklearn for python, and tm, SnowballC, and wordcloud for R software to determine the sentiment analysis on the words in resumes of potential business partners, massage therapists, and client liability form responses on best fit or worst fit to work with.
- Increased positive customer responses by leveraging post-service feedback forms with Jotform.

EDUCATION

MS, Data Science, Lewis University October 22, 2019

- DATA 55000: Machine Learning with Python for NLP, NNs, Gaussian Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Principal Component Analysis, and other machine learning algorithms for classification and regression predictive analytics
- CPSC 54000: Large Scale Data Storage Systems for large scale data in the cloud using AWS, Azure, HDFS, Python, Hive, use of virtual machines in Linux OS with Cloudera, Bitnami, and Ubuntu, cloud computing, encryption, shell security, file transfer protocols, and data compression formats
- BIOL 52100: Research in Biotechnology, presentation development and implementation using PowerPoint, sources of research material to analyze, familiarity with the humans as research subjects type of research work, use of online data repositories to analyze genetic data such as BLAST, encode.org, UCSB data repositories, GEO, Ensembl.org, and more

BS, Math & Economics, University of California San Diego, La Jolla, CA December 12, 2015

- Econometrics 120A:120C: Statistical analysis using STATA on education data using linear, logit, and probit regression analysis for predictive analytics. Statistics evaluating samples generalized to populations with hypothesis testing, confidence intervals, and significance of error measures. Data sets on student classroom size on exam learning outcomes and seatbelts on deaths and survival in auto collisions were used and analyzed with the above predictive analytics models for the class projects in Feb 2014 and Dec 2014 respectively.
- Math 189: Data Analysis and Inference, performed data mining analytics on a large scale data set of songs spanning 100 years, using each song's feature timbres to predict the year a song was released as the target class using R, Excel, STATA, and matlab in this 500 mb file size. This project was completed in May 2015.

DATA SCIENCE PROJECTS

Lewis University, May-July 2019

Meta-Analysis of Genes: Achieved 100% training model accuracy by using 121 gene samples. Analyzed link between 5 research studies with microarray gene expressions and uterine leiomyoma (UL) risk; extracted and processed data from an online data repository and applied latent Dirichlet allocation (LDA), random forest (RF), generalized boosted regression models (GBM), recursive partitioning and regression trees (rpart), as well as a combined model of all algorithms. Identified best-performing data sets.

Lewis University, Jan-Feb 2018

Optimal Arrival & Departure Time Analysis: Compiled, processed, and analyzed 1M+ rows of airport data by using SQL, Linux in virtual machines (VM), and Shell (SSH) in Hadoop Distributed File System (HDFS). Utilized cluster cloud computing. Determined ORC to be the most efficient Hive data storage format for one-column and TextFile for two-column files.

Lewis University, Aug-Sep 2017

Classification of Malignant Tumors: Achieved 96% classification accuracy by using the random forest algorithm to analyze 569 instances of breast cancer data; assessed the utility of ML methods to categorize breast cancers as malignant or benign. Compared results by using multiple algorithms, including Naïve Bayes, AdaBoost, and decision tree. Used Orange BioLab software.

The Massage Negotiator, Dec 2018-Jan 2019

Crime & Real Estate Statistical Analysis: Extracted publicly available data for 14 major metros across the US. Consolidated disparate data sets from multiple sources. Ranked locations based on median pay, crime rate, and population. Used MS Excel, R, Tableau, and data visualization techniques to effectively reach and display conclusions.

The Massage Negotiator, Sep-Oct 2019

Classification of 12 Disease Microarray Samples: Web scraped, downloaded, combined, cleaned, and analyzed 12 different microarray gene studies on uterine leiomyomas, stomach cancer, acne, brain cancer, colon cancer, breast cancer, CBD treated pancreatic cancer, and not diseased sample types using Python machine learning packages sklearn, numpy, pandas, matplotlib, and keras for machine learning classification and one-hot-encoding of those different tissue samples with a 90% classification accuracy given the limited samples of each type of sample. Python and R were used for this project.

The Massage Negotiator, Oct 2019

Text Mining Competitor Reviews: Gathered data from various social media review sites web scraped for content that was then combined, categorized, cleaned with regular expressions, removed of stop words, tokenized by word and root word to find the most compelling reason for consumers of massage therapy services to seek out massage and why they avoid certain massage therapy providers. This aided in developing marketing tools to reach those consumers and grow the massage therapy mobile services business and get more web site activity, bookings of appointments, and referrals by other massage therapist connoisseurs.

The Massage Negotiator, Oct-Nov 2019

Machine Learning on UFC Fighter Actions and Reactions: Compiled, processed, and analyzed six different UFC fighter's previous fights in the UFC for their actions when using features that tracked cumulative hits missed, received, landed against the opponent, and what type of strikes the fighter used for each second with multiple actions for timed second intervals as dummy variables to predict a target of hits landed by that fighter in a simulation with one of the other fighter's actions and reactions with up to 100% accuracy. Python and R packages for machine learning were used for predictive analytics with Ensemble methods, decision trees, random forests, naïve bayes classification, logistic regression, rpart trees, PCA, Kmeans, KNN, LDA, gbm, and glm as a comparison of best performing machine learning algorithm.

The Massage Negotiator, Dec 2019

Kidney Tumor classification of Diabetes and Renal disease microarray samples: This study was done in R that combined the tumor samples from separate studies and analyzed the genes that were most and least expressed in terms of fold change, extracted the top 20 of those genes, then used machine learning's KNN and Random Forest algorithms to classify the tumor sample based on the top 20 gene expression values as advanced, healthy, or early diabetes mellitus with 100% accuracy in classification. Text mining was done on

kidney disease top 20 and bottom 20 genes expressed in the tumors on diabetes mellitus with the gene functional summaries to get an idea of what each of those gene target functions is with a few word clouds that used lemmatization for the word tokens.

The Massage Negotiator, Feb-Mar 2020

Stock analysis of 65-5600 NYSE and NASDAQ stocks, gathered statistical information, wrote automated script in R to grab these same stocks and get the time series lag information from seven days, create ratios, get the cumulative sums of those increasing and decreasing days as counts, group those counts for how long they occurred in the time interval specified at the beginning of the automation script, and create probabilities by the day of what the most likely number of days the current set of days increasing or decreasing will be to buy the stock and also built models predicting 80-95% accuracy on predicting if the next trading day would be an increase of two percent or more.

The Massage Negotiator, Mar-Apr 2020

Developed and tested various massage modality, health and wellness modality recommender systems using R and Python. The modules sklearn, numpy, pandas, nltk, textBlob were used in Python, and tm, tidytext, textstem, caret, dplyr, tidyverse, and stringr in R both implemented models based on text mining, tokenizing benefits and contraindications into lemmatized ngrams then recommending a modality based on benefits after excluding for contraindications for each user input requesting a massage. The machine learning algorithms used gradient boosted trees, random forest, multinomial naive bayes to classify a recommendation with up to 94.5% on the models for wellness and 100% accuracy on the modality recommender.

LINKS

- https://www.linkedin.com/in/janiscoronadba
- https://www.github.com/janjanjan2018
- https://rpubs.com/janisharris

CERTIFICATES

- Linked in Learning-SAS Programming for R Users, Part 1: Certificate Id: Afvx49vCSzXRlFGk7vCSXpb2bwow
- Linked in Learning-NLP with Python for Machine Learning Essential Training: Certificate Id: AYmb0S2tjYewnx5Hf6zMI3PUvLdC
- Linked in Learning-Hadoop for Data Science Tips, Tricks, & Techniques: Certificate Id: Adw5V1XJ_z1FqaDL01MbckYc2ZZ_
- Building Recommender Systems with Machine Learning and AI: Certificate Id: AfYZB2nlw_alwFHfaeS1e9rB4tTy