

# List of Past Analytics Projects

## Statistical Machine Learning

### 1. Water Distribution in the Philippines

**Description:** This involved various analyses of the water distribution rate and wastage rating in the Philippines. The data is a survey of water distribution companies across the country. It is a combination of many categorical variables (e.g. type of municipality, region, etc.) and some numerical variables (area elevation, proximity from the coastline, etc.). The goal of the project was to provide regression and classification models for the water distribution pricing and wastage rating, respectively. Because of the very sparse nature of the data, the highest accuracy rate that has been achieved was 60%.

**Techniques used:** Multiple linear regression, Logistic regression, Generalized Linear Models, CART, Bagging, Random Forest, Boosting, k-Nearest Neighbor, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Ridge Regression, Lasso, Principal Components Regression, Partial Least Squares, General Additive Model, Cubic Splines

**Tools used:** R, RStudio, ggplot

### 2. PhilHealth Acquisition

**Description:** In this project, a financial transaction data has been given. Records consist of types of PhilHealth membership, business franchise, ownership of credit cards, health/financial insurance, mutual funds, savings, stock shares and other financial indicators of users. The goal was to determine the factors that make an individual become a PhilHealth member. It turned out that having social security insurance, personal and microfinance loans lead to acquisition of PhilHealth membership.

**Techniques used:** Association Rule Mining

**Tools used:** RStudio, R

### 3. Analysis of American Universities

**Description:** In this project, a record of universities in the United States has been provided. It consists of admission rates, SAT scores, percentage of degrees awarded, number of students enrolled, average tuition fee, number of full-time faculties and others. The goal was to find latent groupings among these universities. Eight groupings that are related to the field of expertise have been identified.

**Techniques used:** PCA, k-Means Clustering, Hierarchical Clustering

**Tools used:** RStudio, R, Silhouette plot

## Exploratory Data Analysis

### 1. Tagging Weibo Spammers

**Description:** In this project, a data set of users, posts, followers and followees were gathered from the Chinese social network site Weibo. Using the users that have been tagged as spammer or not and other variables like the number of followees or followers, untagged users have been predicted as spammer or not. Result showed that around 30% of Weibo users are spammers.

**Techniques used:** SVM, Random Forest, Logistic Regression

**Tools used:** RStudio, R

### 2. German Cash Loan

**Description:** The dataset classifies people as good or bad credit risks, which are described by a set of attributes. Each observation is loan application where an individual is applying for a loan for various reasons. The goal is to predict whether it will be accepted or rejected. Three methods have been applied and compared in terms of test accuracy. Though not that powerful, the winning method was Random Forest which achieved a 20% error rate.

**Techniques used:** SVM, Random Forest, Logistic Regression, Data Wrangling

**Tools used:** RStudio, R

## Time Series Analysis/Forecasting Analytics

### 1. Analysis and Forecasting of MRT Volume of Passengers

**Description:** In this project, the volume of passengers riding the MRT Line 3 from 2000 to 2010 has been analyzed and forecasted. Comparison of basic seasonal time-series, time-series decomposition, linear regression and exponential smoothing have been used. The seasonal data is best modelled by a seasonal-naïve method with 2% MAPE.

**Techniques used:** Naïve Forecasting, Time-Series Decomposition, X11, SEATS, STL, Linear Regression, Exponential Smoothing

**Tools used:** RStudio, R

### 2. Analysis and Forecasting of Philippine Total Monthly Imports

**Description:** In this project, the total monthly importation of the Philippines, from 2000 to 2010, was analyzed and forecasted. SARIMA and Dynamic Regression models were used. Dynamic regression with Lagged predictors bested the others with 6.2% MAPE.

**Techniques used:** S/ARIMA, Linear Regression with ARIMA, Harmonic Regression, Lagged Predictors

**Tools used:** RStudio, R

### 3. Forecasting the Direction of Bitcoin

**Description:** Using the time-lagged values of features of a sentiment analysis and the Bitcoin price itself, the problem is poised to predict whether the Bitcoin price will go up, go down or stay within a certain bound. Logistic loss  $-\ln Pr(Y|P)$  was used to evaluate the models.

**Techniques used:** Dynamic Regression, Random Forest, LSTM

**Tools used:** R, Python, Google Colab

### 4. Power Plant Sulfur Dioxide Emission

**Description:** Using sensor readings from a power plant's burning system and the limestone feed rate, the amount of sulfur dioxide has been analyzed. It turned out that the amount of limestone being fed to the burning system has very negligible effect to the reduction of emitted sulfur dioxide. Principal Component Analysis and sensitivity analysis confirmed this pattern.

**Techniques used:** Dynamic Regression, ARIMA, VAR, VECM, PCA

**Tools used:** Python, Google Colab, scikit learn

## Bayesian Analytics/Computational Statistics

### 1. Estimating Allele Frequency

**Description:** This involved estimating the frequency  $p$  of an allele in a population that follows the Hardy Weinberg Equilibrium, and an in-breeding coefficient  $f$  for organisms that follow random mating. Result agreed with the values from previous studies.

**Techniques used:** Metropolis-Hastings Algorithm, Gibbs Sampler

**Tools used:** RStudio, R, JAGS, STAN, Latex

### 2. Quinoline Drug on Ames Salmonella Colonies

**Description:** Given the "margolin" data set, the goal of this project was to provide a Bayesian type regression analysis of the effect of the dose ( $D$ ) of quinoline to the bacterial colonies. The final model included covariates  $D$  and  $\log(1 + D)$  and had an agreement with classical Poisson Regression.

**Techniques used:** Bayesian Poisson Regression, Bayesian Hierarchical Models, Metropolis-Hastings

**Tools used:** RStudio, R, STAN, Latex

### 3. Growth Limit of Alicyclobacillus Acidoterrestris in Apple Juice

**Description:** This was implementing the 2011 modelling of a data set involving the growth of bacteria in apple juice using a Bayesian approach. The goal was to fit a Probit model using the data augmentation algorithm of Albert and Chib. The Bayesian approach did not differ considerably from the result of ordinary GLM modeling.

**Techniques used:** Bayesian Probit Regression, Gibbs Sampling, EM Algorithm

**Tools used:** RStudio, R, JAGS, STAN, Latex

#### 4. Flight Departure Delay Prediction: A Comparison of BART and Gradient Boosting

**Description:** This project applied and compared machine learning algorithms, particularly BART and gradient boosting, to predict if a given flight's departure will be delayed or not. The data used to train and test the models were local carrier flights departing from and arriving Manila from March to April 2019. Results showed that BART was as good as gradient boosting in predicting flight delays, with an accuracy of 58% to 70% depending on the Airline.

**Techniques used:** Bayesian Additive Regression Trees, Gradient Boosting

**Tools used:** RStudio, R, XGBoost

#### 5. Optimal Contract Pricing

**Description:** The energy trading team wanted to determine the percentage level and price at which the company is most advantageous at when layout a contract with their customers. Various modelling, optimization and simulations were made for gross margin computations of their contracts. Two other variables were identified as inputs in the simulation: WESM price and Newcastle Coal price. The result were a 99.9% accurate model of their complex gross margin computation and a collection of sampling distributions for each value of contracting level and contract price. The model reduced the computation time of gross margins from 15 minutes to 5 milliseconds.

**Techniques used:** LightGBM, Linear Regression, K-nearest Neighbor, PCA, Monte Carlo, Genetic Algorithm, Particle Swarm, Differential Evolution, Bootstrap sampling

**Tools used:** Jupyter Notebook, Python, Shap

## Deep Learning

#### 1. Binary Image Classification using ANN

**Description:** This project implemented a previous Kaggle competition problem that involved image classification of snakes. Using only MultiLayer Perceptrons, the goal was to identify whether a given image is a python or anaconda. The result achieved a maximum accuracy of 73.33%.

**Techniques used:** Artificial Neural Networks

**Tools used:** Python, Jupyter Notebook, Keras

#### 2. Churn Analysis of Network Subscribers

**Description:** In this project, a data set consisting of 2000 subscribers of a cellular network has been given. The goal was to predict whether a subscriber will churn or not. A churn variable, which signifies whether a customer had left the company two months after observation, had to be classified. The SVM model achieved 70% accuracy whereas an ANN structure got a 72.2% accuracy rate.

**Techniques used:** SVM, Artificial Neural Networks

**Tools used:** RStudio, R, H2O

### 3. Poverty Rate Analysis

**Description:** Given the latest Philippine Family Income and Expenditure Survey data, the objective was to classify each household as poor or non-poor and to estimate the poverty rate index for Region 6. An explainable deep neural network architecture was developed. The model achieved 93% accuracy.

**Techniques used:** Artificial Neural Networks, Wide and Deep Architecture

**Tools used:** Python, Keras, Google Colab

### 4. Philippine Energy Demand Forecasting

**Description:** Given the company's demand data from energy trading, the goal was to provide hourly demand forecasts for 1-day, 3-day and 7-day horizons. Energy demand has been modelled using LightGBM and Convolutional LSTM architectures. Both models achieved 2-3% MAPE.

**Techniques used:** LightGBM, Convolutional LSTM, Encoder-Decoder Architecture

**Tools used:** Python, Keras, Google Colab

## Natural Language Processing

### 1. Japanese Twitter Topic Modelling

**Description:** A set of trending tweets in Japan, with hashtag #ドラゴンボール (DragonBall), was collected in September 2019 and analyzed. Four prevailing topics of discussion that came along with the trend have been extracted: 1) Toy/Action Figures, 2) the DragonBall characters, 3) the Dokkan Battle Event, and 4) Having access to artworks

**Technique used:** Latent Dirichlet Allocation

**Tools used:** Python, Jupyter, Gensim, t-SNE, pyLDAvis, Octoparse

### 2. Text Spam Classification

**Description:** Using the vectorized text message content, length of message and number of punctuation marks as feature variables, the goal of the problem was to classify whether the text message is legitimate or spam. TF-IDF was used for the vectorization and embeddings. Linear regression outperformed Random Forest, Gradient Boosting and Naïve Bayes for the given set of data.

**Techniques used:** Random Forest, Gradient Boosting, Naïve Bayes, Logistic Regression, Lasso, Ridge

**Tools used:** Python, Jupyter, NLTK

## Survival Analysis

### 1. Predicting Risk of Banking Churn Using Survival Analysis

**Description:** Given the debit and credit transactions, the problem was to analyze and predict the risk of churning of customers. Using survival analysis, it turned out that the number of transactions, the nature of maintaining balance and the maximum no-transaction period play a very significant role in predicting the churning of a customer.

**Techniques used:** Kaplan-Meier, Logrank Test, Nelson-Aalen, Cox Proportional Hazards

**Tools used:** Python, Jupyter, Lifeline

### 2. Outage Monitoring and Modeling of Aboitiz Power Plants

**Description:** The stakeholder wished to monitor the behavior of the power plants managed by the company. Using survival analysis, time-based probability distribution and bathtub curves were made. The bathtub curves showed the hazard function of each plants they were monitoring.

**Techniques used:** Kaplan-Meier, Logrank Test, Cox Proportional Hazards

**Tools used:** Python, Databricks, Reliability, Google Data Studio, Tableau