Below is the data science questions relating to the Online Product Sales dataset. Please answer all questions below:

1. Review the dataset and summarize your thoughts on any structural issues:
   1. Is there missing data? Yes, there is missing data.
   2. Is the missing data random or structured: Are some attributes missing more than others? Some missing data is structured, some is random. For instance, Quan\_2 to Quan\_14 is structured missing. Missing at Outcome\_2 to Outcome\_12 is random.
   3. Are any data values glaringly errorneous? It seems that some columns are purposely duplicated. Some columns are purposely made up. Most of those columns are categorical columns. Remove duplicate columns training set has 382 unique columns.
2. Summarize your thoughts in the dataset using exploratory data analysis. Provide a graphic summarizing your results. How do you interpret these results?

Outcomes have exponential distribution.

Some categorical columns are zeros. (artificial data?)

Please see plots attached with email for more details.

1. Select one of the below Machine Learning topics and explain it to:
   1. Someone with significant mathematical experience.
   2. Someone with little mathematical experience.
   3. You can use graphics/ diagrams if you wish

Linear Regression, Logistic Regression, General linear model, Principal Component Analysis, Factor Analysis, K--‐means Clustering, Support Vector Machine, Singular Value Decomposition, Markov Process, Hidden Markov Model, Kalman Filter, Particle Filter, Fourier Transform, Monte Carlo, Markov-chain Monte Carlo, Decision‐tree, Random forest, Kernel density estimation, The curse of dimensionality.

Monte Carlo: Monte Carlo is the art of approximating an expectation by the sample mean of a function of simulated random variables.

b. An example of Monte Carlo is calculating Pi number by using the circle area encircled inside a unit square. The area of the circle is Pi/4. The area of the square is 1. By generating random points in the square, Pi is 4\*(total points inside the circle)/(total points generated)

a. The expected value of a function g of (possible multidimentional) X:

is approximated by

Please allow me to answer this part later.

1. Create several different predictive models using any number of methods you prefer (e.g. regression, clustering, decision trees, etc.). Show their out-of sample performance and compare them using Lift curves, ROC curves, and confusion matrices where appropriate.

Here I used two regressors: Gradient Boosting Regressor and Random Forest Regressor to predict the outcomes of test set. Performance measure was root mean squared error (RMSE). I found that the distribution of outcome is kind of exponentially decaying so that I took the log of the outcome before applying regression. This trick did improved predicting performance. Parameters for each regressor were manually adjusted. This process could be improved by using grid search.

I used weka software to find some optimum features in this modeling. I did not have enough time to apply backward feature selection. Idea is using Gradient Boost Regressor with all features. Then remove one feature at a time to find lowest RMSE. Repeat above steps until one find a subset of features where the RMSE is smallest. One could plot RMSE vs number of features and look for the minimum RMSE.

I think RMSE (or mean absolute error) was better to evaluate the predicting performance in this problem.

I was trying to speed up the modeling by doing a parallelization. But I got some bugs and could not finish it. I could fix it if have more time.

The code is note clean and easy to read yet. I hope that you will consider how I approach the problem instead.