**HW8 : Question 18.1**
**Describe analytics models and data that could be used to make good recommendations to the power company.**
**Here are some questions to consider:**
**The bottom-line question is which shutoffs should be done each month, given the capacity constraints. One consideration is that some of the capacity – the workers' time – is taken up by travel, so maybe the shutoffs can be scheduled in a way that increases the number of them that can be done.**
**….**

I've broken this assessment down to 5 major parts:

1. How to identify paying vs non paying customers
2. How to identify optimum routes to shut the power off
3. Advanced/Miscellaneous aspects
4. What models not to use in this exercise.

In each section I briefly discuss the various options and identify the most optimum option.

Let's dive deeper:

# 1. How to identify paying vs not paying customers

This is a pure classification problem.

- One way to look at this problem is to use a support vector machine approach.
    i. The variables that can be plotted on the axes are: credit scores and number of times they have been delinquent on payments.
    ii. Another SVM model could be run against customers who have paid after a series of non payment cycles. The number of non payments before a payment is received can be plotted on one axis, while the amount paid could be plotted on the other axes on this 2-dimensional graph.
    iii. The goal would be to reduce the number of software failures. So drive $\lambda$ down.

$$\underset{a_0,\ldots,a_m}{\text{Minimize}} \sum_{j=1}^{n} \max\left\{0, 1 - \left(\sum_{i=1}^{m} a_i x_{ij} + a_0\right) y_j\right\} + \lambda \sum_{i=1}^{m} (a_i)^2$$

Support Vector Machine

- Another way to look at it is using logistic regression, and use that to identify paying vs not paying customers.
    - The variables used could be salary, temperature, credit history and 5 consecutive non-payments

1. We could also treat the output of this exericse with a heuristic approach, and get expert opinion on who will pay vs who wouldn't, by asking the customer directly! The customer, if they respond, would naturally commit to paying, but we can ask them how long it would take them to pay.
2. If their estimate matches the average amount of time they take to pay (give and take a standard deviation) we go with that.

## 2. how to identify optimum routes

Now that we have the bad customers, the first step here to map where these bad customers are located. For this we can use k-means clustering to identify the right clusters of target (dud-)customers. Here you identify the optimum clusters, then for each clusters, identify the maximum cost cluster. Using the elbow graph you can figure out the minimum number of clusters required. Overlaying this with on a road map, allows the traffic planner to devise and optimum route to hit these targets within the least amount of time.

## 3. Advanced topics

Here we deal with a fundamental problem: What about the cost when we have to turn service on to customers who paid? We have to trek back to the same site and spend the energy to turn them back on. I believe this specific aspect can be captured also by a linear regression, where the variable is the customer count which was misclassified as duds, but ended up paying, and apply a -ve coefficient. Software can be used to predict the value of the coefficient, and R^2 can then be calculated to gauge accuracy of the model, along with the p-value of each variable to understand impactful vs not variables.

## 4. What not to use:

Equally important is to document what I would _not_ use in the modeling:

1. Weibull or geometric. Both suggest the customer (or cluster of customers) end up paying in the end. However, we are after the customers who NEVER pay, so this will always be an erroneous model to use.
2. Cusum to track change detection for a customer going from paying , or potential paying, to never paying. This is highly unpredictable and cusum totally ignores all other variables impacting the customer, such as major bankruptcy, job loss etc.
3. Personal data: It is costly to use and can be risky per legal requirements. Minimal data is necessary for identifying dud-customers.