In order to find an efficient and accurate solution to this problem, I will employ the use of logistic regression, clustering, and optimization to determine the best course of action for the power company. First, the problem must be broken down into steps that can be solved with an analytic approach:

1. Determine those who will never pay (which household should actually be shut off?). This will be done based on the set of factors below:
    a. Given the following per customer:
        i. Credit Score, Income, Magnitude of Overdue Payments, Time of Payments Overdue, Time as a Customer, & $\frac{Lifetime\ Amount\ Paid\ by\ Customer}{Lifetime\ Amount\ Owed\ by\ Customer}$.
    b. Use:
        i. Logistic regression model
    c. To:
        i. Obtain a response variable ("yes" or "no") of if the customer will never pay or if they will likely pay eventually.

Rationale:

The rationale behind this approach is that by using a logistic regression model, we can mathematically determine the probability that the customer will eventually pay. I have determined the logistic regression model to be the best model to use because of its simplicity to explain to the power company (as opposed to something like CART), and its use based on the linear relationship between the target variable (e.g will the customer pay?) and the predictors. This means that I am assuming that the relationship of each predictor is directly proportional to the target variable, which based on factors like income, credit score, and the others listed, I find this to be a reasonable assumption.

2. Prioritization of the Households to Shut Power Off

   a. Given:
      i. A dataset of only those who are designated not to pay the power company, as per the logistic regression ran in step 1 (but including the addresses of those selected).
   b. Use:
      i. K-Means Clustering Model
   c. To
      i. Geographically group similar customers together to ensure the later optimization step runs at the fastest speed.

Rationale:

The idea to use K-means clustering to geographically group similar customers will be very helpful for the later step of optimizing the routing problem[1]. In order to determine K, we will use the value of the number of available drivers for the clusters. Geographically grouping these non-paying individuals is important because it will minimize the amount of distance needed to travel between each household, if all the shutoffs are done in each cluster first.

3. Optimize the Route Used to Shut Off Power
   a. Given:
      i. Clusters and prior data associated with each cluster.
   b. Use:
      i. An Optimization Model
   c. To:
      i. Minimize the total travel time between each household.

Rationale:

The use of this optimization model would be to minimize the total distance between each cluster, and the travel times associated from cluster-to-cluster[1]. By using data from the clusters generated in the previous step, this step combines the prioritization of customers and optimization of the routes the power company should run to ensure the most efficient solution. The following considerations are taken when performing the optimization model:

Objective Function:

- Minimize the travel times between each cluster and data points inside each cluster

Decision Variables:

- Order of stops the driver must make

Constraints:

- Driver availability, geographical limitations (i.e we must use roads, can't drive straight through), average traffic time for major road networks, & number of trucks.

Discussion:

I chose this method of deducing which household to shut power off for because of its interpretability. By this I mean that the use of logistic regression rather than a more nuanced approach was chosen because it would be much easier to explain to the management at the power company. I also used clustering before optimization as it would prioritize the households based on proximity and magnitude of the payment owed, and it seems to be the standard based off of preliminary research into the subject[1,2]. In addition, predetermining the number of clusters to be equal to the number of drivers would already optimize for each available driver. However, given that we are not certain the size of the dataset I decided to include the number of drivers and the number of trucks in the optimization step. This is because the truck may break down and the driver may call out sick, which means that the number of clusters would not be equal to the number of available drivers, so minimizing the distance between the clusters could still be employed. It is also important to note that this model is likely to not be 100% in ascertaining which households would never pay. This is the case for almost any analytics model as a 100% accurate model is very scarce. So, in order to combat this, I would propose that this method of determining which households should have their power shut off should be run every month. This would update the parameters for each client and as each customer goes further into debt, the linear relationship dictated by the logistic regression would put them closer to the response of never paying their bill.

References:

1. https://www.sciencedirect.com/science/article/pii/S1877050923005604
2. https://www.altexsoft.com/blog/schedule-optimization/