# Problem Statement:

Before we get into what models are best to use for each situation, let’s break down the problem into its various components.

1. Selection of Individuals/Households that are unlikely to ever pay their bill(s) – We would like to filter out those individuals or households unlikely to ever pay their bill. We would like to avoid:
   1. Selecting those individuals who simply forgot to pay their bills.
   2. Selecting those individuals who are not able to pay their bills for the time being because of financial constraints.
2. Selection of which individuals/households to turn off power – We would like to prioritize which households to turn off based on the constraint imposed by the company on number of workers. We will be using vehicle deliver times as well as the constant cost of turning off power.

# My Approach:

As mentioned by Professor Sokol, something we can do is create a model that solves both questions at the same time. However, as also mentioned by Professor Sokol, this solution would likely not be practical and it may be more pragmatic to model each component separately. Let’s examine the first model.

**Model 1 – Selection of Individuals/Households that are unlikely to ever pay their bill(s).**

What data do we need?

In order of what I believe will be most useful include:

* Credit Score
* Income
* Past History of defaults on payment to any company
* Past power bill payment history (examining both interruption in scheduled payments as well as length)
* Value of Home
* Rent or Own
* How long they have lived there
* Single/Married/Number of people in household
* Method of Payment
* Occupation of bill-payer (perhaps certain occupations are more prone defaulting on payments, especially if present economic conditions are not favourable to that occupation. Such individuals may be those who may be able to pay their bills at a later time.)

And we also need to exclude factors such as Age, Race, Sex, Zip Code to prevent discriminatory practices.

Correlated Variables: Many of these variables are likely to be highly correlated. Although at this stage, we are simply trying to ascertain what variables would be of most use to us, we can consider correlation when factoring in the cost of obtaining the data. For example, if the cost of obtaining both income and home value is relatively high, then it might only be worth obtaining one.

Cost of Obtaining Such Data: The cost of obtaining this data is also very important to factor in. And we should have clear reasoning for why such data would be helpful, given that a case would need to be made to the power company for why they should spend on obtaining this data. Certain variables may also cost more to obtain than others. For example, if the power company does not have the credit score of each of its customers, then to pay a third-party company to obtain the credit scores of all of their customers would be very costly.

The best way might be to work to train a model with the data that the power company already possesses. By being quite liberal with who they perceive to be a possible individual/household unlikely to ever pay their bill, they can they do a cost benefit analysis on whether the marginal benefit of obtaining a certain variable’s data for would be greater than the cost of obtaining the data (only for individuals categorized as a possible positive case).

What Model Should We Use?

As described by Professor Sokol, there are a variety of models we could utilize including probability models, classification models or a hybrid approach. And there are a number of pros and cons to all of these approaches.

I am personally in favour of a logistic regression model (or another probability-based model). If we were to use classification for example, then individuals who may only have 60 percent chance of never paying their bills might be lumped in with individuals who have a 90 percent chance of never paying their bills. And there is no way our later models would take into account those probabilities and the levels of risk associated with each customer. Using a probability-based approach, we could use the values generated in later models to help us determine which customers pose the most risk and are worth pursuing.

Conclusion:

Given that the power company has all the data mentioned above, a logistic regression model would be best in determine which households are most likely not going to pay their bill ever.

**Model 2 – Cost Estimation**

There are two things that we should be modelling for this particular component of the problem:

* Cost of power if a non-paying customer’s power is left on
* Cost of turning off power for customers who were willing and able to pay (maybe they forgot or some other reason).

Although Professor Sokol mentioned that we do not need to get into legal or societal costs of shutting down power, I did want to touch upon it. The reason being since the cost of a false positive (incorrectly shutting off power for individual who may have just forgotten, etc.) may be much higher than the cost of classifying an individual as a false negative (not shutting down power for someone who is not likely ever to pay their bill). By examining and quantifying the cost for both, we can create an equation that takes both into account and minimizes total cost.

The reason that I suspect that the cost of a false negative may be high is that potentially lawsuits could be filed against the company if they do not follow a procedure and process that is completely legal. And those costs could also be modelled against each household (high income households may be more of a risk as they may be more willing to file a lawsuit). It may be a wise move for the power company to sent out notices, call on the phone and potentially even send workers on site to speak with the property owner about their bill. This would have the benefit of avoid legal consequences as well as making our first model more accurate as the energy company would be able to filter out people who simply forgot to pay their bills.

Let’s examine what data we need to model the cost of power for non-paying customers.

What data do we need?

Fortunately, the data that is most crucial for this model will already be in the hands of the power company, past usage.

For customers that do not have a long history, in addition to utilizing whatever past usage the company does have for them, they could pull customer credit, financial and payment history data, as well as home value, square footage and household size.

While the company could pull such data for those households with a long history as well, obtaining such data will come at a cost and may only marginally prove the estimate as past usage will probably be a good enough indicator.

What models should we use?

There are once again a variety of models we could utilize. As mentioned by Professor Sokol, the hybrid approach in which we predict total cost associated with all clients is probably not the best. We can be liberal in who we choose to include in our model but including everyone I believe will be unnecessary and inefficient.

For those individuals who have a long history, I think a model like exponential smoothing would probably be most effective in forecasting power usage as it would incorporate trend and seasonality and tends to well on non-stationary data. While the GARCH Model could be utilized, it will only help us determine the variance in our estimates rather than the actual cost. And ARIMA tends to work much better on stationary time series data.

For those individuals who do not have a long history with the company, I think a clustering and regression model would work well. I would imagine that for the data we obtained to use for this specific model (customer credit, financial history, etc.) would cluster well and that the variance in power usage would be relatively low for each of these clusters as well in comparison to a regression-only model. I suspect that power usage is right skewed and in a regression-only model, we may end up overestimating power usage for households with relatively low-power usage. We could use whatever history that the company has and incorporate it through normalization to account for seasonal effects.

Conclusion:

Using the results from Model 1, past energy usage of households, customer credit, financial and payment history data, as well as home value, square footage and household size, use an exponential smoothing model and a clustering/regression model to predict the next month’s energy usage.

**Model 3 – Selection of households**

So far, we should have both the probabilities of those households that are likely not to pay as well as the cost of leaving the power on for those households. Now the question is, which households do we turn off power for. The question really comes down to, which cost is bigger: The cost of leaving the household’s power on or the cost of turning it off. If the latter is bigger, then its not worth turning the power off. If the former is larger it’s probably worth turning off.

Now, if there were not other constraints, we could simply calculate these costs and turn off power for all households where it made sense. However, the constraint is that there may only be a limited number of workers and so households must be prioritized, factoring in the time it might take to turn off each particular household.

Data we will need:

Location, generic drive time estimates, time it takes to shut down power at a location, previous model data

What model should we use?

To calculate the cost of turning off a household, we need to calculate the cost each minute a worker is driving (includes equipment cost, gas, worker wage, etc.) multiplied by the generic driving time to that location and the flat cost of turning off power at a location.

To calculate the cost of leaving power on, we can simply multiply the probabilities obtained in model 1 with the expected power usage. It may be worth factoring in also the cost of misclassifying someone likely not to pay (customer leaves power company forever, etc.) That is beyond the scope of this particular assignment however.

While optimization seems like a great model to use, as I wouldn’t really know how to write the constraints and for all I know it could be an optimization problem that would take very long to solve, I am going to disregard it since a model I could not completely understand is not a model I would want to sell to the power company.

Both clustering and simulation could be used in conjunction with one another. Clustering is quick and easy to implement and by simply color-coding each data point, we might be able to heuristically determine what clusters we should be targeting. We could model several different scenarios and run simulations to determine what households we should be targeting to optimize for cost savings. While optimization might produce the absolute best solution, clustering and simulation would be easy to interpret for a client who may not be well versed in data analytics, would be straight-forward enough to create and should produce a solution that may not be the absolute best, but should be very efficient nonetheless. If the simulations we end up running end up contradicting our hypotheses from the clustering, then we can always choose to opt for an optimization model which may be more time-consuming but would give us the best guaranteed solution.

Regarding the question of whether the power company should hire more workers, the answer comes down to once again – whether the cost of shutting down the power is less than the cost of leaving power on. However, the cost of shutting down power may be more complicated than it seems. The cost of hiring additional workers may be quite a bit initially, as well as the fact that it may take time to get new workers up to speed. They could also take more time in shutting down power at locations or run into certain unforeseeable delays that experienced workers would not run into. Equipment would also need to be allocated for the new workers and all that could bring complications as well as additional costs.

Conclusion:

Given prior data from previous models, locations, generic drive times and time it takes to shut down a location and cost estimates of shutting down a location and per/min costs of driving (inclusive of all costs), use clustering and simulation to prioritize which households to shut down.

lumped in with individuals who have a 90 percent chance of never paying their bills. And there is no way our later models would take into account those probabilities and the levels of risk associated with each customer. Using

Cost of approaching the wrong customers could be potentially very very high. Smaller risk with huge loss. So it may be important to be careful of that. Before just shutting down power, it may be better to create another model that may factor in risk of consequences. Perhaps speaking to them before. Sending notices. Doing everything in a legal way and documenting all of it in case. Doing that process itself may also help ascertain which individuals who cannot pay not and may definitely help to filter out those who simply forgot. Both to avoid legal consequences, and to help better classify those who are deemed a risk. Since there will be more people coming and paying their bills which will make us more likely in determining who is a risk.