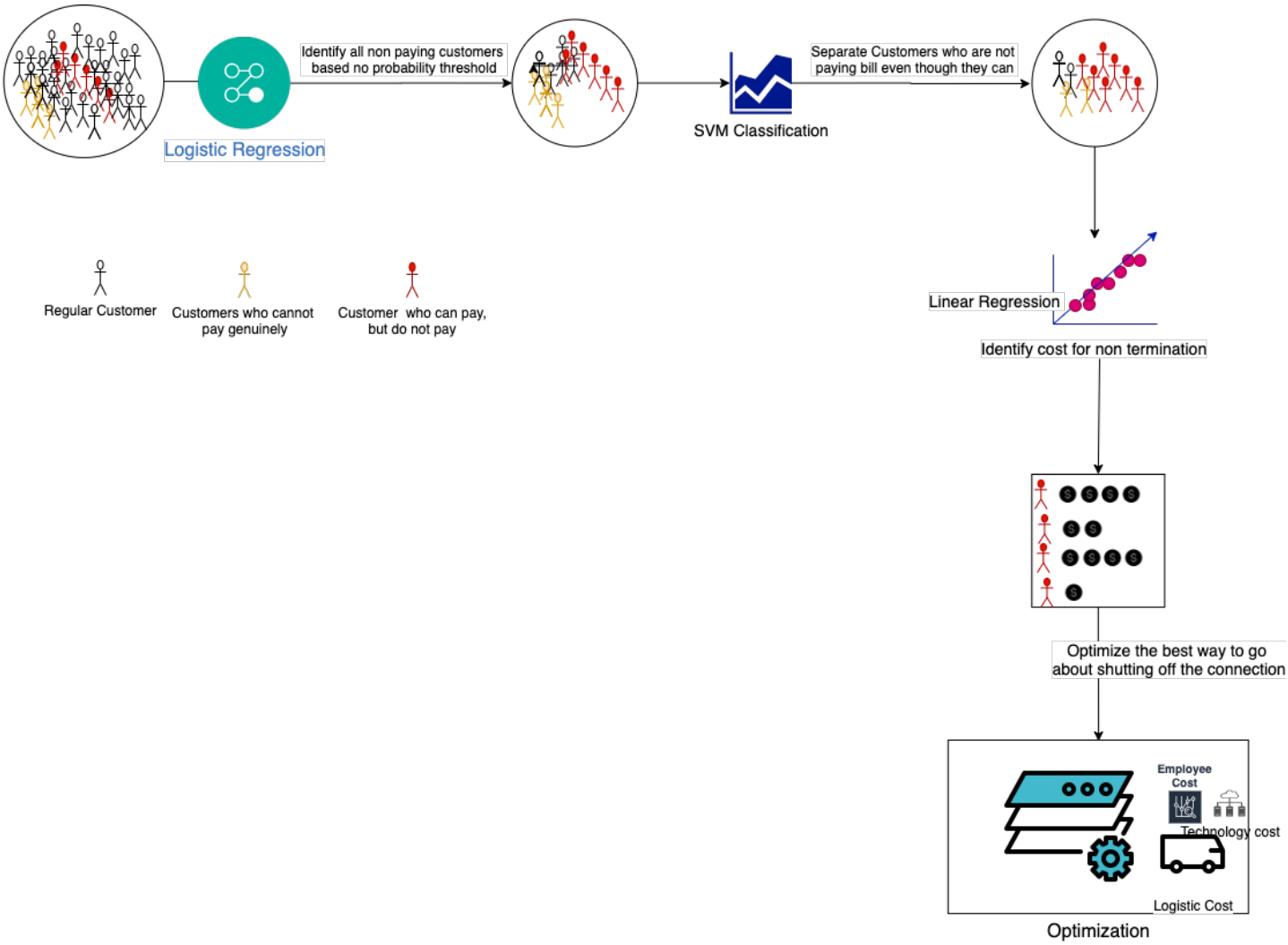
Assignment 8
-Prashant Kubsad

As professor Sokol mentioned in the lecture, am approaching this use case in three steps.

I have put a simple Illustration of the steps I am planning to perform:



<u>Step 1) Identification</u>

Identifying which customers' connection have to be cut. As mentioned in the problem, not all customers who are due to pay their bills are will have their connections terminated. We will have to identify customers who will not pay bill because of genuine economic reasons who will be supported by a fund. Identifying the pool of customers who can pay the bill but will not.

Initial thought was to pick a classification or clustering approach, but I feel like separating the customers who can pay but wont, and customers who are genuinely cannot pay is very tricky. I am thinking we might get higher degree of misclassification. So, am opting to go with logistic regression to predict the probability of the customers who will not pay the bill. Once we have the predictions, we can run classification to identify how man of them are genuine customers who have socio-economic conditions and are unable to pay.

Given the following factors :

- Income
- Credit History
- Credit Score
- Past defaulting history
- Type of house owned : Rent/Own
- Number of years in the current residence
- Past avg energy consumption

Use ***Logistic Regression***
To determine probability of a customer not paying the bill. Using historical data/subject matter expertise we can set the probability threshold beyond which we can consider the customer is not paying the bill. For ex.:0.60, if the probability determined by logistic regression is greater than 0.60, we can mark the customer as a 'Not paying Customer'.

Step 1 a)

Once we have set of customer who are not likely going to pay, we can run a classification model to identify who are intentionally avoiding the payments even though they can. We can use the similar factors as above for classification.

Given the following factors :

- Employment status
- Type of Employment
- Income
- Credit History
- Credit Score
- Past defaulting history
- Type of house owned : Rent/Own
- Number of years in the current residence
- Past avg energy consumption

Use ***SVM classification***
To determine if the customer belongs to a category of people who can pay the bill but are avoiding paying the bill. We could also have used clustering route here but I felt that this a case of supervised classification as we know we want to separate the users between who genuinely cannot pay versus who are not paying even though they can.

Step2) Cost Estimation

The next step in my analysis is to determine the cost of not terminating the power. This can depend on various factors like average energy consumption, number of people in the household and macro weather factors like number of days when the temperature is going to be below 50^o F or above 100^o F. We can use linear regression to predict the cost. Again, we will need subject matter expert and data from past history to determine the cost for each of the influencing factors.

Given the following factors :

Number of people in the household
Size of the house
Avg energy consumption of the household
Avg energy consumption in the neighboring household
Inclement weather risk factor(power company can consult meteorologists to come up with cost factor theta based on nature of weather event, ex.: Snow storm will bring down the temperatures very low and as a responsible company, they would not want to shut off power irrespective of the customer type.)
Avg number of days when temperature will be below 50^o.
Avg number of days when temperature will be above 100^o.

Use ***Linear Regression***

To predict the estimated cost if the power company does not cut the connection for every customer identified the step 1. As one of the peers mentioned in piazza it will be good idea to send out a notice to high risk customers about the power termination and readjust this regression based on the customer's response.

Step 3: Cost of termination

The last step in my analysis would be to estimate cost of terminating power based on the location of customers ordered by high risk to low risk. Using prediction values in the above step, we can sort the customers from highest risk and lowest risk. Once we have this list, we can estimate the cost of terminating the power supply which can be dependent on location of the customers, cost of termination and cost of reconnection if they pay the bill. We can use optimization in combination with other shortest path algorithms that we mentioned in the network lesson.

Given the following factors:

Travel cost
Power supply termination cost
Power supply re connection cost

Probability of payment of dues immediately upon termination
Technology cost

Use optimization
To determine the best way to go about disconnections. I understand that optimization is one of the hardest analytics tool to build, we will have to supplement optimization with subject matter experts on costs and probabilities of repayments. Also, use algorithms like Dijkstra's algorithm to come up with shortest routes and identify travel costs.

Other Observations:

1) As one of the peers suggested in piazza, we can think of using smart meters which reduces logistics cost whenever we have to turn it on or off. But procuring and installing them in the first place itself will incur some cost. To find out at what point we should start investing in these smart meters can also be analyzed using regression to predict when cost of termination increases beyond a point after which installing smart meters makes for sensible approach.

2) We could have gone clustering route to identify the customers who can pay but will not. But if the data is sparse and or not well connected, this approach can give us uninformed outputs.