Given:

- as much historical data as power company has on payment history
    - data will likely be monthly, but we can manipulate it to be one row per bill payer (see some of the calculated fields I would use for clarification)
- as many dimensions of data the power company has about their customers including:
    - Zip Code
    - Age
    - Housing Type (Apt. vs. house)
    - housing address
- Join this data (on zip code, for example) with public or privately acquired data sets, including
    - avg income of zip code
    - property cost
    - crime rate
    - Urbanicity
    - etc
- Calculate some additional dimensions using the above data, including:
    - length of time a tenant has lived in apartment
    - # missed payments / # of months in contract
    - Count number of missed payments
    - Count number of on time payments
    - Count no of missed payments eventually paid late
    - Classification field: three choices - either individual pays on time, late, or never pays
        - We know all of this from the historical payments data

Use:

- **Classification**: Since we are able to deduce the correct classification from the raw data, classification would be the best model to use. I would either use SVM or KNN model to try and see how well we can predict whether or not a person will be in the correct bucket based off of the dimensions and calculated fields. We will definitely need to scale the data here.
- **Validation:** I would split the data into train, test, and evaluate. Additionally, i would use K-Fold cross validation with k=10 parts to try and get the best model possible
- **Optimization: [**When we run the middle in a live scenario] To avoid misclassifying, we can add an optimization term that adds some distance to any point that gets classified as Never Pays. This way, the Minkowski Distance will be artificially inflated for those points and it steers the classification algorithm to classify them as another bucket

To:

- Once this model is created, we can begin to understand which individuals the model classifies as likely to never pay
- I would advise not using the model for a 6-12 month period to let it train even longer on real data and see how accurately it predicts those who will never pay. This will also give time to further optimize the model
- Once we know the individuals the model identifies as never going to pay, we can begin to focus efforts on optimizing how the power company thinks about shutting off power
    - Since you know the address of the individuals, it will not be hard to group individuals by geography
    - You can have a rule about number of months delinquent before shutting off power
    - Send employees out to a concentrated geographic area with the most individuals who are both identified by the model as never going to pay and have passed the threshold of months without paying.