Tuning KNN – Normalizing distance and picking k

Let's say we work at a credit card company and we're trying to figure out if people are going to pay their bills on time. We have everyone's purchases, split into four main categories:

- Groceries
- dining out
- utilities
- entertainment.

What are some ways you might use KNN to create this model?

Possible units would be:

- Groceries - # of items, $/item, frequency of purchase
- dining out – frequency, $/dining event, $/item
- utilities – daily/weekly/monthly consumption, $ per type
- entertainment - # of events, $ per event, frequency

The end goal could be to predict $/day using KNN.

What aspects of KNN would be useful?

1. It would be useful to normalize data using x-scores, since the types of expenditures vary widely.
2. If weighting, less weight could be placed on utilities since these type of expenditures are almost liked fixed costs, similar to groceries, while entertainment and dining out have wider range of choice.
3. The choice of k would depend on unit of time – whether $/day or $/month