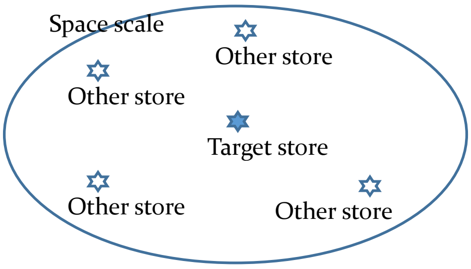
Homework 5: Location for New Target Store

Siqin Yang

Tackle the Problem

In finding out best locations for new stores, space scale is one important factor to consider. Since there will be not only external competitions with competitors, but also cannibalizations with own stores, I have to focus on revenues of stores within a certain space scale, rather than revenue of only the newly open one, to avoid severe internal cannibalizations.

Therefore, each own store would be regarded as the target store in an observation. Other stores within the space scale, no matter they belong to competitors or ourselves, would all be regarded as other stores.



For easy illustration, I would use T for the target store, O for other stores belonging to ourselves, C for those belonging to competitors, and S for the space scale.

Model Design

- Model: A supervised and regression model.
- Input: Three types of data: information about own stores O within the space scale S, information about competition stores C within S, census data within S.
  - Information about O are internal target sales data.
    - For each T, there may be various number of Os within S. In order to make the number of inputs a constant number, I would use weighted average of internal sales data.
      - For each O, their weights is related to their distance to T. Closer Os have bigger weights, which sums up to 1.
  - Information about Cs is their weighted distance to T. Closer Cs have bigger weights, which sums up to 1.
  - For census data within S, there may be multiple zip codes involved for Ts.
    - For each T, there would be a space circle with T as the center. And space circles may cross multiple zip codes. Therefore, I would also use weighted data. Zip regions with larger covered area by S have bigger weights, which sums to 1.
- Output: Weighted average of annual revenue of own stores(T & O) within the space scale.
  - I would give T a relatively big weight and Os small weights, which is

related to their distance to *T*.

- ✧ Observations: Records of all current own stores. Each own store would be treated as target store *T* once, and could be treated as other stores multiple times.
    - ■ I would use last two years' data to track most recently trend.

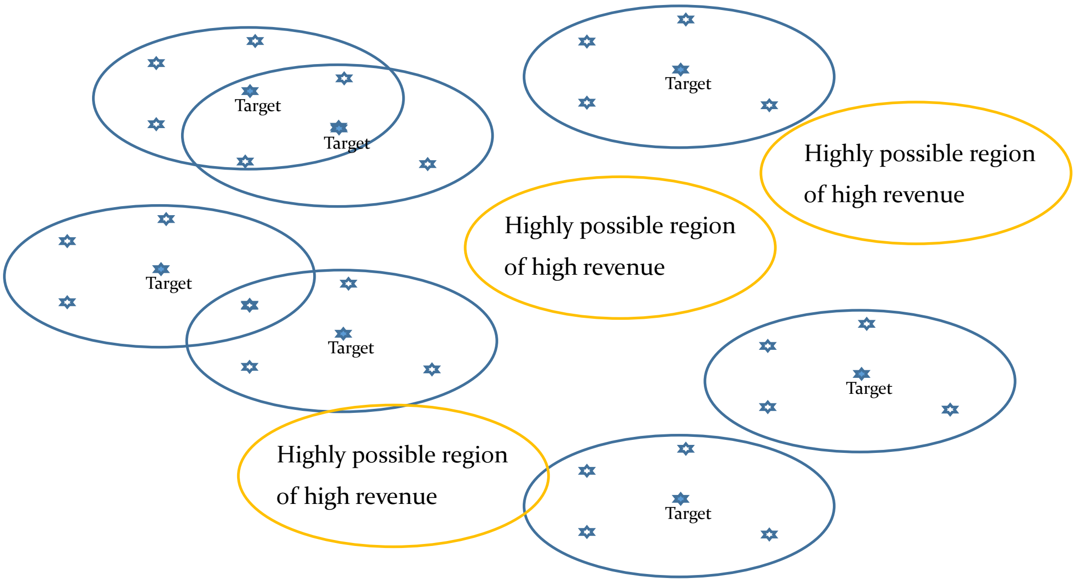
Below is the framework of data.

Target	Inputs (w.r.t space scale)			Output
Store	O(Weighted)	C(Weighted)	Census Data	(Weighted)
1	Z11, ... Z1n	D1	X11, X12, ...	y1
2	Z21, ... Z2n	D2	X21, X22, ...	y2
3	Z31, ... Z3n	D3	X31, X32, ...	y3
......		......		......

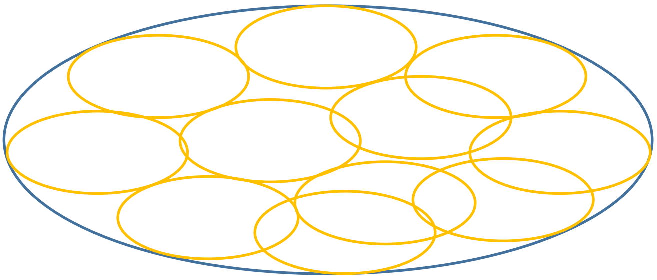
<u>Proposed Framework</u>

- ● Firstly, I would talk to the management to come up with a reasonable space scale *S* to take into account.
- ● With *S*, I could prepare data needed for each *T*.
    - ■ I would randomly split the date to train on 3/4 of observations and test on 1/4.
- ● With inputs and outputs, I could build the supervised regression model. By maximizing accuracy, I would have my model well-tuned and trained.
    - ■ I would use some techniques like feature engineering, feature selection and model tuning, to find out the best model.
- ● Now I could use the model to do prediction.
    - ■ First I would make assumptions to select regions to do prediction on.
        - ◆ I may start with the assumption that it's better to open a new store in a region with less existing other stores, to avoid competitions and/or cannibalizations.
        - ◆ I would talk to the management, to get their professional opinions on making assumptions.
        - ◆ Besides, I would look into the weights of my trained model, to see whether certain factors positively or negatively, and by how much, affects the revenue. For example,
            - ● if information about *C*s has a relatively small absolute weight, then I would not take competitors into account.
            - ● if variable *X1* has a relatively big positive weight, then I would focus more on regions with high *X1*.
            - ● if variable X7 has a relatively big negative weight, then I would focus more on regions with low X7.
    - ■ By making management-supported and data-driven assumptions, I could select regions satisfying the assumption.
        - ◆ For example, I could draw space circles for each existing own stores, and find out regions that are not severely overlapped, i.e. less existing *O*s.

    - Basically, I would get a map like this:



    - The reason I make assumptions is to select a number of highly likely regions, in stead of doing predictions on the universal map. But I could do so by dividing the map into several regions, same scale as the predifined *S*, and do predictions on all regions.
        - On dividing the map, I would trying to cover as many space as possible, and if the area of a uncovered region is bigger than a number *N*, I would make overlapped regions instead.
        - The way of dividing the map into several regions is basically like this:



        - I would probably make the space scale a grid, which would be simpler to divide.

- Now that I have a list of regions *S'* with high possibility of high revenue, I could do prediction within the region.
    - I would use the position of the center as the representative of regions.
    - Predict regions with existing *O*s and *C*s are straightforward.
    - To predict regions without existing *O*s (and/or *C*s), I would do the following steps.
        - Method 1:
            - If there aren't any *O*s within *S*, then I would set *Z*s as zero, to represent no close *O*s.
            - If there aren't any *C*s within *S*, then I would set *D*s as a fix large number, to represent no close *C*s.
        - Method 2:

    - I would use method like KNN, to map census data (and information about Cs, if available) to predicted revenue.
    - Then I could use only census data (and information about Cs, if available) to do the prediction.
- Then I would rank my predicted revenues *R*s and find out the corresponding zip code for the highest ranked position, where I would suggest the management to open the new store.

<u>Alternative Design 1</u>

- Above I use weighted average of various features as my inputs, which may lead to a certain level of error. Therefore, I could use features of individual *O*s and *C*s.
    - ■ To do so, I would first set the amount of *O*s and *C*s, *M1* and *M2*, so as to make the number of my inputs constant.
    - ■ Then I would select the *M1* closest *O*s and *M2* closest *C*s.
        - ◆ If there aren't enough number of *Os* within *S*, then I would set *Z*s as zero, to represent no close *O*s.
        - ◆ If there aren't enough number of *C*s within *S*, then I would set *D*s as a fix large number, to represent no close *C*s.
    - ■ Below is the framework of data.

Target Store	Inputs (w.r.t space scale)							Output (Weighted)
	Os			Cs			Census Data	
	O1	……	O_{M_1}	C1	……	C_{M_2}		
1	Z11, …	……	Z_{1M_1}, …	D1	……	D_{1M_2}	X11, X12, …	y1
2	Z21, …	……	Z_{2M_1}, …	D2	……	D_{2M_2}	X21, X22, …	y2
3	Z31, …	……	Z_{2M_1}, …	D3	……	D_{3M_2}	X31, X32, …	y3
……				……				……

- Other implementation details would be same as the framework above.

<u>Alternative Design 2</u>

- Alternatively, I could focus on zip5 level and therefore provide a list of ranked zip codes. The model design is therefore no longer targeted at target store, but at zip codes.
- First I would calculate the average annual revenue for zip codes where there're existing own stores, and rank them.
- Then, I would make following changes to the model:
    - ■ Observations of the model would be zip codes. One observation is one zip code.
    - ■ Inputs are still information about own stores *O* within space scale, information about competition stores *C* within space scale, census data within space scale. However, the space scale is no longer predefined but the region according to the zip code. Besides,
        - ◆ internal target sales data about O are no longer weighted but just the average.

        - distances of *Cs* is no longer weighted but just the average, too.
        - census data is data about that zip code.
    - Output is no longer weighted but just the average of those have the same zip code.
- I could build the model and make predictions to new regions where there are no existing ones.
    - I would use the above two method to deal with the problem of no data for *O*s.
- With calculated revenue and predicted revenue, I could rank them and therefore recommend the management to open a store in regions with high revenue.