Big Data HW1

Data mining
in real estate sale and media data

Instruction

- According to the case study, **formulate** a forecasting model.
- **Explain** and **Justify** your modeling procedures step by step
  - General statement of the problem
    - What is your definition of best allocation?
  - Collection of data relevant to the model
    - What is the dependent variable in your model ?
    - What are the independent variables ?
    - Do you include other sources of open data? And why?
  - Selection of the model
    - Formulate your forecasting model, be it linear regression model or other non-linear forms.
  - Statistical evaluation of the model
    - Propose the diagnostics used in your forecasting model.

Submission

- Form groups of 1 - 4 members for this homework and fill in the group members in hw1_group_sheet
- Note the group need not be the same as the group in final project
- The requirement is the same for each group

Submission

- Upload the file as follows to ceiba
    - <group_name>.zip (e.g. example.zip)
        - report.pdf
        - readme (optional)
- Upload by one group member is enough

Submission

- The evaluation will be based on model feasibility and depth of detail in the report
- Encourage literature survey and creative but reasonable model construction
- Due at 04/13/2015 22:00
- Late Penalty: 10% per day

學生問題:

1. 宣傳支出, 成交收入 have negative numbers, is this an error or is it a real transaction?
2. The homework is about money allocation for marketing resources, but 宣傳支出 has no information about where the money was spent. It seems like **客戶參訪** could contain the information about allocation, but it has no money column.
3. In 對照.txt there are names associated with numbers, which I assume to be the index of the clients, however there are many indexes that are missing from the text file. Significantly, 228937730 is the most common index, but this index does not appear in the text file.
4. It would be very helpful to have more information about this data, where it comes from, what it is used for, what each column means, how it is collected, etc. Also, I can understand if the source of the data remain confidential, but it would be helpful to know what type of company it is and what is its business model so that we can make analysis that would be helpful.

學生問題:

請問這份作業具體是要做什麼呢？
從資料看起來
1.只有一年的資料
2.廣告投放只有地區跟金額, 無法知道這筆廣告影響到的是哪些客戶跟收入
我們想過把資料切上下半年度或分地區, 但都不太可行
model應該要長怎樣?input跟output?
能不能給個範例或模板, 這份作業真的有點抽象
謝謝

Real world data issues

1. 在 Real World 大數據問題的確會遇到資料錯誤以及企業為了保護隱私而做 data massage 的動作。
2. 在本案例中，資料經過線性變換以保護隱私資料。

2015/03/15（Analysis.txt）（對照.txt）

2015/04/13（20150413 data.csv）（20150413 對照.txt）

Real world data issues

1. 資料來源:營建業兼房地產商提供的營運資料, 包括從宣傳廣告的投放, 到客戶來訪, 到房地產成交的量化時間序列數據。
2. 類別
    a. 宣傳支出:所有針對房地產銷售專案的支出皆根據費用類別收集, 並且按照費用發生時間建檔。
    b. 客戶參訪:所有房地產銷售的客戶來訪皆根據參訪時間建檔。
    c. 成交收入:所有房地產銷售的成交收入皆根據成交時間建檔。
3. 宣傳媒介 (media)
    a. 宣傳支出:紀錄支出費用的類別, 包括平面廣告、網路媒體、簡訊廣告、銷售活動費、業務獎金......等等。
    b. 客戶參訪:所有參訪客戶皆填寫問卷, 紀錄該客戶所接觸的宣傳媒介。
    c. 成交收入:所有成交客戶皆填寫問卷, 紀錄該客戶所接觸的宣傳媒介。

Real world data issues

4. 銷售地區(location):紀錄該房地產子專案所屬的地區。
5. 日期:可能會受到員工資料輸入的時間影響。
   a. 宣傳支出:紀錄費用發生時間。
   b. 客戶參訪:紀錄客戶來訪時間。
   c. 成交收入:紀錄客戶成交時間。
6. 金額
   a. 宣傳支出:紀錄費用金額。
   b. 客戶參訪:每筆資料代表一個客戶參訪, 因此金額欄位為0。
   c. 成交收入:紀錄成交金額。

Example（）

- According to the case study, **formulate** a forecasting model.
- **Explain** and **Justify** your modeling procedures step by step
  - General statement of the problem
    - What is your definition of best allocation?

e.g.

定義 ”最佳” 行銷預算分配為「最高的銷售金額」

rationales: [......]

Example

- Collection of data relevant to the model
    - What is the dependent variable in your model ?
    - What are the independent variables ?
    - Do you include other sources of open data? And why?

e.g.

dependent variable (y) = daily transaction amount

independent variable 01 (x1) = daily expense of media 1

independent variable 02 (x2) = daily expense of media 2

independent variable 21 (x21) = daily expense of location 1

independent variable 22 (x22) = daily expense of location 2

rationales: [......]

Example

- Selection of the model
  - Formulate your forecasting model, be it linear regression model or other non-linear forms.

e.g.

linear regression model with 30-day time lag

$y\,(t) = a + b1{*}x1(t-30) + b2{*}x2(t-30) +\ldots+ b21{*}x21(t-30) + b22{*}x22(t-30) +\ldots+ \mu$

rationales: [......]

Example

- Statistical evaluation of the model
  - Propose the diagnostics used in your forecasting model.

e.g.

linear regression model with 30-day time lag

$y\,(t) = a + b1{*}x1(t-30) + b2{*}x2(t-30) + \ldots + b21{*}x21(t-30) + b22{*}x22(t-30) + \ldots + \mu$

Use adjusted R-squared for general model evaluation

Use p-value < 0.05 for individual variable evaluation

rationales: [......]