Homework 6: Portfolio Forecast Using Account level Data

Siqin Yang

Tackle the Problem

The bank would like to make use of not only portfolio level data, but also monthly customer summary data to predict future losses. And the bottleneck here is how to encode customer level data into portfolio level ones. To do so, I would use statistics of distributions to represent the monthly behavior data of 30 million customers.

Since customer behavior, as well as the economic environment, is constantly changing, I would like to set a fixed time window *T* so that I would only take data within *T* into account when I build models on any time point *P*.

Model Design

- Model: A supervised and regression model.
- Input: Portfolio level data as well as encoded customer data.
  - Both portfolio level data and encoded customer data are on a month level, but only about months within *T*.
  - Encoded customer data are statistics about distributions.
    - By distribution I mean the distribution, for each month, of all 30 million customers' behavior data.
    - For example, for month 1 and field 1 of the behavior data, I could calculate its central tendency(e.g. mean, median, sum), variation(e.g. standard deviation, range) and shape(e.g. skewness, kurtosis) of 30 million records and use them as my inputs.
- Output: Total portfolio losses over next 6 months of each time point *P*.
- Observations: Records about each time point *P*.
  - My time point *P* would be on a month level. One observation is about one month.
  - Several months (e.g. May 2020) could not be used as my observations, since I do not have their future 6-month losses as outputs.

Below is the framework of data.

	Inputs (within time window)				Output
Time Point	Portfolio Data (for each month)	Encoded Customer Data (for each month)			(Total portfolio losses over next
		Field 1	...	Field m	6 months)
Apr. 2020	Z11, Z12, ...	F111, F112, ...	...	Fm11, Fm12, ...	y1
Mar. 2020	Z21, Z22, ...	F121, F122, ...	...	F121, Fm22, ...	y2
Feb. 2020	Z31, Z32, ...	F131, F132, ...	...	Fm31, Fm32, ...	y3
……		……			……

Proposed Framework

- First I would talk to the management to decide the time window *T*.

- Then I would prepare data needed for each time point *P*.
    - Output for each observation could be calculated easily.
    - Portfolio data is already available for each month within *T*.
    - I would calculate the statistics for each field of customer data, for each month within *T*.
        - I would calculate as many statistics as possible at this stage.
    - Eventually I would have a considerably wide data.
        - For example, if I have 20 fields for customer data, and I calculate 10 statistic for each field of each month. And there are 12 months within the time window.
        - Then for each *P*, I would have 12 * 2 portfolio inputs, and 12 * 20 * 10 encoded customer inputs. That is total 2,424 columns.
- Therefore, I would do feature selection to reduce the dimension of my inputs.
- I would randomly split the date to train on 3/4 of observations and test on the rest 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.
- With model and available inputs (since they are about past data), I could predict the total losses of next 6 months and let the bank know.

<u>Alternative Design</u>

The framework above use data of each months (within time window) as inputs. An alternative way is to use a weighted average of all past months' data.

- Specifically, inputs are still portfolio level data and encoded customer data.
- However, portfolio level data and encoded customer data are no longer on a month level, but calculated as the weighted average of all past months' data.
    - Weights are based on for how long each month is away from the time point *P*. Closer times have bigger weights.
    - Encoded customer data are still statistics about distributions, but averaged according to weights.
        - For example, for field 1 of the behavior data, I could calculate its central tendency(e.g. mean, median, sum), variation(e.g. standard deviation, range) and shape(e.g. skewness, kurtosis) of 30 million record. I would calculate these statistics for each month, and then average them to one number based on weights.

Below is the framework of data.

	Inputs (weighted average of all months before time point)				Output
Time Point	Portfolio Data (weighted)	Encoded Customer Data (weighted)			(Total portfolio losses over next 6 months)
		Field 1	…	Field m	
Apr. 2020	Z11, Z12, …	F111, F112, …	…	Fm11, Fm12, …	y1
Mar. 2020	Z21, Z22, …	F121, F122, …	…	F121, Fm22, …	y2
……		……			……

- The main difference is that, in this way, for each *P*, I would have 2 portfolio inputs, and 20 * 10 encoded customer inputs. That is total 222 columns, instead of 2,424 columns.

- The advantage of this design is that I could make use of all past data, not only those within time window.