Kyle Lim
Cpt_S 451

Project Milestone 2
7/4/24

Business Metrics of Classification

In Project Milestone Two, there are several metrics to classify a business; however, before proposing these metrics, a summary of the yelp input data will be provided for reference - only the most pertinent information will be displayed. Firstly, the yelp_business.JSON file provides the following information:

Summary of Business

- Business location (address & zip code)
- Average star rating
- Total reviews
- Restaurant price range
- Check-ins (with data)
- A list of attributes:
  - Ambiance
  - Categories

Summary of Zip Codes

- Zip code
- Median income
- Mean income

Given this list of information, there are grounds for several different contexts. These contexts will be listed here:

1. Average and median income for a zip code.
2. Volume of recent reviews can be derived from other yelp input files.
3. Categories indicate services a business provides.
4. Ambiance categorizes the environment of a business.
5. Recent average reviews, derived from yelp input files, can provide current business's performance.

Popularity Classification:

To classify a business as popular, businesses in every zip code are selected. These businesses are partitioned by zip code. Because these businesses were partitioned with descending check-in values beforehand row numbers are created with order by fk_zipcode, num_checkins desc. The

Success Classification:

To classify a business as a success, every value above the median for (checkin count, zip code) in the business relation is found. Once businesses in the top 50th percentile are retrieved and partitioned, the businesses deemed failures (in the bottom 50th percentile for total check-ins) are pruned. Then of the remaining businesses, the algorithm determines whether the business has an average star rating >= 4. That is the success metric.

Popular Classification SQL:

```
with partition as (
    select
        name,
        business_id,
        num_checkins,
        fk_zipcode,
        city,
        row_number() over (partition by fk_zipcode order by
fk_zipcode, num_checkins desc) as rn
    from business
)
select business_id, name, num_checkins, city, fk_zipcode
from partition
where rn <= 10
```

Success Classification SQL:

```
with partitioned_vals as (
    select
        fk_zipcode,
        business_id,
        num_checkins,
        stars,
        row_number() over (partition by fk_zipcode order by
num_checkins desc) as rn,
        count(*) over (partition by fk_zipcode) as business_count
    from business
), median_half as (
    select
        fk_zipcode,
        business_id,
        num_checkins,
        stars,
        rn,
        count(*) over (partition by fk_zipcode) as business_count
    from partitioned_vals pv
    where rn <= (business_count + 1) / 2
)
select
    b.name,
    b.fk_zipcode,
    b.stars,
    b.num_checkins,
    case
        when mh.business_id is not null and mh.stars >= 4.0 then
'Success'
        else 'Failure'
    end as did_succeed
from business b
left join
median_half mh on
b.business_id = mh.business_id
order by did_succeed desc
```