15.072: The Analytics Edge **Fall 2021**
Short Homework 4: From predictions to prescriptions

Out: November 01; Due: November 11, 11:59 pm.

Distribution Planning

A retail company has been facing increasingly frequent stockouts, leading to missed revenue opportunities and poor customer satisfaction. The root cause is the combination of strong growth in online sales and limited capacity of the company's distribution centers (DCs). To address the situation, the company has contracted you as a consultant to assess its operations and recommend a strategy to expand its DC footprint. The events take place in January 2015, and any new DC would become operational in the second half of 2015.

Specifically, the engagement has two objectives: (i) forecasting online demand for each county in the Northeast region on a weekly basis from mid-2015 through to the end of 2017, and (ii) optimizing which DCs to build in order to meet future demand in the most economical way.

To support the analysis, you have access to the following datasets on Canvas:

- `Dartboard_historical.csv`: This file spans the three-year period 2012-2014. Each row corresponds to a week and a county in the Northeast region determined by its Federal Information Processing Series (FIPS) code. The dataset consists of 11 fields: County FIPS Code; state name; county name; county latitude; county longitude; year; county income per capita in dollars; county population; number of weeks elapsed since the beginning of 2012; and Dartboard's actual online sales amount, in dollars.
- `Dartboard_future.csv`: This file spans the "future" 2.5-year planning horizon, from mid-2015 to end-2017. The dataset has the same structure as `Dartboard_historical.csv` except that sales figures are not known. Population and income figures were obtained from official governmental projections.
- `Dartboard_DCs.csv:` the location of 20 DC sites, including 3 existing DCs and 17 potential ones. For each existing DC (Providence, Richmond, Youngstown), you have access to its capacity. For each potential DC, you have access to the fixed cost of construction (in $), the variable cost of construction (in $ per sq.ft.), and the maximum capacity (in sq.ft.).

In order to relate sales to DC capacity, we make the following assumptions. First, each pallet carries SKUs with a retail sales value of $1,000 on average. For instance, a sales volume of $500,000 requires approximately 500 pallets. Second, each pallet measures 48" in length and 40" in width. Therefore, each pallet requires a footprint of 13.33 sq.ft., which is rounded up to 13.5 sq.ft. to account for placement tolerances. Third, all DCs are standardized to allow five vertical tiers of storage. Third, products are transported by truck, and each truck holds 20 pallets. The average operating cost of a truck is $1.55 per mile, which includes driver's salary, diesel fuel, lease and maintenance costs, etc. Finally, the inventory turnover rate is 8 weeks; therefore, at any point in time, the inventory corresponds to the sales in the subsequent eight-week period.

Demand Forecasting

Based on exploratory analysis, you decide to model weekly sales at the county level as follows. First, you focus on sales per capita, as opposed to total sales. Second, you define three features: income, time and seasonality. Third, you aim to capture the facts that sales tend to increase by a given *percentage* week over week and that *relative* changes in income tend to translate into *relative* changes in sales. Therefore, you consider the following regression specification:

$$\log\left(\frac{d_{tc}}{pop_{tc}}\right) = \beta_0 + \beta_1 \log(inc_{tc}) + \beta_2 \cdot t + \sum_{s=1}^{13} \gamma_s u_{ts},$$

where:

- d_{tc} denotes the demand in week t from county c, in dollars;
- pop_{tc} denotes the population in week t from county c;
- inc_{tc} denotes the income per capita in week t from county c, in dollars;

- t denotes the number of weeks elapsed since the start of year 2012;
- u_{ts} is a binary indicator equal to 1 if week t falls in "season" s, and 0 otherwise.

a. Split the historical data into a training set, comprising all observations in 2012-2013, and a test set, comprising all observations in 2014. Fit the regression model to the training set. [**20 pts**]

(i) Report the coefficients of your model.

(ii) Interpret the coefficients of your model.

(iii) Report out-of-sample performance of your model (R^2, the MAE and the RMSE).

b. Project your demand forecast on the "future" 2.5-year planning horizon. [**20 pts**]

(i) Plot the projected inventory requirements against existing DC capacity, in pallets.

(ii) What is the expected capacity shortage over the last 8 weeks of 2017, in % of DC capacity?

Optimizing the Distribution Footprint

Based on your sales predictions, you now aim to determine (i) which of the 17 potential DCs to build, (ii) their physical capacities, in square feet, and (iii) which counties should be served by each (existing or new) DC. For simplicity, we assume that each county is served by a single DC. The key requirement is to eliminate stockouts through the planning horizon; to simplify further, we simply ensure that sufficient capacity should be available to cover demand over the last 8 weeks of 2017. The company's objective is to minimize the total cost of distribution over the planning horizon spanning mid-2015 to end-2017, including upfront DC construction costs and the transportation costs incurred throughout the horizon.

c. Formulate an optimization model to solve the problem. [**20 pts**]

(i) Provide a mathematical formulation of the model. Make sure to clearly define your inputs, decision variables, objective function and constraints.

d. Implement and solve your optimization model. [**20 pts**]

(i) Report the new DCs (if any).

(ii) Report the number of counties served by each DC.

(iii) What is the cost of DC construction and the total cost of transportation?

Hint: You will need to compute distances between counties and DCs. We will consider here the haversine distance, available in the **`Distances`** package. You can compute the distance between a point of latitude **`lat1`** and longitude **`long1`** and a point of latitude **`lat2`** and longitude **`long2`** as follows:

```
haversine((lat1,long1),(lat2,long2), 3958.8)
```

where 3958.8 is the Earth's radius in miles.

e. To evaluate your solution, propose a sequential benchmark as follows. First, find the cheapest set of DCs (and their capacities) that enable the company to avoid stockouts over the last 8 weeks of 2017. Second, given that set of DCs, optimize which counties are served by which DC. [**20 pts**]

(i) What are the costs of DC construction and distribution under this alternate approach?

(ii) Comment briefly.