

15.095 Homework 4

Due: Nov 28, 2018



In this assignment, you will be working with data adapted from the Kaggle competition <https://www.kaggle.com/c/favorita-grocery-sales-forecasting>. The data came from a large South American grocery chain, and the prediction task was to forecast sales for over 200,000 products sold in 55 grocery stores in Ecuador. Using this data set, we will consider the single-period Newsvendor problem with multiple items. Our objective is to determine how much of each product 1-100 to stock in grocery store #1 for the next day, August 15, 2017. For this problem, we will make the following assumptions:

- We restock the grocery store daily.
- The grocery store has a maximum capacity of Q units that we can stock.
- We can stock at most $\frac{1}{10}Q$ units of any particular product.
- The price of a product on promotion is 20% less than the base price.
- If a product is perishable and we don't sell our entire quantity, then we lose money equal to the cost of product times the number of extra units.
- If a product is non-perishable and we don't sell our entire quantity, then we can save the extra units to sell another day and do not lose money. (For revenue purposes, you can assume we can sell the extra units back to the supplier at the end of the day at full cost).

We will provide you with the following data files to use for the exercises:

- **items.csv**: The subset of 100 grocery store products that we will consider for this problem, with data fields including each product's ID, category, subcategory, base price, cost, and whether the product is perishable (1) or not (0).
- **sales.csv**: Historical sales data for these 100 products at grocery store #1 from January 2, 2017 to August 15, 2017. Data fields include date, product ID, number of products sold, whether the product was on promotion (1) or not (0), and whether the product was on display (1) or not (0).
- **sideinformation.csv**: Extra information from the past year which may help to predict sales data, including daily oil prices and national holidays of Ecuador. Some values for the daily oil prices that were missing have been imputed for you.



1. Formulate an optimization problem to maximize profit for grocery store #1 for the next day (August 15, 2017) given the provided historical data and assumptions. Describe all of the variables in the model and specify which are the decision/auxiliary variables, uncertain parameters, and constants. If you need to make any additional assumptions, explain these as well. 
2. Describe three approaches to solve the prescriptive optimization problem that you have proposed. What are the pros and cons of each approach?
3. Train a machine learning model to predict the product sales for August 15, 2017, and compare your model against a simple baseline (e.g. predict the sales from the previous day). What is the MAE of each method? 
4. Implement your best prescriptive method from part 2 incorporating the predictive model from part 3 and compare the results to a naïve baseline for different values of Q . What is the optimality gap between each method and the oracle solution (which can forecast demand perfectly in the future)? What is the percentage gain in profit from using the best prescriptive approach compared to the naïve baseline method?