

Data Analytics for a Grocery Store

You got an analyst position for an undisclosed grocery store in Edmonton metro area. The first day of work, you find out that your supervisor is on leave for a month and that nobody else is willing to supervise you in the meantime. The only things you have been given are some data (see 'grocery.accdb' and 'data_dictionary.pdf' on eClass) and a computer with a few software packages. You decide to spend your time on a small project.

Write-up due: December 9 at 11pm

Your Task

The goals of the project are:

Goal 1: Explore the data and discover two interesting findings.

- You may use any approach we learned in exploratory data analysis lectures.
- Here are some examples of findings (no penalty for using one of these).
 - i. Ex 1: A relationship between purchasing patterns and customer characteristics
 - ii. Ex 2: A relationship between purchasing patterns and the days of a week
 - iii. Ex 3: Profiling of customers or baskets

Goal 2: Build a prediction model that can help increase profit or efficiency.

1. Set up a business objective. Here are some examples (no penalty for using one of these).
 - Ex 1. Predict which customer will be profitable – Imagine that customers will fill out a short survey about themselves (note that our data is limited in customer characteristics) and their shopping behavior when they become our member. Based on the response, we want to predict whether a customer will be profitable or not, and if predicted to be profitable, we will suggest the customer to be a gold member. Note that you need to define what 'profitable' means.
 - Ex 2. Predict who is more likely to purchase meat (or any kind) next week – Based on purchasing history of past weeks and customer characteristics, we want to predict if a customer is likely to buy meat next week, so that we can customize mail-in coupons.
 - Ex 3. Predict if a customer is willing to add battery (or any item) to the current basket – These days, some grocery stores have a promotion in which the checkout personnel asks if a customer is interested in adding a certain item to their basket. It is often called 'the deal of the week'. Based on characteristics of a basket, we want to predict whether the customer wants to add a certain item. Note that this example makes a prediction for a basket, while Ex 1 and 2 make a prediction for a customer.
2. Build multiple prediction models for the objective.
 - Try at least three different types of models (KNN, LDA, Tree, Random Forest, etc).
 - For each model type, generate multiple models by using different sets of input variables.
 - Also, generate multiple models by changing model parameters (if there is any).

3. Evaluate the generated models and make a comparison. Choose the best one, explain your reasoning, and report its final performance on a test set.

The process will involve trial and error. You may have to change even your business objective. Do not expect to get an interesting finding or a good prediction model at your first try.

Do not forget that data preparation is the most important and time-consuming step in any data mining project. The given data set is a real one. You will find many missing/inconsistent/extreme data. You may use your judgement in determining how to prepare the data, but make sure you explain your reasoning in the write-up. You may also consult with me to make those decisions.

You are not limited to those techniques, models, and software you learn in this course. However, if you use something not covered in this course, make sure you explain how it works and how you used it, so that I can understand what you did.

Presenting Your Work

You will present your work in two forms, reflective presentation and a written report.

Reflective Presentation

- Each group has 10 minutes for presentation.
- Discuss your overall approach in handling this project. For example, what are your findings? What are the procedures you used in deriving the findings? Your observations on the dataset? How many methods did you apply, and what are their corresponding results? What are the prediction models you created? Which model you think is the best? What additional information you may need to improve the prediction ability? What was the most difficult part in this project, and how did you overcome it? You do not need to mention all of the above – they are examples.
- There is no need to show all of your detailed analyses/results in the presentation. You may present one part of your analyses as an example – it is up to your choice.
- The presentation should be professional and efficient. The target audience is a manager of the grocery store who has taken this course.
- **The presentation will be evaluated on organization, presentation skills, evidence of in-depth analyses, and time management. These four criteria will be equally weighted.**

Project Report

- Submit a report in PDF format, showing all your work done for this project.
- Do not exceed 7 pages (excluding appendix). Use font size 11 and 1 inch margins. It may be double- or single-spaced. You should keep it as concise as possible, but at the same time include all the discussions, charts, and findings you think are necessary.
- Appendix may include only the final form of your codes and screenshots of the results generated by the codes.
- You may also include screenshots of the results in the body text of the report, so choose where to include them as appropriate.

- Do not submit any original script file. Show all the necessary results in the report.
- The written report may contain different results that were not included in the presentation. The business objective in the write-up may be different from that in the presentation.
- Regarding Goal 1, present each of your findings as follows:
 - A well-written (and as short as possible) statement that describes your finding
 - A plot(s) or a table(s) that supports the validity of your finding. This plot or table must be easy to understand.
 - A brief explanation of how you found it and why it is important for this grocery
- Regarding Goal 2, present your prediction analysis as follows:
 - A well-written (and as short as possible) sentence that describes your business objective
 - A brief explanation of your models and how you created them
 - Evaluation of your models and comparison
 - Business implications for this grocery
- The report will be evaluated on the following. These four criteria will be equally weighted.
 - Interestingness: high if the potential impact of your finding or prediction is big
 - Validity: high if your evidence and analysis is convincing, low if it's not
 - Technique: high if it demonstrates your proficiency on data mining techniques
 - Writing: structure, logic, and writing skills

Bonus mark – Each group can receive some bonus marks by raising questions for other groups' presentation(s). Marks may vary depending on the quality of the questions.