# MKTG 6620: Final Project
# Check Canvas for Submission Date and Time

In order to complete the project successfully, please keep the following in mind:

1. This is a group project. Each group can have 2 members. Please take prior permission to form a 3 persons group. You are also free to submit this project as an individual submission.
2. The document has four parts
   I. The details of the business situation
   II. The submission format and instructions to complete the report,
   III. Data description in Appendix 1
   IV. Preprocessing requirements for different methods in Appendix 2.

## I. Business situation

This is your first day as data scientist at a grocery store chain. You look at your schedule and find that you have a meeting with both the brand manager and sales manager. At the meeting they tell you that they want to know how to make the Orange Juice category perform better than what it does currently. You learn that the grocery store chain sells two brands of orange juice Citrus Hill (CH) and Minute Maid (MM). MM gets higher margins than CH[1].

Listening to their description of the situation and their requirements it seems like the Brand Manager is interested in finding out what variables influence a customer's probability of buying MM. In other words, what can he do to increase customers' probability of buying MM? On the other hand, the sales manager is interested in building a predictive model in which he can simply predict the probability of a customer purchasing MM.

They look to you for providing a solution to their specific problems. You tell them that although both relate to MM, the nature of questions they need answered makes the analysis process very different. They tell you to use different analyses if needed but to inform them what you used to answer their question, why this method is better, and importantly what are your specific recommendations. They ask you to present a written report in which you clearly explain the method and your recommendations in a user-friendly manner.

You are given a dataset that contains 1070 purchases in which the customer either purchased Citrus Hill (CH) or Minute Maid (MM) Orange Juice [see Appendix 1 for a description of this dataset].

[1] Higher margin means they make more money on per unit sale of MM than CH

Considering the different goals of the brand and sales manager, you first need to figure out whether using the same method can help you answer both their questions or would you have to use different methods. Although you would prefer to use the same method, you realize that answering their specific queries as completely as possible is more important. To understand the problem better you write down the questions asked by each manager.

Question of the Brand manager

1. What predictor variables influence the purchase of MM?
2. Are all the variables in the dataset effective or are some more effective than others?
3. How confident are you in your recommendations?
4. Based on your analysis what are the specific recommendations you have for the brand manager?

Question of the Sales manager

1. Can you build a predictive model that can inform him the probability of customers buying MM?
2. How good is the model in its predictions?
3. How confident are you in your recommendations?

Recommended Analysis Steps:

1. Run both a logistic regression and gradient boosted trees before providing your recommendations. Provide a clear justification why you choose one model over the other. The justification should be based on the results you obtain from both logistic and gradient boosted tree models.
2. Check for correlation among variables. Some variables are highly correlated. Take steps to handle multicollinear data. This is especially essential for logistic regression (see appendix 2 for other preprocessing steps necessary to perform logistic regression).
3. When using gradient boosted trees, you will have to use an XAI tool PDP in your analysis to understand how the predictor variables are affecting the outcome variable.

## <u>II. Submission Format and Instructions to complete the written report</u>

Create your report in R Markdown and submit both the .rmd file as well as the .pdf file[2]. The written report should consist of the following sections.

[2] If you have never used R markdown, please see the :

1. Brief Introduction: https://rmarkdown.rstudio.com/lesson-1.html
2. Detailed Introduction: https://bookdown.org/yihui/rmarkdown/

**Cover Page**: Your Name (mention the name of all group members) and the Title of your report

**Define the problem**:

In this section describe what problems are being faced by the sales manager and the brand manager. Importantly what is your objective? What do they expect from you (this should not exceed 1 page).

**Define method(s)**:

Which method(s) of analysis are you planning on using and why? Defend your choice by explaining how the method(s) will help you answer the specific questions that have been asked. Would one method suffice or would you need more than one method[3].

- Did you scale/standardize variables? What type of preprocessing did you perform?
- What efforts did you make to reduce overfitting (i.e., train/test split, cross-validation etc.)?
- Describe the data and variables that you used in your analyses.
- What happens when you include all variables?
- Is it better to not include some variables? Why? Provide methodological rationale for your responses and any criteria you used to include or exclude variables. Did multicollinearity play a role in your decision to exclude some variables.
- Was the predictive performance of the Logistic and Boosted Trees model comparable when using metrics such as accuracy, ROC-AUC ?
- When you use the XAI method PDP with the Gradient Boosted Trees, did you observe the same pattern of influence of predictor variables on the outcome as you see in the logistic regression (e.g., PDP shows predictor variable X positive affects Outcome Y. Does Logistic regression also show positive influence of X on Y?).
- Explain in detail the analyses you conducted and any assumptions you made.

**Results and Conclusion**

In this section you would be using the output from the analyses to generate your recommendation. Explain in detail what your analyses suggest. The output from the method(s) that you used would help you in your responses. It is important to explain the R output in an understandable manner. Don't just copy paste your R output. Make sure to interpret and explain it.

In R markdown as well as the pdf file, your code and output should be visible. This way we can replicate your analysis.

[3] Hint: Use methods discussed in class so far.

Provide your recommendation to each of the managers. If you are unable to answer any specific question of theirs, be upfront about it and explain clearly why. Don’t blame the data. If you want to include results graphically feel free to do so. However, remember to explain the graphs clearly.

Important points to keep in mind for the report

1. Please clearly mark each section and subsection. Merging everything into long, unmarked paragraphs is unhelpful.
2. Ensure that you are providing proper reference and including the reference in a reference section.
3. Table and graphs should be clearly numbered with titles and should be self-explanatory. If you are adding them at the end of the report, please reference them appropriately.
4. Please edit the analyses output you include in the appendix so that it is understandable to any reader. As mentioned above, merely pasting R output would not suffice.
5. Check for spelling and grammatical mistakes. A methodologically sound report riddled with language errors suffers overall.
6. Revise, edit, and read the report carefully a few times before submitting.
7. Turnitin and other digital verifications will be used to check for plagiarism and similarity of your submission to any other submitted report.

## III. APPENDIX 1 – Data Description

Use the following code to download data in R and convert the variables to appropriate numeric and factor types.

```
OJ<-read.csv(url("http://data.mishra.us/files/project/OJ_data.csv"))

OJ[2:14] <- lapply(OJ[2:14], as.numeric)
OJ$Purchase <- as.factor(OJ$Purchase)
sapply(OJ,class)
```

OJ is a data frame with 1070 observations for the following 14 variables.

Purchase

> A factor with levels 0 and 1 indicating whether the customer purchased Citrus Hill (1) or Minute Maid Orange Juice (0).

PriceCH

Price charged for CH. Also called List Price for CH

PriceMM

Price charged for MM. Also called List Price for MM

DiscCH

Discount offered for CH

DiscMM

Discount offered for MM

SpecialCH

Indicator of special on CH. Special can be a free gift, loyalty points etc.

SpecialMM

Indicator of special on MM. Special can be a free gift, loyalty points etc.

LoyalCH

Customer brand loyalty for CH. That is, probability to buy CH (over MM) based on prior purchase behavior.

SalePriceMM

Sale price for MM. This is the difference between the list price and discount.

SalePriceCH

Sale price for CH. This is the difference between the list price and discount.

PriceDiff

Sale price of MM less sale price of CH

PctDiscMM

Percentage discount for MM

PctDiscCH

Percentage discount for CH

ListPriceDiff

List price of MM less list price of CH

# IV. APPENDIX 2 - Preprocessing

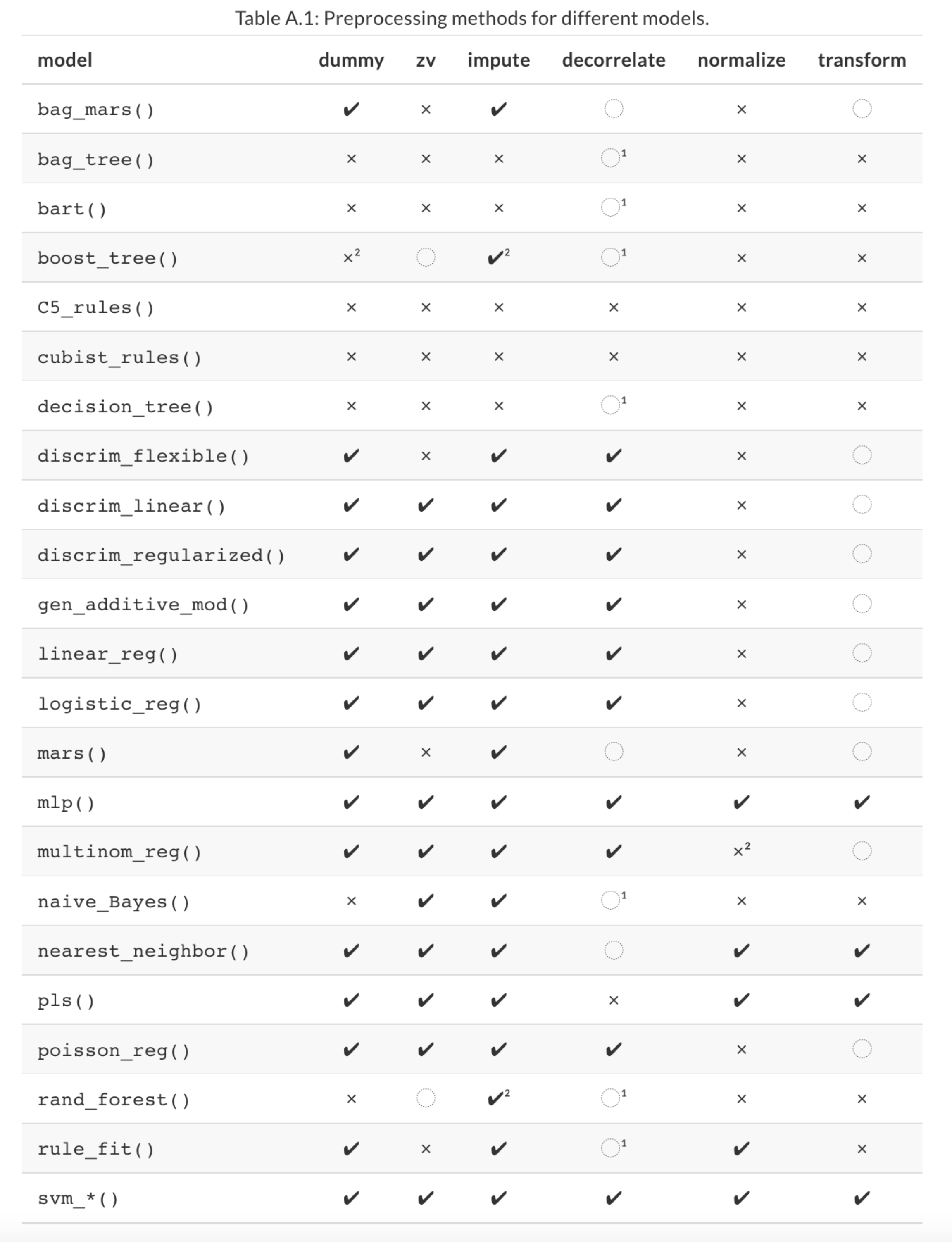
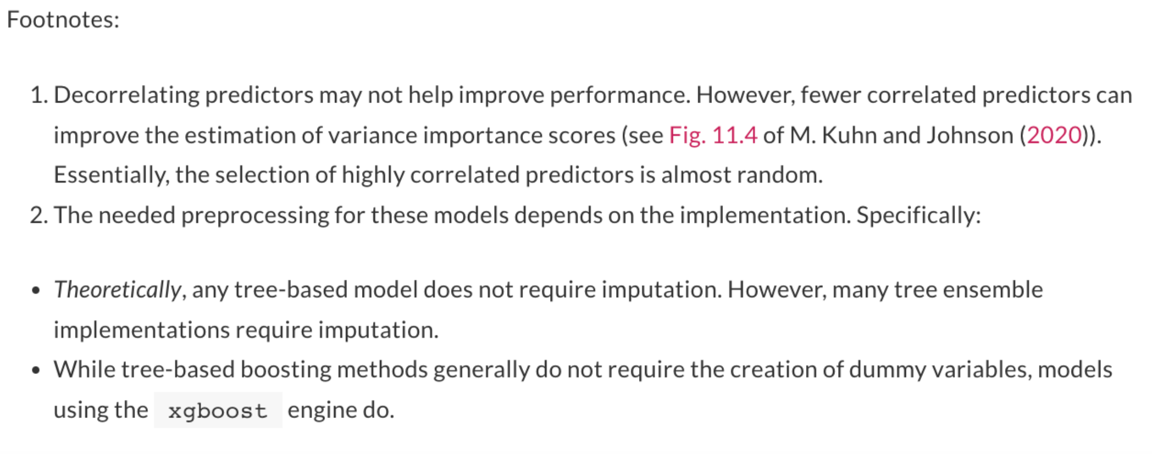


Table A.1: Preprocessing methods for different models.

| model | dummy | zv | impute | decorrelate | normalize | transform |
|---|---|---|---|---|---|---|
| `bag_mars()` | ✔ | × | ✔ | ◌ | × | ◌ |
| `bag_tree()` | × | × | × | ◌[1] | × | × |
| `bart()` | × | × | × | ◌[1] | × | × |
| `boost_tree()` | ×[2] | ◌ | ✔[2] | ◌[1] | × | × |
| `C5_rules()` | × | × | × | × | × | × |
| `cubist_rules()` | × | × | × | × | × | × |
| `decision_tree()` | × | × | × | ◌[1] | × | × |
| `discrim_flexible()` | ✔ | × | ✔ | ✔ | × | ◌ |
| `discrim_linear()` | ✔ | ✔ | ✔ | ✔ | × | ◌ |
| `discrim_regularized()` | ✔ | ✔ | ✔ | ✔ | × | ◌ |
| `gen_additive_mod()` | ✔ | ✔ | ✔ | ✔ | × | ◌ |
| `linear_reg()` | ✔ | ✔ | ✔ | ✔ | × | ◌ |
| `logistic_reg()` | ✔ | ✔ | ✔ | ✔ | × | ◌ |
| `mars()` | ✔ | × | ✔ | ◌ | × | ◌ |
| `mlp()` | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| `multinom_reg()` | ✔ | ✔ | ✔ | ✔ | ×[2] | ◌ |
| `naive_Bayes()` | × | ✔ | ✔ | ◌[1] | × | × |
| `nearest_neighbor()` | ✔ | ✔ | ✔ | ◌ | ✔ | ✔ |
| `pls()` | ✔ | ✔ | ✔ | × | ✔ | ✔ |
| `poisson_reg()` | ✔ | ✔ | ✔ | ✔ | × | ◌ |
| `rand_forest()` | × | ◌ | ✔[2] | ◌[1] | × | × |
| `rule_fit()` | ✔ | × | ✔ | ◌[1] | ✔ | × |
| `svm_*()` | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |



Footnotes:

1. Decorrelating predictors may not help improve performance. However, fewer correlated predictors can improve the estimation of variance importance scores (see Fig. 11.4 of M. Kuhn and Johnson (2020)). Essentially, the selection of highly correlated predictors is almost random.
2. The needed preprocessing for these models depends on the implementation. Specifically:

- *Theoretically*, any tree-based model does not require imputation. However, many tree ensemble implementations require imputation.
- While tree-based boosting methods generally do not require the creation of dummy variables, models using the `xgboost` engine do.

**Source: https://www.tmwr.org/pre-proc-table.html**