

**SCHOOL OF INFORMATICS & IT**

## Machine Learning for Decision Making

## Estimating the expected expenditure of a customer for Q4

Student Name (Matric Number): Javen Lai Le Yu (2202934B)

Tutorial Group: P03

Tutor: Zaw Htet Wai

Submission Date: 7/12/2023

# Declaration of Originality

I am the originator of this work and I have appropriately acknowledged all other original sources used as my references for this work.

I understand that Plagiarism is the act of taking and using the whole or any part of another person’s work, including work generated by AI, and presenting it as my own.

I understand that Plagiarism is an academic offence

and if I am found to have committed or abetted the offence of plagiarism in relation to this submitted work, disciplinary action will be enforced.

**Declaration on the use of Generative AI tools for assignments**

|  |
| --- |
| Describe how you have used Generative AI tools such as ChatGPT or Dall.E-2 in your assignment.  Show snapshots of the conversations with the AI tool (i.e., the prompts you used and the response you get from the AI tool). |
| PDCA + Coding:  <https://chat.openai.com/share/4207c097-4921-43ce-9b3a-a3e6523cbf3a>  <https://chat.openai.com/share/141fa7be-eeeb-4895-8b66-36aae1ff3a9d>  <https://chat.openai.com/share/8be6cac6-6c72-49af-85fa-1aa825ffaa0e>  REPORT:  <https://chat.openai.com/share/17ba933c-e697-4718-a226-57c7a0e3eee1>  I used ChatGPT to generate code under ‘Do’ for some PDCAs. I would write my intended outcome and what code I want, followed by a snippet of a sample code from Part 1 (to create new features):    I also used ChatGPT for debugging code and troubleshooting.:    I also asked it to rephrase some explanation to be clearer, fluent, and concise for easier readability and understanding:    Report writing:      I use chatGPT to generate ideas on how model can be used after deployment. I also ask it to give me ideas on how to phrase the ideas in concise, clear, and succinct manner. |
| How do you indicate the reference?  The content generated by AI tools are not retrievable except by the user who generated them, so they are considered non-recoverable sources. Although non-recoverable data or quotations in APA Style papers are usually cited as personal communications, with ChatGPT-generated text there is no person communicating. Quoting text from ChatGPT chat is therefore more like sharing the output of an algorithm, with a reference list entry and the corresponding in-text citation.  According to the official APA Style site, ChatGPT references should be cited as:  E.g. OpenAI. (2023). *ChatGPT* (Sep 25 version) [Large language model]. <https://chat.openai.com/chat> |

**Content Page**

Introduction....................................................................................................................................6

Preview of Model............................................................................................................................6

Model Interpretation...................................................................................................................6-8

Customer Study..............................................................................................................................8

- Low-Spending

- High-Spending

- Insights

Recommendations........................................................................................................................ 9

Model Development.................................................................................................................... 10

PDCA 1-5...................................................................................................................................... 11

Decision Tree Regressor & Classifier............................................................................................ 12

Random Forest............................................................................................................................. 13

Model Deployment................................................................................................................. 14-15

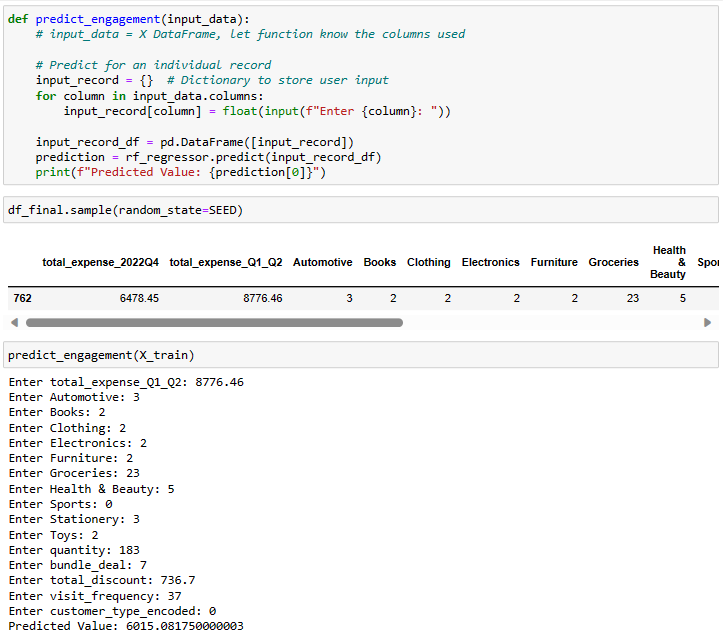
* Deployment Plan
* Decision-Making
* Business Planning
* Implications

References....................................................................................................................................16

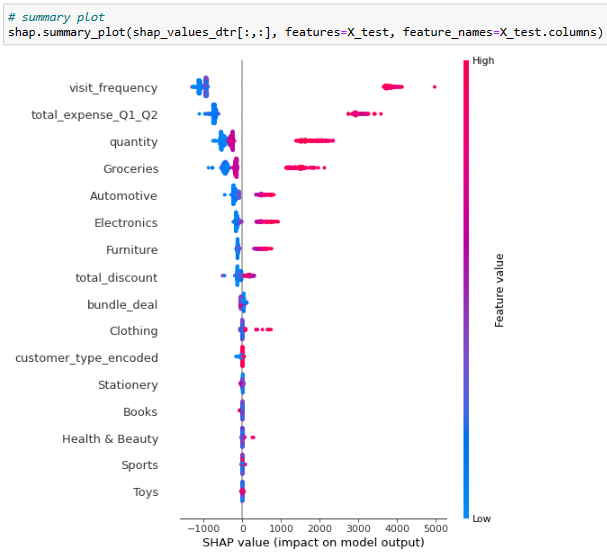
**Introduction**

The Random Forest Regressor is trained on Retail Hypermarket’s 2022 transactional data to predict a customer’s expected Q4 expenditure based on Q1-Q2 behavior.

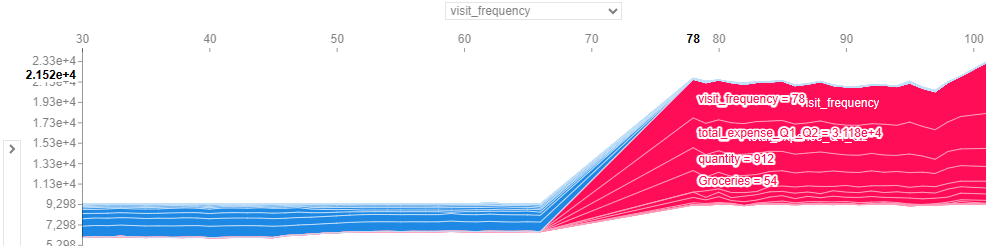
**Preview of Model**



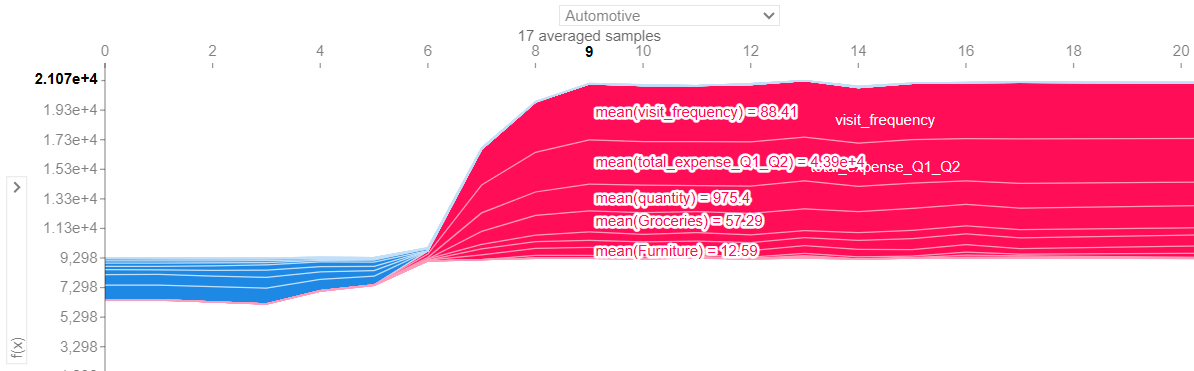
**Model Interpretation**

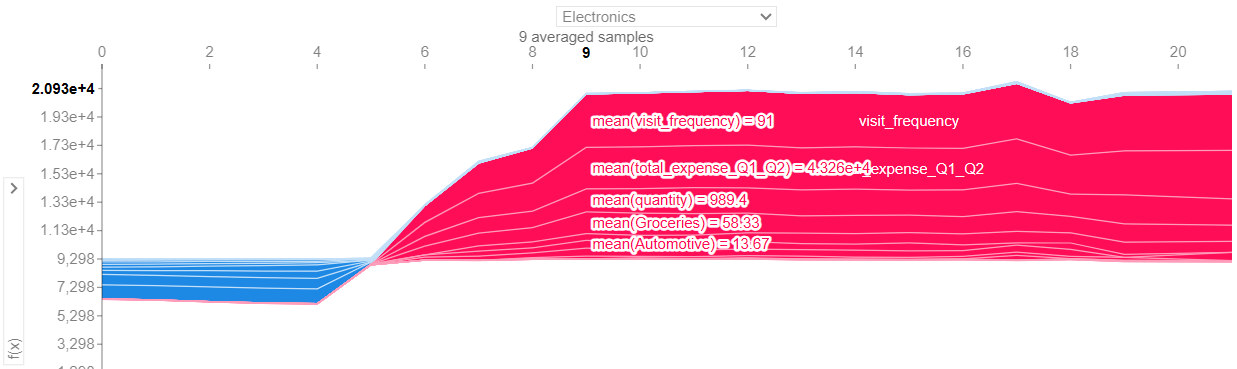


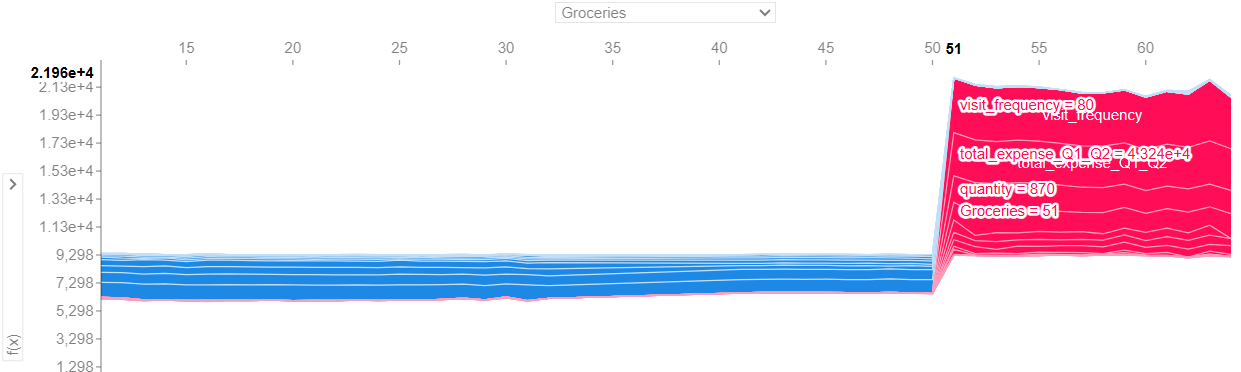
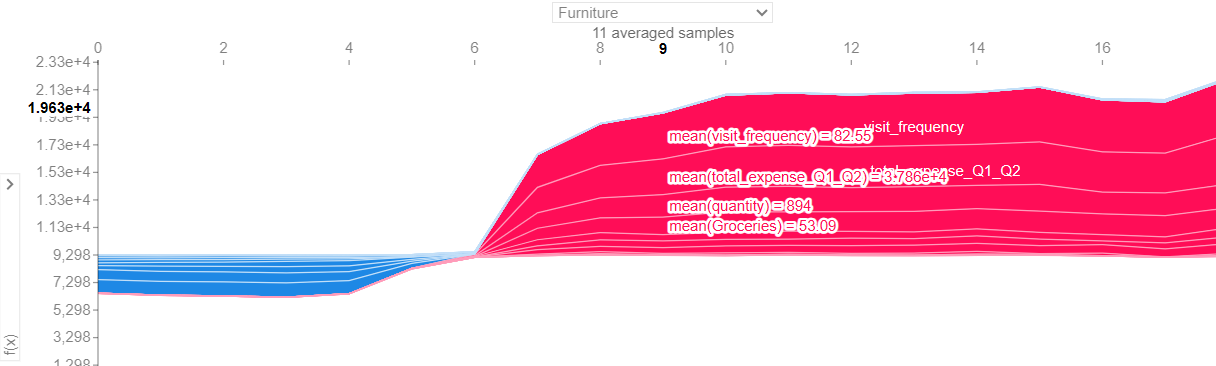
1. High visit\_frequency = higher Q4 spending:



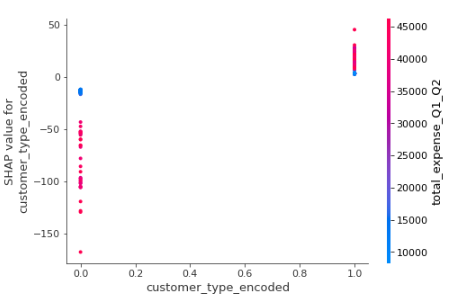
1. Engagement of at least 9 automotive, electronics, furniture, or 51 groceries leads to higher spending:







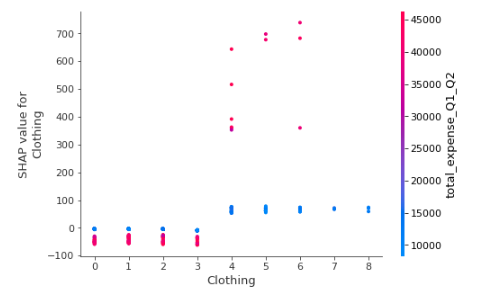
1. High-spending customer\_type B spent less during Q4, **suggesting potential loss of interest in shopping at Retail Hypermarket**:



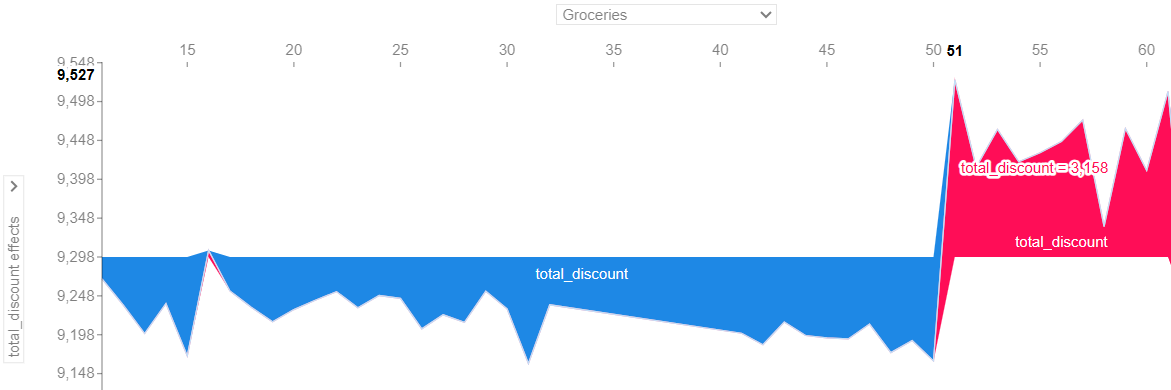
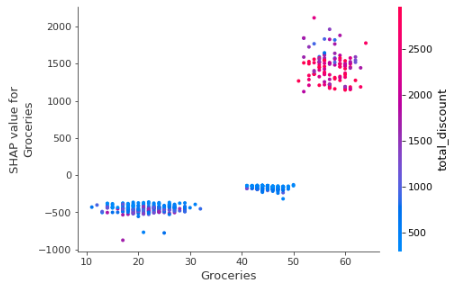
1. There are **2 types of clothing customers**:

* Branded consumers (focus on 1-3 brands)
* Thrift consumers (fast fashion)

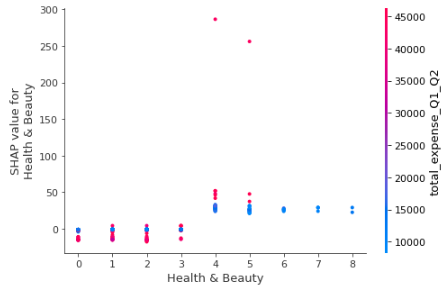
**Targeting Thift consumers is more profitable** than focusing on certain brands/designs:



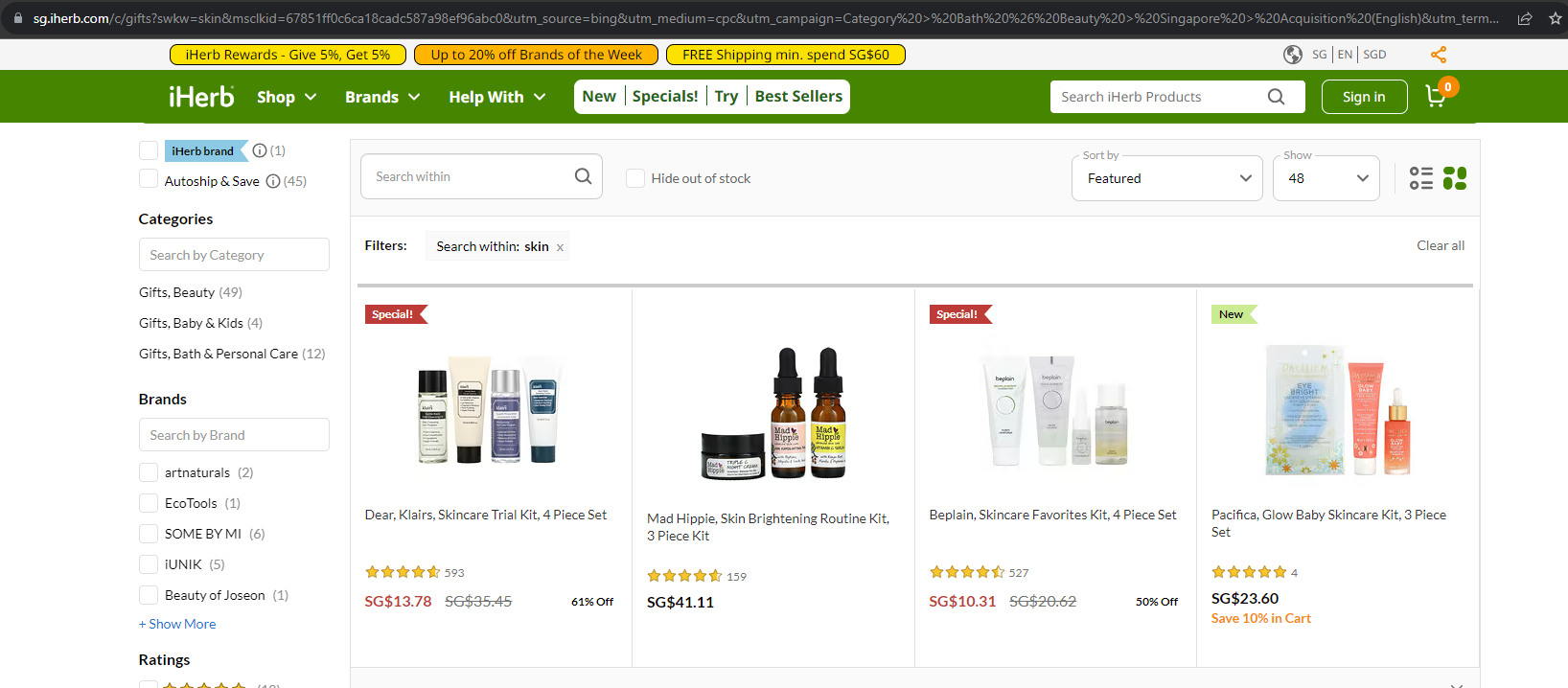
1. **Discounts could be effective** in improving sales of groceries because customers engaged in groceries received a lot more discounts:



1. Customers buying at least 4 **Health/Beauty products may potentially spend more**, revealing a cross-selling opportunity where products could be bundled to encourage customers to buy them together:



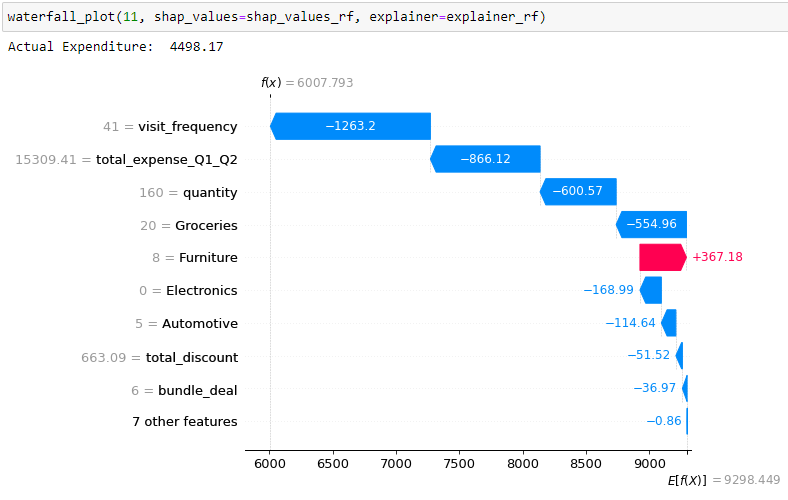
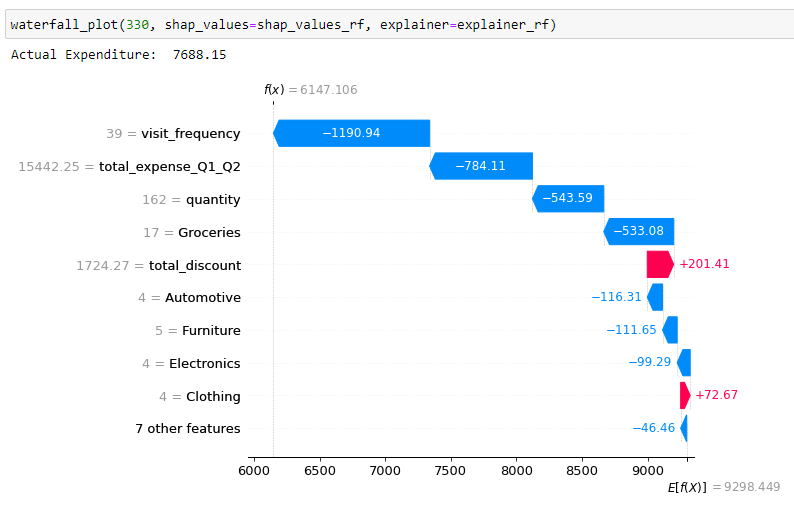
Skincare routines require a set of products:



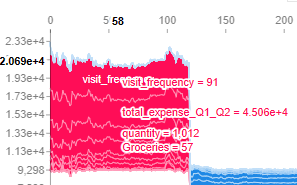
**Customer Study**

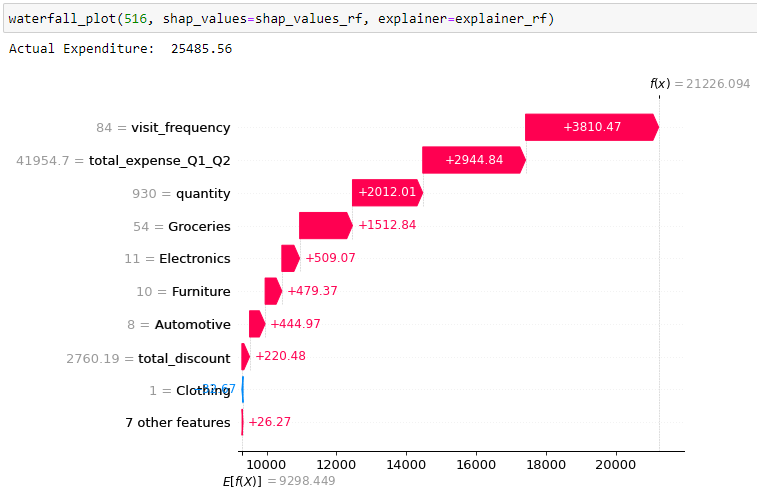
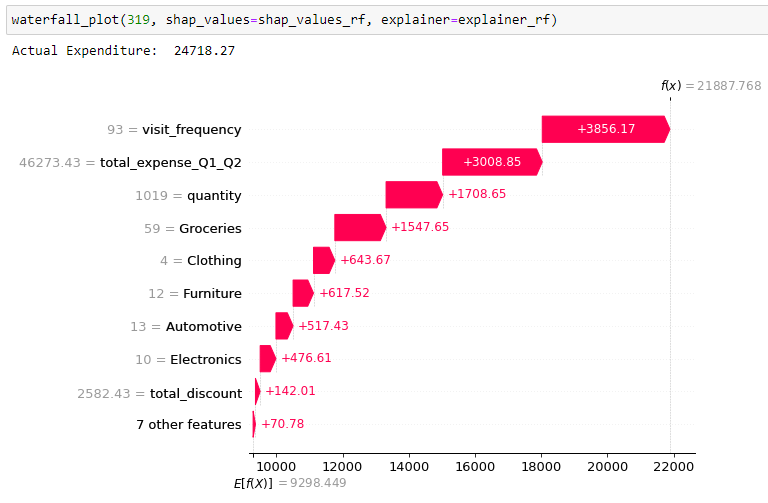
**Low-spender:**

****



**High-spender:**

****

**** ****

**Insights:**

- Customers who visited below 78 days are expected to spend less.

- Higher engagement in groceries, electronics, furniture, automotive, and clothing leads to increased spending.

**Recommendations**

1. **Build customer loyalty:**

Loyalty programs that encourage customers to shop at least 156 days a year for exclusive benefits (gifts/discounts).

1. **More variety of groceries and fashion/clothing:**

Improve engagement as customers have more choices, leading to increased spending.

1. **Focus marketing on automotive, electronics, and furniture:**

Customers who are engaged in lucrative categories spend more. Enticing customers to buy more of such products using compelling promotions could improve revenue.

1. **Bundle relevant beauty products:**

Ensure customers won’t miss out on a product from their skincare routine because they don’t have to manually find each product if products are bundled.

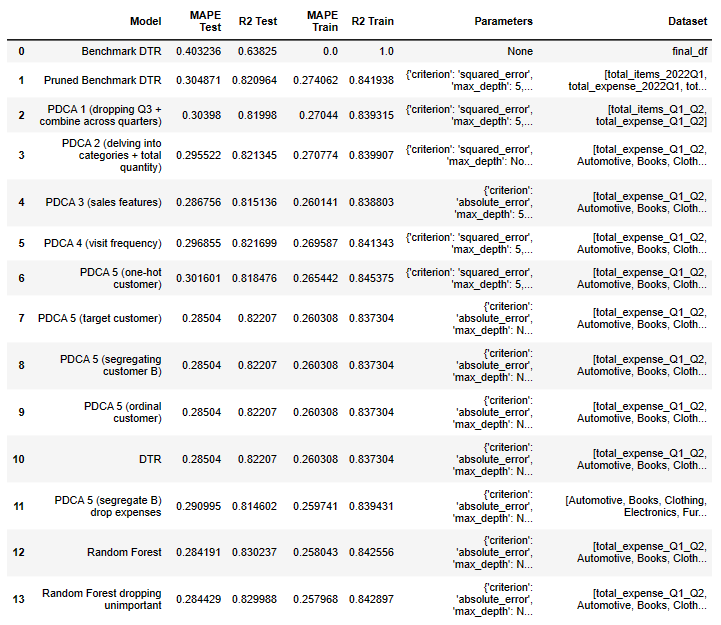
1. **Discount on slow-moving groceries:**

Discounts are effective in boosting sales for groceries, hence could be used to get rid of poor-performing groceries that are hard-to-sell.

1. **Rekindle engagement of Customer Type B:**

Conduct surveys to understand changing preferences and implement campaigns that attract them.

**Model Development**



**PDCA1:**

1. Rounded-off expenses to 2d.p. to match conventional understanding of prices.
2. Dropped Q3 features as total\_items and total\_expense across Q1-Q3 have high collinearity. This makes the model usable directly after Q2, giving the business the duration of Q3 to approve and implement strategies formulated using the model.



1. Lastly, total\_items and total\_expense for Q1-Q2 are combined to provide overview interpretation of how total expenses or items for Q1-Q2 influence predictions.

**PDCA2:**

1. To understand how engagement for various categories influences potential spending, I will count how many products a customer purchased for each category.
2. I will replace total\_items with quantity to represent amount of items the customer bought.

**PDCA3:**

To understand how common sales strategies influence customer spending.

1. Implemented discounts by deducting the unit\_price the customer bought a product for from its highest Q1-Q2 pricing.
2. Introduced Bundles/Cross-Selling that counts occurrences where customer buys multiple products a day.

**PDCA4:**

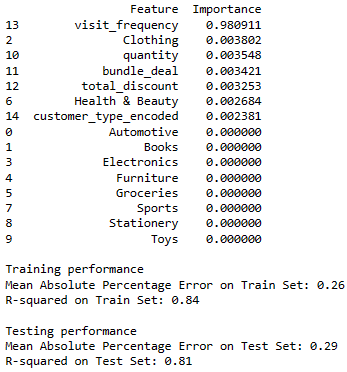
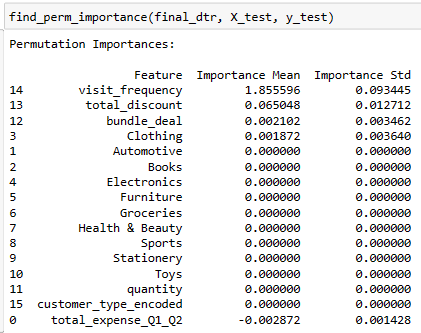
To enable data-driven loyalty programs, I added visit\_frequency as a metric to assess customer loyalty based on the number of days they shopped.

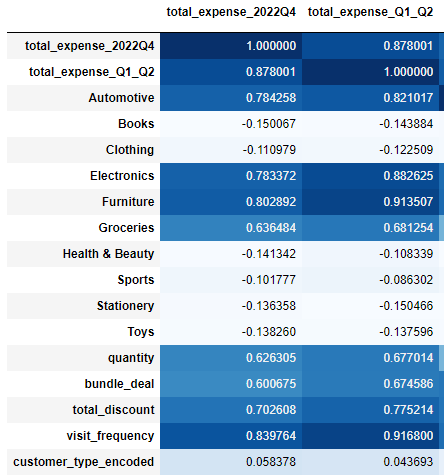
**PDCA5:**

Segmented customers by encoding 0 for customer\_type B and 1 for others based on the first letter of customer\_id, because type B is weakest spender across all quarters.

**DTR:**

Cannot contain total\_expense\_Q1\_Q2 due to negative importance, which indicates negative impact on model performance.

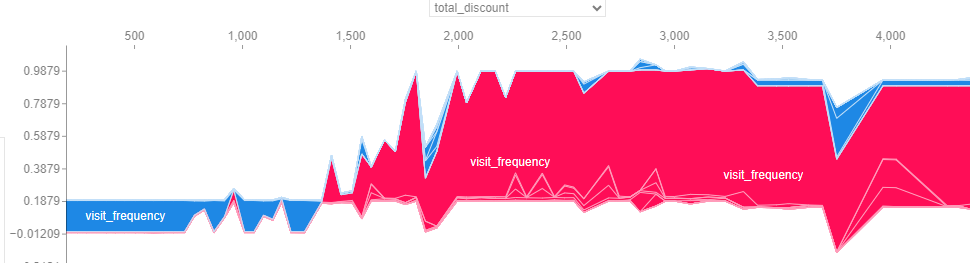
 

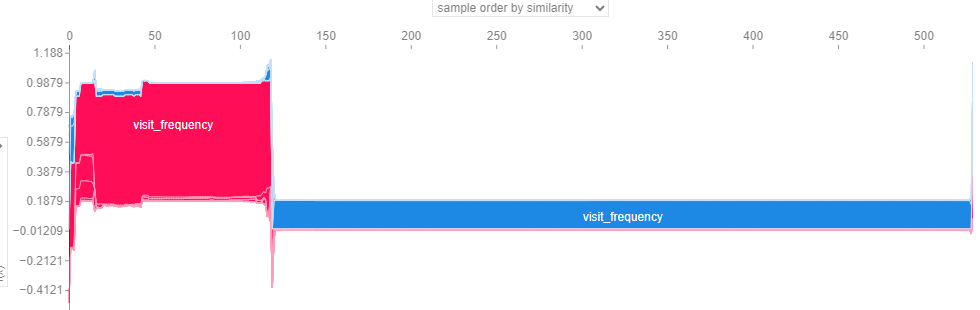


Additionally, total\_expense has multicollinearity with many features, suggesting information overlaps with other features which could affect interpretability.

**DTC:**

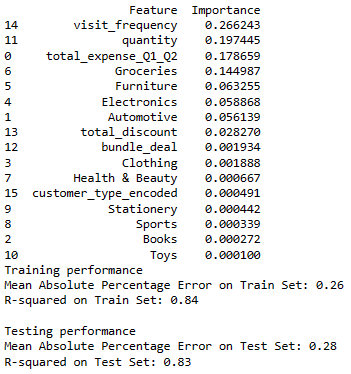
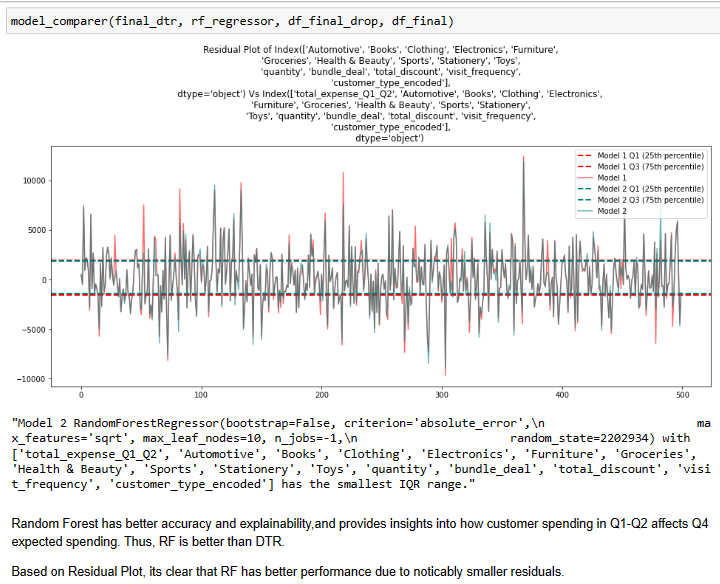
**Rejected** due to poor splits and heavy reliance on “visit\_frequency”; cannot learn other features’ impact.

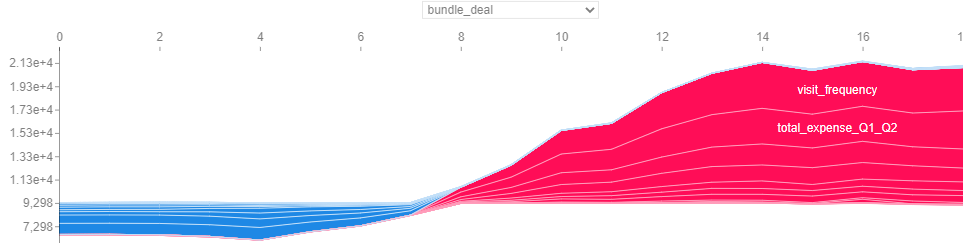


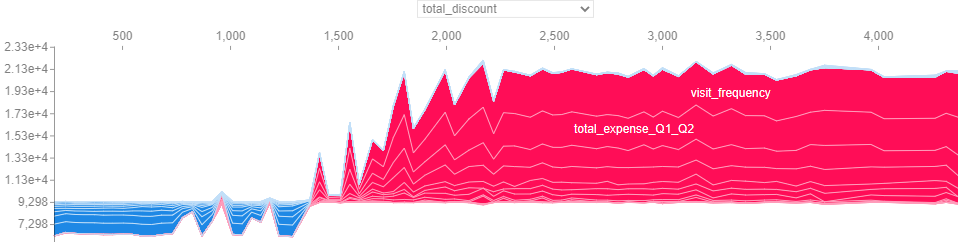


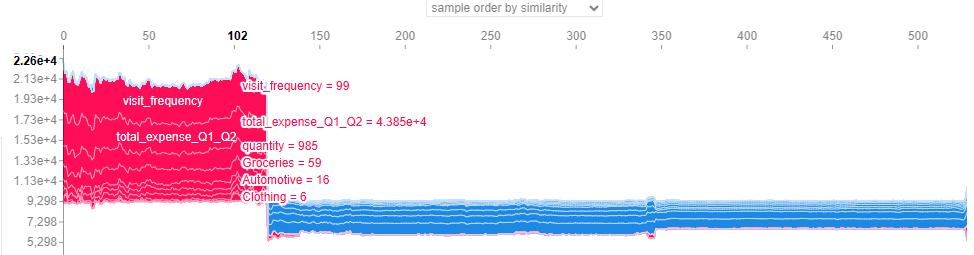
**RF (CHOSEN)**:

As [RF is robust against multicollinearity](https://www.researchgate.net/post/Are-Random-Forests-affected-by-multi-collinearity-between-features), **total\_expense can be used**.

 ****





****

1. Best-performing Regressor.
2. Clean splits in features, clearer interpretation.
3. Multiple features assessed to derive prediction, allowing comprehensive insights.

**Model Deployment**

**Deployment Plan:**

The model can be hosted on scalable cloud services like AWS or Azure to ensure business can expand instances if needed. After the last day of Q2, the business’ transactional dataset will be loaded into an ETL pipeline to transform the data for model usage, to predict each customer’s expected Q4 spending.

**Decision-Making:**

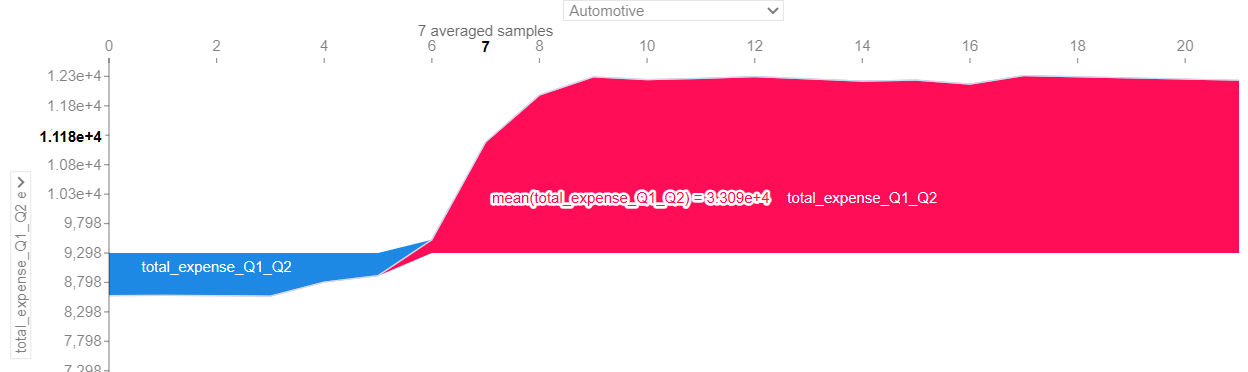
1. Summing predicted spending of customers to **forecast the business’ Q4 revenue** enables **informed investment decisions and resource allocation** without jeopardizing financial stability.
2. Experiment with strategies in Q1-Q2 and compare the predicted Q4 spending with previous years to **assess the effectiveness of campaigns**. This aids in **refining future** **strategies** for optimal results.
3. **Validate the feasibility of new products/ventures**.

E.g. Check if diversifying clothing products improved revenue and decide on action for next quarter:

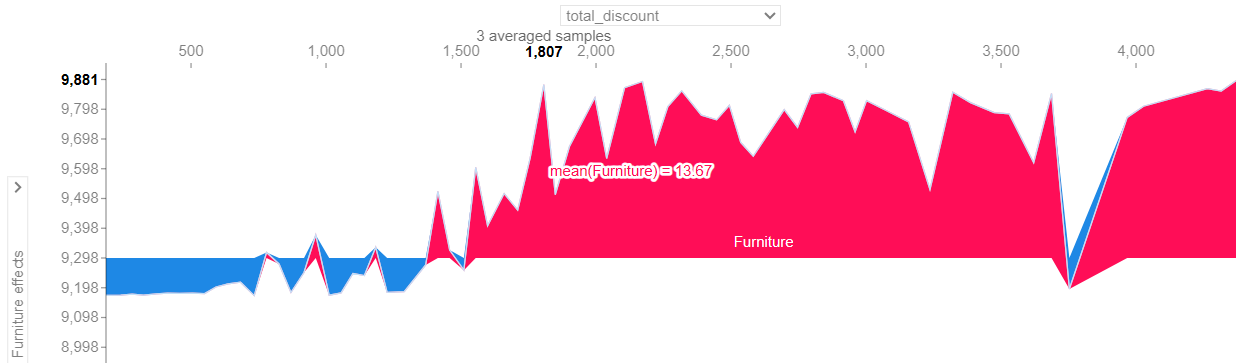


**Business Planning:**

1. Identify trending categories that attract higher spending, allowing businesses to **focus resources on optimizing and expanding successful products**:

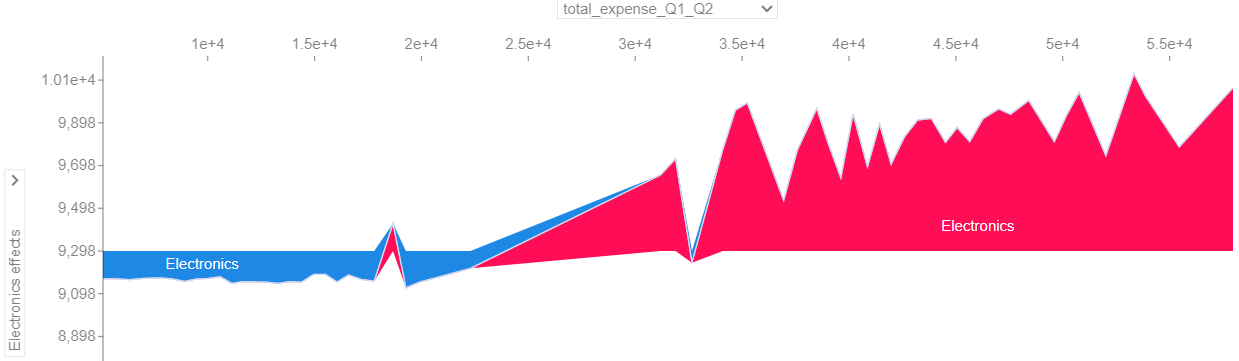


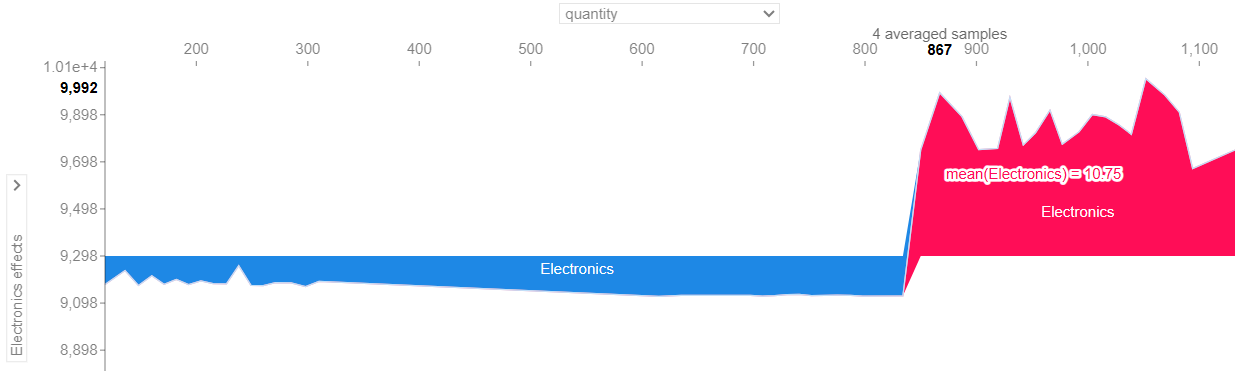
1. Analyze the interaction of discounts and bundled products across various categories to **uncover the most profitable strategies** for each category; Implement **data-driven pricing and bundling strategies**:



1. Optimize inventory to **ensure sufficient stock for products without excess**.

* e.g. customer expected to spend $XXXX on electronics for Q4 based on Q1-Q2 behavior, so the business should prepare around $XXXX worth of electronics for this customer to avoid stockouts or overstocking:

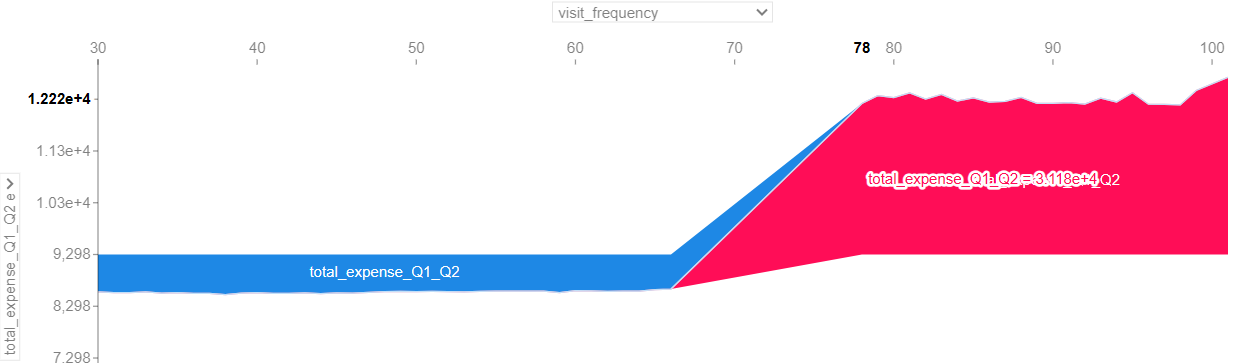


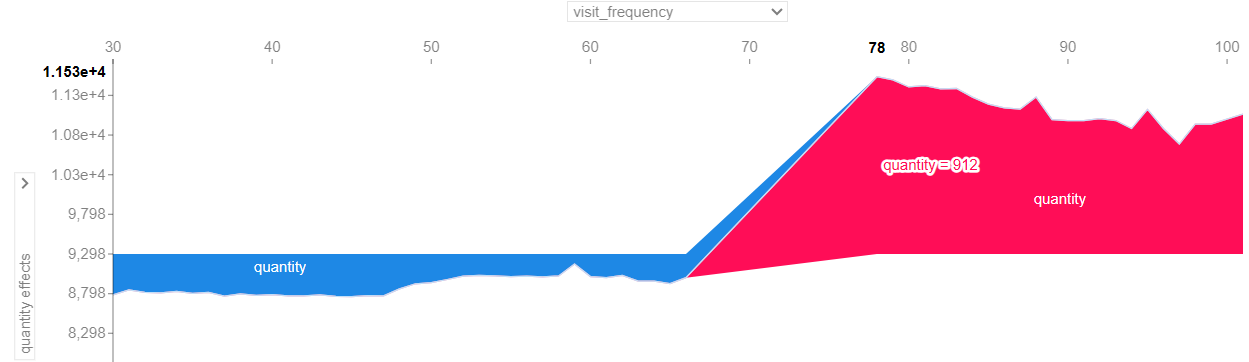


1. **Identify customers showing signs of disengagement** with Retail Hypermarket.

* early prevention against customer loss by re-engaging customers using campaigns that target their needs.

1. Implement **data-driven loyalty programs** that entice customers shop for XX number of days for benefits like discount on the next purchase, since customers usually spend more when they visit XX days a year:





**Implications:**

1. Random Forest is computationally expensive due to its robustness in evaluating the effects of each factor. Hence, more computational power and time (for model to run) is required when RF is deployed.
2. Intricacies of customer engagement may change over time and differ from the pattern the RF uses to derive predictions. It’s recommended to re-train the model annually using the previous year’s Q1-Q4 data.
3. Unanticipated factors like economic downturn, health crisis, etc. are not factored by model. Business should account for external factors as they could cause discrepancies in actual Q4 spending.

**End of Report**

**Word Count: 999** (excluding cover page, declaration, and content page)

**References**

Skincare routine: <https://sg.iherb.com/c/gifts?swkw=skin&msclkid=67851ff0c6ca18cadc587a98ef96abc0&utm_source=bing&utm_medium=cpc&utm_campaign=Category%20%3E%20Bath%20%26%20Beauty%20%3E%20Singapore%20%3E%20Acquisition%20(English)&utm_term=skin%20care%20gift%20set&utm_content=Category%20-%20Bath%20%26%20Beauty%20-%20Gift%20Sets&gclid=67851ff0c6ca18cadc587a98ef96abc0&gclsrc=3p.ds>

Random Forest with Multicollinearity:

[Are Random Forests affected by multi-collinearity between features? | ResearchGate](https://www.researchgate.net/post/Are-Random-Forests-affected-by-multi-collinearity-between-features)