Stat 451 Group Project

By: Madeline Bittman, Leah Cape, Honor Durham, Carissa Kinnart, and Zinnia Nie

Introduction to our Dataset

- We looked at consumer data that was collected on shoppers at Gap Inc. for market analysis and for tailoring advertisement methods
- 3900 customers' purchases were recorded as part of this dataset

Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes
2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes
3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes
4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes
5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes
6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes
7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2	Yes
8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes
9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6	Yes
10	57	Male	Handbag	Accessories	31	Missouri	M	Pink	Spring	4.8	Yes
11	53	Male	Shoes	Footwear	34	Arkansas	L	Purple	Fall	4.1	Yes
12	30	Male	Shorts	Clothing	68	Hawaii	S	Olive	Winter	4.9	Yes
13	61	Male	Coat	Outerwear	72	Delaware	M	Gold	Winter	4.5	Yes
14	65	Male	Dress	Clothing	51	New Hampshire	M	Violet	Spring	4.7	Yes
15	64	Male	Coat	Outerwear	53	New York	L	Teal	Winter	4.7	Yes
16	64	Male	Skirt	Clothing	81	Rhode Island	M	Teal	Winter	2.8	Yes
17	25	Male	Sunglasses	Accessories	36	Alabama	S	Gray	Spring	4.1	Yes
18	53	Male	Dress	Clothing	38	Mississippi	XL	Lavender	Winter	4.7	Yes
19	52	Male	Sweater	Clothing	48	Montana	S	Black	Summer	4.6	Yes
20	66	Male	Pants	Clothing	90	Rhode Island	M	Green	Summer	3.3	Yes
21	21	Male	Pants	Clothing	51	Louisiana	M	Black	Winter	2.8	Yes
22	31	Male	Pants	Clothing	62	North Carolina	M	Charcoal	Winter	4.1	Yes
23	56	Male	Pants	Clothing	37	California	M	Peach	Summer	3.2	Yes
24	31	Male	Pants	Clothing	88	Oklahoma	XL	White	Winter	4.4	Yes

Important Variables In The Dataset

Customer ID: A unique identifier assigned to each individual customer, facilitating tracking and analysis of their shopping behavior over time.

Age: The age of the customer, providing demographic information for segmentation and targeted marketing strategies.

Gender: The gender identification of the customer, a key demographic variable influencing product preferences and purchasing patterns.

Item Purchased: The specific product or item selected by the customer during the transaction.

Category: The broad classification or group to which the purchased item belongs (e.g., clothing, electronics, groceries).

Purchase Amount (USD): The monetary value of the transaction, denoted in United States Dollars (USD), indicates the cost of the purchased item(s).

Location: The geographical location where the purchase was made, offering insights into regional preferences and market trends.

Color: The color variant or choice associated with the purchased item, influencing customer preferences and product availability.

Season: The seasonal relevance of the purchased item (e.g., spring, summer, fall, winter), impacting inventory management and marketing strategies.

Review Rating: A numerical or qualitative assessment provided by the customer regarding their satisfaction with the purchased item.

Subscription Status: Indicates whether the customer has opted for a subscription service, offering insights into their level of loyalty and potential for recurring revenue.

Discount Applied: Indicates if any promotional discounts were applied to the purchase, shedding light on price sensitivity and promotion effectiveness.

Promo Code Used: Notes whether a promotional code or coupon was utilized during the transaction, aiding in the evaluation of marketing campaign success.

Previous Purchases: Provides information on the number or frequency of prior purchases made by the customer, contributing to customer segmentation and retention strategies.

Frequency of Purchases: Indicates how often the customer engages in purchasing activities, a critical metric for assessing customer loyalty and lifetime value.

Questions We Evaluated

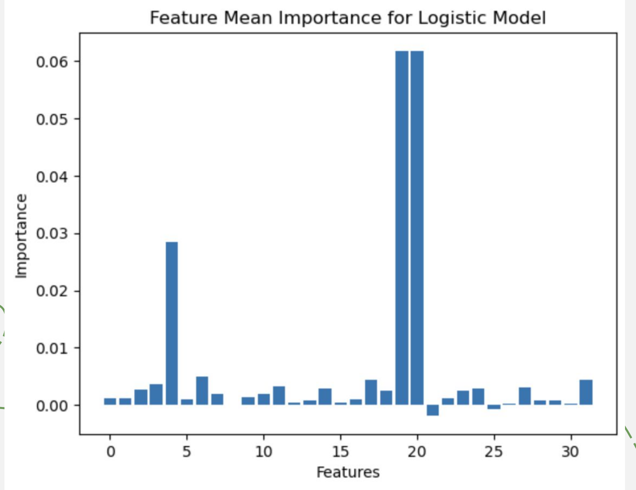
1. What factors are most influential on whether or not a Gap customer subscribes to Gap's subscription service?
2. How does seasonality and region affect how frequently a customer makes a purchase?

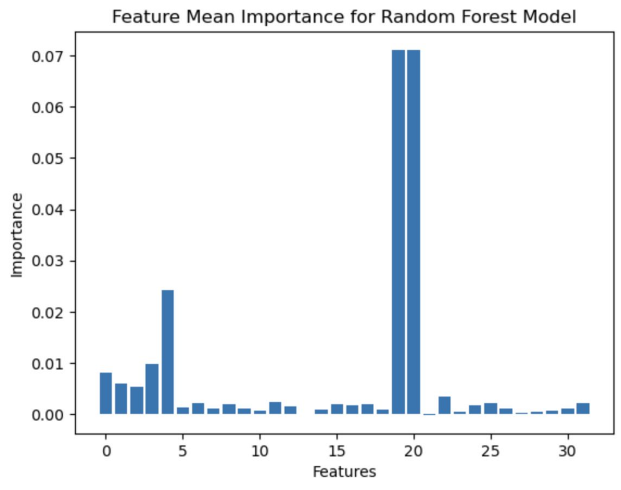
Predicting Subscription Status

- WHY?
  - To give marketing managers at Gap insight on what type of customer is most likely to be subscribed
  - Can decide who to cater marketing towards based off of findings
- HOW?
  - Permutation Feature Importance
    - Depends on both the data and the model
    - Different models have different importances

Methods

- Split data to 70% train and 30% test
- Trained a simple Logistic Regression model, and a Random Forest Classifier
  - Logistic Regression accuracy: 0.843
  - Random Forest accuracy: 0.852
- Then run permutation importance on each with 30 repeats



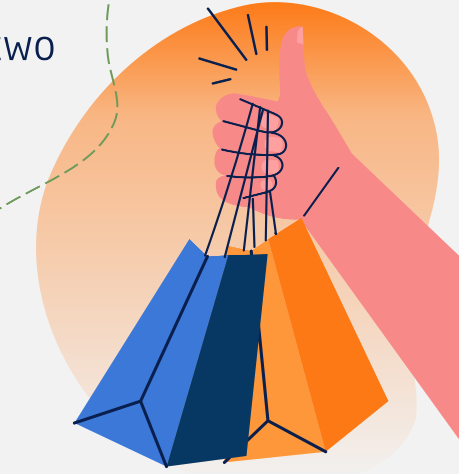


```
Promo Code Used_Yes 0.063 +/- 0.007
Discount Applied_Yes 0.063 +/- 0.007
Gender_Male 0.032 +/- 0.005
Category_Clothing 0.005 +/- 0.002
Frequency of Purchases_Quarterly 0.004 +/- 0.002
Previous Purchases 0.004 +/- 0.001
Season_Fall 0.004 +/- 0.001
Shipping Type_Standard 0.003 +/- 0.001
Payment Method_Credit Card 0.002 +/- 0.001
Category_Footwear 0.002 +/- 0.001
Size_S   0.002 +/- 0.001
Size_M   0.001 +/- 0.001
Payment Method_Cash 0.001 +/- 0.000
```

```
Discount Applied_Yes 0.069 +/- 0.006
Promo Code Used_Yes 0.069 +/- 0.006
Gender_Male 0.022 +/- 0.005
Previous Purchases 0.009 +/- 0.001
Age      0.008 +/- 0.001
Purchase Amount (USD) 0.006 +/- 0.001
Review Rating 0.005 +/- 0.001
Payment Method_Cash 0.004 +/- 0.001
Season_Fall 0.002 +/- 0.001
Shipping Type_Express 0.002 +/- 0.001
Frequency of Purchases_Quarterly 0.002 +/- 0.001
Category_Clothing 0.002 +/- 0.001
Payment Method_PayPal 0.002 +/- 0.001
Size_L   0.002 +/- 0.001
Payment Method_Debit Card 0.002 +/- 0.001
Shipping Type_Free Shipping 0.002 +/- 0.001
Frequency of Purchases_Annually 0.001 +/- 0.001
Season_Spring 0.001 +/- 0.001
```

Predicting Subscription Status Results

- Both models had the same top 3 most important features
  - Whether a promo code was used
  - Whether a discount was applied
  - Gender of the customer
- Made a new dataset with only the top 3 features
- Repeated the process of splitting and training the two models
  - Logistic Regression accuracy: 0.852
  - Random Forest accuracy: 0.852



Predicting Purchase Frequency

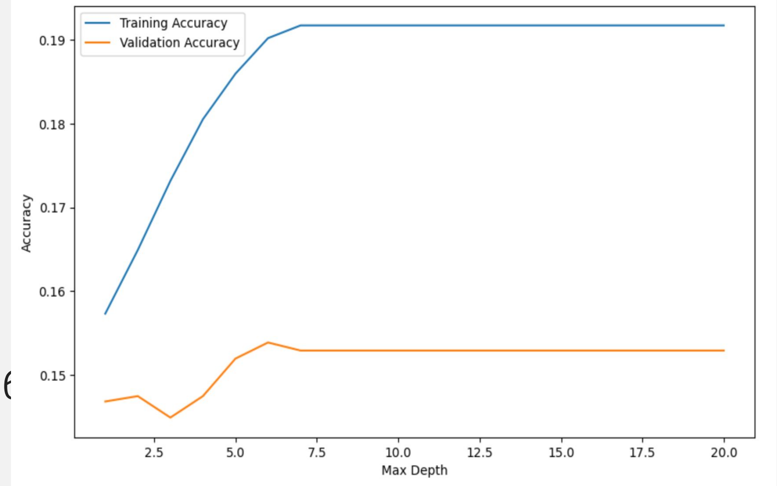
- WHY?
  - Identify if seasonality and customers in varying regions have a correlation to purchase frequency.
  - Can decide who and when to cater marketing towards based off of findings
- HOW?
  - Train Decision Tree Classifier
  - Use Ensemble Learning to improve performance over a basic decision tree

Methods

1. **Use a Random Grid Search CV to determine best max depth for decision tree classifier**
- Split data 80% train 20% test
- Fit on train data with the best max depth
2. **Try each of bagging, random forest, and gradient boosting to see whether they improve performance over a basic decision tree.**
- Of remaining 20%, split in half to get 10% validation, 10% test
- Fit on train data but test on validation

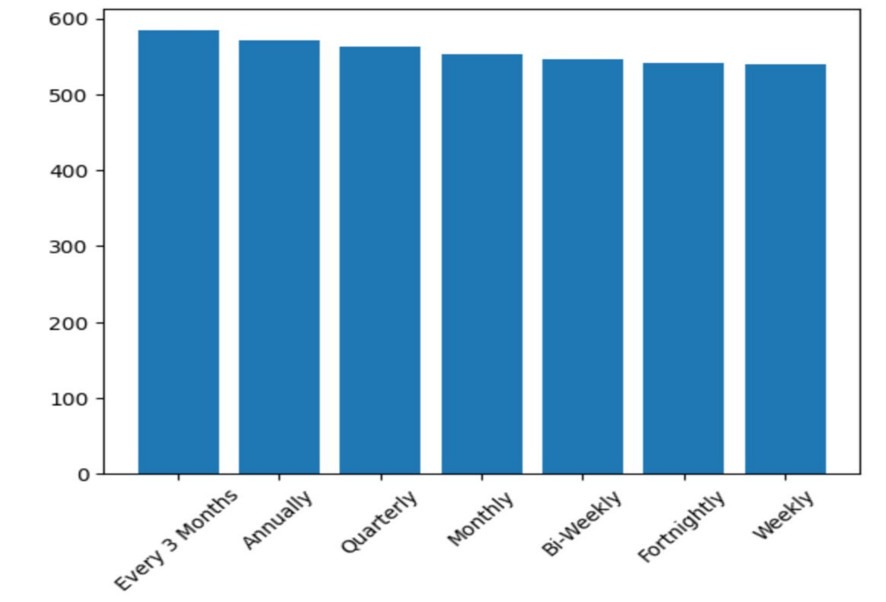
Results

- Best max depth resulted to 6
- Out of all models used for testing, Gradient Boost performed the best
    - Decision tree accuracy = 0.145
    - Bagging accuracy = 0.136
    - RandomForest accuracy score= 0.136
    - GradientBoosting accuracy score= 0.16



What's Happening to the Accuracy?

1. Predicting 7 categories, so performance accuracy around 1/7 is expected
2. Fake data?
    a. Several categories seemingly uniform
    b. Hard to draw relationships from these variables if they share this distribution



Conclusion

- **Promo codes, discounts, and gender columns** had the most impact on subscription status
    - Often subscription members get exclusive discounts, which then leads to **more spending** so this is logical
    - Companies can **use this strategy** to offer more discounts which could lead to more subscriptions which can lead to even more spending
- Seasonality and customers in varying regions and their correlation to purchase frequency is **less predictable** than we assumed
    - With predicting 7 categories, the lack of accuracy could make sense
    - With the discovery of the **uniform data**, it makes sense that there isn't a more logical correlation between region, season, and spending frequency
    - In real data, we could assume that companies would use this data to further market to regions that were **underperforming in purchase frequency** in certain seasons
- Further research could include more **demographic data**, looking at who to invest marketing into