MSiA 400 Team 9 Group Project

Yiyang Cao, Brian Hong, Samuel Swain, Robert Chen

The Client

- Upscale American department store
- Founded in 1938
- 2020 revenue of $6.34 billion
- NYSE: DDS

Dillard's

The Style of Your Life.

https://careers.dillards.com/Careers/information/history
https://www.macrotrends.net/stocks/charts/DDS/dillards/revenue

Background & Business Question

- Retail stores struggle to stock the perfect combination of items
- Some need extra incentive - **discounts** - to sell

We aim to make this process smoother by **predicting the potential required discount** of items, informing businesses on stocking items before they make the investment.

Getting Data Into The Cloud

- Checked and re-ordered feature variables in TRNSACT, to make uploadable to the MSiA PostGRESQL cloud server
- Cleaned SKUINFO csv file
    - 9000 rows with errors, caused by extra commas from vendor names

```
drops = []

for i, v in enumerate(df[0]):
    if df.iloc[i, 0].count(',') > 10:
        drops.append(i)

len(drops)
```

```
8148
```

```
df = df.drop(drops)
```

- sku
- store
- register
- trannum
- interid
- saledate
- style
- quantity
- orgprice
- amt
- amt2
- seq
- mic
- unknown

EDA

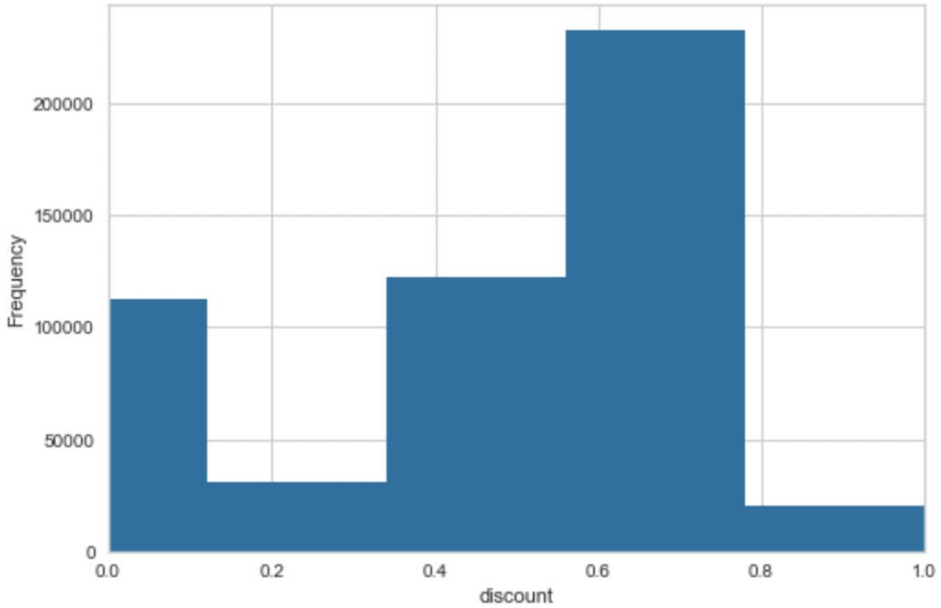
- New “discount” variable
  - Percentage of discount / 100
- Drop the irrelevant columns
  - “unnamed: 0”
  - “style”
  - “size”
  - “unknown”

```
df = psql.read_sql(""" with discount as (
select trnsact.sku, ROUND(AVG(orgprice), 2) AS orgprice, round(avg((orgprice-retail)/orgprice), 2) as discount
from trnsact join sksinfo Using(store, sku)
where orgprice != 0
group by 1
)
select *
from discount join skuinfo using(sku) """, conn)
```

```
df = df.drop(columns = ['Unnamed: 0', 'style', 'size', 'unknown'])
df.head()
```

	sku	orgprice	discount	dept	classid	upc	color	packsize	vendor	brand
0	3	440.00	0.00	6505	113 ...	400000003000	WHISPERWHITE ...	1	5119207	TURNBURY ...
1	69	12.25	0.00	7106	915 ...	400000069000	RED CALF ...	1	4816434	ME TOO ...
2	73	12.00	0.75	6107	001 ...	400000073000	WHITE ...	6	2017178	NOBILITY ...

Histogram of the “discount” percent column



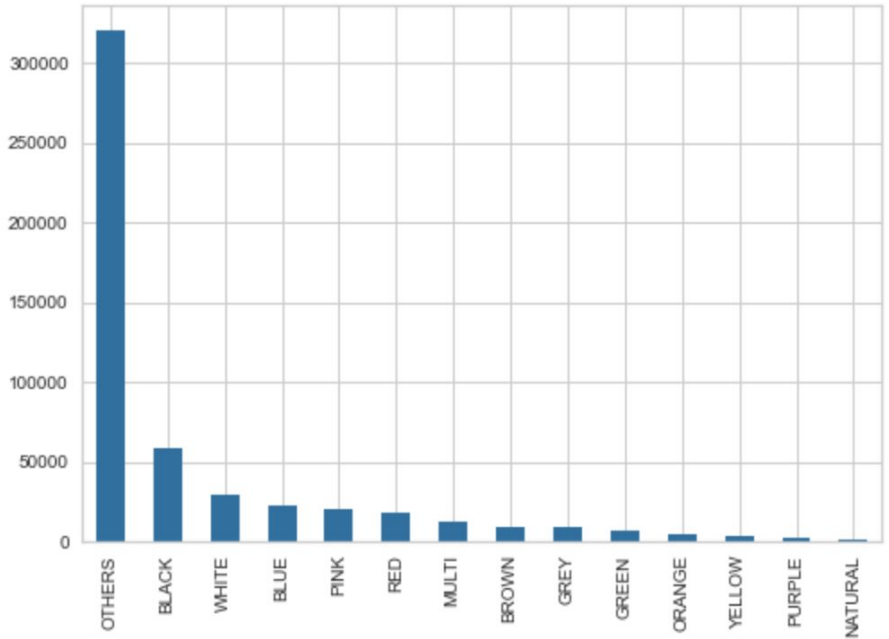
Data Cleaning

- Create “color group”
  - Designed color list containing the 14 most popular groups
- Group brands into 5 levels by average price

```
color_lst = [['WHITE', 'SILVER', 'BONE', 'IVORY', "WHT", "CREAM", "WH"],
             ['BLACK', 'BLK','ONYX',"OBSIDIAN","BLA", 'MIDNIGHT'],
             ['GREEN', 'OLIVE', 'AQUA', 'TURQUOISE', 'LIME', 'TURQ', "SAGE","TEAL"],
             ['NATURAL','TAUPE', 'KHAKI', "CAMEL", "SAND", "STONE", "BEIGE", "KH"],
             ['PURPLE', 'LILAC', 'PLUM', 'LAVENDER'],
             ['RED', 'BURGUNDY', 'WINE', 'CARDINAL'],
             ['BLUE', 'NAVY', 'INDIGO', 'DENIM'],
             ['MULTI', 'ASSORTED', "ASST", "ROYAL"],
             ['PINK', 'ROSE', 'FUCHSIA', "FUSCHIA"],
             ['BROWN','CHOCOLATE', 'TAN'],
             ['GREY', 'CHARCOAL'],
             ['YELLOW','GOLD'],
             ['ORANGE',"CORAL"],
             ["UNDEFINED", "CLARIFY",'NO COLOR', "NA"]]
```

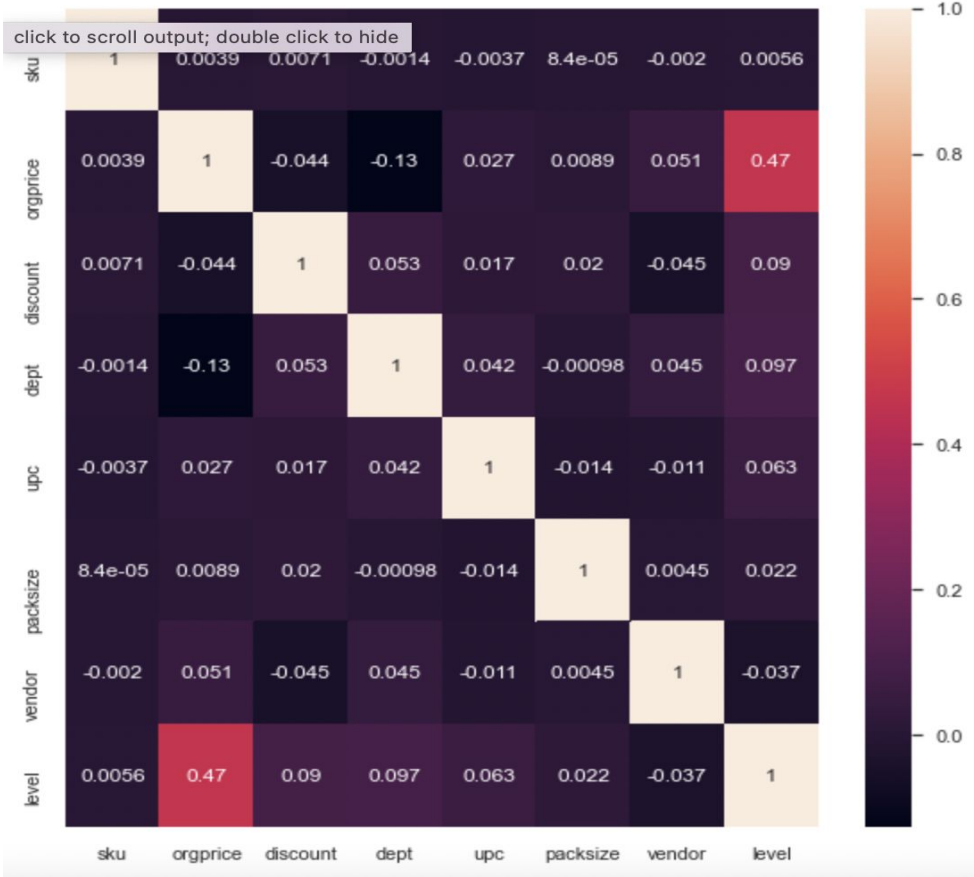
	sku	orgprice	discount	dept	classid	upc	color	packsize	vendor	brand	color group	level
0	3	440.00	0.00	6505	113 ...	400000003000	WHISPERWHITE	1	5119207	TURNBURY ...	WHITE	5.0
1	69	12.25	0.00	7106	915 ...	400000069000	RED CALF	1	4816434	ME TOO ...	RED	4.0
2	73	12.00	0.75	6107	001 ...	400000073000	WHITE	6	2017178	NOBILITY ...	WHITE	4.0
3	78	59.04	0.17	5301	001 ...	400000078000	BROWN	1	9912767	M.M. & R ...	BROWN	4.0
4	91	68.00	0.75	9801	004 ...	400000091000	CORAL	1	9212766	GABAR IN ...	OTHERS	5.0

Visualization of “color group” distribution



Correlation Plot

- The correlation among variables are very low (except orgprice, but we will not include it in the model)
- The multicollinearity would not be a problem.

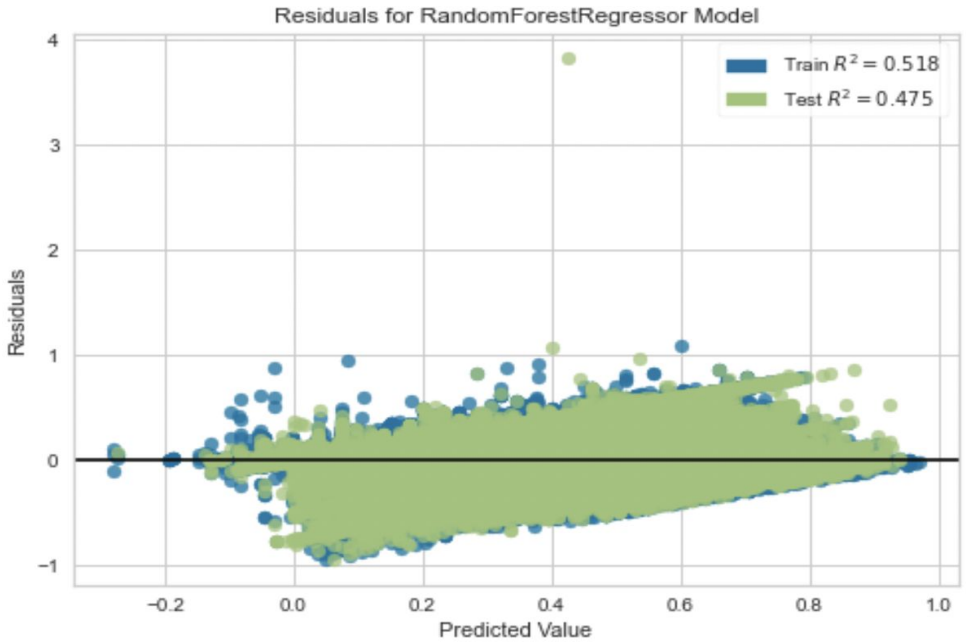


Feature Engineering & Modeling

- Features: dept, classid, color group, vendor, level
- Feature Engineering:
  - One hot Encoding: dept, classid, color group, vendor
  - Ordinal Encoding: level

Model Evaluation	Linear Regression	Random Forest Regressor	KNeighborsRegressor
Test R^2	0.0114	0.4753	0.3795
Test RMSE	0.2904	0.2115	0.23

Residual plot for RandomForest Regressor



ROI Analysis

Aggregate Values	
Sales (Volume)	120,000,000.00
Sales (USD)	$7,500,000,000.00
Sales at Discount (%)	53.00%
Sales at Discount (Volume)	63,600,000.00
Sales at Discount (USD)	$2,749,493,549.59
Sales at Full (%)	47.00%
Sales at Full (Volume)	56,400,000.00
Sales at Full (USD)	$4,750,572,000.00
Merchandise Inventories (%)	18.48%
Merchandise Inventories (Volume)	27,200,000.00
Merchandise Inventories (USD)	$1,700,000,000.00
Cost of Sales (USD)	$5,000,000,000.00
Cost of Sales at Discount (USD)	$2,650,000,000.00
Cost of Sales at Full (USD)	$2,350,000,000.00

Average Values	
Average Sale Price	$62.50
Average Discount Amount	48.68%
Average Discount Sale Price	$43.23
Average Full Sale Price	$84.23

Baseline Case (no predicted discounts)	
Profit Margin of Sales at Discount	3.62%
Profit of Sales at Discount	$99,493,549.59
Profit Margin of Sales at Full	50.53%
Profit of Sales at Full	$2,400,572,000.00
Total Profit Margin	33.33%
Total Profit	$2,500,000,000.00

Model Benefits (predicted discounts, low case)	
Reduction in Sales at Discount	0.01%
Sales at Discount (%)	52.99%
Sales at Discount (USD)	$2,748,974,777.23
Sales at Full (%)	47.01%
Sales at Full (USD)	$4,751,582,760.00
Total Profit	$2,500,557,537.23
Change from Baseline Case (%)	**0.02%**
Change from Baseline Case (USD)	**$557,537.23**

Model Benefits (predicted discounts, high case)	
Reduction in Sales at Discount	0.03%
Sales at Discount (%)	52.97%
Sales at Discount (USD)	$2,747,937,232.49
Sales at Full (%)	47.03%
Sales at Full (USD)	$4,753,604,280.00
Total Profit	$2,501,541,512.49
Change from Baseline Case (%)	**0.06%**
Change from Baseline Case (USD)	**$1,541,512.49**

Investment Cost	
Duration (Days)	270.00
Salary (Data Scientist)	$120,000.00
Number of Data Scientists	4.00
Total Data Scientist Cost	$355,068.49
Computing Time (Hours)	6,480.00
Computing Cost (Cloud, USD)	$0.48
Total Computing Cost	$3,110.40
Total Investment Cost	**$358,178.89**

ROI Analysis	
Low Case (USD)	**$199,358.33**
High Case (USD)	**$1,183,333.60**
Low Case (%)	**55.66%**
High Case (%)	**330.38%**

https://www.annualreports.com/HostedData/AnnualReportArchive/d/NYSE_DDS_2004.pdf

Conclusion

- ROI
  - 55%+ in low case, with 0.01% improvement in discount avoidance
  - 330%+ in high case, with 0.03% improvement in discount avoidance
- Due to significant differences in profit margins:

marginal reduction in stocking products that require discounts

=

positive ROI