MLDS 400 Team 13 Group Project

Yi (Grace) Xie, Inu Tenneti, Xinyang (Oliver) Zhou, Lishan (Lisa) Gao

Background

- Dataset: Dillard’s
- Founded in 1938
- Upscale American department store chain with approximately 282 stores in 29 states
- 5 tables

Business Question

- *Difference brands, stores, and states can all affect the profit*
- *Which specific store can have high profit?*

We want to build models to solve:

Will the store has high profit or low profit across different brands, stores, states, and years?

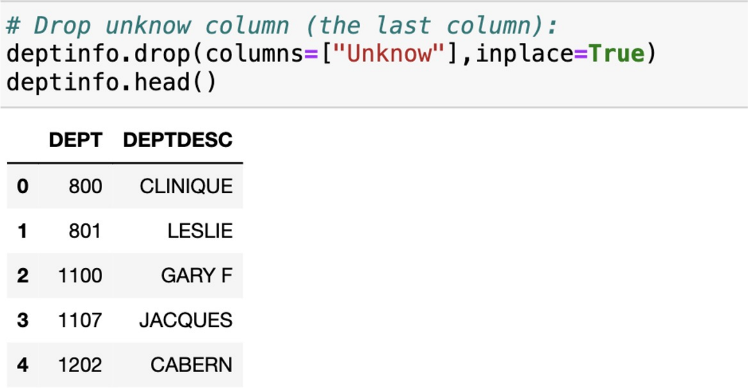
Preparing the Data

- Upload tables to PostgreSQL Server

- Read into pandas

- Clean the data

- Remove the last columns from all tables



```
# Drop unknow column (the last column):
deptinfo.drop(columns=["Unknow"],inplace=True)
deptinfo.head()
```

	DEPT	DEPTDESC
0	800	CLINIQUE
1	801	LESLIE
2	1100	GARY F
3	1107	JACQUES
4	1202	CABERN

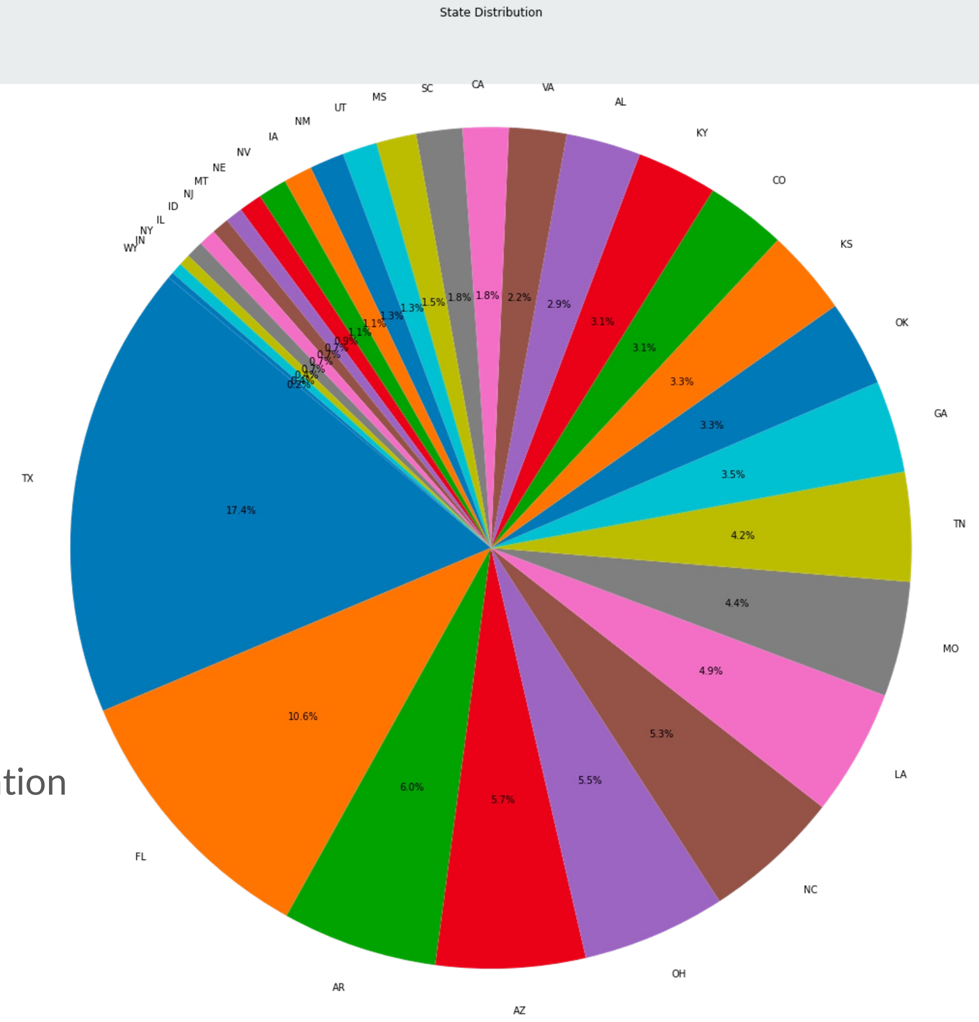
EDA

- Check missing value

- Plot distributions of columns

- This is a visualization of the store location

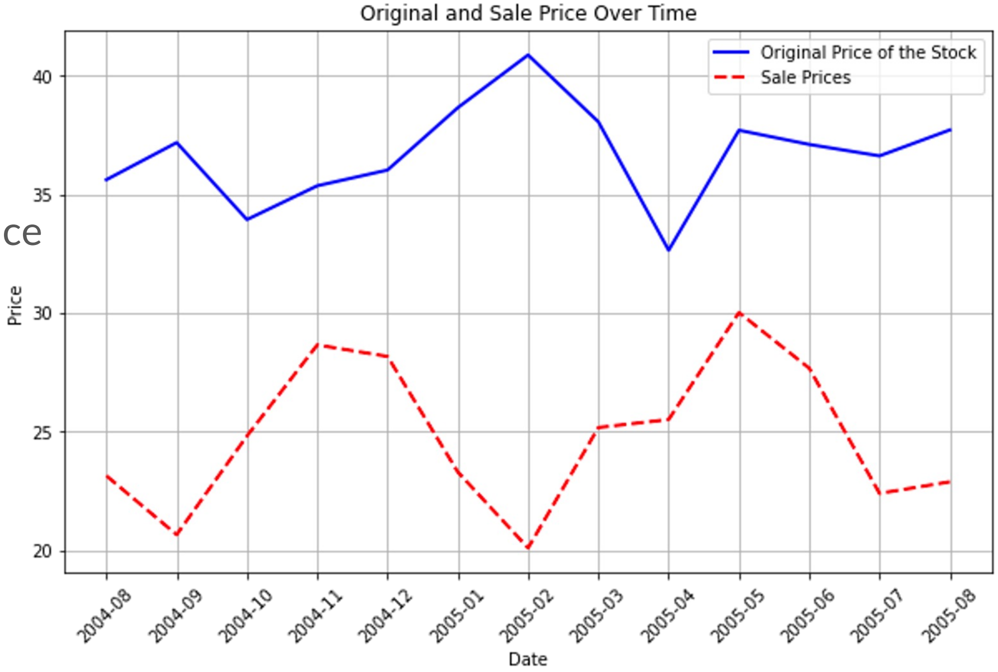
- TX takes 17.4% of the stores



EDA

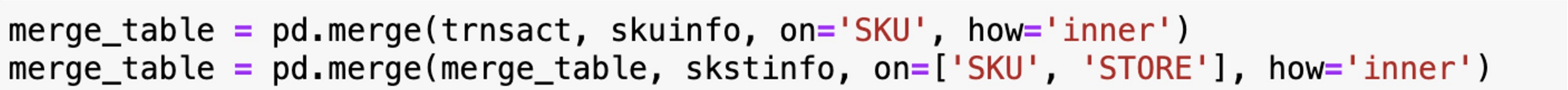
- Big gap between original and sale price

- The gap fluctuate by time

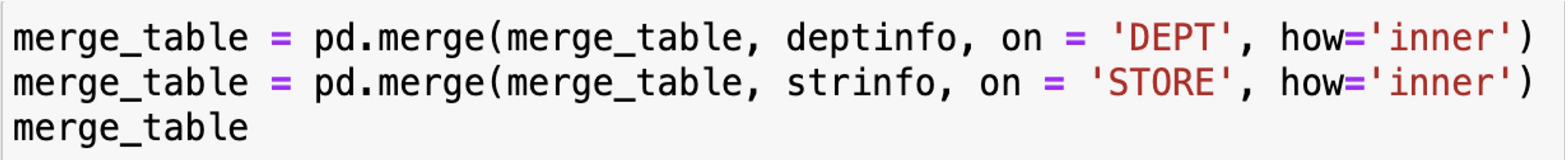


Merge Table

- Combined all 5 tables through the identifier between each of them



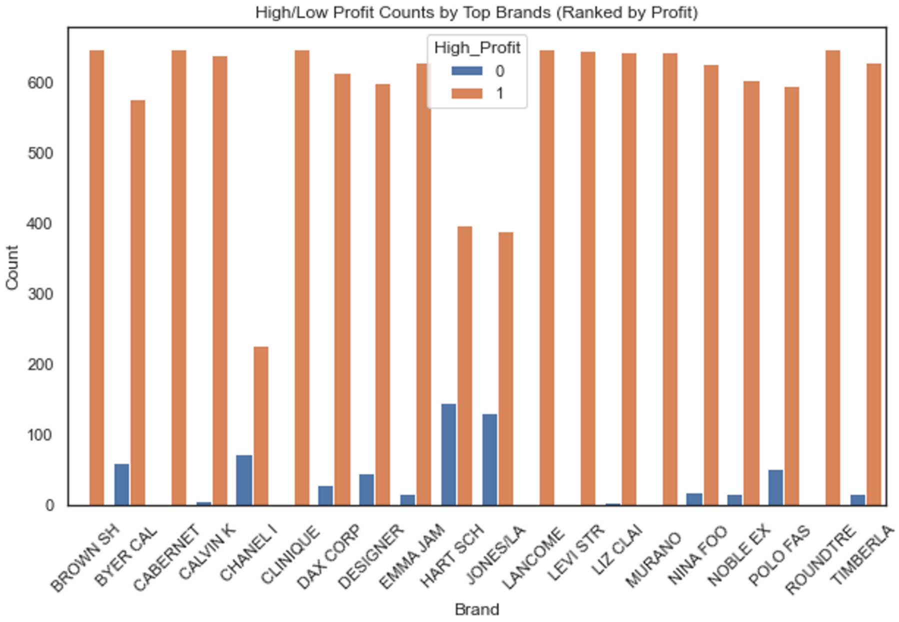
```
merge_table = pd.merge(trnsact, skuinfo, on='SKU', how='inner')
merge_table = pd.merge(merge_table, skstinfo, on=['SKU', 'STORE'], how='inner')
```

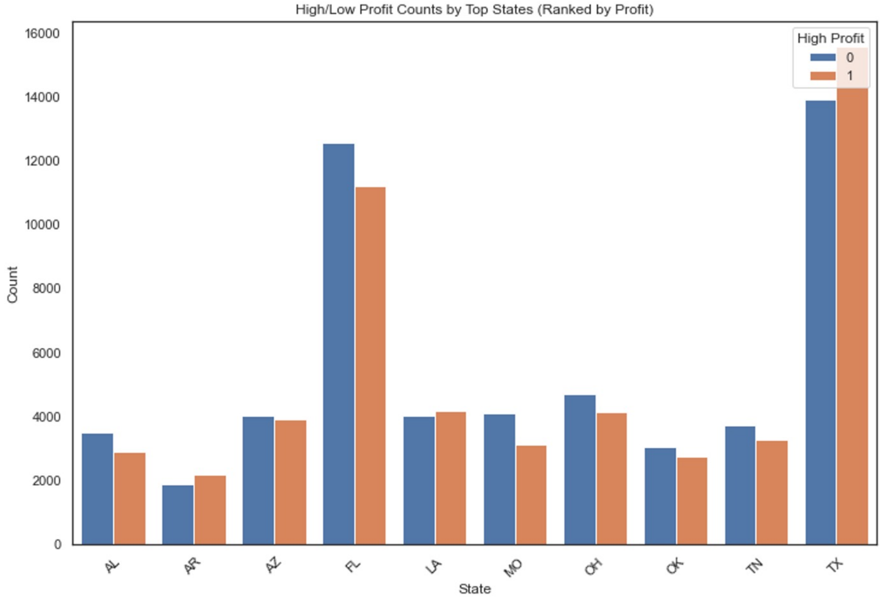


```
merge_table = pd.merge(merge_table, deptinfo, on = 'DEPT', how='inner')
merge_table = pd.merge(merge_table, strinfo, on = 'STORE', how='inner')
merge_table
```

EDA

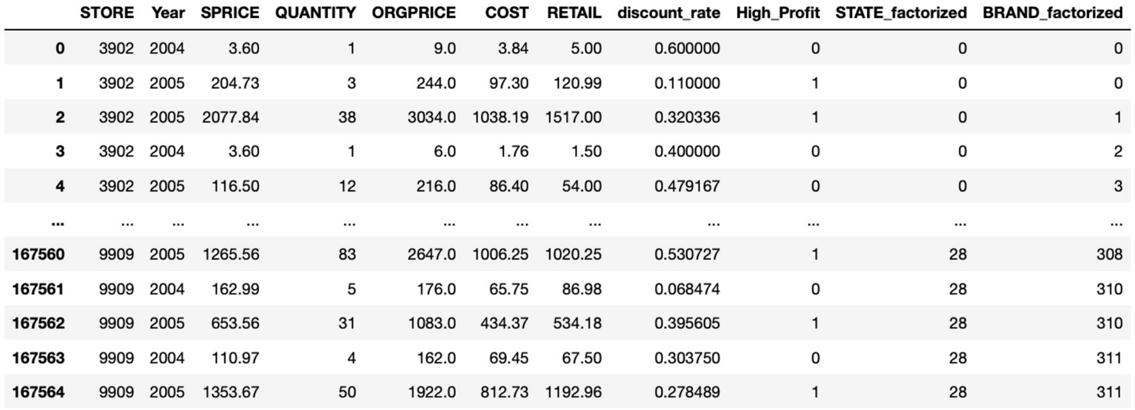
- High profit/low profit count by states and brands (ranked by profit)





Feature Engineering

- **Features**: STATE, STORE, BRAND, Year, SPRICE, QUANTITY, ORGPRICE, COST, RETAIL, discount_rate, High_Profit
- **Response Variable**: High_Profit (1 if > 100, 0 if <= 100)
- **Feature Engineering**: Factorize STATE and BRAND → STATE_factorized, BRAND_factorized



	STORE	Year	SPRICE	QUANTITY	ORGPRICE	COST	RETAIL	discount_rate	High_Profit	STATE_factorized	BRAND_factorized
0	3902	2004	3.60	1	9.0	3.84	5.00	0.600000	0	0	0
1	3902	2005	204.73	3	244.0	97.30	120.99	0.110000	1	0	0
2	3902	2005	2077.84	38	3034.0	1038.19	1517.00	0.320336	1	0	1
3	3902	2004	3.60	1	6.0	1.76	1.50	0.400000	0	0	2
4	3902	2005	116.50	12	216.0	86.40	54.00	0.479167	0	0	3
...	...	...	...	...	...	...	...	...	...	...	...
167560	9909	2005	1265.56	83	2647.0	1006.25	1020.25	0.530727	1	28	308
167561	9909	2004	162.99	5	176.0	65.75	86.98	0.068474	0	28	310
167562	9909	2005	653.56	31	1083.0	434.37	534.18	0.395605	1	28	310
167563	9909	2004	110.97	4	162.0	69.45	67.50	0.303750	0	28	311
167564	9909	2005	1353.67	50	1922.0	812.73	1192.96	0.278489	1	28	311

Modeling - Logistic Regression

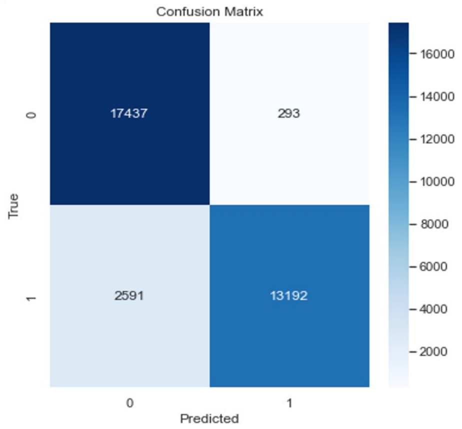
Parameters:

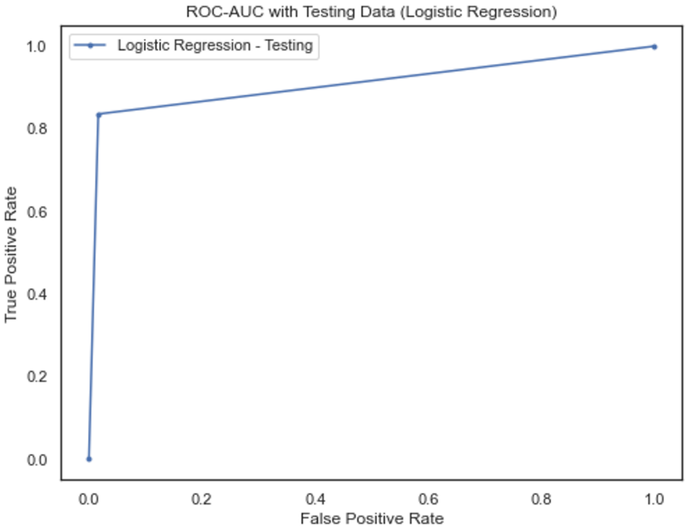
- Regularization: L1 penalty
- Solver: liblinear

AUC (testing): 0.9096551816927398

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.98	0.92	17730
1	0.98	0.84	0.90	15783
accuracy			0.91	33513
macro avg	0.92	0.91	0.91	33513
weighted avg	0.92	0.91	0.91	33513


Confusion Matrix

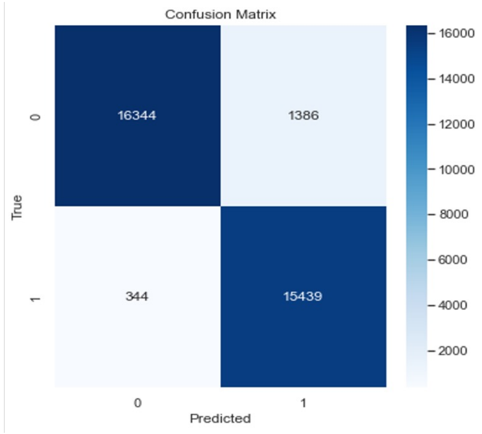

ROC-AUC with Testing Data (Logistic Regression)

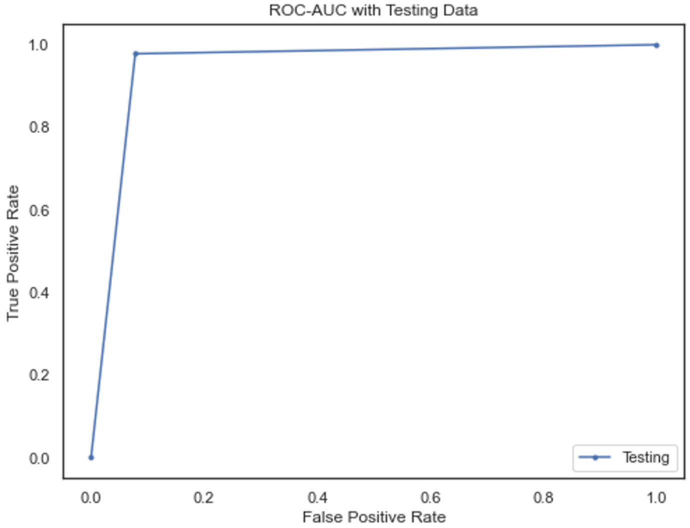
Modeling - Decision Tree

Parameter: max_depth: 3
AUC (testing): 0.9500159041518359

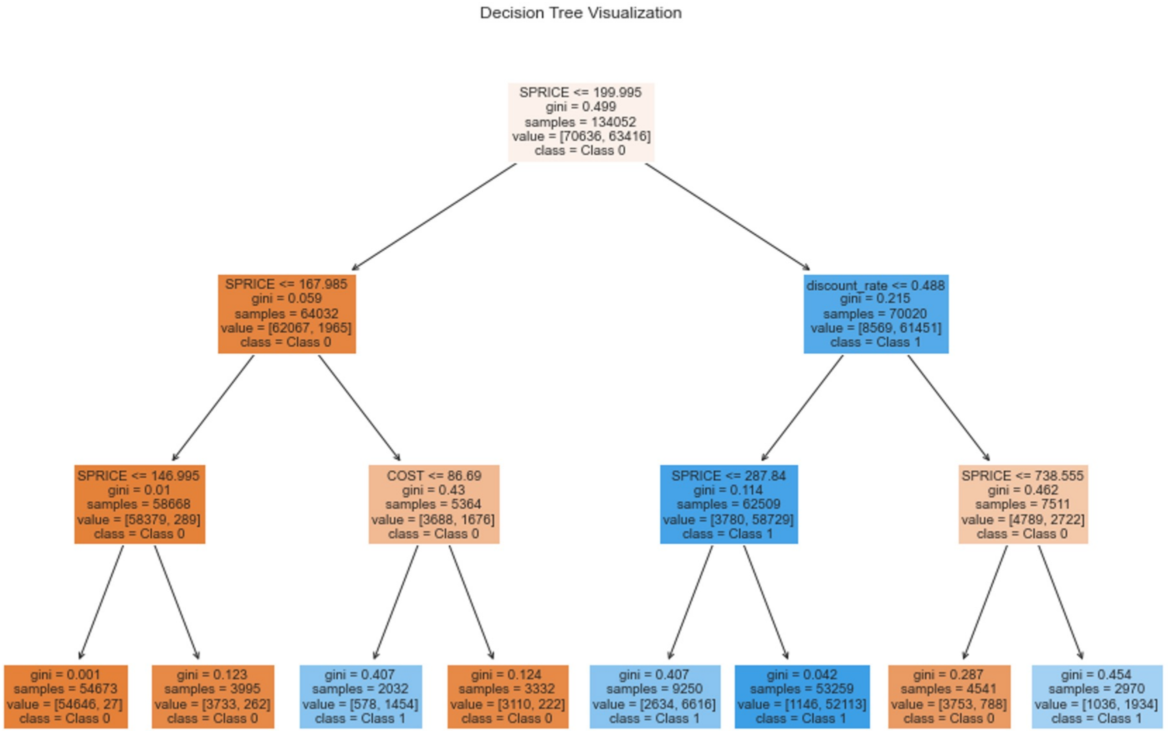
Classification Report:

	precision	recall	f1-score	support
0	0.98	0.92	0.95	17730
1	0.92	0.98	0.95	15783
accuracy			0.95	33513
macro avg	0.95	0.95	0.95	33513
weighted avg	0.95	0.95	0.95	33513





Decision Tree Plot



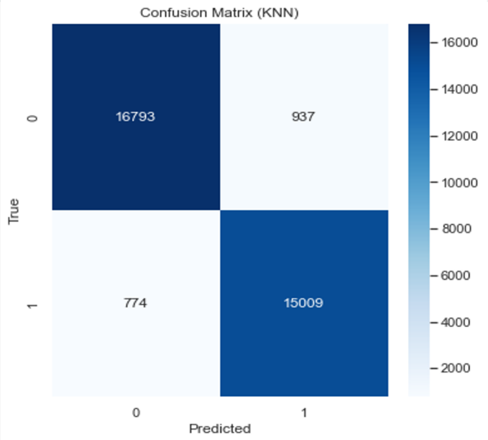
Modeling - KNN

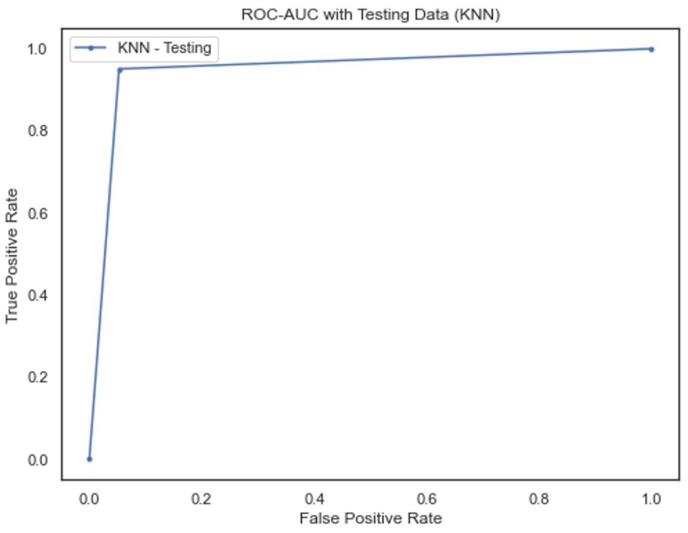
Parameter: n_neighbors = 5

AUC (testing): 0.9490558069022625

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.95	0.95	17730
1	0.94	0.95	0.95	15783
accuracy			0.95	33513
macro avg	0.95	0.95	0.95	33513
weighted avg	0.95	0.95	0.95	33513



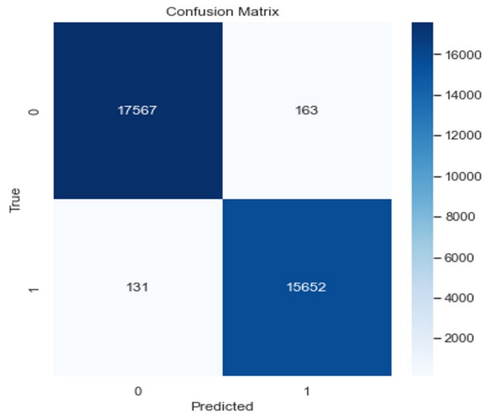


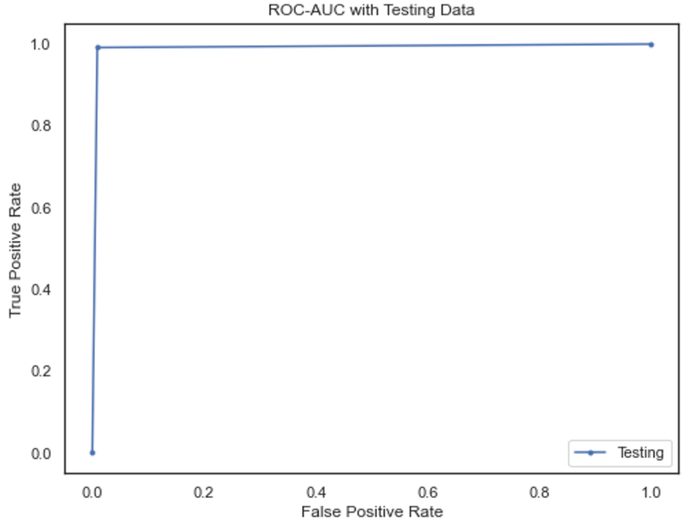
Fine Tune Decision Tree Model

Parameter: max_depth: 10
AUC (testing): 0.9913590818710571

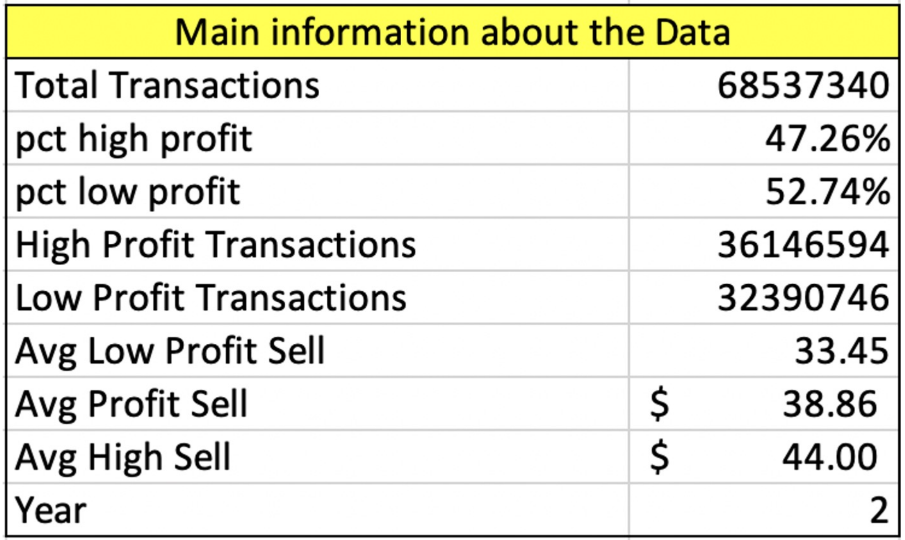
Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	17730
1	0.99	0.99	0.99	15783
accuracy			0.99	33513
macro avg	0.99	0.99	0.99	33513
weighted avg	0.99	0.99	0.99	33513

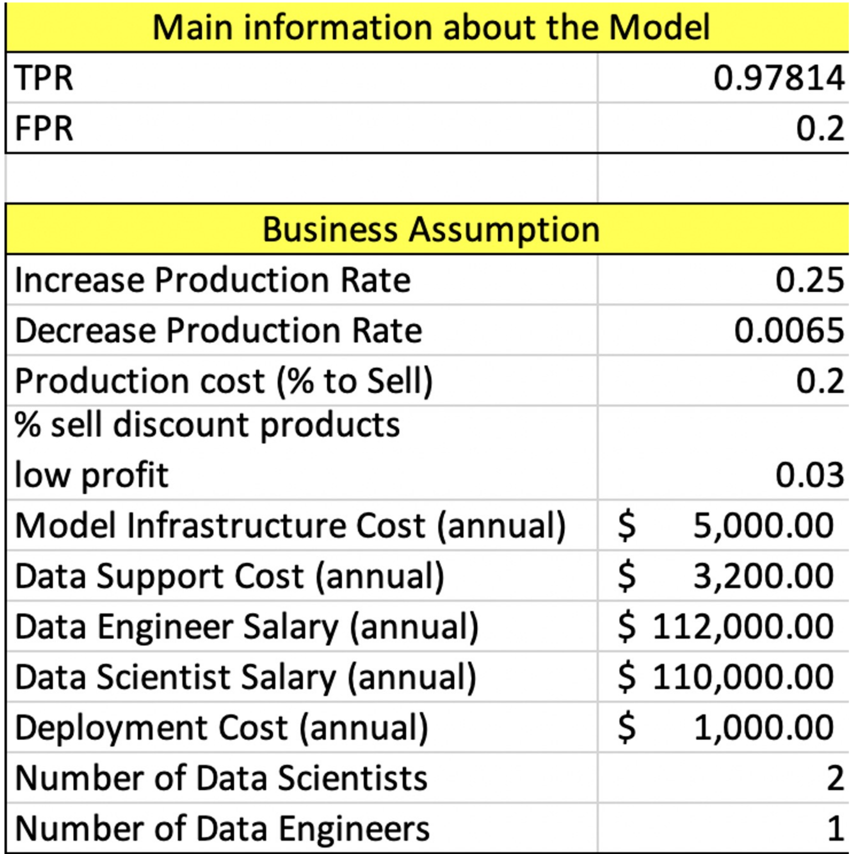




ROI Analysis



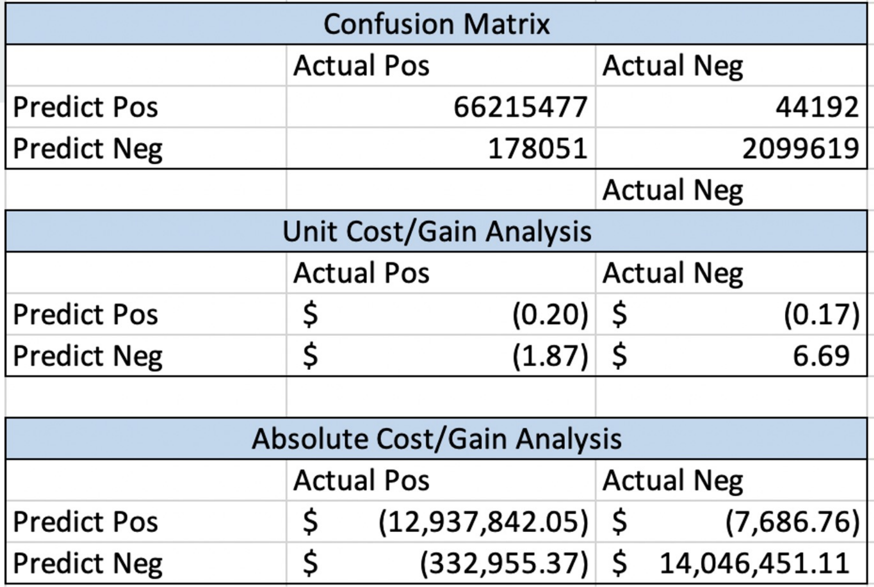
Main information about the Data	
Total Transactions	68537340
pct high profit	47.26%
pct low profit	52.74%
High Profit Transactions	36146594
Low Profit Transactions	32390746
Avg Low Profit Sell	33.45
Avg Profit Sell	$ 38.86
Avg High Sell	$ 44.00
Year	2



Main information about the Model	
TPR	0.97814
FPR	0.2

Business Assumption	
Increase Production Rate	0.25
Decrease Production Rate	0.0065
Production cost (% to Sell)	0.2
% sell discount products low profit	0.03
Model Infrastructure Cost (annual)	$ 5,000.00
Data Support Cost (annual)	$ 3,200.00
Data Engineer Salary (annual)	$ 112,000.00
Data Scientist Salary (annual)	$ 110,000.00
Deployment Cost (annual)	$ 1,000.00
Number of Data Scientists	2
Number of Data Engineers	1

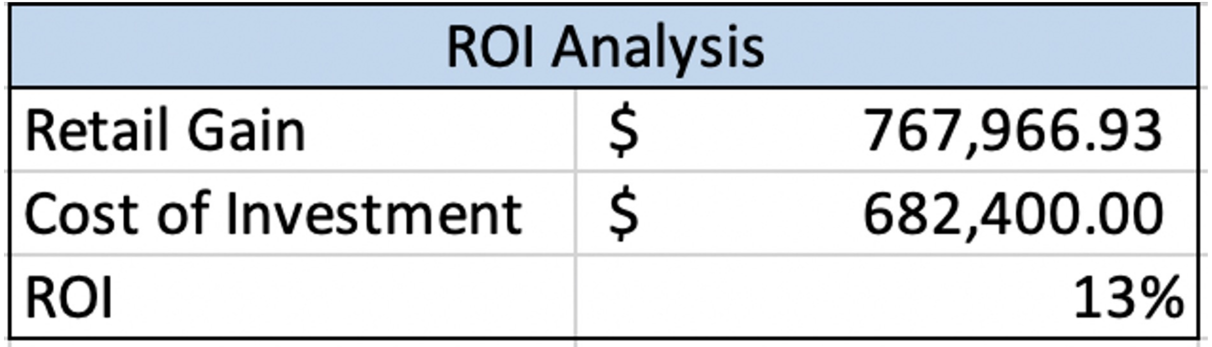
Result



Confusion Matrix		
	Actual Pos	Actual Neg
Predict Pos	66215477	44192
Predict Neg	178051	2099619
		Actual Neg

Unit Cost/Gain Analysis		
	Actual Pos	Actual Neg
Predict Pos	$ (0.20)	$ (0.17)
Predict Neg	$ (1.87)	$ 6.69

Absolute Cost/Gain Analysis		
	Actual Pos	Actual Neg
Predict Pos	$ (12,937,842.05)	$ (7,686.76)
Predict Neg	$ (332,955.37)	$ 14,046,451.11



ROI Analysis	
Retail Gain	$ 767,966.93
Cost of Investment	$ 682,400.00
ROI	13%

The End
Questions?