



Project 4 Presentation

Shopping Mall Sales

Group 2: Bryan Denq, Brendan Trafford, Danielle Torpy, and Rothana Kim



Road Map

01

**Background
Information**

02

**Supervised
Learning**

03

**Neural
Network**

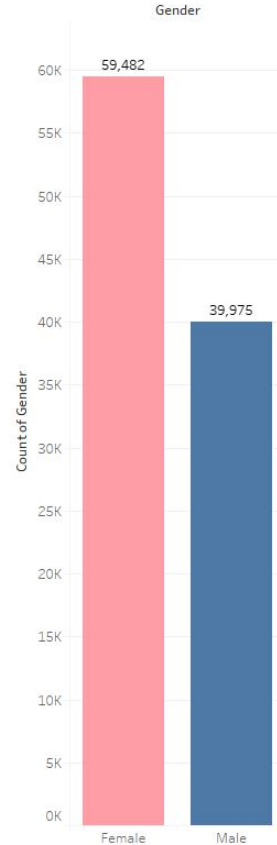
04

**Concluding
Statements**

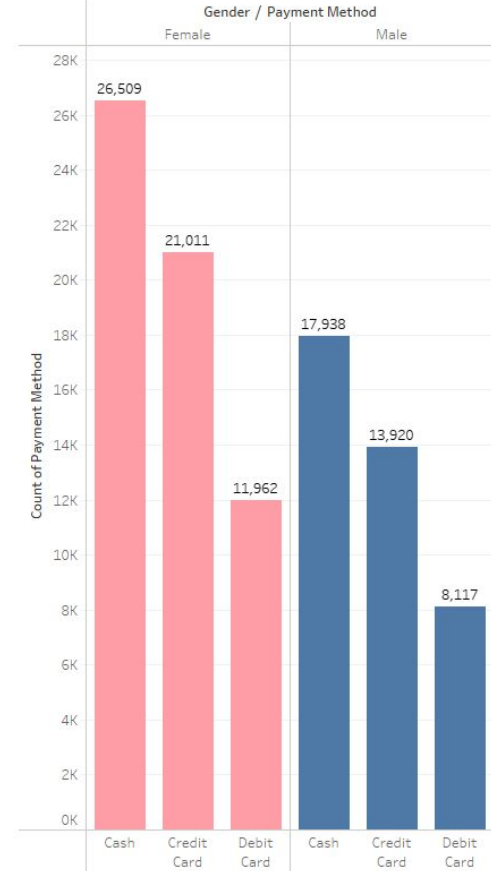
Background Graphs

- Dataset retrieved from Kaggle
 - <https://www.kaggle.com/datasets/mehmettahirasan/customer-shopping-dataset>
 - "Shopping information from 10 different shopping malls [in Istanbul, Turkey] between 2021 and 2023."
- Tableau Graphs
 - # of Females > # of Males

Total count of Males/Females



Payment method count by Gender





Question in mind

Can we predict the gender of a shopper with the given information in this data set?

Answering this question can provide insights into:

- What products to sell to certain genders
 - Males could be more likely to buy Toys and Technology
- What day of the week would be busiest?
 - Families could come in during the weekends while Single females may come in during the week
- Who is buying more things or expensive things?
 - One gender could be buying more things or buy more expensive things



Cleaning the Data

- Removal of unnecessary columns such as customer_id and invoice_id
- Changed date format using datetime from DD/MM/YYYY to YYYY-MM-DD
 - This allowed us to convert each date to a day of the week in new column labeled “Day of week”
 - Used replace function to convert days of week from numbers to actual day names
- Used pd.get_dummies to convert all strings to one hot encodings
 - Converted columns were: “Days of week”, “payment_method”, “Gender”, “Category” and “shopping_mall”

Cleaning the Data: Examples

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	Unnamed: 9	shopping_mall
0	I138884	C241288	Female	28	Clothing	5	1500.40	Credit Card	5/8/2022	NaN	Kanyon
1	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	NaN	Forum Istanbul
2	I127801	C266599	Male	20	Clothing	1	300.08	Cash	9/11/2021	NaN	Metrocity
3	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	NaN	Metropol AVM
4	I337046	C189076	Female	53	Books	4	60.60	Cash	24/10/2021	NaN	Kanyon

	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall	Day of week
0	Female	28	Clothing	5	1500.40	Credit Card	2022-05-08	Kanyon	Sunday
1	Male	21	Shoes	3	1800.51	Debit Card	2021-12-12	Forum Istanbul	Sunday
2	Male	20	Clothing	1	300.08	Cash	2021-09-11	Metrocity	Saturday
3	Female	66	Shoes	5	3000.85	Credit Card	2021-05-16	Metropol AVM	Sunday
4	Female	53	Books	4	60.60	Cash	2021-10-24	Kanyon	Sunday

```
new_shop = pd.get_dummies(shopping)
```



Road Map

01

**Background
Information**

02

**Supervised
Learning**

03

**Neural
Network**

04

**Concluding
Statements**



Supervised Learning: Logistic Regression

Training Score	0.4513
Testing Score	0.4529

We went with Logistic Regression because this method estimates the probability of the shopper's gender based on all variables such as price, category, age, quantity, day of week, and payment method.



Supervised Learning: Bag_and_Boost

Training Score	0.5989
Testing Score	0.5954

We choose to try Bag and Boost as AdaBoost is specifically good at binary classification and is good at creating a strong classifier out of multiple weak classifiers.



Supervised Learning Poor Performers

	Training Score	Testing Score
Random Forest	0.7626	0.5251
Extra Trees	0.4037	-0.4528
K-Nearest Neighbor (KNN)	0.1673	-0.2032



Road Map

01

**Background
Information**

02

**Supervised
Learning**

03

**Neural
Network**

04

**Concluding
Statements**



Neural Network

Loss	0.6781
Accuracy	0.5917

Started with 2 Hidden Levels

Tired to optimize with a a 3rd Hidden level, and changed the units but each try was worse than the original



Road Map

01

**Background
Information**

02

**Supervised
Learning**

03

**Neural
Network**

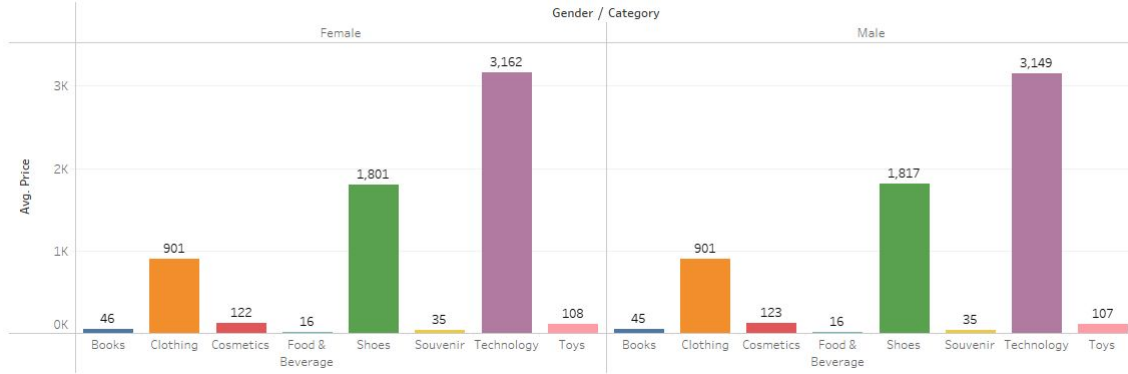
04

**Concluding
Statements**

Conclusion

- Results:
 - Bag and Boost
- Limitations:
 - Data is strange
 - Cultural differences
- Looking Ahead
 - Check Data Collection
 - More Data

Avg. Price by Category



Avg. Price of Goods bought by Day of the week

