In Drive Coupon Recommendation

Project Presentation

Written by **Rania Hana Rafifa**

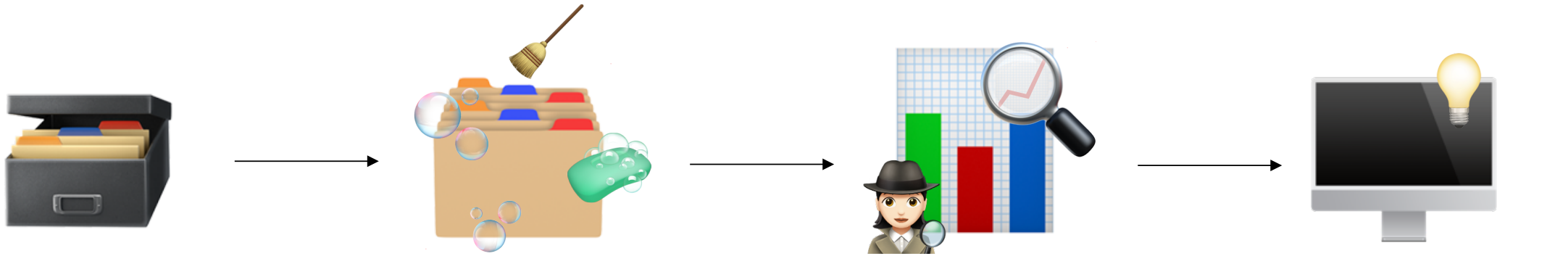
💭 Business Problem

> “Coupons are second only to word-of-mouth when it comes to influencing customer purchasing decisions.”
>
> *Gallimore Industries*

❓ How can we predict whether a customer will accept or reject the offered coupon?

This prediction aims to help companies in offering a correct coupon so that more customers will redeem it and the business’s sales

🗄️ Data Pipeline



Raw Data → Data Preparation → EDA & Model Training → Insights & Recommendation

🗃️ Raw Data

💥 **Data was obtained from Kaggle which collected via a survey on Mechanical Turk of 12,000+ respondents which describe different driving scenarios,** including the user's destination, weather, passenger, coupon attributes, user attributes, and contextual attributes

	destination	passanger	weather	temperature	...	toCoupon_GEQ25min	direction_same	direction_opp	Accept(Y/N?)
0	No Urgent Place	Alone	Sunny	55	...	0	0	1	1
1	No Urgent Place	Friend(s)	Sunny	80	...	0	0	1	0
2	No Urgent Place	Friend(s)	Sunny	80	...	0	0	1	1
3	No Urgent Place	Friend(s)	Sunny	80	...	0	0	1	0
4	No Urgent Place	Friend(s)	Sunny	80	...	0	0	1	0
...	...	...	...	...	...	...	...	...	...
12679	Home	Partner	Rainy	55	...	0	1	0	1
12680	Work	Alone	Rainy	55	...	0	0	1	1
12681	Work	Alone	Snowy	30	...	0	1	0	0
12682	Work	Alone	Snowy	30	...	1	0	1	0
12683	Work	Alone	Sunny	80	...	0	1	0	0

- The data consists of **12,684 rows** and **25 columns**.
- Among those columns, there are : **8 numerical** and **17 categorical variables**.
- The **target variables** is **'Accept(Y/N?)'.**
- If customers **accept the coupon** it is labelled as **1**, **otherwise** it is labelled as **0**.
- This problem is **binary classification problem**

🗃️ Raw Data – Column Description

1. Destination: Destination of users — No Urgent Place | Home | Work
2. Passanger: Who are the users in the car — Alone | Friends | Partner | Kids
3. Weather: — Sunny | Snowy | Rainy
4. Temperature: — 80 | 55 | 30
5. Coupon: Type of coupons — Coffee House | Restaurant (<20) | Carry out & Takeaway | Bar | Restaurant (20-50)
6. Expiration: Validity of Coupon — 1d | 2h
7. Gender: — Male | Female
8. Age: — Below 21 | 26 | 31 | 36 | 41 | 46 | 50 plus
9. Marital Status: — Married Partner | Single | Unmarried Partner | Divorced | Widowed
10. Has Children: — 0 | 1
11. Education: — Some college | No degree | Bachelor degree | Graduate degree | Associate degree | High school | Some high school
12. Occupation: — Unemployed | Student | Business & FInance | Healthcare support | Legal | Retired | ... And so on
13. Income: — <$12500 | $12500-$24999 | $25000-$37499 | $50000-$62499 | ... | >$100000

Raw Data – Column Description

16. Car: Vehicle driven by users — Mazda5 | Do not drive | Scooter | Crossover | Any old cars
17. Bar: how often do the users go to a bar every month? — Never | Less 1 | 1~3 | 4~8 | gt8
18. Coffee House: how often do the users go to a coffee house every month? — Never | Less 1 | 1~3 | 4~8 | gt8
19. Carry Away: how often do the users get take-away food every month? — Never | Less 1 | 1~3 | 4~8 | gt8
20. Restaurant Less Than 20: how often do the users go to a under $20 restaurant every month? — Never | Less 1 | 1~3
21. Restaurant 20 To 50: how often do the users go to a $20-$50 restaurant every month? — Never | Less 1 | 1~3 | 4~8
22. toCouponGEQ_5mins: is driving distance to the restaurant greater than 5 mins? — 0 | 1
23. toCouponGEQ_15mins: is driving distance to the restaurant greater than 15 mins? — 0 | 1
24. toCouponGEQ_25mins: is driving distance to the restaurant greater than 25 mins? — 0 | 1
25. Direction_same: Is the restaurant in the same direction as the user's current destination? — 0 | 1
26. Direction_opp: Is the restaurant in the opposite direction as the user's current destination? — 0 | 1
27. Accept(Y/N?): Will user accept the coupon? — 0 | 1

Data Cleaning – Remove Duplicates and Missing Values

- Drop duplicate rows and keep the first ones
- Check for missing values

	MissingValueCount	MissingPercentage
car	12287	99.144678
Bar	106	0.855322
CoffeeHouse	215	1.734850
CarryAway	148	1.194223
RestaurantLessThan20	128	1.032841
Restaurant20To50	188	1.516985

Drop 'car' columns

✓ Drop columns with up to 80% missing values
✓ As for the rest of missing columns, perform mode imputation.

Data Cleaning – Check features correlations

- Check the correlation using Correlation Matrix

	temperature	has_children	toCoupon_GEQ5min	toCoupon_GEQ15min	toCoupon_GEQ25min	direction_same	direction_opp
temperature	1.000000	-0.016963	NaN	-0.141124	-0.230067	0.088885	-0.088885
has_children	-0.016963	1.000000	NaN	0.078686	-0.011651	-0.032276	0.032276
toCoupon_GEQ5min	NaN	NaN	NaN	NaN	NaN	NaN	NaN
toCoupon_GEQ15min	-0.141124	0.078686	NaN	1.000000	0.321919	-0.297284	0.297284
toCoupon_GEQ25min	-0.230067	-0.011651	NaN	0.321919	1.000000	-0.190759	0.190759
direction_same	0.088885	-0.032276	NaN	-0.297284	-0.190759	1.000000	-1.000000
direction_opp	-0.088885	0.032276	NaN	0.297284	0.190759	-1.000000	1.000000
Accept(Y/N?)	0.064074	-0.044889	NaN	-0.086050	-0.107855	0.016356	-0.016356

✓ Drop **toCoupon_GEQ5min** because it shows **no correlation** with another columns.

✓ **'direction_same' is inveresely correlated with 'direction_opp'.** There is no need for 2 perfectly inverse correlated features, thus we can **remove one of the features**. Here, I **remove 'direction_opp'.**

Data Cleaning – Feature Engineering

Feature engineering : The process that takes raw data and transforms it into features that can be used to create a predictive model using machine learning or statistical modelling

toCoupon_GEQ15min	toCoupon_GEQ25min
0	0
0	0
1	0
1	0
1	0
...	...
0	0

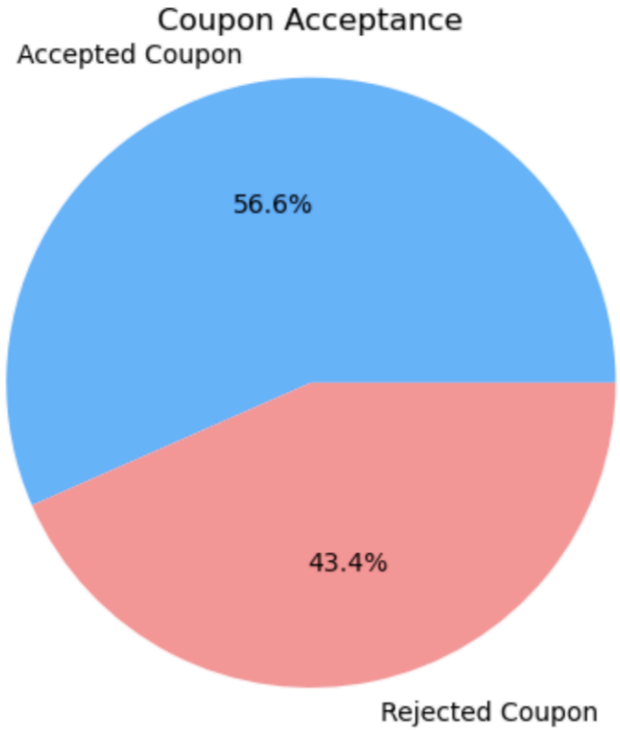
- Create new column called ‘tocoupon’
- If the driving distance <15 mins -> 0
- If the driving distance <25 mins -> 1
- If the driving distance 25 or more -> 2

✓ Drop ‘toCoupon_GEQ15min’,’toCoupon_25min’

✓ That has changed from **12,684 rows** and **25 columns** to **12393 rows and 21 columns**

EDA

Let’s look at target variable distribution

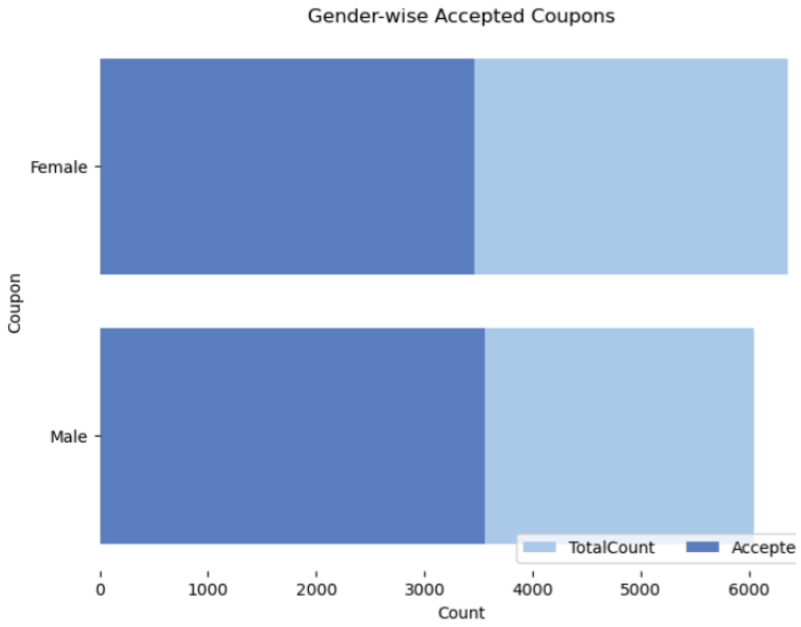


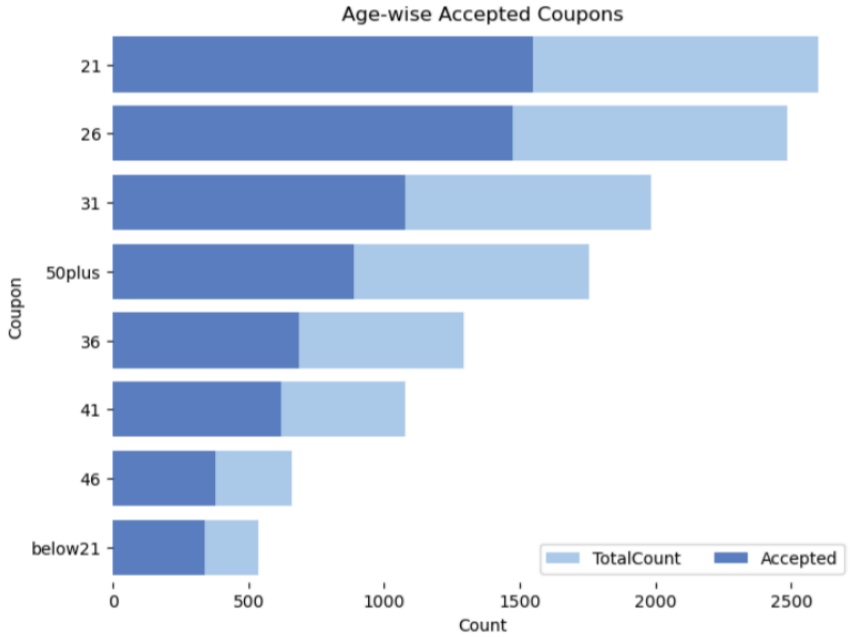
- By using Univariate Analysis, what are the criteria for identifying customers with the highest tendency to redeem the coupon?
- How did the coupons perform?

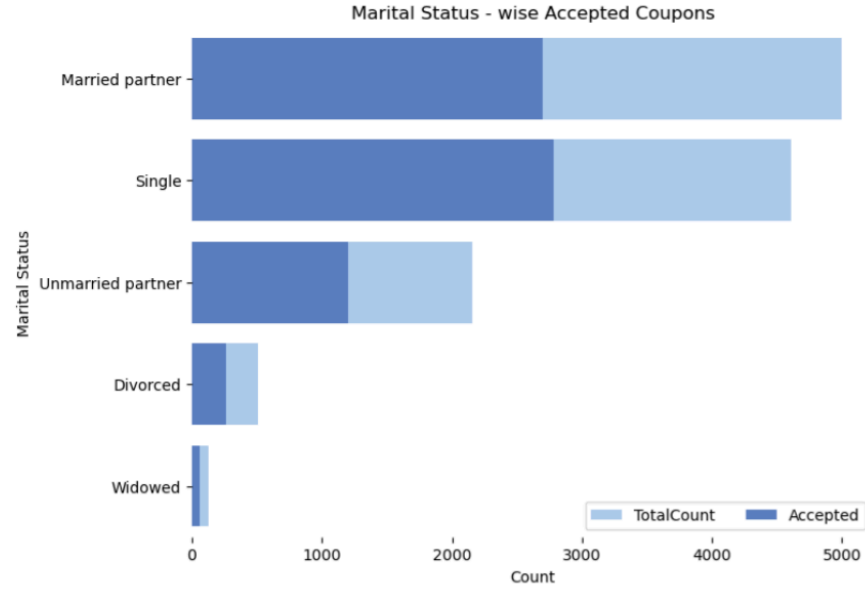
Target variable has roughly balanced class distribution, with **7012 (56.6%) accepted coupons** and **5381 (43.4%) Rejected coupons.**

EDA

By using Univariate Analysis, what are the criteria for identifying customers with the highest tendency to redeem the coupon?

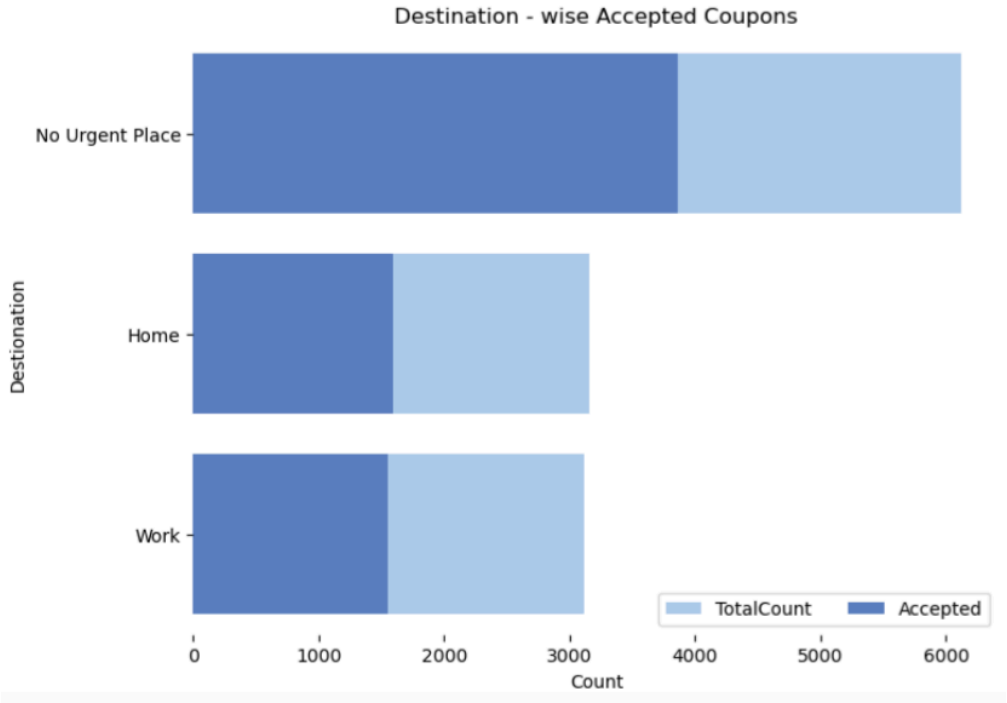


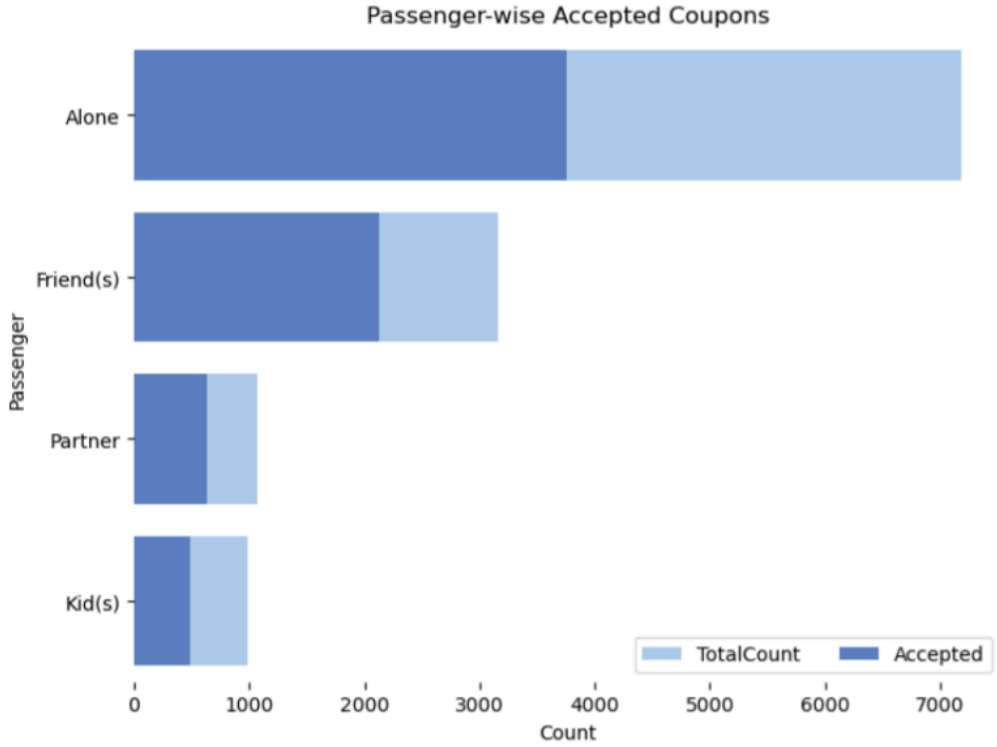




EDA

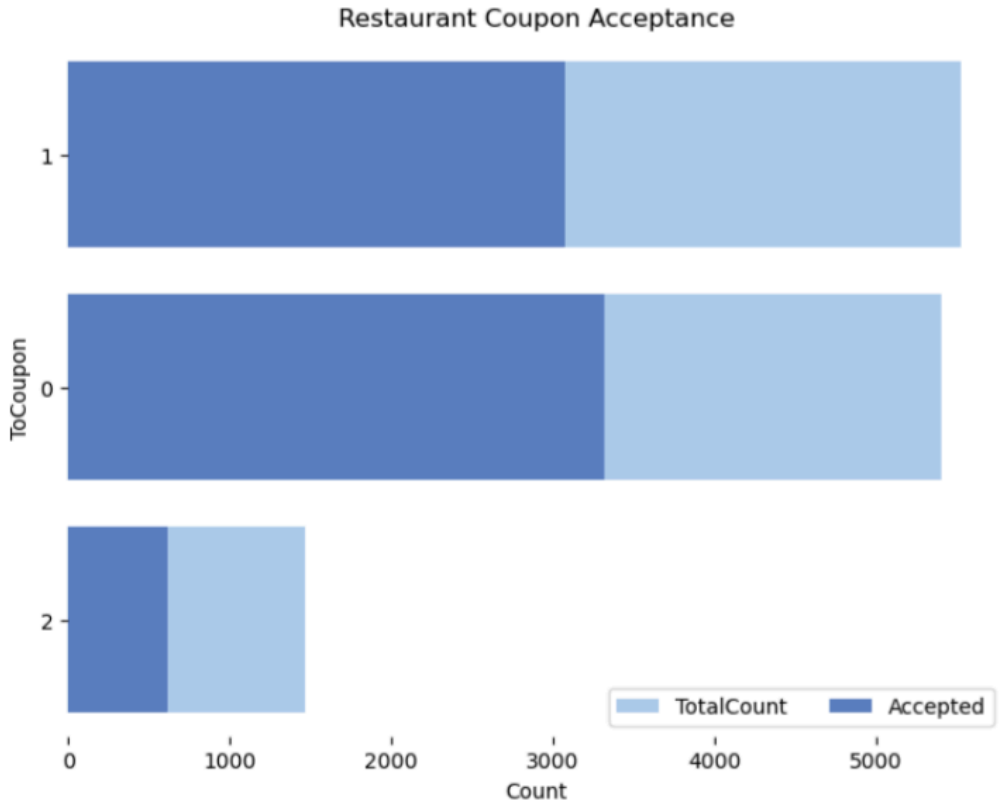
By using Univariate Analysis, what are the criteria for identifying customers with the highest tendency to redeem the coupon?

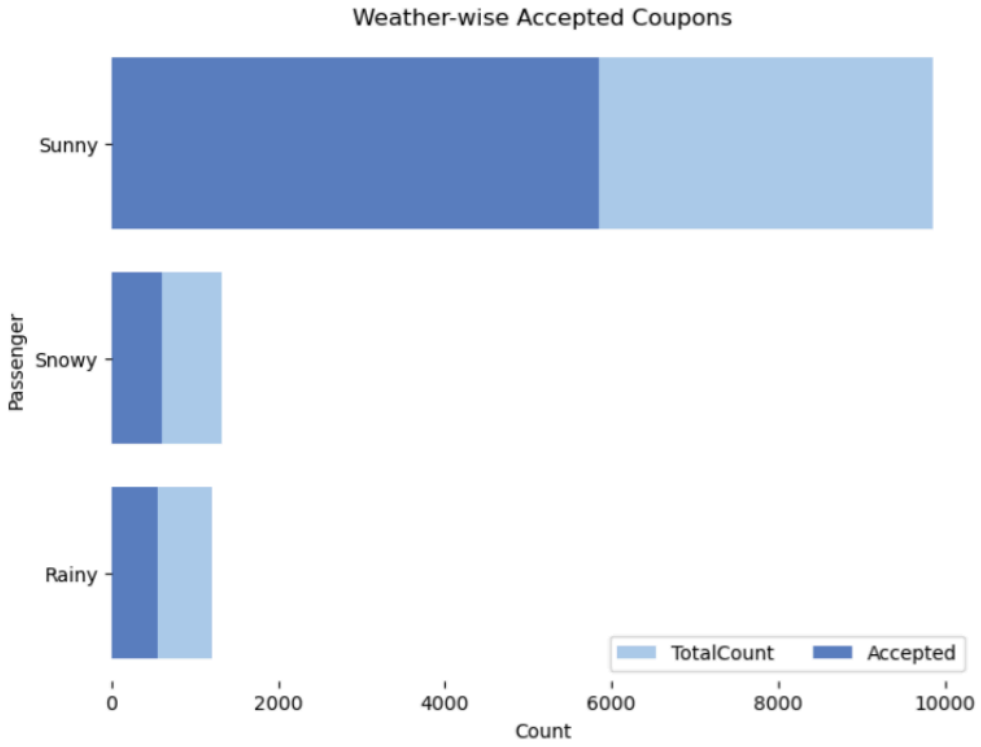




EDA

By using Univariate Analysis, what are the criteria for identifying customers with the highest tendency to redeem the coupon?

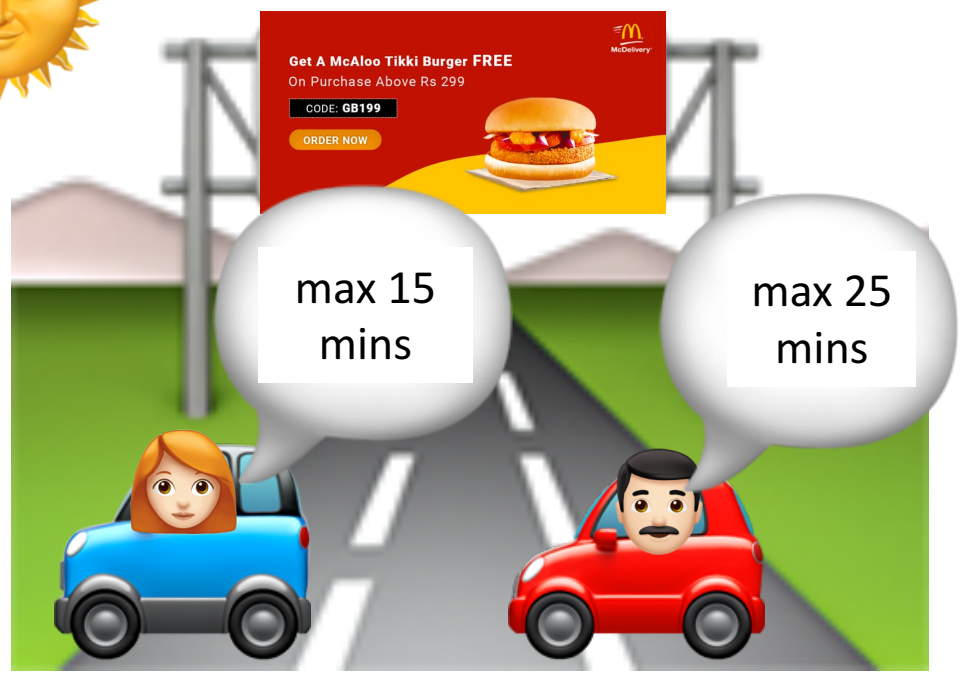




EDA

By using Univariate Analysis, what are the criteria for identifying customers with the highest tendency to redeem the coupon?

26.7℃



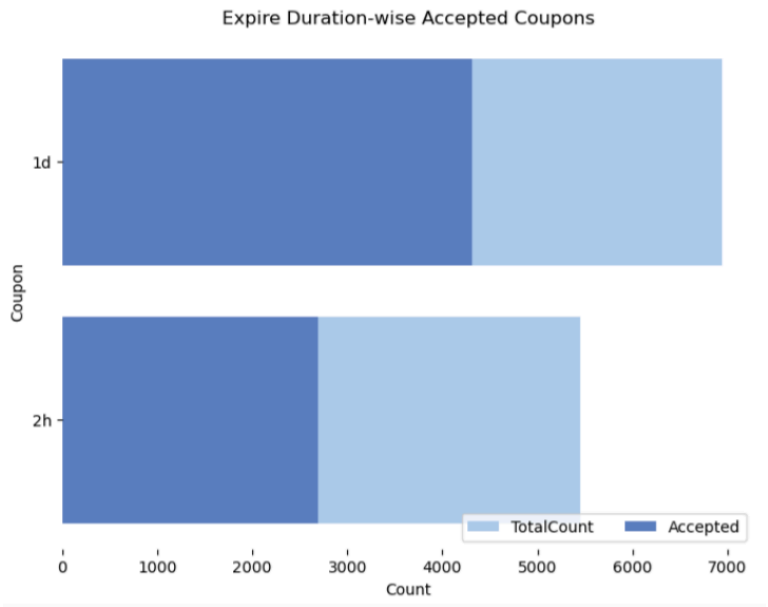
Gani (21, fresh graduate)
- Female
- Single
- Education: Law degree
- Hangout (no urgent places)

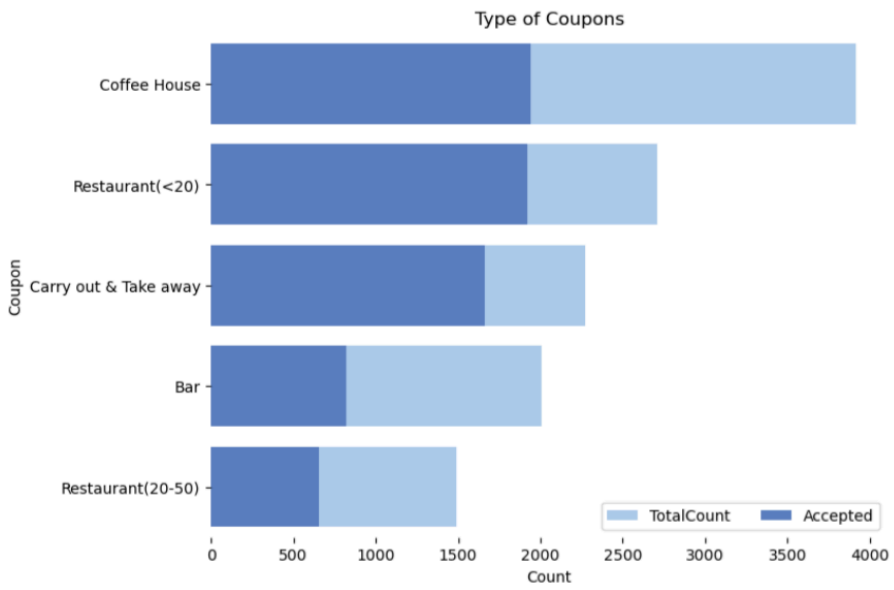
Cliff (26, Mathematician)
- Male
- Married
- Education: Math degree
- Destination: Library (no urgent places)

Note: direction does not affect coupon acceptance

EDA

How did the coupons perform?

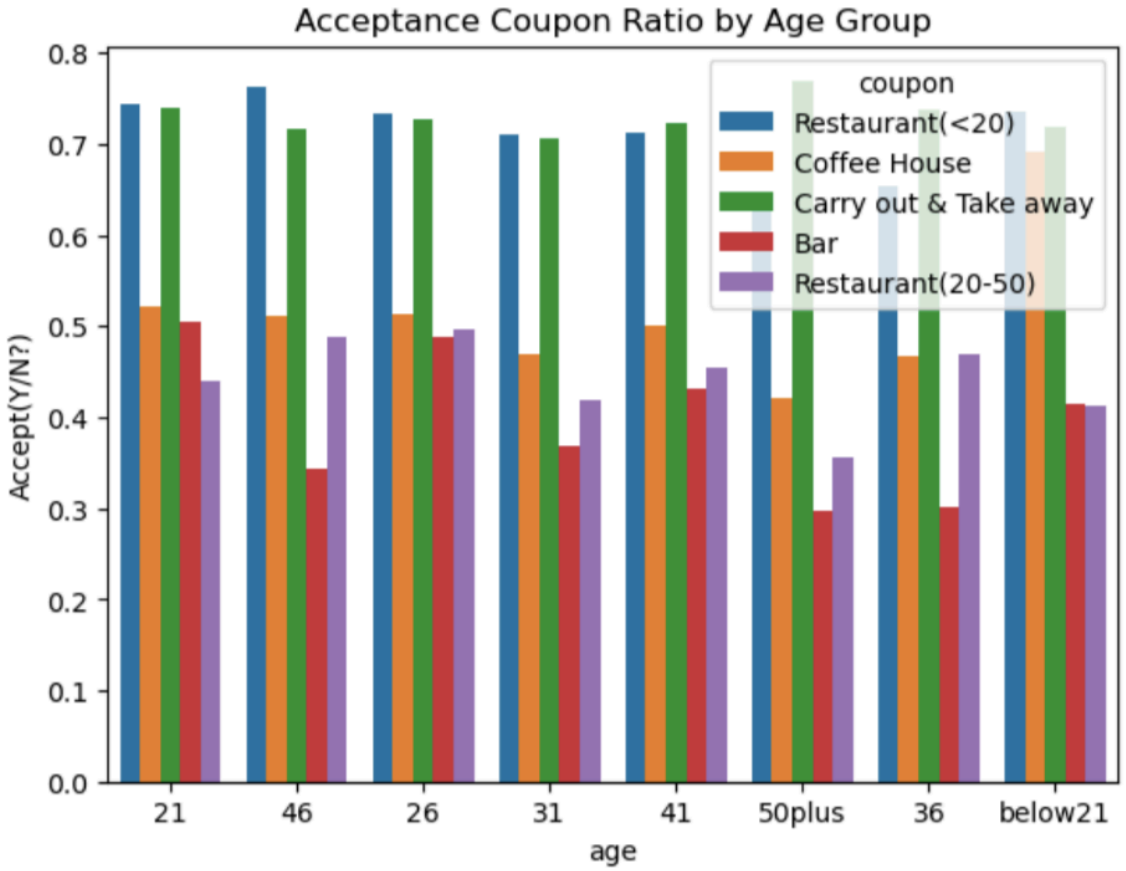




- 1-day expire coupons are the most distributed coupons, followed by their high acceptance ratio.
- Coupons offered by low-priced restaurants and Carry out & take away are the most likely to be accepted, while bar coupons have the lowest acceptance ratio.

EDA

How did the coupons perform?



- In every age groups, people tend to accept Restaurant(<20) and Carry Out & Take Away coupons.
- Unlike all other age groups, people below 21 are likely to accept coffee house coupons with an acceptance ratio of approximately 70%.
- Customers aged 50 and above have the lowest bar coupon acceptance ratio while having the highest acceptance ratio for takeaway coupons.

EDA

How did the coupons perform?

Restaurant (<$20/each)

Liked by all age groups, esp. by people in late 40s

Coffee House

Mostly liked by the minors (below 21)

Take away & Carry Out

Liked by all age groups, esp. by people above 50s

Bar

- Lowest acceptance ratio
- mostly accepted by people in their 20s

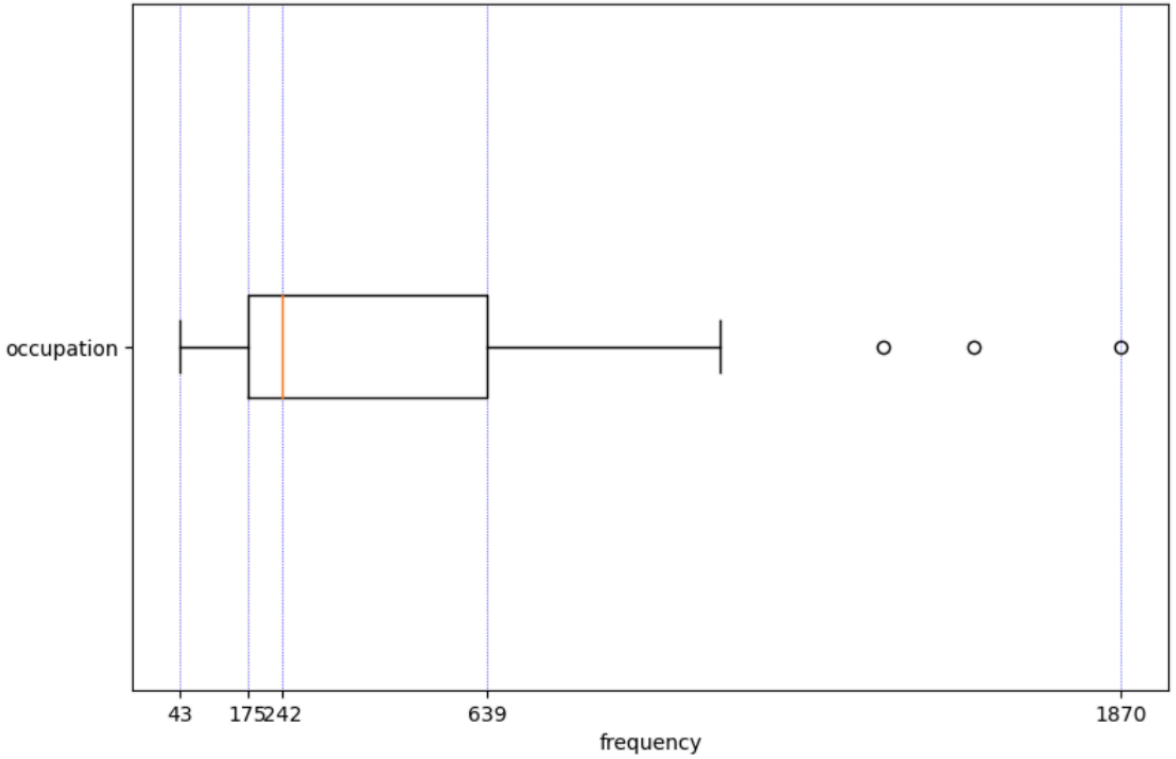
Restaurant ($20-$50/each)

- Low acceptance rate
- Mostly accepted by people in late 20s and 40s

Modelling Preparation

To perform machine learning (ML), models require numerical input data. However, raw data comes in various formats which are not directly useable by ML models. To convert non-numeric data into numerical format suitable for ML, process called 'encoding' is used. Here, ordinal encoding is used.

Prior to ordinal encoding, we need to perform feature engineering on occupation.



Previously, there are 25 types of occupation. To simplify the encoding process, I classify those jobs into 4 groups:

1. Low acceptance -> <175
2. Medium low acceptance -> 175 – 241
3. Medium high acceptance -> 242 – 638
4. High acceptance -> >638

Modelling

For binary classification problems, the model used:

1. Logistic Regression
2. KNN
3. Decision Tree
4. SVM
5. Random Forest
6. Bagging
7. XGBoost

Randomized search is also used for hyperparameter tuning.

Model performance is the evaluated using ROC AUC scores and log loss functions

📂 Model Outputs

	Model	Hyperparameter1	Train_loss	Train_AUC	Test_loss	Test_AUC
0	Logistic Regression	0.01	0.114	1.000	0.114	1.000
1	KNN	24.00	0.406	0.934	0.433	0.908
2	DecisionTree	10.00	0.406	1.000	0.433	1.000
3	SVM	0.10	0.406	1.000	0.433	1.000
4	Random Forest	20.00	0.406	1.000	0.433	1.000
5	Bagging	500.00	0.000	1.000	0.000	1.000
6	XGBoost	20.00	0.000	1.000	0.000	1.000

Bagging and **XGBoost** exhibit the highest model accuracy, as indicated by their ROC AUC scores and log loss functions.

🗂️ Insight & Recommendation

People tend to accept the offered coupons when they are alone

Place the bar coupon billboard nearby the central business district and campus

Place Coffee House coupon billboards close to High School

Make sure that the maximum distance from billboard to the store is 25-mins drive

Experiment of Longer Duration Coupons (e.g. 1 week / month)

THANK YOU