Project Summary

Project Title	Enhancing Coupon Recommendation System Using DTC and MLP
Date	June 2023
Project Objective	Performance Comparison of the Decision Tree (DTC) and Multilayer Perceptron (MLP) on a dataset, and choosing the best-performing model with supporting evidence
Dataset	Dataset: in-vehicle-coupon-recommendation Retrieved from: https://archive.ics.uci.edu/ml/machine-learning-databases/00603/in-vehicle-coupon-recommendation.csv
About Dataset	Amazon Mechanical Turk poll survey describes different driving scenarios including the destination, current time, weather, passenger presence, etc., and then asks participants whether they would accept the coupon if were the driver
Data Features	➢ 12684 rows, 26 columns ➢ Target (0 and 1) is well-balanced ➢ Majority of features are categorical
Cleaning and Preprocessing steps	➢ Dropped column with >5% missing entries ➢ Removed 74 duplicates ➢ Encoded using OrdinalEncoder. ➢ Split train and test 70% ➢ Scaled using StandardScaler (only for MLP)
Model 1: DTC	Accuracy: 70%, Parameters: max_depth=6, max_leaf_nodes=22
Model 2: MLP	Accuracy: 65%, Parameters: hidden_layer_sizes=(10,)
Best Model and reason	Decision Tree Classifier (DTC), because: ➢ Higher precision (71%) and recall score (78%) ensure that customers receive recommendations that are relevant to their needs and preferences ➢ It is well-suited for datasets comprised of discrete or categorical variables ➢ Faster to compute

Tools used

Python: Scikit-learn (for MLP, data scaling, and evaluations), Category Encoders and OrdinalEncoder (for data preprocessing and feature engineering), Seaborn, Matplotlib, and Plotly Express (for visualisation).

Contents

1. Introduction

The objective of this report is to utilize machine learning techniques, specifically Multi-Layer Perceptron (MLP) and Decision Tree Classifier (DTC), to enhance customer satisfaction, engagement, and contribute to overall company growth. We will explore and compare these models to identify which one, or a combination thereof, best enhances our coupon recommendation system.

2. Data Exploration

The 'in-vehicle coupon recommendation Dataset' was gathered from a survey on Amazon Mechanical Turk poll. The survey asked about different driving situations to predict if drivers would accept coupons.

The dataset includes 12,684 entries with 26 attributes.

Feature	Description	Data Type
destination	**Driver's destination:** "No Urgent Place", "Home", "Work"	C
passenger	**The passenger(s) in the car:** "Alone", "Friend(s)", "Kid(s)", "Partner"	C
weather	**The weather when the driver is driving:** "Sunny", "Rainy", "Snowy"	C
temperature	**The temperature when the driver is driving (in °F):** "55", "80", "30"	C (continuous data treated as categorical)
time	**The time at which the driver is driving:** "2PM", "10AM", "6PM", "7AM", "10PM"	C/numeric categorical (continuous data treated as categorical)
coupon	**Type of coupon that will be accepted:** "Restaurant(<20)", "Coffee House", "Carryout & Take away", "Bar, Restaurant(20-$50)"	C
expiration	**The expiration date of the coupon:** "1d", "2h"	C/numeric categorical (continuous data treated as categorical)
gender	**Driver's gender:** "Female", "Male"	C
age	**Driver's age:** "21", "46", "26", "31", "41", "50plus", "36", "below21"	C/numeric categorical (continuous data treated as categorical)
maritalStatus	**Driver's marital status:** "Unmarried partner", "Single", "Married partner", "Divorced", "Widowed"	C
has_Children	**Whether the driver has child(ren) or not:** 0: no, 1: yes	C (binary variable)
education	**Driver's educational background:** "Some college - no degree", "Bachelors degree", "Associates degree", "High School Graduate", "Graduate degree (Masters or Doctorate)", "Some High School"	C
occupation	**Driver's occupation:** "Unemployed", "Architecture & Engineering", "Student", "Education&Training&Library", "Healthcare Support", "Healthcare Practitioners & Technical", "Sales & Related", "Management", "Arts Design Entertainment Sports & Media", "Computer & Mathematical", "Life Physical Social Science", "Personal Care & Service", "Community & Social Services", "Office & Administrative Support", "Construction & Extraction", "Legal", "Retired", "Installation Maintenance & Repair", "Transportation & Material Moving", "Business & Financial", "Protective Service", "Food Preparation & Serving Related", "Production Occupations", "Building & Grounds Cleaning & Maintenance", "Farming Fishing & Forestry"	C
income	Driver's income: "$37500 - $49999", "$62500 - $74999", "$12500 - $24999", "$75000 - $87499", "$50000 - $62499", "$25000 - $37499", "$100000 or More", "$87500 - $99999", "Less than $12500"	C/numeric categorical (continuous data treated as categorical)
Car	**Car model driven by the driver:** "Scooter and motorcycle", "crossover", "Mazda5"	C
Bar	**The frequency of restaurant visits per month:** "never", "less1", "13", "gt8", "nan48"	C
CoffeeHouse	**Frequency of cafe visits per month:** "never", "less1", "48", "13", "gt8", "nan"	C/numeric categorical (continuous data treated as categorical)

CarryAway	**Frequency of takeaway food consumption per month:** "n48", "13", "gt8", "less1", "never"	C/numeric categorical (continuous data treated as categorical)
RestaurantLessThan20	**Frequency of restaurant visits per month, where the average expense per person is less than $20:** "48", "13", "less1", "gt8", "never"	C/numeric categorical (continuous data treated as categorical)
Restaurant20To50	**Frequency of restaurant visits per month, where the average expense per person is between 20–50:** "13", "less1", "never", "gt8", "48", "nan"	C/numeric categorical (continuous data treated as categorical)
toCoupon_GEQ5min	**Open to travelling beyond a 5-minute distance to use the coupon:** 0: no, 1: yes	C (binary variable)
toCoupon_GEQ15min	**Open to travelling beyond a 15-minute distance to use the coupon:** 0: no, 1: yes	C (binary variable)
toCoupon_GEQ25min	**Open to travelling beyond a 25-minute distance to use the coupon:** 0: no, 1: yes	C (binary variable)
direction_same	**Whether the restaurant or cafe mentioned in the coupon is in the same direction as drivers' current destination:** 0: no, 1: yes	C (binary variable)
direction_opp	**Whether the restaurant or cafe mentioned in the coupon is in the opposite direction as drivers' current destination:** 0: no, 1: yes	C (binary variable)
Y	**Whether the driver will accept the coupon or not:** 0: no, 1: yes	C (binary variable)

3. Data Cleaning

Missing Values

The feature 'car' has the highest missing value rate of 99% (12576). As per general guidelines, missing values that are less 5% have minimal impact on the overall analysis.

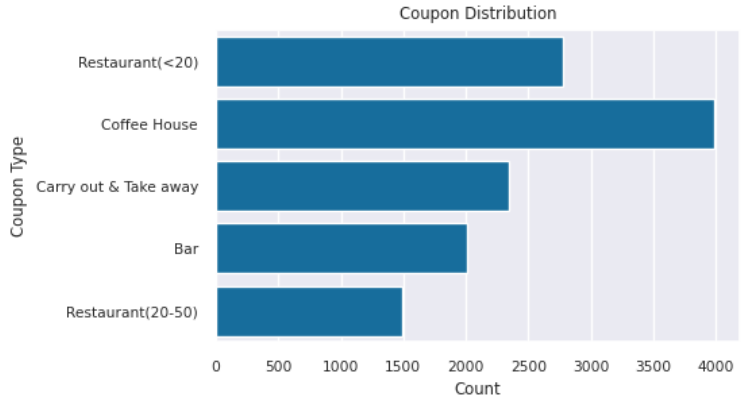
	missing_count	missing_percentage
car	12576	99.148534
Bar	107	0.843582
CoffeeHouse	217	1.710817
CarryAway	151	1.190476
RestaurantLessThan20	130	1.024913
Restaurant20To50	189	1.490066

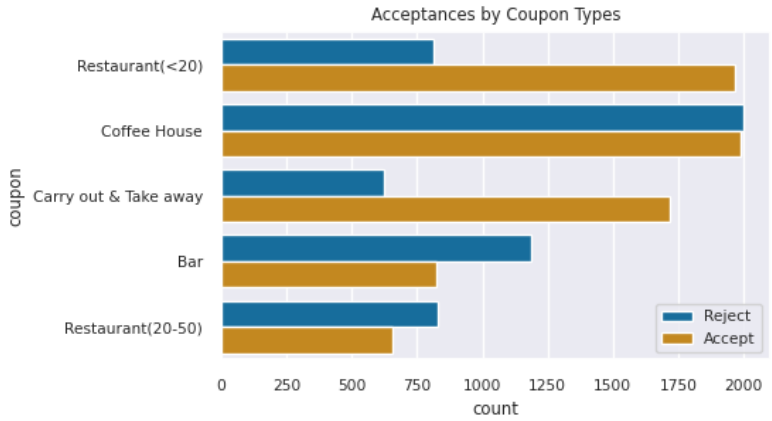
Duplicates

74 duplicates have been removed.

4. Exploratory Data Analysis (EDA)

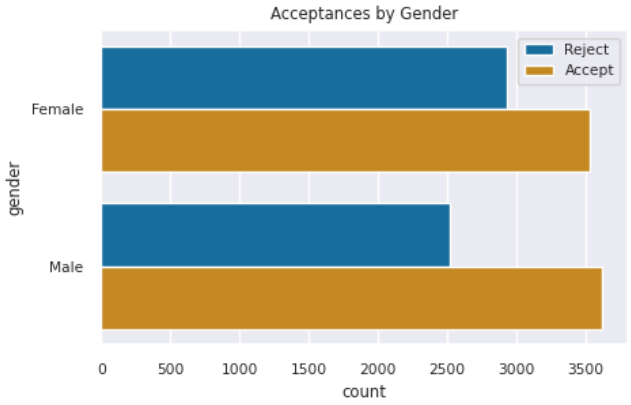
Distribution of coupons



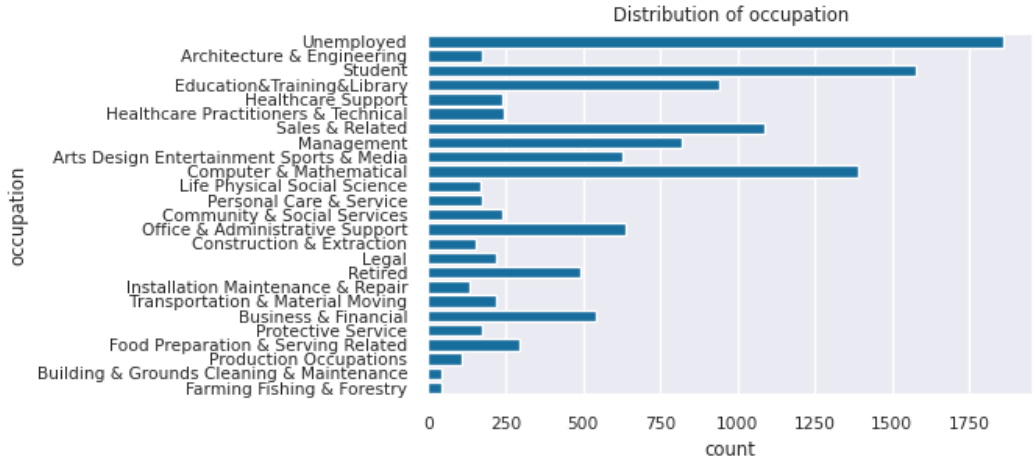


- Coffee House distributes the most coupons. It shows the highest acceptance rate but also experiences the highest rejection rate.
- Coupons for light meals like restaurants(<20 seats) and takeaways tend to have higher acceptance rates than rejection rates, whereas coupons that have the lowest distribution rate like bars and restaurants (20-50 seats) typically face higher rejection rates.

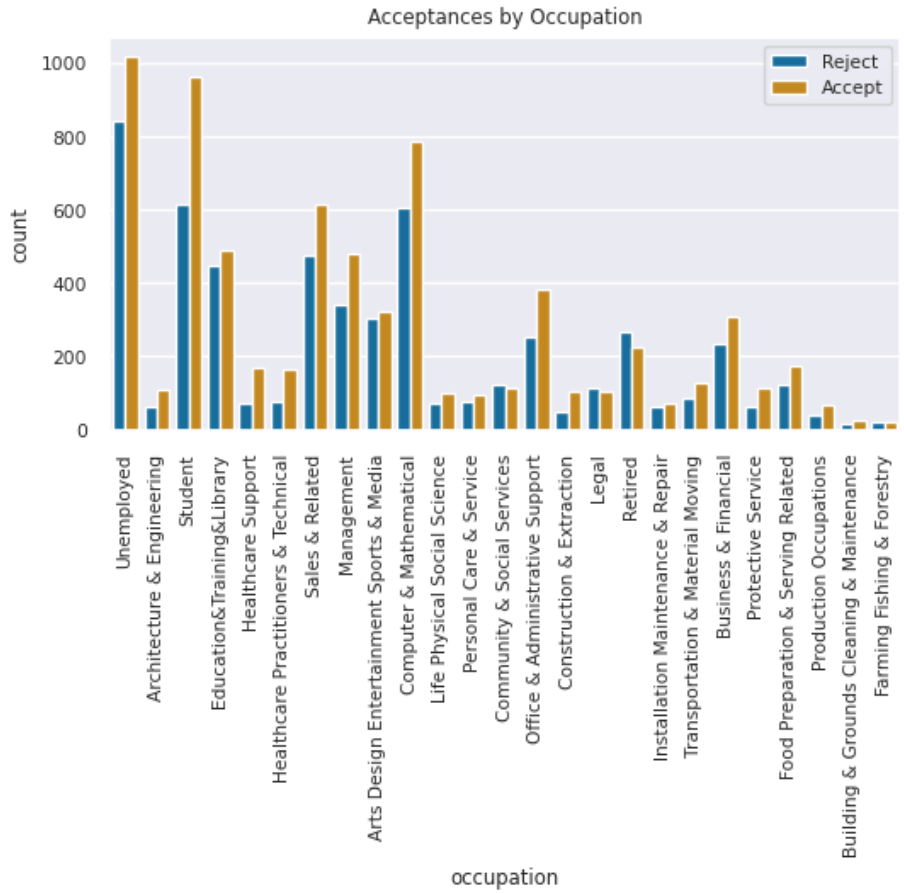
a. Gender / Occupation & Salary

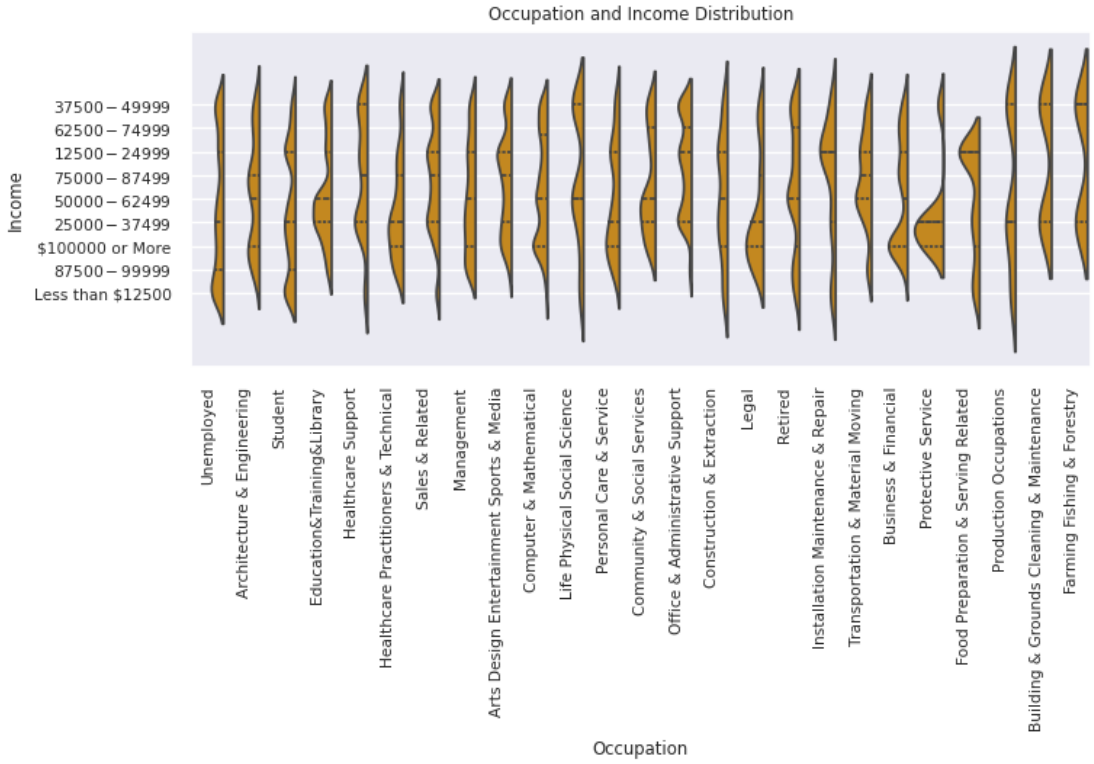


- Both males and females have a similar acceptance rate. A larger proportion of females tend to accept the coupon compared to males.



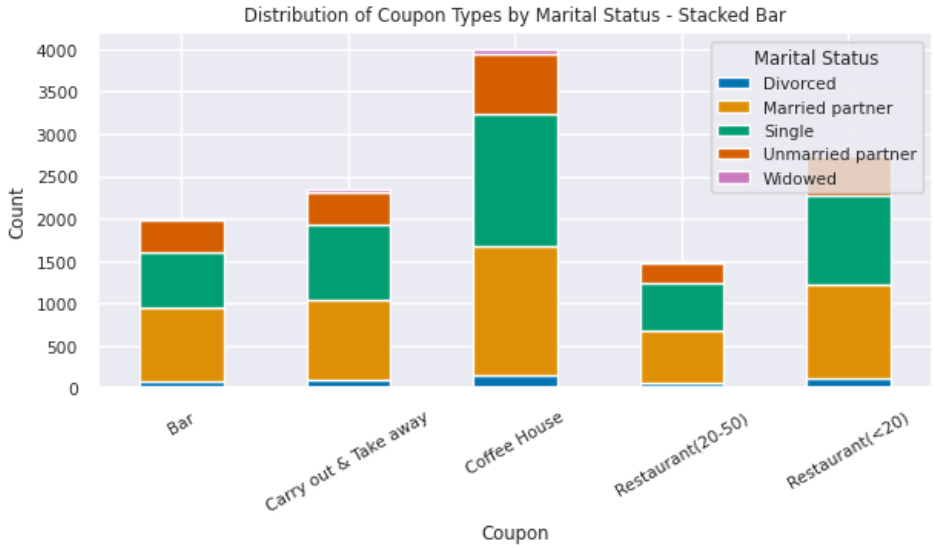
- There is a high ratio of drivers who fall under the categories of Unemployed, Student, Education&Training&Library, Sales & Related, and Computer & Mathematical occupations.

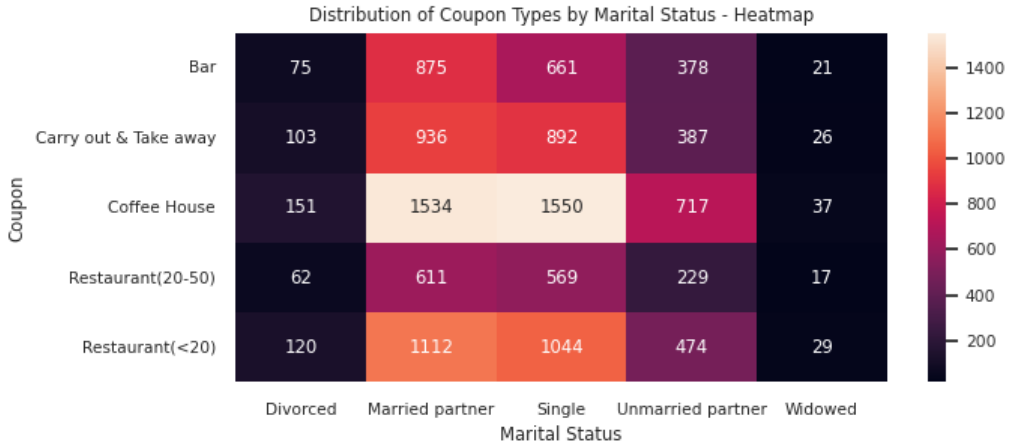




- Categories that have higher rejection rates are Community/Social Service, Legal, and Retired.

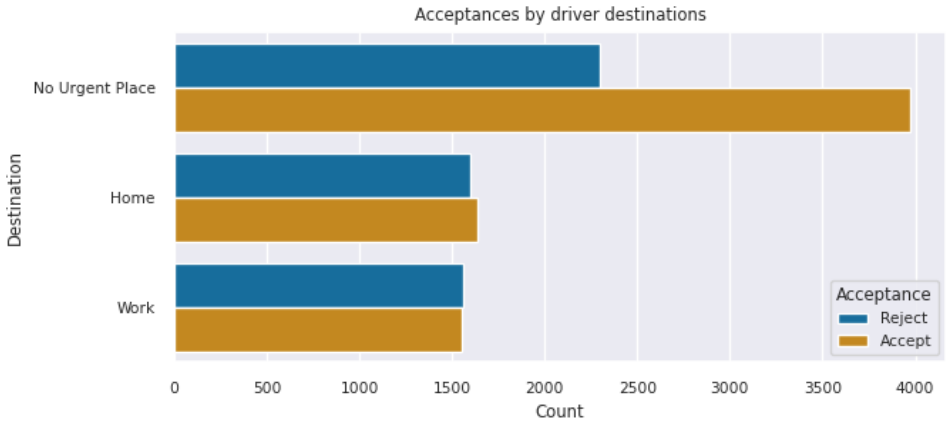
b. Marital Status

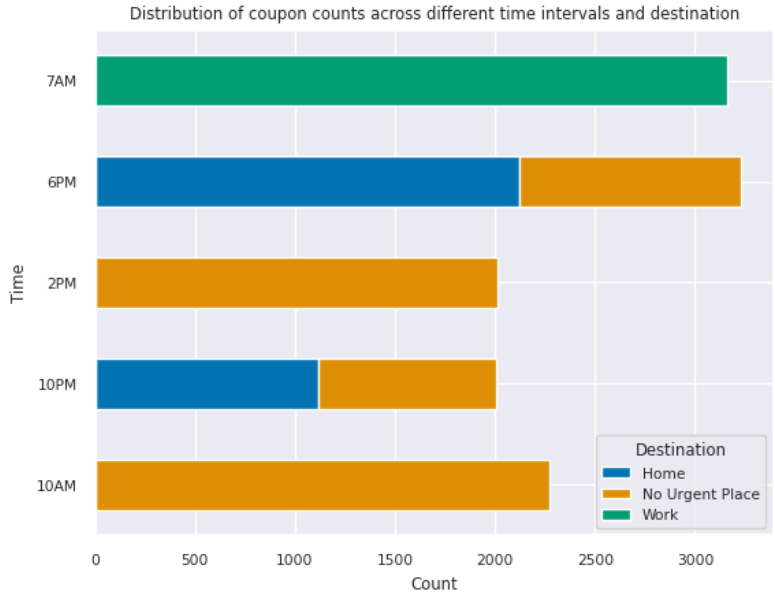


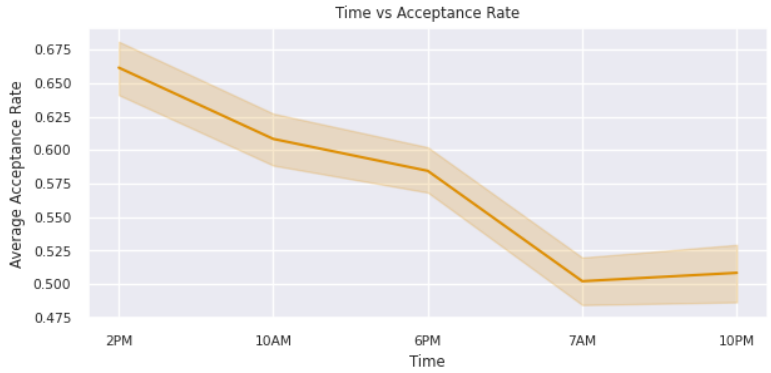


- Coupons are distributed widely among drivers classified as 'single' and 'married partners' across all categories.

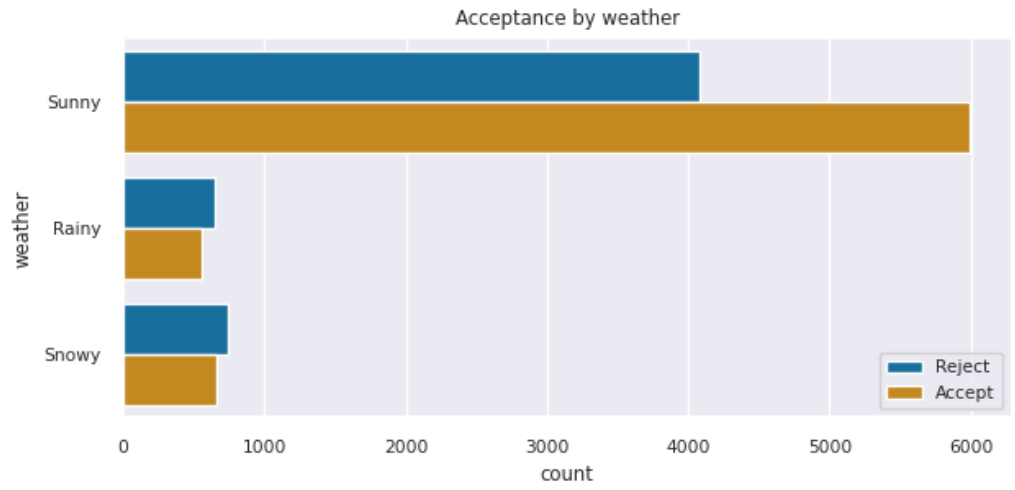
c. Contextual: Drivers Destination and Time

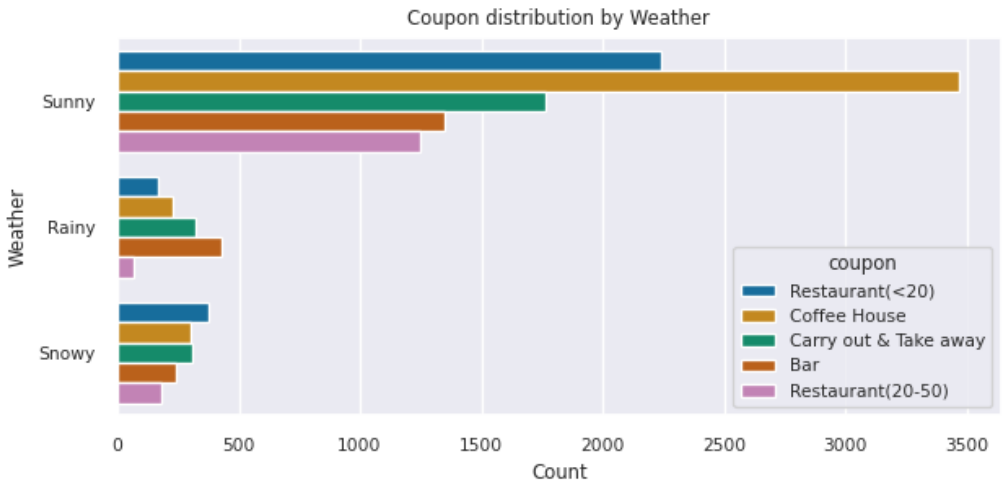






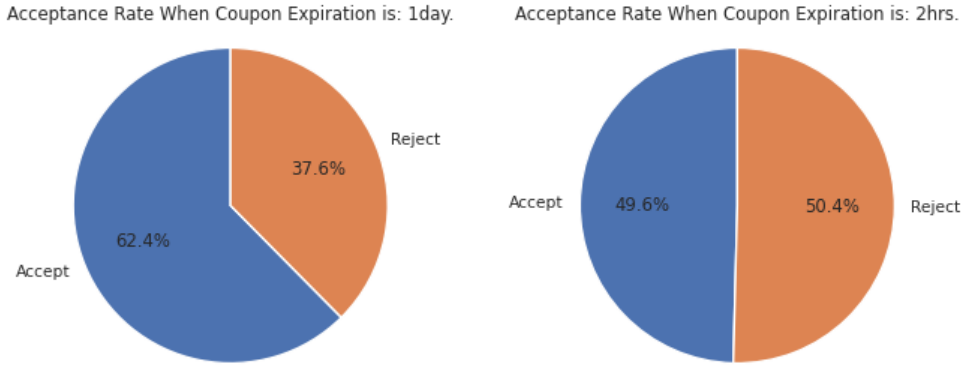
- More drivers would accept when they are not in a hurry to reach a destination. Acceptance rates are higher outside of rush hours, with a peak around 2pm, after lunchtime. Less popular times for accepting coupons are around 7am, during morning commute hours, and at 10pm, closer to bedtime.



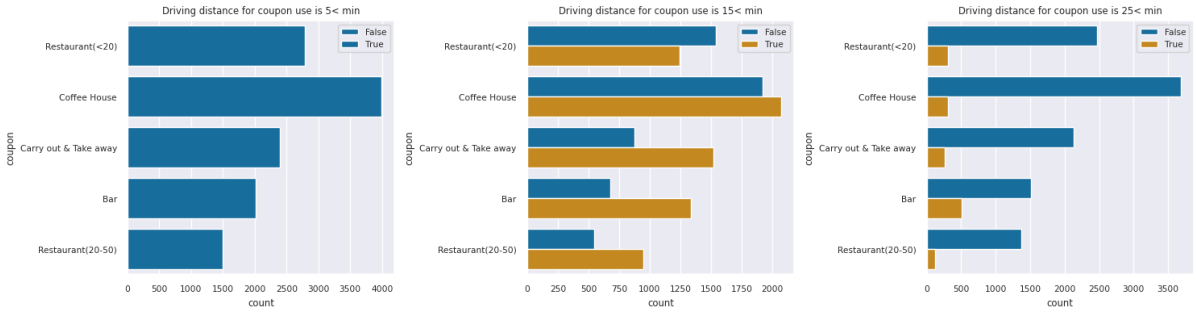


- Coupons are more frequently distributed on sunny days compared to cold and rainy days. People are more likely to reject coupons when the weather is rainy or cold.

d. Situational: Coupons Expiration Dates, Drivers destination, Weather/Season



- Acceptance rates are higher when drivers are offered coupons with longer expiration dates.



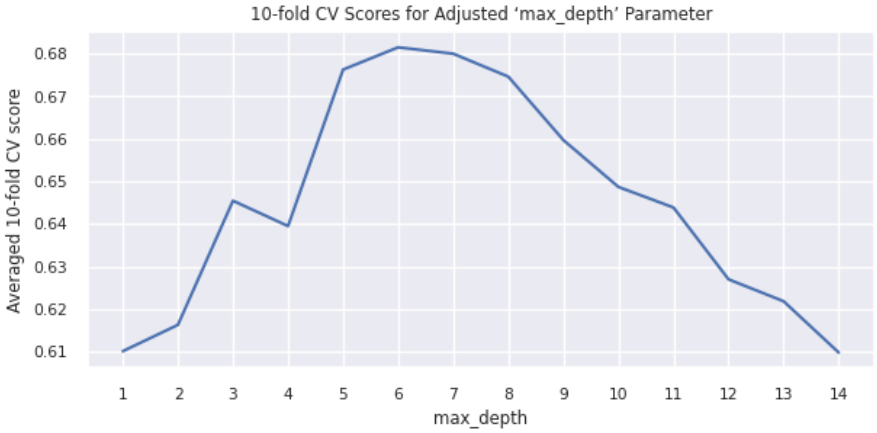
- Drivers tend to decline the coupon as the distance to drive grows longer.

5. Model 1: Decision Tree Classifier

Hyperparameter Tuning

As a baseline, I used default parameters, resulting in a decision tree consisting of 4667 nodes and an accuracy score of 0.69 (2dp).

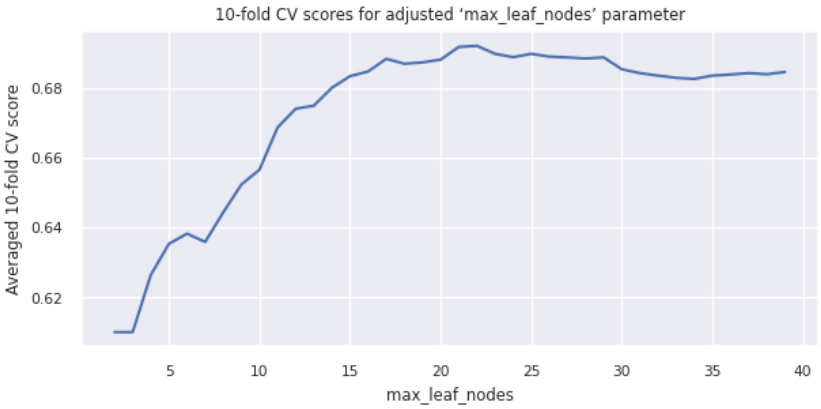
10-fold CV scores for adjusted 'max_depth' parameter



- Adjusted the max_depth parameter to limit the depth of the decision tree, testing different values of max_depth from 1 to 14 using 10-fold cross-validation for each value.
- Diagram below displays the average scores from 10-fold cv for various max_depth values. The optimal max_depth is 6.
- After adjusting the max_depth parameter, nodes in the tree decreased to 127, and the accuracy score improved to 0.70 (2dp)

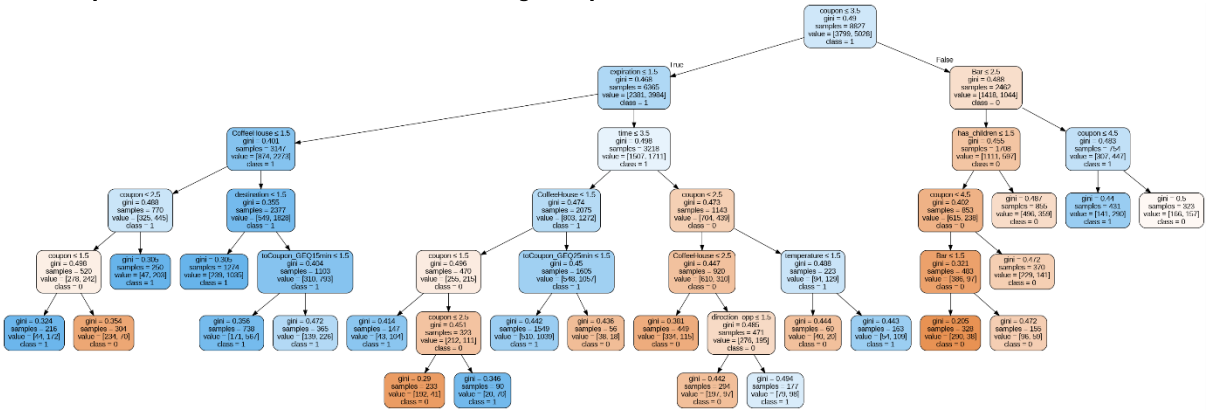
Refinement

10-fold CV scores for adjusted 'max_leaf_nodes' parameter



- I focused on adjusting the max_leaf_nodes parameter, as it limits the max number of leaf nodes in the decision tree.
- Used max_depth value of 6 as a starting point, then examined different max_leaf_node values from 2 to 39. Through this process, we determined that a max_leaf_node value of 22 resulted in the highest score from our 10-fold cross-validation.
- The model achieved an accuracy score of 0.70 (2dp). Adjustments did not significantly improve the accuracy of DTC.

Final optimised classification tree using Graphvis



- This binary tree comprises a total of 47 nodes - 24 internal nodes with 22 leaf nodes that make final classification decisions.
- The optimised tree has a depth of 6 levels. This depth value was identified as optimal during the decision tree's parameter optimisation (depth indicates the longest path from the initial node to any leaf node in the tree).

DTC Evaluation: Confusion Matrix

1661 instances were correctly classified as 'accept' (TP).
970 instances were correctly classified as 'reject' (TN).
684 instances were incorrectly classified as 'accept' when they were actually 'reject' (FP).
468 instances were incorrectly classified as 'reject' when they were actually 'accept' (FN).

DTC Evaluation: Model summary report

	precision	recall	f1-score	support
0	0.67	0.59	0.63	1654
1	0.71	0.78	0.74	2129
accuracy			0.70	3783
macro avg	0.69	0.68	0.68	3783
weighted avg	0.69	0.70	0.69	3783

- Precision 0: Model predicts a customer will respond negatively to a coupon, it is correct 67% of the time.
- Precision 1: Predicts a customers that will respond positively to a coupon, it is correct 71% of the time.

- Recall 0: Out of all the customers who truly would not respond positively to a coupon, the model correctly predicts 59% of them.

- Recall 1: out of all the customers who truly would respond positively to a coupon, the model correctly predicts 78% of them.
- F1-score 0: Out of all instances predicted as 'reject' by the model, 63% are actually 'accept'.
- F1-score 1: 74 means the model has a good overall performance in correctly identifying customers who will respond positively to a coupon.
- Accuracy: 70% of the predictions made by the model are correct (predictions across both 1 and 0 classes).

Summary
The model performs adequately in predicting coupon acceptance and rejection. Error rate of 30% indicates there are still opportunities for enhancement, especially in reducing false predictions. Further optimisation could focus on improving precision for 'reject' predictions and reducing overall prediction errors.

6. Feature importance

Feature importance (based on the final classification model)

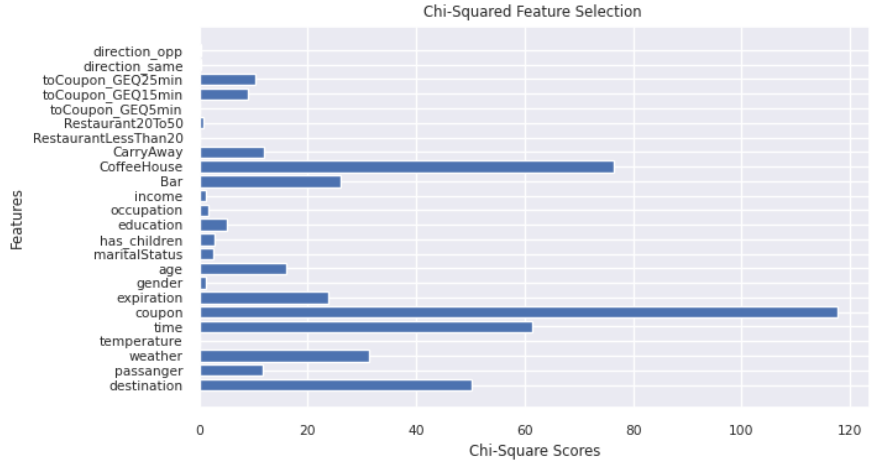
	Feature	Importance
5	coupon	0.477
6	expiration	0.142
15	CoffeeHouse	0.103
4	time	0.095
14	Bar	0.094
10	has_children	0.021
21	toCoupon_GEQ25min	0.016
23	direction_opp	0.014
20	toCoupon_GEQ15min	0.013
0	destination	0.013
3	temperature	0.012
9	maritalStatus	0.000
11	education	0.000
1	passanger	0.000
13	income	0.000
8	age	0.000
7	gender	0.000
16	CarryAway	0.000
17	RestaurantLessThan20	0.000
18	Restaurant20To50	0.000
19	toCoupon_GEQ5min	0.000
2	weather	0.000
22	direction_same	0.000
12	occupation	0.000

- Each score under ‘Importance’ indicates the significance of input features in predicting the target variable, whether the driver will accept the coupon or not.
- The feature 'coupon' holds a substantially higher importance score of 0.477 plays the most significant role in influencing the model’s decision-making process.
- Following 'coupon', factors such as 'expiration', 'coffee house', 'bar', and 'time' may also play a notable role in influencing the model's decision-making process.

Feature Importance: Chi-Square

To improve the performance of classification models, we eliminate features (attributes) that do not significantly impact the model's decision-making process. This time, we will try using the Chi-Squared Method.

Feature selection using Chi-Squared Method



- The top 5 significant features: Coupon (117.77), CoffeeHouse (80.34), time (76.41), destination (50.17), and expiration (23.77).

Comparing the 2 feature selections

Top 5 features from Chi-Square method (desc)	Top 5 features from DTC model (desc)
Coupon	Coupon
CoffeeHouse	expiration
time	CoffeeHouse
destination	Bar
expiration	Time

- 'Coupon' shows strong potential for influencing / predicting the target variables.
- 'CoffeeHouse,' 'time,' and 'expiration' appearing in both the Chi-Square and DTC model results suggest a strong significance. When the same feature shows up in both lists, it indicates a high level of relevance.
- We observe certain features that exist in the Chi-Square method but not in the other method, and vice versa (like 'Bar' and ‘Destination’).

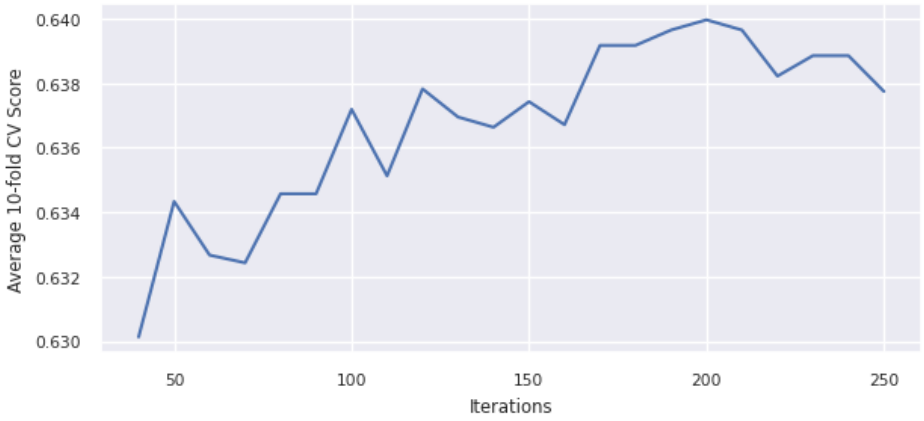
7. Model 2: Multi-Layer Perceptron (MLP)

Testing model: single layer, hidden_layer_size =10, 10-fold-cv, 140 iterations = cv-score 0.64(2dp)

Initial parameter: hidden_layer_size=10 (single hidden layer with 10 neurons).

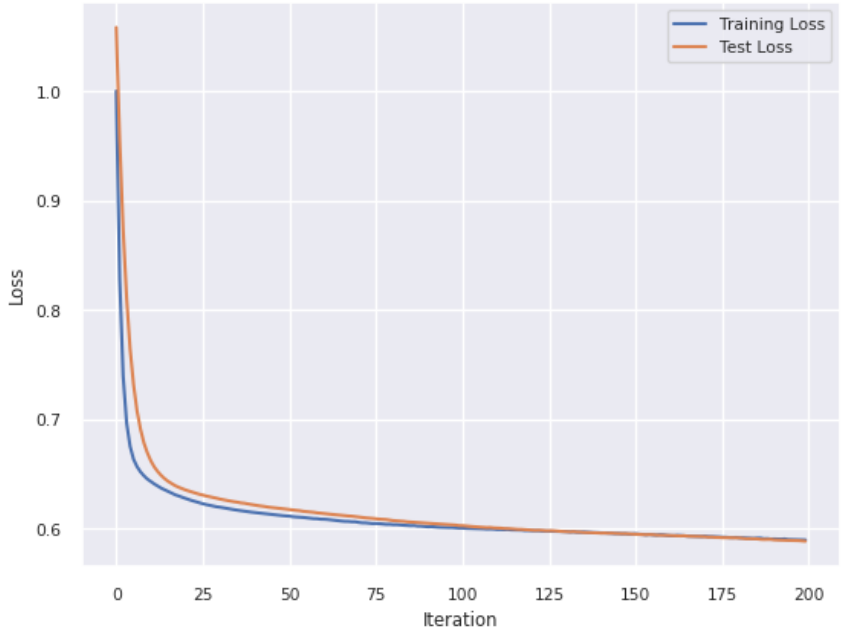
To find the optimal number of iteration, I tested a range of maximum iterations in increments of 10, ranging from 40 to 250.
200 iterations produced the highest average 10-fold cross-validation score of 0.64(2dp).

Testing the model using a 10-fold CV test



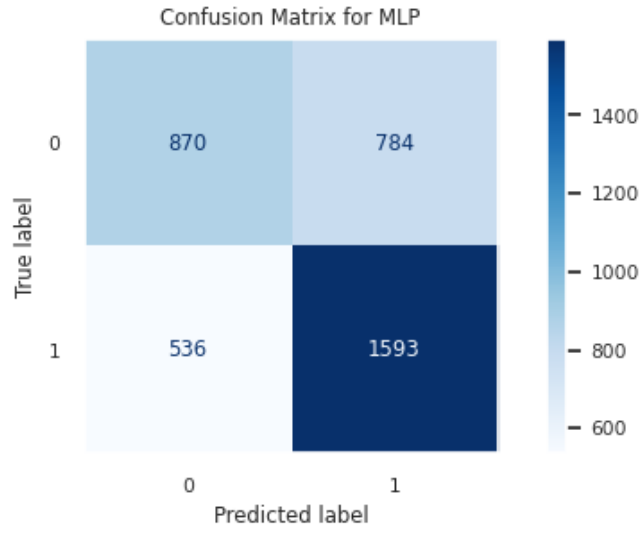
Evaluations

Evaluation: MLP Loss and error values



- The closely aligned curves show a corresponding decrease, meaning that an appropriate number of iterations has been used for the model.

MLP Evaluation: Confusion Matrix



MLP Evaluation: Model summary report

	precision	recall	f1-score	support
0	0.62	0.53	0.57	1654
1	0.67	0.75	0.71	2129
accuracy			0.65	3783
macro avg	0.64	0.64	0.64	3783
weighted avg	0.65	0.65	0.65	3783

7. Performance Comparison: DTC vs MLP

Performance metrics: DTC vs MLP

	precision	recall	f1-score	support
0	0.67	0.59	0.63	1654
1	0.71	0.78	0.74	2129
accuracy			0.70	3783
macro avg	0.69	0.68	0.68	3783
weighted avg	0.69	0.70	0.69	3783

	precision	recall	f1-score	support
0	0.62	0.51	0.56	1654
1	0.67	0.75	0.71	2129
accuracy			0.65	3783
macro avg	0.64	0.63	0.63	3783
weighted avg	0.64	0.65	0.64	3783

- Precision: Reliability and correctness of the collected data. It's important metric to make informed and improved business decisions. The decision tree classifier holds a higher score than the MLP classifier, indicating that DTC does a better job in correctly identifying true positives while minimising false positives.
- Recall/Sensitivity: Metric to determine the likelihood of the machine recommending/offering irrelevant coupons to customers. DTC holds a higher recall score of 78%, indicating that the model does a better job of reducing the chances of falsely labelling true positives as negatives compared to the MLP classifier.

Which model is better?

DTC performs better in terms of reducing false positives and negatives. Additionally, it is more suitable for this particular dataset (that consist of discrete or categorical variables) and its capability of handling these variables.

DTCs are also well known for their simplicity and interpretability. The decision-making process is transparent (often called “white-box”). Moreover, due to its simplicity, they are faster and don’t require excessive computational resources compared to MLP Classifiers.