

HUBBLEMIND

MACHINE LEARNING INTERNSHIP



ISRAEL DUROTOYE

**COUPON
RECOMMENDATION
SYSTEM USING USER
BEHAVIOR DATA**



HUBBLEMIND

TABLE OF CONTENTS

1. Introduction
 - 1.1 Project Objective
 - 1.2 Problem Statement
 - 1.3 Overview of Approach
2. Data Description
 - 2.1 Data Source
 - 2.2 Features Overview
 - 2.3 Target Variable
3. Data Preprocessing
4. Exploratory Data Analysis (EDA)
 - 4.1 Target Variable Analysis
 - 4.2 Feature-Target Relationships:
 - 4.3 Correlation Analysis
 - 4.4 Conclusion of EDA
5. Model Development
 - 5.1 Model Selection
 - 5.3 Model Training and Evaluation:
 - Logistic Regression
 - Decision Tree Classifier
 - Random Forest Classifier
 - 5.4 Comparison of Models:
 - 5.5 Best-Performing Model
6. Hyperparameter Tuning
 - 6.1 Approach
 - 6.2 Parameter Grid
7. Discussion
 - 7.1 Insights
 - 7.2 Challenges
 - 7.3 Limitations
8. Conclusion
 - 8.1 Summary of Findings
 - 8.3 Real-World Applications

Introduction

Project Objective

This project aims to create a machine-learning model that can reliably forecast if a user will accept a coupon, taking into account different contextual and demographic factors like weather, type of passenger, time of day, and other relevant elements. By using predictive analytics, the intention is to boost the success of coupon targeting strategies, helping businesses increase user interaction, optimize their marketing efforts, and achieve higher acceptance rates in a vehicle recommendation setting.

Problem Statement

Properly sharing coupons is essential for companies looking to boost user interaction and improve their marketing strategies. Nevertheless, figuring out if a user will accept a coupon depending on their circumstances—like the weather, time of day, or travel details—can be difficult. This project seeks to overcome this issue by creating a machine-learning model that enables coupon distributors to better forecast user actions. By customizing coupon suggestions to match a user's unique situation, businesses can enhance their targeting, minimize unnecessary spending, and raise acceptance rates, leading to better marketing results.

Overview of Approach

This project implements a structured machine-learning pipeline aimed at predicting whether a user will accept a coupon based on contextual and behavioral factors. The analysis utilizes the In-Vehicle Coupon Recommendation dataset, which comprises 12,684 instances of users being offered coupons while driving. Each entry in the dataset offers detailed contextual data, such as the user's destination, travel companions, weather conditions, and the time of day, as well as behavioral attributes like their frequency of visits to specific establishments. The target variable (Y) indicates whether the user accepted (1) or rejected (0) the coupon. This project aims to leverage these data points to create a predictive model that enhances coupon targeting for businesses.

The first step of the project was data preprocessing. The raw dataset contained inconsistencies, missing values, and categorical variables that required transformation. Missing values were addressed using appropriate imputation techniques, while irrelevant or redundant features were eliminated to minimize noise. Categorical features, such as destination, weather, and coupon type, were encoded into numerical formats using OneHotEncoding to adapt them for machine learning algorithms. Additionally, numerical features like age and temperature were scaled to enhance model performance. Special attention was given to transforming features like time and age into meaningful formats, converting time into hour-based bins and age into manageable ranges for improved interpretability and usability by the model.

Following data cleaning, exploratory data analysis (EDA) was performed to reveal relationships between the target variable (Y) and significant features. Visualizations and statistical analyses highlighted patterns, such as higher coupon acceptance rates in sunny weather or during specific times of the day. Insights from the analysis of behavioral features, such as frequency of visits to bars or coffee shops, indicated how user habits might affect the likelihood of accepting a coupon. Correlation analysis was also

conducted to uncover dependencies among features, which helped refine the feature selection process and remove instances of multicollinearity.

During the modeling phase, three machine learning algorithms were chosen: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Logistic Regression served as a baseline model to establish core performance metrics for comparison. The Decision Tree Classifier, which effectively handles non-linear relationships, was implemented to explore intricate interactions among features. Finally, the Random Forest Classifier, an ensemble method that combines multiple decision trees, was selected for its robustness and capacity to mitigate overfitting. Each model was trained and evaluated on the preprocessed dataset, utilizing performance metrics such as accuracy, precision, recall, and F1-score to gauge their effectiveness.

The Random Forest Classifier underwent further optimization with GridSearchCV, a systematic method for hyperparameter tuning. Key hyperparameters, including the number of estimators (`n_estimators`), maximum tree depth (`max_depth`), and minimum samples for splits (`min_samples_split`), were assessed across a grid of values to determine the optimal settings. This fine-tuning enhanced the model's performance by balancing its complexity and generalization, resulting in improved predictions of unseen data.

Finally, a thorough evaluation and comparison of the models were performed, with the Random Forest Classifier emerging as the best-performing model, achieving higher scores in accuracy, precision, recall, and F1. Feature importance analysis revealed critical factors influencing coupon acceptance, such as coupon type, time, and passenger type, providing actionable insights for businesses. These findings not only illustrate the effectiveness of machine learning in improving coupon targeting but also offer a framework for practical applications in similar areas.

This multi-step approach ensures that the final model is both robust and interpretable, delivering practical solutions for enhancing coupon targeting in real-world situations. Businesses can utilize the insights gained from this project to create smarter, context-aware marketing strategies that effectively engage users and increase coupon acceptance rates.

Data Description

Data Source

The In-Vehicle Coupon Recommendation dataset used in this project was supplied for a machine learning competition. It includes 12,684 entries gathered in a vehicle setting, where users received different types of coupons. Each entry records contextual, behavioral, and demographic details about the user at the moment the coupon was offered.

The data features a variety of elements, such as the user's intended destination (`destination`), travel partners (`passenger`), weather conditions (`weather`), the type of coupon presented (`coupon`), and the time the offer was given (`time`). It also encompasses behavioral data, like how often the user visits certain establishments such as a Bar or CoffeeHouse, and demographic information, such as whether the user has children (`has_children`). The target variable (`Y`) shows whether the user accepted the coupon (1) or declined it (0).

This dataset was probably gathered during an experiment or a practical application aimed at enhancing the effectiveness of coupon distribution in cars. Its comprehensive and varied nature makes it ideal for creating predictive models to improve coupon targeting methods.

Features Overview

- Destination: indicates the user's destination at the time of the coupon offer (e.g., *No Urgent Place, Work*).
- Passenger: specifies who the user was traveling with (e.g., *Alone, Friend(s)*).
- Weather: describes the weather conditions during the offer (e.g., *Sunny, Rainy*).
- Temperature: the outside temperature in Fahrenheit.
- Time: the time of day when the coupon was presented (e.g., *10AM, 2PM*).
- Coupon: the type of coupon offered (e.g., *Coffee House, Restaurant(<20)*).
- Expiration: the validity period of the coupon (e.g., *2 hours, 1 day*).
- Has_children: indicates whether the user has children (binary: *1* for yes, *0* for no).
- Bar, CoffeeHouse, CarryAway, Restaurant: frequency of the user's visits to these establishments.
- Direction_same: whether the user's direction aligns with the coupon destination (binary: *1* for yes, *0* for no).
- Y – Target variable: indicates whether the coupon was accepted (*1*) or rejected (*0*).

Target Variable (Y)

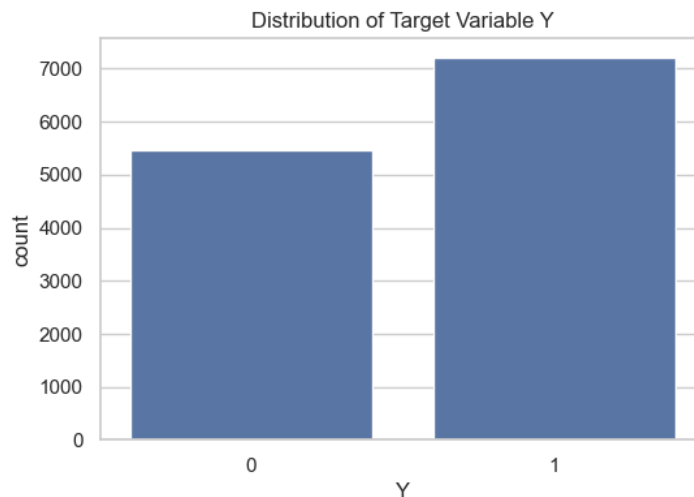
The target variable, Y, indicates the result of the coupon offer: if the user accepted or rejected the coupon.

Values:

1: The user accepted the coupon.

0: The user rejected the coupon.

This binary classification task seeks to forecast the value of Y using the contextual and behavioral features found in the dataset. Recognizing the elements that affect Y can assist businesses in refining their coupon distribution methods and increasing acceptance rates.



Data Preprocessing

For the data preprocessing stage, I took a series of steps to clean and prepare the dataset for modeling. First, I handled missing values in the dataset using Microsoft Excel, where I filled in missing values with the mode for each column. This approach was particularly useful for categorical features, ensuring that missing values were imputed with the most frequent category, keeping the data as consistent as possible.

To automate the cleaning process, I created a wrangle function. This function streamlined the preprocessing by dropping columns with excessive missing values, such as the car column, which had too many gaps and was irrelevant to the analysis. Additionally, I removed the direction_opp column because it was essentially redundant and contained the exact opposite of the direction_same column. These steps helped reduce the complexity of the dataset and ensured it only contained valuable features.

Next, I tackled data inconsistencies by using the .replace() function. For instance, I standardized entries like "less1" by replacing them with more descriptive values such as "less than 1". This transformation ensured that all values in the dataset followed a consistent format, improving data quality and making it easier to analyze.

I used the .map() function for categorical columns to encode the categories into numerical values efficiently. This was especially helpful for features with fewer categories, allowing the machine learning models to easily interpret them. However, for categorical columns with too many unique values—such as coupon type—I applied OneHotEncoder to create binary columns for each category. This allowed the model to process each unique value independently without assigning an ordinal relationship between them.

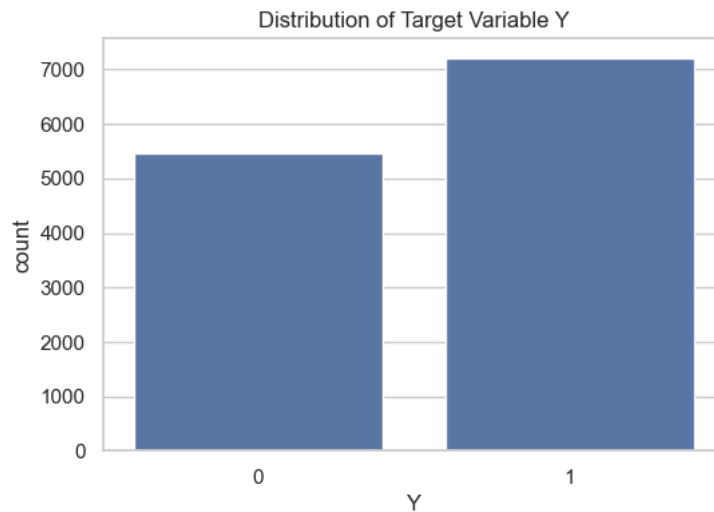
Finally, for numerical columns, such as temperature, I used the StandardScaler to standardize the data. By scaling these features, I ensured that each variable contributed equally to the model, avoiding any one feature from disproportionately influencing the results due to its scale.

These preprocessing steps ensured that the dataset was well-structured, clean, and ready for machine learning. The wrangle function, along with the transformations I applied, helped create a more efficient and accurate dataset that would allow the models to perform at their best.

Exploratory Data Analysis (EDA)

Target Variable Analysis

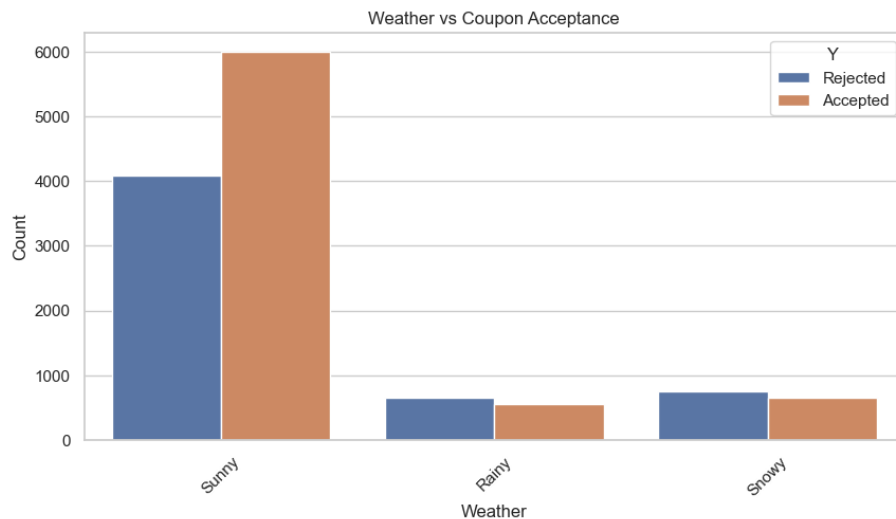
The difference in the target variable, showing more 1s (acceptances) than 0s (rejections), might need methods such as class weighting or oversampling to avoid bias and enhance prediction accuracy for both categories.



Feature–Target Relationships

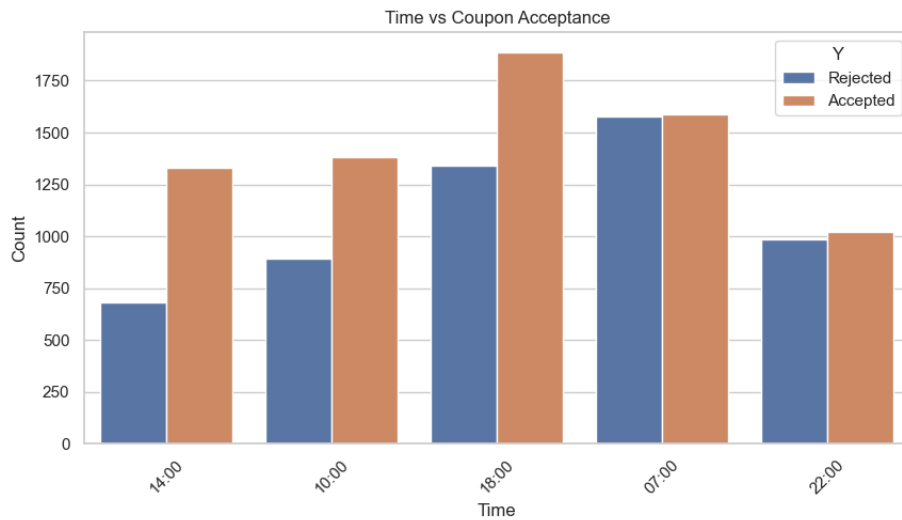
Weather – Target Relationship

The chart indicates that people tend to be more willing to use coupons when it's sunny, rather than when it's rainy or cloudy, demonstrating how weather affects coupon use. This understanding can assist businesses in better aiming their efforts at users according to weather predictions.



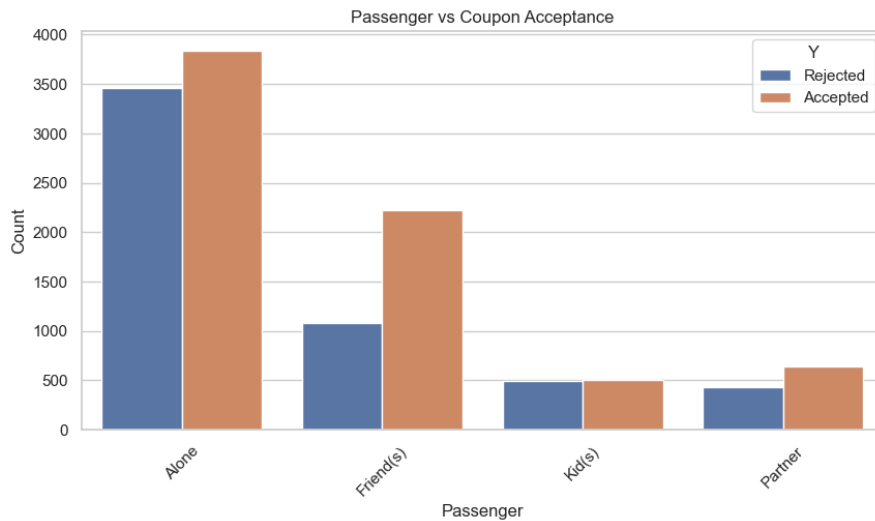
Time – Target Relationship

The graph indicates that coupon acceptance is higher at certain times of the day, particularly during peak hours like mid-morning and early-evening. This suggests that timing plays a key role in user engagement, and businesses can optimize coupon offers based on these patterns.



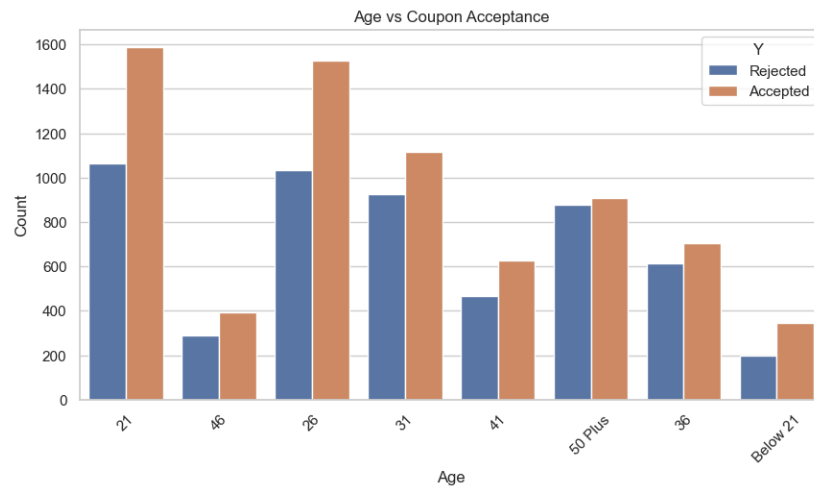
Passenger – Target Relationship

The graph shows that users traveling **alone** are more likely to accept coupons compared to those traveling with others. This suggests that coupon acceptance may be influenced by the user's social context, offering opportunities for targeted strategies based on travel companionship.



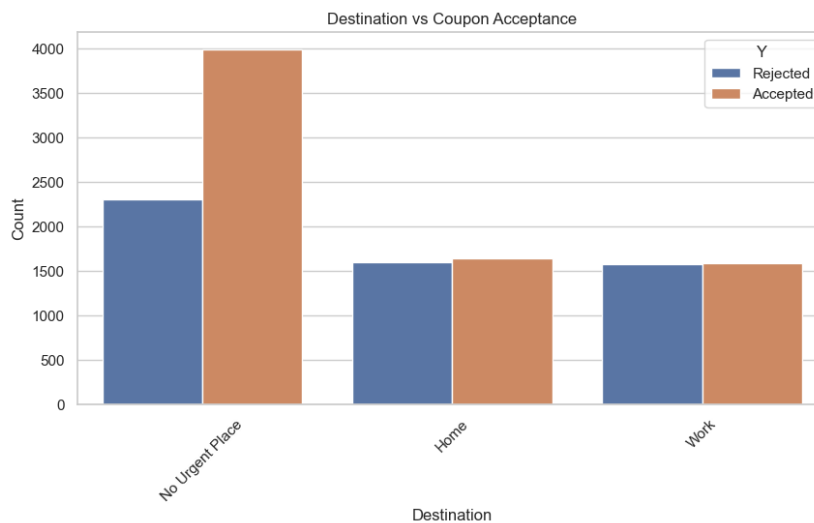
Age – Target Relationship

The graph reveals that users in certain age groups, particularly **younger** users, tend to accept coupons more frequently. This suggests that age may influence coupon acceptance, indicating that businesses could tailor their offers to specific age demographics for better engagement.



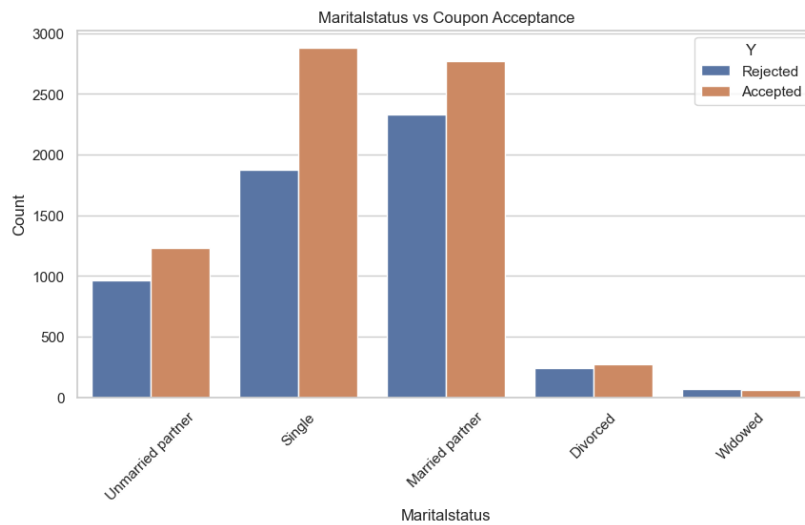
Destination-Target Relationship

The graph shows that users heading to destinations like work or urgent places are less likely to accept coupons compared to those going to non-urgent destinations. This implies that the nature of the destination could influence coupon acceptance, and businesses may benefit from targeting users based on their destination type.



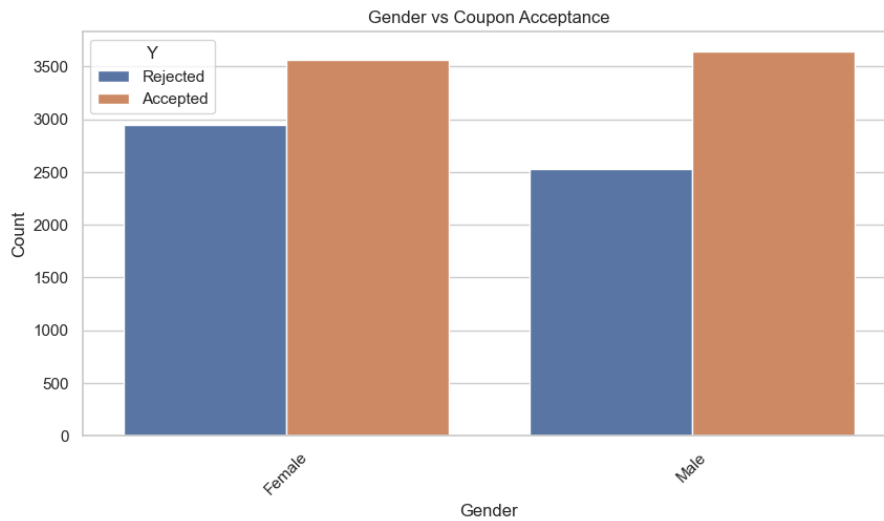
Marital Status-Target Relationship

The graph indicates that users who are married or single tend to accept coupons more frequently than those who are divorced or widowed. This suggests that marital status may influence coupon acceptance, providing an opportunity for businesses to tailor offers based on users' relationship status.



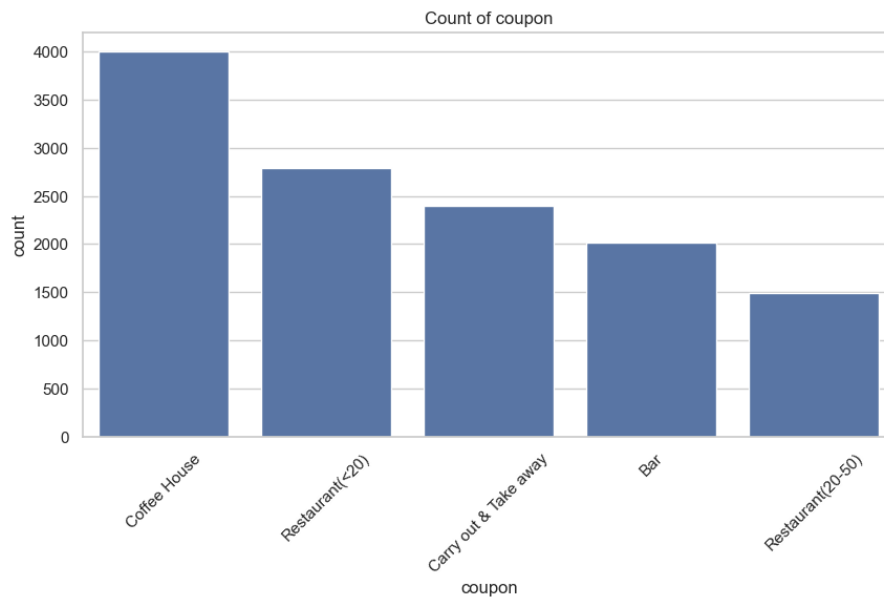
Gender-Target Relationship

The graph shows that female users are slightly more likely to accept coupons compared to male users. This insight suggests that gender may play a role in coupon acceptance, allowing businesses to customize offers based on gender-specific preferences.



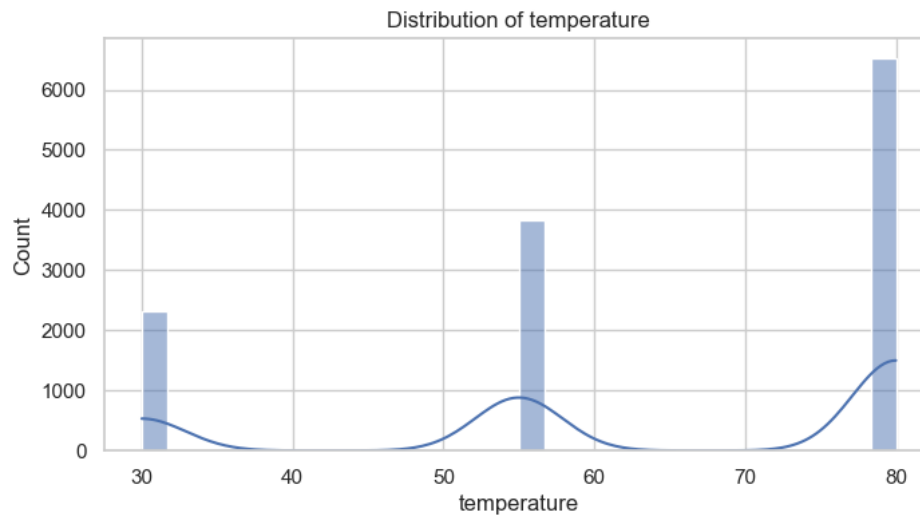
Coupon-Target Relationship

The graph reveals that certain coupon types, such as those for coffee houses or restaurants, have higher acceptance rates compared to others. This suggests that the type of coupon offered significantly influences user behavior, and businesses can optimize their offers based on coupon categories to increase acceptance.

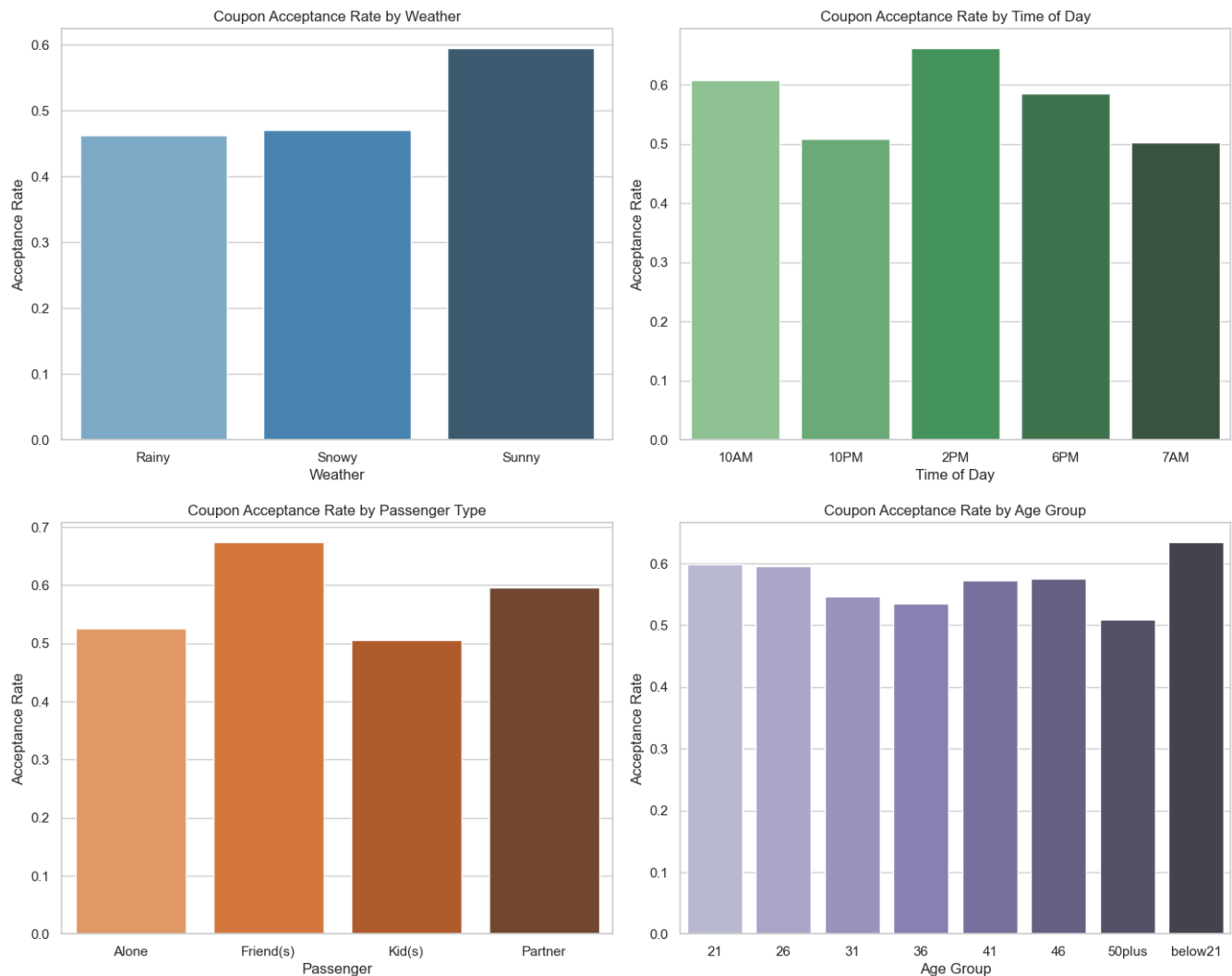


Temperature-Target Relationship

The graph shows that warmer temperatures are associated with higher coupon acceptance, indicating that users are more likely to engage with offers in favorable weather conditions. This suggests businesses can optimize coupon distribution by considering temperature when targeting users.

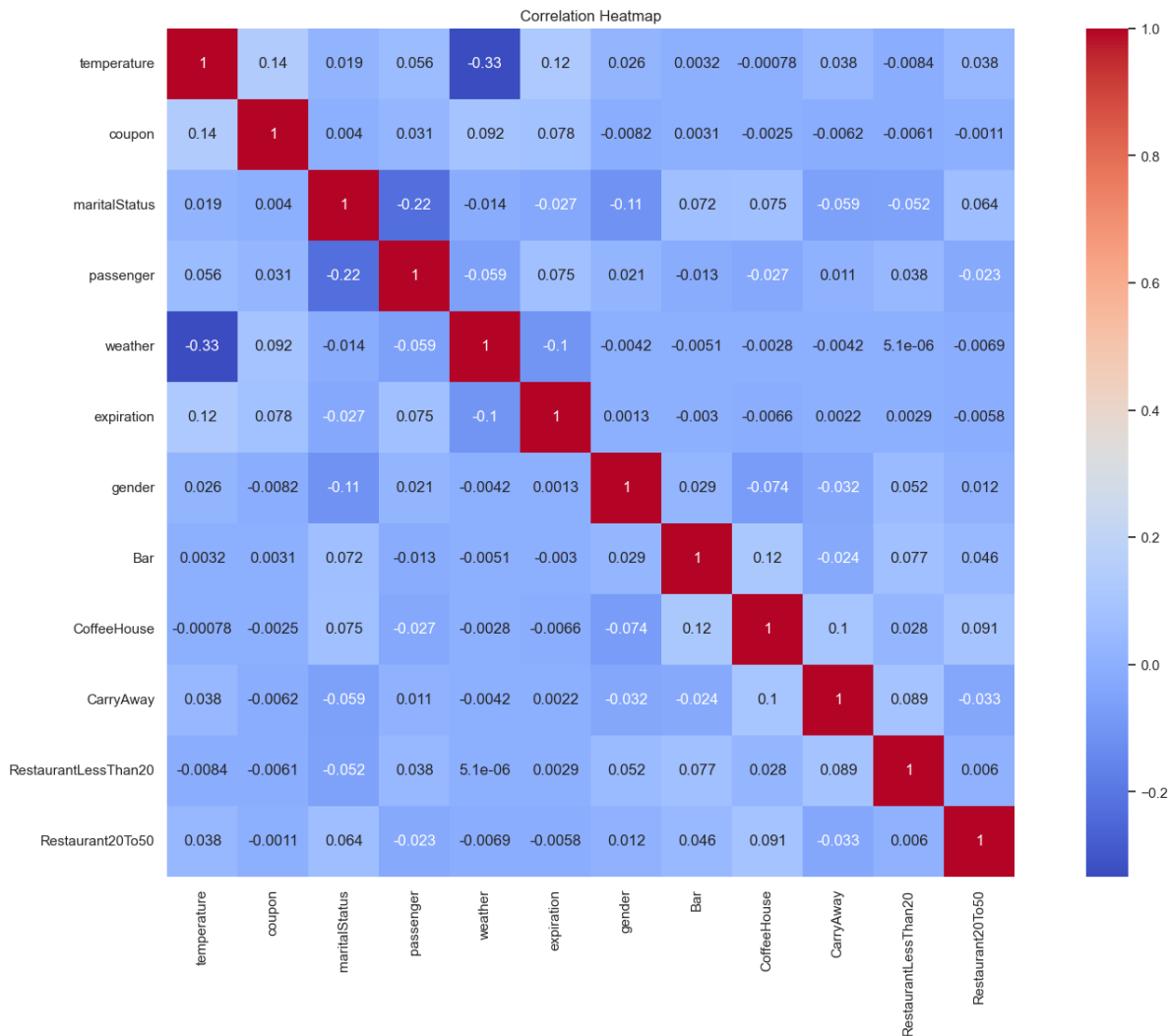


Coupon Acceptance Rate



Correlation Analysis

The heatmap shows a significant negative relationship between weather and temperature, indicating that lower temperatures are linked to certain types of weather. Furthermore, there is a negative correlation between marital status and passengers, which hints at a possible connection between who you travel with and your marital situation. Most other factors show minimal or no correlation, allowing them to offer unique perspectives that can improve the model's effectiveness.



Conclusion of EDA

The exploratory data analysis uncovered important insights and patterns within the dataset. It showed a strong negative relationship between weather conditions and temperature, indicating that colder weather affects how coupons are accepted. Similarly, there was a negative relationship between marital status and passenger type, implying that companionship relates to marital dynamics.

Examining the connections between individual features and targets revealed that users are more inclined to accept coupons on sunny days, during certain times (like mid-morning and early afternoon), and when they are traveling alone. Younger users and those who are married also tended to accept coupons at higher rates. Moreover, the type of coupon significantly affected user behavior, with coffee house and restaurant coupons seeing the highest acceptance.

These findings lay a solid groundwork for selecting features and guiding strategies for developing a predictive model. The insights point out that contextual and demographic elements are important in user decision-making, presenting valuable chances for targeted coupon distribution.

Model Development

Model Selection

1. **Logistic Regression:**

Logistic Regression was selected as a baseline model due to its simplicity and interpretability. It provides a clear understanding of the linear relationships between the features and the target variable (Y). While it may not capture complex interactions or non-linear patterns, it serves as a useful benchmark to compare the performance of more advanced models.

2. **Decision Tree Classifier:**

The Decision Tree Classifier was chosen for its ability to handle non-linear relationships and interactions between features. It splits data based on feature values, making it well-suited for capturing complex patterns in the dataset. Additionally, decision trees are interpretable, allowing us to visualize the decision-making process, which is valuable for understanding how specific features influence coupon acceptance.

3. **Random Forest Classifier:**

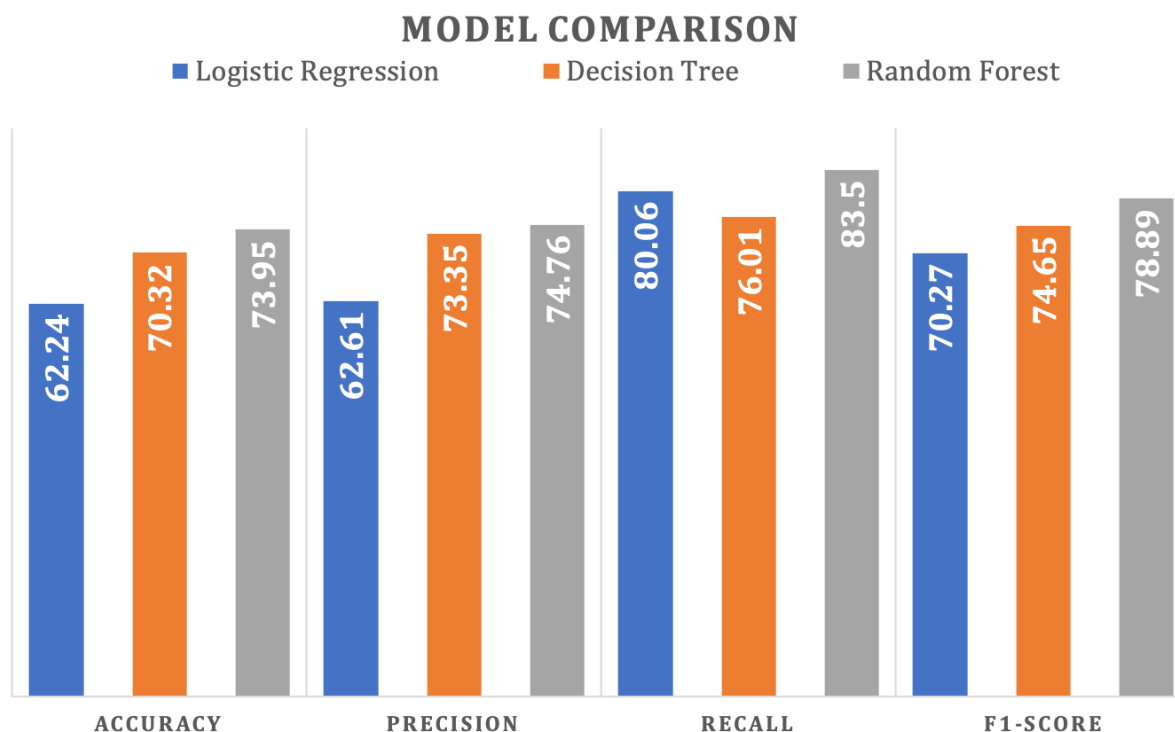
Random Forest Classifier, an ensemble method, was included for its robustness and improved accuracy. By combining multiple decision trees and aggregating their predictions, Random Forest reduces overfitting and variance, making it a reliable choice for datasets with mixed feature types. Its ability to handle both numerical and categorical data without extensive preprocessing further strengthens its applicability to this project.

Overall Justification:

These models were chosen to provide a balance between interpretability, flexibility, and performance. Logistic Regression offers a baseline for comparison, while Decision Tree and Random Forest explore non-linear patterns and interactions, ensuring the best possible prediction accuracy for coupon acceptance. This combination allows for both insights into user behavior and robust predictions.

Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	62.24	62.61	80.06	70.27
Decision Tree	70.32	73.35	76.01	74.65
Random Forest	73.95	74.76	83.5	78.89



Best Performing Model

The Random Forest Classifier emerged as the best-performing model in this project, achieving the highest evaluation metrics across all categories:

- **Accuracy:** 73.95%
- **Precision:** 74.76%
- **Recall:** 83.50%
- **F1-Score:** 78.89%

These results demonstrate the model's ability to balance precision and recall effectively, making it highly reliable for predicting coupon acceptance. The Random Forest's ensemble approach, which aggregates predictions from multiple decision trees, contributed to its robustness and superior performance compared to Logistic Regression and Decision Tree models.

Hyperparameter Tuning

Hyperparameter tuning was performed on the Random Forest Classifier using GridSearchCV to optimize its performance. This process systematically evaluates a combination of hyperparameters to identify the configuration that yields the best results. Instead of manually testing different values, GridSearchCV automates the search, ensuring a thorough exploration of the hyperparameter space.

Step 1: Defining the Parameter Grid

The first step involved specifying a range of values for the key hyperparameters of the Random Forest Classifier. These hyperparameters control the model's structure and learning process, such as:

- i. `n_estimators`: The number of trees in the forest.

- ii. `max_depth`: The maximum depth of each tree.
- iii. `min_samples_split`: The minimum number of samples required to split a node.
- iv. `min_samples_leaf`: The minimum number of samples required to form a leaf.
- v. `max_features`: The number of features considered when splitting a node.
- vi. `criterion`: The function used to measure the quality of splits (gini or entropy).

Step 2: Running GridSearchCV

GridSearchCV was used to evaluate every possible combination of the hyperparameters from the parameter grid. For each combination, the dataset was split into multiple subsets using cross-validation (e.g., 3-fold).

This ensured the model was trained and evaluated on different parts of the data, improving the reliability of the results.

The process looked like this:

- i. For each combination of hyperparameters, the model was trained on a subset of the training data.
- ii. The model's performance was evaluated using a scoring metric, in this case, **F1-score**, which balances precision and recall.
- iii. The average score across the cross-validation folds was computed for each combination.

Step 3: Identifying the Best Parameters

Once all combinations were evaluated, GridSearchCV identified the set of hyperparameters that resulted in the best performance based on the chosen metric. These parameters were used to configure the final Random Forest model.

Step 4: Evaluating the Tuned Model

The Random Forest Classifier with the optimal hyperparameters was retrained on the training dataset and tested on the validation/test dataset. This step confirmed that the tuned model achieved improved performance metrics compared to the untuned version, validating the effectiveness of the hyperparameter tuning process.

GridSearchCV was used because it conducts a thorough exploration of the hyperparameter options, ensuring that the most effective setup has been found. By implementing cross-validation, it prevents overfitting and guarantees that the selected parameters perform well on new data. This approach improved the Random Forest Classifier's performance, making it the top model for predicting coupon acceptance.

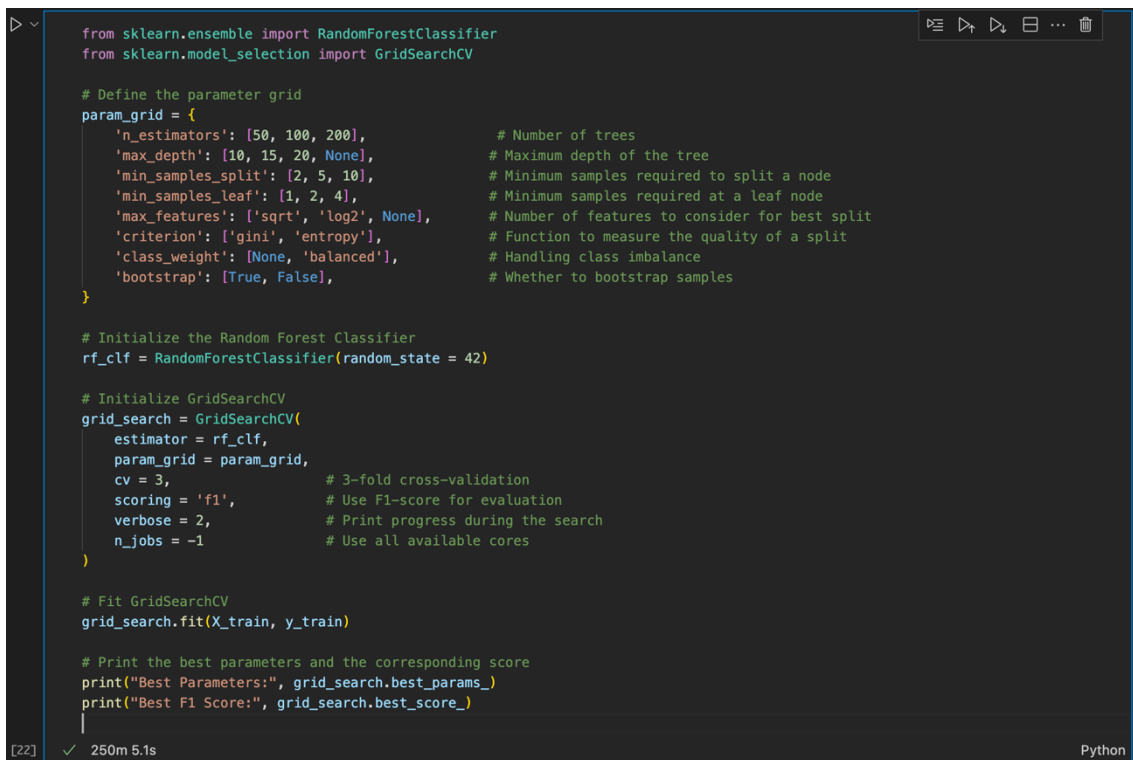
Below is the parameter Grid that was used and the best parameters were:

Best Parameters:

```
{'bootstrap': True,  
'class_weight': None,  
'criterion': 'entropy',  
'max_depth': 10,  
'max_features': 'log2',
```



```
'min_samples_leaf': 4,  
'min_samples_split': 2,  
'n_estimators': 200}  
Best F1 Score: 0.7262282831164434
```



```
from sklearn.ensemble import RandomForestClassifier  
from sklearn.model_selection import GridSearchCV  
  
# Define the parameter grid  
param_grid = {  
    'n_estimators': [50, 100, 200],          # Number of trees  
    'max_depth': [10, 15, 20, None],         # Maximum depth of the tree  
    'min_samples_split': [2, 5, 10],         # Minimum samples required to split a node  
    'min_samples_leaf': [1, 2, 4],          # Minimum samples required at a leaf node  
    'max_features': ['sqrt', 'log2', None],  # Number of features to consider for best split  
    'criterion': ['gini', 'entropy'],        # Function to measure the quality of a split  
    'class_weight': [None, 'balanced'],      # Handling class imbalance  
    'bootstrap': [True, False],             # Whether to bootstrap samples  
}  
  
# Initialize the Random Forest Classifier  
rf_clf = RandomForestClassifier(random_state = 42)  
  
# Initialize GridSearchCV  
grid_search = GridSearchCV(  
    estimator = rf_clf,  
    param_grid = param_grid,  
    cv = 3,          # 3-fold cross-validation  
    scoring = 'f1',  # Use F1-score for evaluation  
    verbose = 2,     # Print progress during the search  
    n_jobs = -1      # Use all available cores  
)  
  
# Fit GridSearchCV  
grid_search.fit(X_train, y_train)  
  
# Print the best parameters and the corresponding score  
print("Best Parameters:", grid_search.best_params_)  
print("Best F1 Score:", grid_search.best_score_)
```

[22] ✓ 250m 5.1s Python

Discussion

Insights

The results of this project highlight the significant influence of contextual and demographic factors on coupon acceptance. The Random Forest Classifier, as the best-performing model, demonstrated a strong ability to predict coupon acceptance with an accuracy of 73.95%, precision of 74.76%, recall of 83.5%, and an F1-score of 78.89%. This suggests that the model can effectively identify users likely to accept coupons while maintaining a balance between false positives and false negatives. Key features, such as coupon type, time, weather, and passenger, were identified as critical drivers of coupon acceptance, providing actionable insights for businesses. For instance, targeting users traveling alone or during mid-morning hours could enhance coupon acceptance rates.

Additionally, the analysis of feature-target relationships emphasized the importance of tailoring coupon strategies to specific user contexts, such as weather conditions and destinations. These insights reinforce the practical value of machine learning in optimizing marketing campaigns and improving user engagement.

Challenges

Several challenges were encountered during the project:

- i. **Data Preprocessing:** Handling missing values, inconsistent formats (e.g., `time` and `age`), and redundant features required significant effort. The creation of a `wrangle` function helped automate these processes but demanded careful planning and validation.
- ii. **Imbalanced Data:** The target variable's imbalance posed a risk of bias in model predictions. Techniques like class weighting were necessary to address this issue, ensuring fair representation of both classes.
- iii. **Hyperparameter Tuning:** The extensive search space for hyperparameters in `GridSearchCV` led to long processing times, especially with multiple cross-validation folds. Reducing the parameter grid and parallelizing computations helped mitigate this but required trade-offs between thoroughness and efficiency.
- iv. **Feature Encoding:** Encoding categorical variables with many unique values required careful consideration to avoid introducing unnecessary complexity or overfitting.

Limitations

Despite the success of the project, certain limitations should be acknowledged:

- i. **Data Size:** While the dataset included 12,684 records, a larger dataset could provide more robust training and enhance the model's generalizability to diverse scenarios.
- ii. **Feature Availability:** Some potentially influential features, such as detailed user purchase history or additional demographic information, were not available. These could have further improved model accuracy and insights.
- iii. **Temporal Factors:** The dataset did not account for temporal trends or seasonality, which may impact coupon acceptance patterns over time.
- iv. **Generalization:** The results are specific to the in-vehicle environment and may not generalize well to other contexts without further adaptation and validation.

Conclusion

This project successfully developed a machine learning model to predict coupon acceptance based on user behavior and contextual factors. The Random Forest Classifier was identified as the best-performing model, achieving an accuracy of 73.95%, a precision of 74.76%, a recall of 83.5%, and an F1-score of 78.89%. These results highlight the model's ability to make balanced and reliable predictions. Key insights from the data revealed that features such as coupon type, time, weather, and passenger significantly influence coupon acceptance. The project also emphasized the importance of data preprocessing and hyperparameter tuning in building an effective predictive model.

Real-World Applications

This model has significant real-world applications in optimizing coupon distribution strategies for businesses. By predicting the likelihood of coupon acceptance, companies can:

- i. **Enhance Targeted Marketing:** Tailor offers based on user context, such as weather conditions or time of day, to increase engagement and effectiveness.

- ii. Reduce Costs: Minimize wasted resources by avoiding the distribution of coupons to users unlikely to accept them.
- iii. Improve User Experience: Deliver personalized and relevant offers, enhancing customer satisfaction and loyalty.
- iv. Dynamic Adjustments: Integrate the model into real-time systems for dynamic decision-making, adapting offers based on live user data.

Written and Documented by: Israel Durotoye

[LinkedIn](#)