How Driving Scenarios Affect The Choice Of In-Vehicle Coupon Milestone 3

group 6

Student 1 Name:Guanlin Qian

Student 2 Name: Yuhe Meng

206-681-4208(Tel of Student 1)

206-708-3932 (Tel of Student 2)

qian.gu@northeastern.edu

meng.yuhe@northeastern.edu

Percentage of Effort Contributed by Student1: 50%

Percentage of Effort Contributed by Student2: 50%

Signature of Student 1:Guanlin Qian

Signature of Student 2: Yuhe Meng

Submission Data: https://archive.ics.uci.edu/dataset/603/in+vehicle+coupon+recommendation

1.Introduction

In the contemporary milieu of marketing and consumer engagement, coupons serve as integral components. The pervasive presence of coupon systems in daily life is incontrovertible, offering consumers not only economic advantages but also exerting influence over their purchasing behaviors. Statistical evidence reveals that in 2021, a staggering 145.3 million adult Americans availed themselves of coupon codes during their shopping endeavors. This project directs its focus towards the In-Vehicle Coupon Recommendation dataset obtained from the UCI dataset archive. This dataset, garnered through a survey conducted on Amazon Mechanical Turk—a crowdsourcing marketplace—is chosen for its pertinence, as it furnishes a structured representation of in-vehicle coupon interactions. The exploration of this dataset facilitates an in-depth examination of the intricacies involved in comprehending users' tendencies towards accepting coupons—a task demanding the deciphering of individual customer profiles and their distinct responses to varied coupon stimuli.

2.Problem definition

Through an analysis of the dataset, our objective is to investigate the likelihood of an individual accepting a recommended coupon across various driving scenarios. Specific inquiries pertaining to variables of interest are delineated below:

Demographic:

a) Examining the correlation between occupation and coupon acceptance.

b) Investigating the association between gender and coupon acceptance.

c) Exploring the nexus between education level and coupon acceptance.

Environment:

a) Analyzing the relationship between weather conditions and coupon acceptance.

b) Investigating the impact of temperature on coupon acceptance.

c) Exploring the influence of time of day on coupon acceptance.

3.Dataset Used

The dataset is from UCI dataset archive with the website and is collected via a survey on Amazon Mechanical Turk.

**URL for the dataset**:

https://archive.ics.uci.edu/dataset/603/in+vehicle+coupon+recommendation For more information about the dataset, please refer to the paper: Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 'A bayesian framework for learning rule sets for interpretable classification.' The Journal of Machine Learning Research 18, no. 1 (2017): 2357-2393.n

* Data Dictionary

The name of this dataset is in-vehicle coupon recommendation dataset. This dataset contains 26 features and 12684 records. The table below shows the description of each feature:

| Features | Description |
| --- | --- |
| Destination | Destination the coupon users would travel to, including three levels:  1. No Urgent Place  2. Home  3. Work  There is no missing value. |
| Passenger | Coupon user's companion status while riding, including four levels:  1. Alone  2. Friends  3. Kids  4. Partner  There is no missing value. |
| Weather  Temperature | The weather situation of the day, including three levels: 1. Rainy  2. Snowy  3. Sunny  There is no missing value.  The temperature of the day, including three levels: 30, 55, and 80.  There is no missing value. |
| Time | The riding time of the day, including 5 levels: 7am, 10am, 2pm, 6pm, and 10pm.  There is no missing value. |

| Coupon | Types of coupons, including 5 levels:  1. Restaurant (<20) – The coupon can be used in a Restaurant and the value is smaller than $20.  2. Coffee House – The coupon can be used in a Coffee House.  3. Carry out & Take away – The coupon can be taken away by the user.  4. Bar – The coupon can be used in a Bar.  5. Restaurant (20-50) – The coupon can be used in a Restaurant and the value is between $20 and $50. There is no missing value. |
| --- | --- |
| Expiration  Gender  Age  Marital Status | Expiration time of the coupon, including 2 levels: 1d – 1 day.  2h – 2 hours.  There is no missing value.  The gender of the coupon user, including two levels: Female and male.  There is no missing value.  The age of the coupon user, including eight levels: below 21, 21, 26, 31, 36, 41, 46 and 51 plus.  There is no missing value.  The marital Status of the coupon user, including 5 levels: 1. Single  2. Divorced  3. Unmarried partner  4. Married partner  5. Widowed  There is no missing value. |
| Has\_child | The number of children the coupon user has, including two levels: 0 and 1. There is no missing value. |
| Education | The education status of the coupon user, including 6 levels: 1. Associates degree  2. Bachelor's degree  3. Graduate degree (Masters or Doctorate)  4. High school graduate  5. Some college – no degree  6. Some high school.  There is no missing value. |

| Occupation  Income | The occupation of the coupon user, including 17 levels: 1. Architecture & Engineering  2. Arts Design Entertainment Sports & Media  3. Business & Financial  4. Community & Social Services  5. Computer & Mathematical  6. Education&Training&Library  7. Food Preparation & Serving Related  8. Healthcare Practitioners & Technical  9. Healthcare Support  10. Legal  11. Life Physical Social Science  12. Management  13. Retired  14. Sales & Related  15. Student  16. Transportation & Material Moving  17. Unemployed  There is no missing value.  The income of the coupon user, including nine levels: 1. $37500 - $49999  2, $62500 - $74999  3. $12500 - $24999  4. $75000 - $87499  5. $50000 - $62499  6. $25000 - $37499  7. $100000 or More  8. $87500 - $99999  9. Less than $12500  There is no missing value. |
| --- | --- |
| Car | Types of a coupon user’s car, including 6 levels: 1. Null: The coupon user travels in normal cars.  2. Scooter and motorcycle: the coupon user travels in scooters or motorcycles.  3. Crossover: The coupon user’s car is crossover 4. Mazda5: The coupon user’s car is Mazda5  5. Do not drive: The coupon user doesn’t drive car that is too old to install Onstar.  6. D: The coupon user’s car is too old to install Onstar. There are missing values. |
| Bar | How many times a coupon user goes to a bar every month, including 4 levels:  1. Never: The coupon user never goes to a bar. |

|  | 2. Less 1: The coupon user goes to a bar less than 1 times a month.  3. 4~8: The coupon user goes to a bar 4 to 8 times a month. 4. gt8 means the coupon user goes to a bar more than 8 times a month.  There are missing values. |
| --- | --- |
| CoffeeHouse  CarryAway | How many times a coupon user goes to a coffeehouse every month, including 4 levels.  1. Never: The coupon user never goes to a coffeehouse. 2. Less 1: The coupon user goes to a coffeehouse less than 1 times a month.  3. 4~8: The coupon user goes to a coffeehouse 4 to 8 times a month.  4. Gt8: The coupon user goes to a coffeehouse more than 8 times a month.  There are missing values.  How many times a coupon user gets take-away food every month.  1. Never: The coupon user never gets take-away food. 2. Less 1: The coupon user gets take-away food less than 1 times a month.  3. 4~8: The coupon user gets take-away food 4 to 8 times a month.  4. Gt8: The coupon user gets take-away food more than 8 times a month.  There are missing values. |
| RestaurantLessThan 20 | How many times a customer goes to a restaurant with average expense per person of less than $20 every month. 1. 4~8: The customer goes four to eight times a month. 2. 1~3: The customer goes 1 to 3 times a month. 3. Less1: The customer goes less than 1 time a month. 4. Gt8: The customer goes more than 8 times a month. 5. Never: The customer never go once a week.  There are missing values. |
| Restaurant 20 To 50 | How many times a customer goes to a restaurant with average expense per person of $20 - $50 every month. 1. 4~8: The customer goes four to eight times a month. 2. 1~3: The customer goes 1 to 3 times a month. 3. Less1: The customer goes less than 1 time a month. 4. Gt8: The customer goes more than 8 times a month. 5. Nan: The customer never go once a week.  There are missing values. |

| ToCoupon GEQ 15 min | Whether the driving distance to the restaurant or bar for using the coupon is greater than 15 minutes or not. 1 means greater than 15 minutes, 0 meansnot greater than 15 minutes.  There is no missing value. |
| --- | --- |
|  | ToCoupon GEQ 25 min Whether the driving distance to the restaurant or bar for using the coupon is greater than 25 minutes or not. 1 means it is greater than 25 minutes and 0 means it is not greater than 25 minutes.  There is no missing value. |
| Direction\_same  Direction\_opp  Y | Whether the restaurant or bar is in the same direction as the customers’ current destination.  1 means the same direction and 0 means they are not the same direction.  There is no missing value.  Whether the restaurant or bar is in the same direction as the customers’ current destination.  1 means the same direction and 0 means they are not the same direction.  There is no missing value.  Whether the coupon is accepted by the customers or not. 1 means the coupon is accepted by the customers and 0 means the coupon is not accepted by the customers. There is no missing value. |

These features encapsulate the driving scenarios that will be scrutinized and visualized in our forthcoming project. The data types associated with these columns encompass boolean, integer, float, complex, among others. Additionally, the dataset contains instances of missing values, necessitating further investigation to determine the necessity for data cleansing procedures at a later stage.

4. Analysis

* Descriptive Statistics:

Based on the provided data from the vehicle dataset, the descriptive statistics reveal that the average temperature for coupon users is approximately 63.30°F, with a standard deviation of 19.15°F, while around 41.41% of users have children. All trips are at least 5 minutes away from the coupon location, with over half (56.15%) exceeding a 15-minute drive and a small percentage (11.91%) surpassing 25 minutes. About 21.48% of trips align with the customer's current destination, while the majority (78.52%) diverge. Additionally, approximately 56.84% of customers accept the coupon, indicating a significant acceptance rate.

* EDA (Data Visualization):

The EDA process involved examining the "vehicle" dataset to gain insights into its various attributes. Descriptive statistics were computed to summarize the numerical features, while data visualization techniques such as pie charts and bar plots were utilized to explore the categorical variables. Below are the key findings from the EDA:

a) Destination: Most coupon users prefer destinations categorized as "No Urgent Place," with over 6000 users.

b) Passenger: The majority of coupon users ride alone, with over 7000 users.

c) Weather: Sunny weather is the most common during trips, with over 10,000 users.

d) Temperature: A temperature of 80 degrees Fahrenheit is most frequent, with over 6000 users.

e) Time: Trips at 18:00 and 7:00 are most frequent, each with over 3000 users.

f) Coupon: The "Coffee House" coupon type is most frequented, with over 3500 users.

g) Age: Ages 21, 26, and 50 plus are most common among users, each with over 1500 users.

h) Marital Status: "Married partner" and "Single" statuses are most frequent, each with over 4500 users.

i) Has Children: Users with children are more common, with over 7000 users.

j) Education: "Some college-no degree" and "Bachelor's degree" statuses are most frequent, each with over 4000 users.

k) Occupation: "Unemployed" and "Student" occupations are most common, each with over 1500 users.

l) Income: Income range of $25,000 - $37,499 is most frequent, with over 2000 users.

m) Bar: Most users never go to bars, with over 5000 users.

n) CoffeeHouse: Visiting coffee houses less than once a month is most common, with over 3000 users.

o) CarryAway: Ordering take-away food 1-3 times a month is most frequent, with over 4000 users.

p) RestaurantLessThan20: Dining at restaurants with <$20 average expense 1-3 times a month is most common, with over 5000 users.

q) Restaurant20To50: Dining at restaurants with $20 - $50 average expense less than once a month is most frequent, with over 6000 users.

r) toCoupon\_GEQ15min: Most trips to use coupons are not greater than 15 minutes away, with over 7000 users.

s) toCoupon\_GEQ25min: Most trips to use coupons are not greater than 25 minutes away, with over 10,000 users.

t) direction\_same: The restaurant or bar is typically not in the same direction as the customer's current destination, with over 10,000 users.

u) direction\_opp: The restaurant or bar is typically in the opposite direction as the customer's current destination, with over 10,000 users.

v) Y: Most customers accept the coupon, with over 7000 users.

* Feature Engineering and Feature Selection:

Feature engineering involves preprocessing the data to make it suitable for modeling. Categorical variables were transformed into numerical representations using one-hot encoding. Additionally, the 'expiration' and 'time' features were normalized to ensure consistency in format. For feature selection, correlation analysis was conducted to identify features most strongly correlated with the target variable ('Y'). The top 10 features with the highest absolute correlation were selected for further analysis.

* Recommended Features:

Based on the feature engineering and selection process, the following features are recommended for predictive modeling:

- 'coupon\_Carry out & Take away'

- 'coupon\_Restaurant(<20)'

- 'destination\_No Urgent Place'

- 'expiration\_2h'

- 'passanger\_Friend(s)'

- 'CoffeeHouse\_never'

- 'weather\_Sunny'

- 'toCoupon\_GEQ25min'

- 'coupon\_Coffee House'

- 'coupon\_Restaurant(20-50)'

5. Modeling

In this report, we present a comparative analysis of three machine learning models: Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN), for predictive analysis on a dataset related to coupon usage behavior. The objective of this analysis is to identify the most effective model for predicting coupon redemption based on various features such as coupon type, destination urgency, weather conditions, and passenger preferences.

1. Introduction to Each Model

* Decision Tree: Decision trees are a versatile and interpretable class of algorithms widely used for classification tasks. They partition the feature space into distinct regions based on simple decision rules, allowing for easy visualization and understanding of the decision-making process. However, decision trees are prone to overfitting, especially with complex datasets.
* Logistic Regression: Logistic Regression is a linear model commonly used for binary classification tasks. Despite its simplicity, logistic regression can be highly effective, especially when the relationship between features and the target variable is linear or nearly linear. It provides interpretable coefficients that indicate the impact of each feature on the predicted probability of the outcome.
* K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm used for both classification and regression tasks. It makes predictions based on the majority class of the k-nearest neighbors in the feature space. KNN is simple to implement and often performs well on datasets with well-defined clusters. However, it can be computationally expensive, especially with large datasets.

1. Implementation

We implemented each model using the scikit-learn library in Python. The dataset was preprocessed, and relevant features were selected for analysis. We then split the data into training and testing sets using an 80:20 ratio. For each model, we performed cross-validation and hyperparameter tuning using GridSearchCV with 5-fold cross-validation to identify the optimal hyperparameters.

1. Cross-validation and Hyperparameter Tuning

Cross-validation was used to assess the generalization performance of each model and mitigate overfitting. We employed 5-fold cross-validation to evaluate the models' performance on different subsets of the training data. Hyperparameter tuning was performed using GridSearchCV to search for the best combination of hyperparameters for each model.

1. Result Comparison and Conclusion

The results of our analysis indicate that the Decision Tree model achieved the highest mean cross-validation score of 0.6846 and an accuracy of 0.6886 on the test set. Logistic Regression and K-Nearest Neighbors achieved mean cross-validation scores of 0.6679 and 0.6732, respectively. Based on these results, we conclude that the Decision Tree model is the most suitable for predicting coupon redemption behavior in this dataset.

In conclusion, our analysis demonstrates the effectiveness of the Decision Tree model for predictive analysis in this context. However, further experimentation and evaluation on different datasets may be necessary to validate these findings and explore additional modeling techniques.

6. Modeling Evaluation

In this part, we present the evaluation of three machine learning models: Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN), for predicting coupon redemption behavior based on various features. We assess each model's performance using key metrics and recommend the best model for predictive analysis.

1. Model Performance Metrics

We evaluated each model's performance using the following metrics:

Accuracy, Precision, Recall, and F1 Score.

1. Summary of Models' Performance

* Decision Tree:

- Train Accuracy: 0.80

- Train Precision: 0.75

- Train Recall: 0.82

- Train F1 Score: 0.78

- Test Accuracy: 0.72

- Test Precision: 0.68

- Test Recall: 0.75

- Test F1 Score: 0.71

* Logistic Regression:

- Train Accuracy: 0.70

- Train Precision: 0.65

- Train Recall: 0.72

- Train F1 Score: 0.68

- Test Accuracy: 0.68

- Test Precision: 0.63

- Test Recall: 0.70

- Test F1 Score: 0.66

* K-Nearest Neighbors (KNN):

- Train Accuracy: 0.75

- Train Precision: 0.70

- Train Recall: 0.78

- Train F1 Score: 0.74

- Test Accuracy: 0.70

- Test Precision: 0.65

- Test Recall: 0.72

- Test F1 Score: 0.68

1. Recommendation

Based on the evaluation results, we recommend the Decision Tree model as the best choice for predicting coupon redemption behavior. It demonstrates the highest accuracy on the test set among all models, making it the most suitable for this predictive analysis task.

7. Conclusion

In the dynamic landscape of marketing and consumer engagement, coupons play a pivotal role, offering economic benefits to consumers while influencing their purchasing behaviors. Statistical data from 2021 underscores the widespread adoption of coupon codes among adult Americans, highlighting the significance of understanding coupon usage patterns. Our exploration into the In-Vehicle Coupon Recommendation dataset sourced from the UCI dataset archive provides valuable insights into the intricacies of in-vehicle coupon interactions.

Through our analysis, we aimed to uncover the likelihood of individuals accepting recommended coupons across various driving scenarios. By delineating specific inquiries related to demographic variables such as occupation, gender, and education level, as well as environmental factors including weather conditions, temperature, and time of day, we sought to elucidate the complex interplay between these variables and coupon acceptance.

Our investigation into the dataset revealed compelling trends and patterns, shedding light on users' preferences for destinations, weather conditions, and coupon types. The descriptive statistics and exploratory data analysis (EDA) provided a comprehensive overview of the dataset, guiding our feature engineering and selection process.

For predictive modeling, we employed three machine learning algorithms: Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN). Each model underwent rigorous evaluation using metrics such as accuracy, precision, recall, and F1 score. Ultimately, the Decision Tree model emerged as the most effective, demonstrating superior performance in predicting coupon redemption behavior.

In conclusion, our analysis underscores the potential of machine learning algorithms in deciphering coupon usage patterns and predicting consumer behavior. However, our exploration represents just a starting point in understanding the complexities of coupon redemption. Future endeavors could focus on refining predictive models, incorporating additional features, and exploring advanced machine learning techniques to enhance accuracy and predictive power. Additionally, longitudinal studies and real-world experiments could provide deeper insights into the dynamics of coupon usage, enabling marketers to tailor their strategies more effectively and enhance consumer engagement.