St. Francis Xavier University
CSCI 528: Advanced Data Analytics - Project Report
Rahul Dev Bodiga (202207298), Meghana Damarakonda (202207299)

In-Vehicle Coupon Recommendation

Abstract

In this project, we have chosen an in-vehicle coupon recommendation dataset from UCI Repository which is a binary classification task. The major focus of the work we made is on analysis and visualization of the dataset we chose. While performing data visualization we found a lot of meaningful insights and most of those insights have been described in this report. We have analyzed the features that are impacting to accept the coupon when it is recommended to a user and found the order of importance of each feature on which a business party should focus on while targeting the customers so that they can be profited.

1. Introduction

We use coupons more frequently and these coupons provide a win-win situation to both the customers and the business parties. To gain profits from these coupons any business should not target every customer with every coupon, instead they need to follow better strategies to offer coupons to right customers. Hence, we have to create better recommendation systems using the data present with us and by extracting meaningful insights from the data.

The manufacturers who produce cars are tied up with local businesses, to profit their business by recommending coupons to customers who purchase the cars from various manufacturers. These recommendation systems are being embedded into the car such that customers will be offered various kinds of coupons.

2. Data Analysis and Visualization

Our dataset consists of 12684 samples and 26 features in which one of the features is a binary target class with values 0 and 1 where 1 represents that survey respondents have accepted the coupon and 0 represents that they have rejected the coupon.

First, we tried to analyse what are the type of attributes present in our dataset where we found that there were around 70% percent of categorical features and the rest of them were numerical including the target. We made a visual analysis to see if there is any missing data present in our dataset and found that car feature was having around 99% of missing values and features such as Bar, Coffee House, Carry Away, ResturantLessThan20, Resturant20to50 were having values missing between 1-2 percent. As the feature car was not very informative, we excluded that feature for our further analysis. Instead of imputing the missing values in rest of the features, we dropped those samples such that it will not bias our inferences.

Now we checked all the number of unique values in all features and found that there were 25 different occupation values which was making the data points very sparse. Based on the type of values present it was clear that some features were user specific such as income of the customer, his age and marital status, etc. Some other features were context-based such as what was customers destination, temperature and weather while travelling, etc. Two other attributes were coupon specific, such as type of coupon, and expiration time of the coupon.

2.1 Univariate Analysis

To see how the categories in feature were distributed we have performed univariate analysis by visualizing each of the categorical features first. It was found that most of the survey respondent's destination was not to any urgent place, and rest of them were either travelling to office or home. When a coupon was offered most of the passengers were alone and some of them were with friends, and very less were with their kids. For most of the customers, it was a sunny weather and most of them were offered coupon in the evening around 6PM and early in the morning around 7AM. Majority of the customers were either married or single. We also observed that most of the customers who were offered a coupon have an age between 21 to 31 and most of them were having low range of salaries.

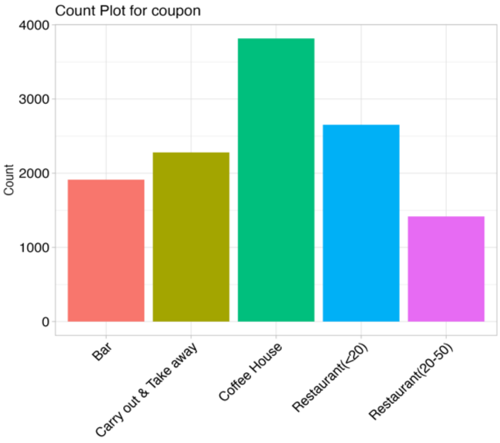

Figure 1: Frequency distribution of categories in coupon feature

In figure 1, we can see that most of the coupons that were offered to the customers are Coffee House coupons, followed by the coupons of cheap restaurants, and then carry out and take away coupons. The coupons offered were expiring either in one day or two hours and mostly the ones that are expiring in a day were offered to the customers. Later, we tried to visualize the number of visits made by the customers to the specific coupon-based locations. It was found that most of the customers visit a coffee house less than 1 time per month. While the customers who visit a cheap restaurant more than 8 times were very less in number.

Then we have performed univariate analysis for numerical features. From Figure 2, we can observe that the features direction-opposite and direction-same are significantly oppositely correlated with each other, hence we have excluded direction-opposite feature. The features toCoupon_GEQ5min (toCoupon_GEQ15min, toCoupon25min) means that a coupon-based location or venue is at least 5min (15 min, 25 min, respectively.) away from the driver's location. There is no variance in the value of the feature toCoupon_GEQ5min, which means that all venues are at least 5 minutes away from the driver's location. Hence, we have excluded that specific feature. We have decided to combine the features toCoupon_GEQ15min and toCoupon_GEQ25min and convert them into categorical feature named Drive Time, which describes the time taken (within 15 min, between 15-25 min, more than 25 min) to reach those venues. Finally, if we observe temperature has only 3 discrete values, we have decided to map those numeric values into categoric values such as low, medium and high temperature. And most of them who were offered a coupon were travelling at high temperature.

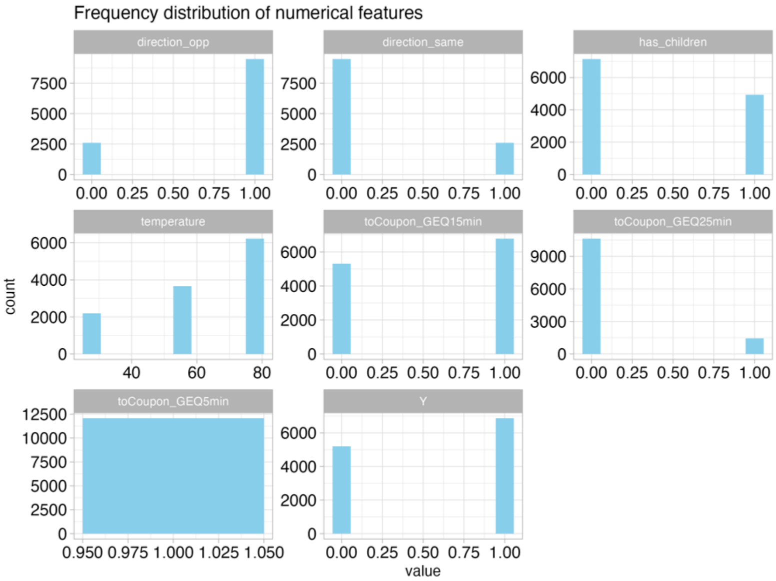

Figure 2

2.2 Bivariate Analysis

We have reduced the number of features from 26 to 22. Now we wanted see how would every feature impact the acceptance of a coupon, hence we have performed a bi-variate analysis between every feature and target class. We found that, customers who were not travelling to any urgent place were having high chance of accepting the coupon which is around 55%. If the customers were travelling with their friends, the chance of accepting a coupon was 30.96% and if those who were travelling alone have majorly rejecting the coupon. People travelling at high temperature and sunny weather are more likely to accept a coupon. If a coupon is offered to the customers in the evening most of them were accepting and it was mostly rejected by customers in the morning. People with no children were highly accepting the coupon and all of the people with various occupations are accepting the coupon and among them, students have high acceptance rate. Among the various type of coupons offered, carry out & take-away coupons and cheap restaurant coupons have high acceptance rate. Coffee house coupons had high rejectance rate.

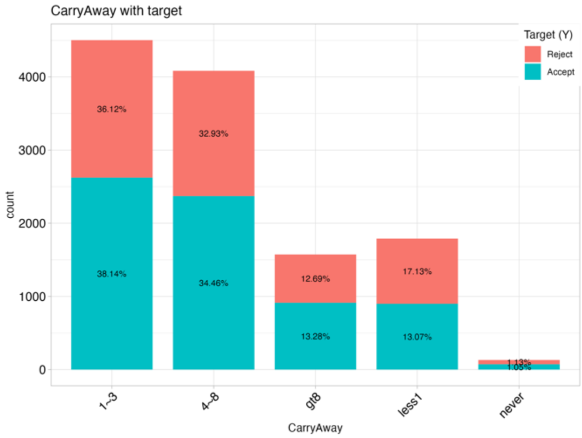

Figure 3

It was also observed that coupon with more expiry time has high acceptance of the coupon. And also, from figure 3 we can see that people who take-away more at least one time a month are highly accepting the coupon. Even though people who were travelling in the same direction of the venue were less, the chance of accepting a coupon if offered was high. People were definitely accepting the coupon if the time to reach the venue is less than 15 minutes from users current location.

2.3 Categorical vs Categorical Analysis

We wanted to analyse how would two combined features effect the acceptance of a coupon by the customer. We found that people were accepting coupon if the climate was hot and sunny because most of them would travel in that climate. People who were travelling along with their friends to no urgent place were having high acceptance of coupons around 17%. It was found if a coffee house coupon is offered to the customers around 6PM they will mostly reject it. If a Bar coupon is offered to the customers in the morning around 10AM they are highly rejecting it. Interestingly it was found that irrespective of the time, if a cheap restaurant coupon is offered to the customer, they were highly accepting it.

We saw that when a coffee house coupon is offered to a customer and if it expires within 2 hours there is very high chance of rejecting it and vice versa. From this we can conclude that most of the coffee house coupons were expiring within 2 hours. In case of take-away and cheap restaurants even if the coupon is expiring in 2 hours or in one day customers are accepting it.

3. Results

From all these insights gained we wanted to see which feature would be considered as the main attribute for splitting the data when we run a decision tree algorithm for the all the samples. It was found that Coupon feature was considered as main attribute for initial splitting of the data as it was having high information gain value which was 0.57. Later we wanted the order of importance of each feature, so we ran a random forest algorithm and extracted the feature importance and scaled the values between 0 and 1.

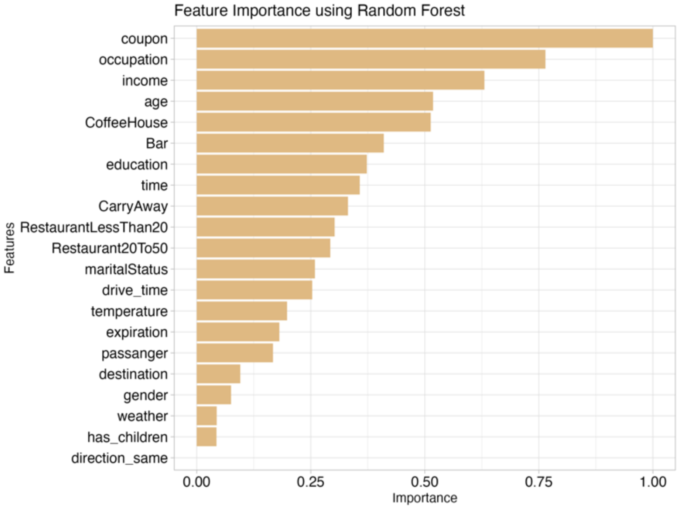

Figure 4

From figure 4 we can see that coupon is having the highest importance, followed by occupation. It can be seen that age and visiting to coffee house are having almost equal importance.

4. Conclusion and Future work

We would like to conclude that we have done the analysis on what features are impacting for the customer to accept a coupon and local businesses should focus on these features while targeting the users. From, this point on these important features can be used in creating better recommendation systems that will profit their business.

5. References

1. Wang, T., Rudin, C., Liu, Y., Klampfl, E. & Macneille, P. A Bayesian Framework for Learning Rule Sets for Interpretable Classification. Journal of Machine Learning Research **18**, 1–37 (2017).

2. Kapil, A.R. (2022) What is Univariate Analysis? Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. Available at: https://www.analytixlabs.co.in/blog/univariate-analysis/ (Accessed: 30 November 2023).

3. Missing-data imputation - department of statistics. Available at: http://www.stat.columbia.edu/~gelman/arm/missing.pdf (Accessed: 30 November 2023).

4. Sinclair, A. J. Univariate Analysis. Handbook of Exploration Geochemistry 2, 59–81 (1983).

5. Overview of data visualization - springer Available at: https://link.springer.com/content/pdf/10.1007/978-981-15-5069-0_2.pdf. (Accessed: 30th November 2023)