# IDS 702 - Final Project: Who is more likely to accept a coupon? - In Vehicle Coupon Acceptance

Aarushi Verma

12/11/2021

## Summary

In the project, we use logistic regression to model the odds of a driver accepting a coupon based on a range of characteristics. The goal of the project is to identify important characteristics that are associated with and affect a driver's decision to accept a coupon and quantify these relationships. From the results, we find most variables in the dataset have an impact on the odds. Specifically, variables such as gender, coupon, weather, expiration have noticeably large effects. However, we must ponder on the validity of the inferences given a few limitations of the data.

## Introduction

A survey was carried out on Amazon Mechanical Turk (Crowd sourcing market place) to record responses to various conditions based on which a driver may decide to accept a coupon or not. The data from this survey is being used to build a Bayesian Framework for Machine Learning Research for classification problems. For this project, we are using this data to understand what are the most important factors that may influence a driver's decision to accept a coupon. Using logistic regression to quantify the factors that affect a driver's decision we are interested to see whether the response variable varies across demographics, coupon specific and driver specific factors.

## Data

### Data Pre-Processing

The dataset is obtained from UC Irvine's Machine Learning Repository. It contains 12,684 observations for 26 variables. The response variable is a binary variable indicating whether a driver accepted a coupon offered to them or not. Other variables include all categorical variables which can be bifurcated into 4 categories:

- Demographic (Age, Gender, Education etc.)
- Driver Specific (Destination, Passenger etc.)
- Coupon Specific (Type, Expiration)
- External (Weather, Temperature, Time of day)

Overall, the dataset is not extremely unbalanced with 5,474 respondents who accepted the coupon and 7,210 who did not accept the coupon.

Before proceeding with the analysis, we inspected the dataset and observed several issues that needed to addressed before going forward.

The first issue with the data is of missing values. For one of the columns - `car` ~99% of the values are missing. There is no way to impute such large amount of missing data since it would not be representative of the actual scenario. Secondly, we do not have enough information available to impute and therefore we drop this column. We have 5 more columns with some missing values - `Bar`, `CoffeeHouse`, `CarryAway`, `RestaurantLessThan20` and `Restaurant20To50`. However, the percentage of missing values for each of these variables is less than 5%, therefore we drop these observations.

Since all the variables are factor variables, the second issue we observed was of high cardinality or multiple unique levels for a single variable. For example, `occupation` has 26 levels. Too many levels may result in fewer observations for each category and this may affect the inferences made based on the model. To rectify this issue, we combined levels of jobs into similar categories based on external research and reduced the number of levels to 8.

The third issue was of correlation between variables. Two variables in the data `direction_same` and `direction_opp` represent if the restaurant coupon offered to the driver is in the same direction of their destination or opposite direction. These variables have a correlation of -1 indicating that if one variable is true for one respondent, the other will be false. Since these variables

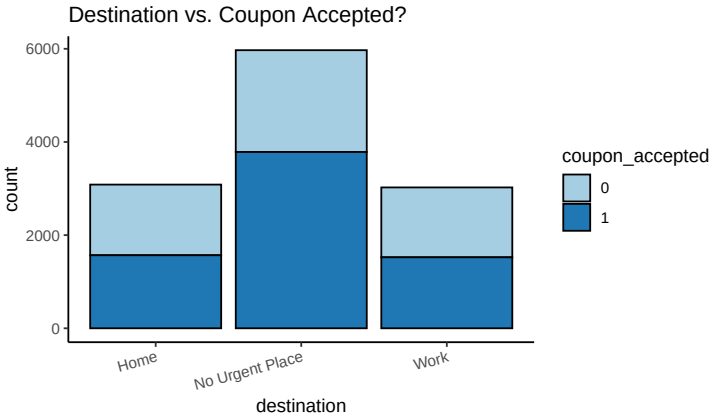
provide us with the same information we dropped one of them. The variable `weather` and `temperature` also seem to add the same information to the data. Each have 3 levels which provide us the same information. Therefore we drop the `temperature` column. Similarly, the variable $toCoupon_GEQ5min$ which represents if the restaurant is 5 mins away from a driver's house, only has values 1 for all the observations. This does not provide any information to explain the variability in the response variable and hence it is dropped. We must note that since all the variables are factors, computing correlation is tough. (We could only check for binary coded variables)

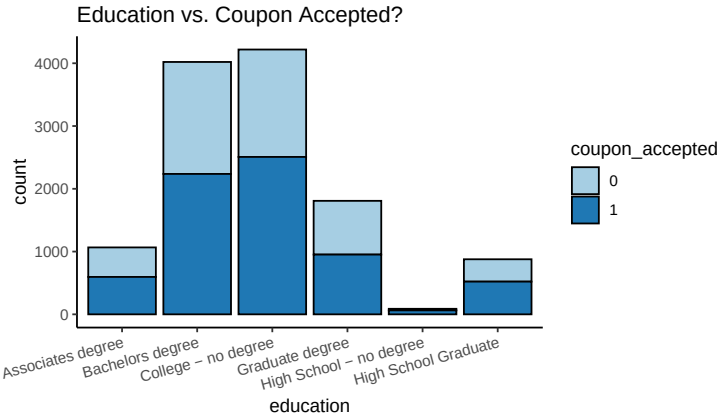
After these processes the predictors are reduced to 21 and number of observations to 12,079.

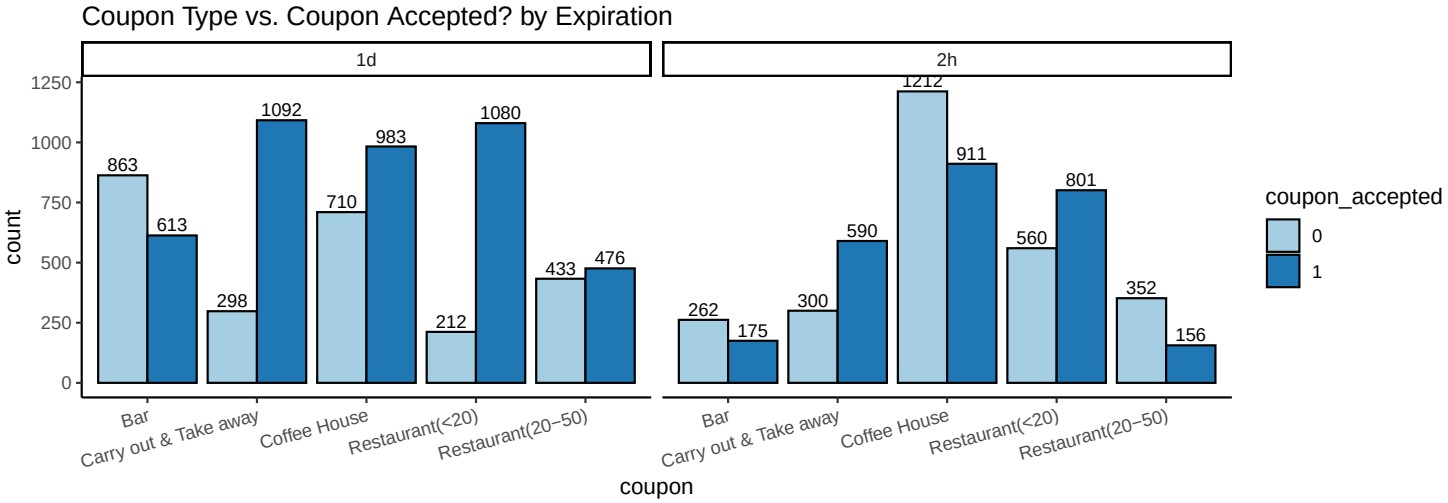
**Exploratory Data Analysis**

Before the modelling process, data exploration was undertaken to elucidate the relations and associations among the possible predictors. A large number of plots and summary statistics were generated to identify associations between the variables and make strong judgement about which predictors were likely to prove important in the modeling process. Additionally, these plots and statistics were used to identify potential concerns within the data that may impact the analysis. In particular, as one of the objectives of this analysis is to determine if gender, occupation, and education are important predictors, their plots were studied closely.

Due to the nature of the dataset, many predictors have interesting interaction effects that are worth exploring. The graphs below show a sample of spread of predictors and potential interaction effects between predictors over the response variable.




The plot "Destination v.Coupon Accepted?" and "Education v.Coupon Accepted?" shows the relationship between coupon acceptance the two variables. We can observe that the number of people who have accepted coupons varies wherein we have a maximum number of people accepting coupons when their destination is *No Urgent Place* and if they are people in *college with no degree*

The plot "Coupon Type v.Coupon Accepted by Expiration" shows that the relationship between the probability of accepting coupon and the coupon type differ depending on the whether coupon expires within 2 hours or 24 hours. Specifically, a coupon for a Coffee House is more likely to be decline if it expires within 2 hours as opposed to when it is valid for 24 hours. However, a coupon for a bar is more likely to be decline when it is valid for 24 hours.

While there are multiple interesting interactions, it is important to note that in the proceeding model building section, not all of potential interactions are included in the final model owing to lack of observations in each category and ease of interpretation of findings.

## Model

To build our model, we first built our baseline logistic regression model which included all our main effects. Next we used step wise selection using BIC to generate our final logistic model with only main effects.

The initial model is given as the following:

$$y_i|x_i \sim Bernoulli(\pi_i)\ log(\frac{\pi_i}{1-\pi_i}) = x_i\beta,$$

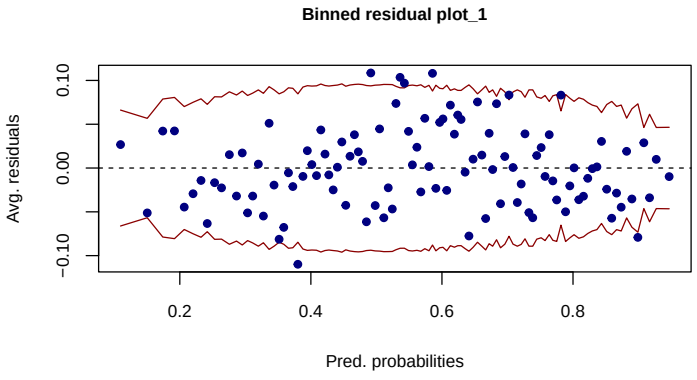
where $yi$ is the binary variable indicating whether a driver accepts a coupon and $xi$ includes all predictors as main effects. Using backward selection with the BIC criterion our model has 9 main effects that were significant. To assess the model further, we observe binned plots of the residuals against the fitted values and the continuous predictors. We check for randomness in these plots to ensure that our model satisfies the independence of errors assumption. The binned plot for residuals against fitted values seems random, except for a couple of points which are outside of the 95% confidence intervals. Since our data only has factor variable we cannot check the binned plot for any specific variables to assess any need for transformations

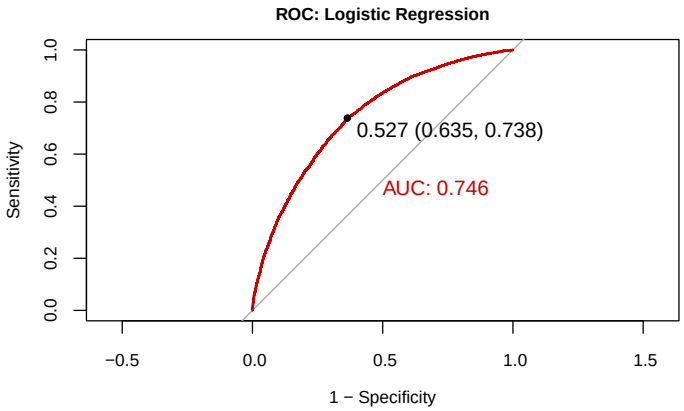
To further explain the variation in our model,we then included some interaction effects which we thought to be significant, or they answered questions with respect to the study. With the help of anova tests we assessed if these interaction were significant to our model.

We incorporated the following interaction effects, based on EDA and questions of interest:

- `gender` * `age`
- `gender` * `education`
- `education` * `income`
- `coupon` * `expiration`
- `coupon` * `gender`

We then performed stepwise model selection, using AIC as well as BIC as the criteria for variable selection. The full model resulted in improvement in the model fit (better AIC), but only improved the binned residual plot sightly. Since our data has high number of observations, we decided to proceed with backward selection using BIC. As per BIC, only the interaction between coupon and expiration is retained in our model. To further assess, which interactions should be incorporated in the model, we conducted an anova Chi-squared test and compare the p values. Given the nature of the data, there were multiple interactions that come out to be statistically significant. However, for the purpose of our analysis, we only kept the interaction between coupon and gender to simplify interpretations of our model.




The final model contains all main effects from the base logistic regression model, and the following interactions: * `coupon` * `expiration` * `coupon` * `gender`

The binned residual plot for the final model also shows similar patterns as the full model. Specifically, the points are randomly distributed with no discernible pattern. For this model, there are fewer outliers although, it is only a marginal difference. These patterns imply that logistic regression assumption are not potentially violated. All the attributes in our data are factor variables and therefore we cannot calculcate the VIF scores to assess multicollinearity issues.

To assess our final model's performance, we also observe the RoC curve.Initially, using mean as the cut-off threshold, a driver is predicted to be accept the coupon if the predicted probability is greater than or equal to mean, otherwise, decline. The

model achieves 68% accuracy, 0.67 sensitivity, and 0.69 specificity. This means that the model predicted 68% of the data correctly. 0.67 sensitivity means that given a driver accepted the coupon, the model has 67% probability of predicting coupon was accepted. 0.69 specificity means that given coupon was declined, the model has 69% probability of predicting it was declined. In addition, the model also achieves the AUC score at 0.75. However we see that out model does a poor job in classifying when the coupon was accepted.We adjust the threshold as per the ROC curve , which improves accuracy to 69% and sensitivity to 73%.

## Results

The standard estimates of all coefficients in our model were statistically significant. We exponentiate the standard estimates and their confidence intervals to interpret them on the odds scale. The model results are shown below.

We can observe that all four categories of factors (Demographic, Coupon Specific, Driver Specific, and External) are statistically significant in influencing a driver's decision to accept a coupon at the 95% confidence level.

- Keeping all else equal, compared to male drivers, female drivers have higher odds by 1.99 times of accepting a coupon, which is almost 100% higher. The odds of coupon acceptance for drivers under the age of 21 is found to be the highest which is 1.03 times or 3% more compared to age group 21-30. Compared to driver's with an associate degree, driver still in high school without a graduate degree have the highest odds of accepting a coupon, which 97% higher than the former.
- The odds of accepting a coupon differ by type of coupon as well as its expiration. Keeping all else constant compared to a coupon for a Bar, the odds of accepting a coupon for Carry Out and Take Away is almost 7 times higher. On the other hand, counter intuitively, the odds of accepting a coupon that expires in 2 hours compared to one that expires in 24 hours is 0.68 times or 31% lesser.
- Observing the interaction between coupon and expiration, there are many factors that are statistically significant. One of the interpretations for these combinations is that the odds of accepting a coupon for a CoffeeHouse that expires in 2 hours is 0.75 times, or decreases by 25%, compared to a coupon for a Bar that expires in 24 hours. Similarly looking at the interaction between coupon and Gender, we can interpret one of the combinations as that, keeping everything else the same, the odds of male driver accepting a coupon for Carry out and Take away is 0.65 times or decreases by 35% in comparison to a female driver accepting a coupon to a Bar.

## Limitations and Conclusions

There are a few potential limitations in our model. Firstly, we removed the car variable due to missing data. While we cannot use missing value imputation methods in this case, it is possible that car variable may explain an essential amount of randomness in our model. The second major limitation is that the data used is not reliable as it is crowd sourced and any one can put in any value for the responses such as incorrect age and income. The third limitation is the nature of the data, since all the factors are categorical, it makes interpretation challenging. There are variables for which continuous data could be collected, such as age, time of day, coupon expiry. This will allow for better model fitting based on review of deviance and binned residuals of individual variables.

To conclude, we note that the likelihood of accepting coupons, varies greatly with respect to demographic factors with Females accepting higher number of coupons compared to males. We also see, age group, income as well as education playing a role in whether a driver accepts a coupon or not where in we see that the levels for each of these variables is statistically significant. Acceptance of coupons is higher when the weather is sunny, and the driver's destination is no urgent place. The type of coupon seems to be a variable that is influencing a driver decision the most, wherein compared to a coupon to a Bar, Carry Out and Takeaway is preferred.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.59 | 0.19 | 3.17 | 0.00 |
| destinationNo Urgent Place | 0.68 | 0.07 | 9.12 | 0.00 |
| destinationWork | -0.08 | 0.06 | -1.40 | 0.16 |
| passangerFriends | 0.18 | 0.07 | 2.42 | 0.02 |
| passangerKids | -0.44 | 0.09 | -4.81 | 0.00 |
| passangerPartner | 0.08 | 0.09 | 0.90 | 0.37 |
| weatherSnowy | -0.22 | 0.09 | -2.33 | 0.02 |
| weatherSunny | 0.40 | 0.07 | 5.59 | 0.00 |
| couponCarry out & Take away | 2.00 | 0.11 | 17.76 | 0.00 |
| couponCoffee House | 0.82 | 0.10 | 8.37 | 0.00 |
| couponRestaurant(<20) | 2.16 | 0.12 | 18.49 | 0.00 |
| couponRestaurant(20-50) | 0.80 | 0.12 | 6.69 | 0.00 |
| expiration2h | -0.38 | 0.12 | -3.19 | 0.00 |
| genderMale | 0.69 | 0.10 | 6.95 | 0.00 |
| age31-40 | -0.11 | 0.06 | -2.03 | 0.04 |
| age41-50 | 0.03 | 0.07 | 0.40 | 0.69 |
| age50plus | -0.22 | 0.07 | -3.04 | 0.00 |
| agebelow21 | 0.04 | 0.12 | 0.30 | 0.77 |
| educationBachelors degree | -0.09 | 0.08 | -1.10 | 0.27 |
| educationCollege - no degree | 0.03 | 0.08 | 0.41 | 0.68 |
| educationGraduate degree | -0.31 | 0.09 | -3.37 | 0.00 |
| educationHigh School - no degree | 0.68 | 0.28 | 2.44 | 0.01 |
| educationHigh School Graduate | 0.15 | 0.11 | 1.47 | 0.14 |
| income$12500 - $24999 | -0.16 | 0.08 | -1.99 | 0.05 |
| income$25000 - $37499 | 0.03 | 0.08 | 0.44 | 0.66 |
| income$37500 - $49999 | -0.04 | 0.08 | -0.44 | 0.66 |
| income$50000 - $62499 | 0.14 | 0.08 | 1.75 | 0.08 |
| income$62500 - $74999 | -0.31 | 0.10 | -3.19 | 0.00 |
| income$75000 - $87499 | -0.20 | 0.10 | -2.06 | 0.04 |
| income$87500 - $99999 | -0.36 | 0.10 | -3.76 | 0.00 |
| incomeLess than $12500 | -0.13 | 0.10 | -1.33 | 0.18 |
| Bar4-8 | -0.08 | 0.09 | -0.90 | 0.37 |
| Bargreater than 8 | -0.36 | 0.14 | -2.64 | 0.01 |
| Barless than 1 | -0.20 | 0.06 | -3.22 | 0.00 |
| Barnever | -0.25 | 0.06 | -4.09 | 0.00 |
| CoffeeHouse4-8 | -0.01 | 0.07 | -0.21 | 0.83 |
| CoffeeHousegreater than 8 | -0.35 | 0.08 | -4.29 | 0.00 |
| CoffeeHouseless than 1 | -0.47 | 0.06 | -8.11 | 0.00 |
| CoffeeHousenever | -0.95 | 0.06 | -15.36 | 0.00 |
| direction_opp1 | -0.49 | 0.06 | -8.07 | 0.00 |
| occupation_classOthers | -0.43 | 0.16 | -2.68 | 0.01 |
| occupation_classRetired | -0.89 | 0.16 | -5.47 | 0.00 |
| occupation_classSocial | -0.75 | 0.17 | -4.56 | 0.00 |
| occupation_classStudent | -0.70 | 0.13 | -5.21 | 0.00 |
| occupation_classTrade Workers | -0.62 | 0.14 | -4.39 | 0.00 |
| occupation_classUnemployed | -0.72 | 0.13 | -5.71 | 0.00 |
| occupation_classWhite Collar | -0.71 | 0.12 | -6.03 | 0.00 |
| couponCarry out & Take away:expiration2h | -0.56 | 0.16 | -3.57 | 0.00 |
| couponCoffee House:expiration2h | -0.28 | 0.14 | -1.99 | 0.05 |
| couponRestaurant(<20):expiration2h | -0.95 | 0.15 | -6.19 | 0.00 |
| couponRestaurant(20-50):expiration2h | -0.67 | 0.17 | -3.92 | 0.00 |
| couponCarry out & Take away:genderMale | -0.43 | 0.14 | -3.05 | 0.00 |
| couponCoffee House:genderMale | -0.61 | 0.12 | -5.10 | 0.00 |
| couponRestaurant(<20):genderMale | -0.52 | 0.14 | -3.82 | 0.00 |
| couponRestaurant(20-50):genderMale | -0.48 | 0.15 | -3.19 | 0.00 |

Table 1: Logistic Regression Results (Log Odds Scale)