Chrystalla Havadjia, Maria Barrera Valdez, Dara Hashemi

Accept or Reject: In-Vehicle Coupon Recommendation for a Local Low-cost Restaurant

Work Division

Given that each team member has experience in both data analytics and coding, the project was evenly split in those areas. As a result, we worked together to clean, transform, analyze, and report the findings of the dataset as a group. Maria has experience with marketing and will use that domain knowledge to help find patterns and prune factors if they are redundant in the data.

Project Definition

Service oriented platforms frequently utilize coupons to entice customers and increase the likeness of a purchase. According to Juniper Research (2018), mobile coupon users will pass 1 billion by the year of 2019, driven by increased consumer engagement as retailers reach out through more mobile channels (Ahmed). However, rather than sending promotions collectively, machine learning can assist in tailoring coupon suggestions based on customer wants and profiles. To accomplish this goal, we will use classification to build and find the best model that most accurately predicts the set of factors which will lead to a customer using a coupon for a low-cost restaurant (less than $20 per person).

Description of Background

Big Data has allowed the growth of leveraging machine learning techniques and data science to enhance marketing across all industries. It is no secret that coupons bring streams of business to a company, but by targeting individuals who are more likely to accept the coupons based on their actions or preferences, businesses can save time and money by pushing out coupons straight to consumers with preferences which match the company's product. This data is crucial to a company's success because it helps it identify their target audience and allows them to gain insight into the purchasing behavior of their customers (Johnson).

Description of Dataset

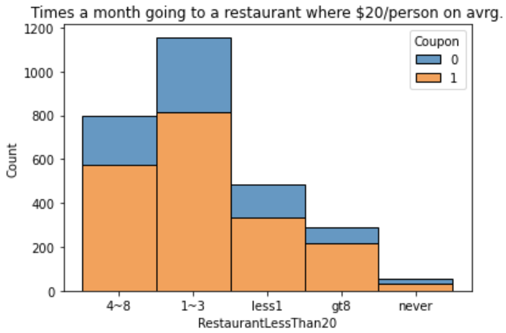
The dataset from the UCI Machine Learning Repository includes information about customer behavior collected through a survey using Amazon Mechanical Turk (Wang, 22). The coupons are in-vehicle coupons that are displayed on the car's infotainment system. The driver can choose to use the coupon before it expires. The raw data has 23 attributes and 12,684 instances but contains missing values. The attributes of the data only includes categorical variables and offers five kinds of coupons; coffee houses, Restaurant<20 (less than $20 per person) or

Restaurant20To50 ($20-$50 per person), bars, and carry out restaurants (Wang, 32). There are features about the driver such as their destination (if it is urgent), gender, age, marital status, if they have children, their education level, occupation, income, how often the customer visits the location in a month. It also includes conditional attributes such as the kind of passengers in the car, weather, temperature, current time, coupon expiration (one day or in two hours), if the location is in the same or opposite driving direction, and how far of a drive it is from customers current location. The target variable is a binary attribute, whether the customers choose to accept the coupon (1) or decline the coupon (0). To build a better model, we focused on a single coupon type; therefore, we investigated whether a customer will accept a coupon from a restaurant that is less than $20 per person or not.

The quality and quantity of data is important to ensure accurate predictions and good decision making. This dataset has both good quantity and quality of data. The data has completeness, where essential fields such as coupons, expiration time, and customer destination have no missing values. The data has some formatting issues, where age is considered categorical data and does not follow a typical distribution in addition to some features having redundancy, such as direction_same and direction_opposite. However, this was easily resolved as mentioned below.

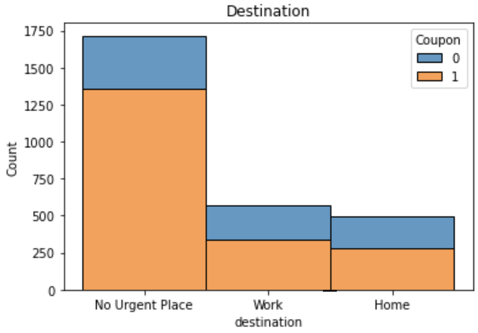
Data Cleaning

Since we are only interested in one type of coupon, Restaurant(<20), we removed the other coupon types from the dataset including Bar, CarryAway, and Restaurant20To50. As a result, out of the total 12,684 instances our dataset was reduced to 2,786 instances. Additionally, we realized that the 2,786 instances did not represent the unique number of users; instead, the same user surveyed multiple times with 652 unique instances. The direction_same and direction_opposite to the restaurant features are equal opposites, so we only kept direction_same.



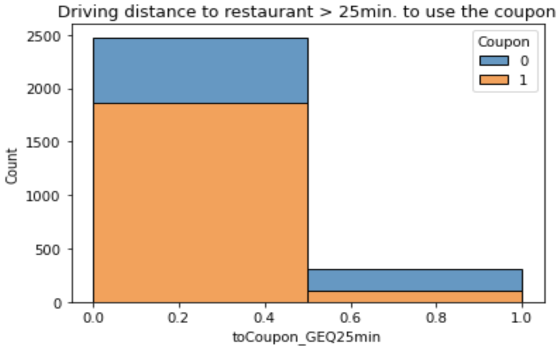
Data exploration

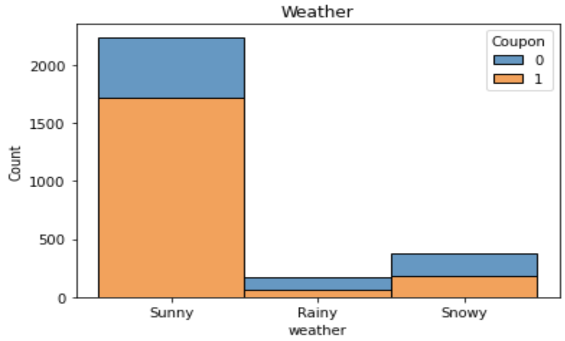
We began by finding features with missing values, visualizing the distributions, and determining obvious features which should be excluded based on our coupon type. Features with missing values were the following: the customer's car type, the distance the



customer is from the restaurant in time (e.g., toCoupon_GEEQ5min), the features that describe how often a customer will visit different locations (Restaurant<20 ext.), and the columns direction_same, direction_opp which determine which direction the customer is driving in relation to the restaurant. Upon further exploration, toCoupon_GEQ5 had only one value and the customer car feature had the majority of the values missing instances; therefore, we removed these values from the dataset.

Each feature was visualized using different bar graphs. The independent feature 'Y' had 70.7% 'yes' instances and 29.3% 'no' instances, which implies that most individuals accepted the coupon. Of the total instances, 62% of the customers' final destination was 'No Urgent Place' and 52% were driving 'Alone'. Additionally, 80.4% of the coupons were sent on a sunny day, with the majority accepted, as opposed to being rejected half the time on rainy and snowy days. We also observed that most individuals who took the survey purchased food from a low-cost restaurant 1-3 times a week, with only 2% never purchasing.





Also, coupons that expired within a day were more likely to be accepted compared to those that expire in two hours. Though not included in the visualization, gender was almost equally distributed between males and females, and they accepted or rejected the coupon on similar instances as well.

Data Preprocessing

The Restaurant(<20) attribute had 27 missing values and CoffeeHouse had 47 missing values, therefore, we imputed them using the mode. Most of the values for the categorical features were inconsistent and we changed the format. For example, Age had six distinct numerical values, apart from below21 and 50plus; therefore, we transformed the values into ranges: <21,21-25, 26-30, 31-35, 36-40, 41-45, 46-50 , and >50. Similarly, we did the same for CoffeeHouse and RestaurantLessThan20, which had numerical values, text, and special characters.

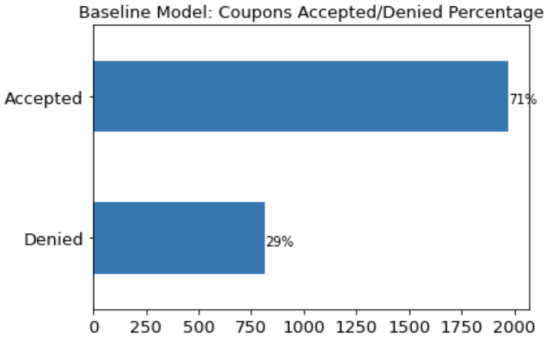
Feature Engineering

To further dive into the data, the next step is to test significance of different features the dataset offers to whether or not a customer will accept a specific coupon offering. Testing significance by calculating p-values helps find the variables which have the strongest relationship with each other, making it easier to filter which features would be best for a classification model. Because the data consists only of discrete and/or categorical type, a series of chi-squared tests on variables with 3 or more levels for the predictor and whether or not the consumer accepts the coupon as the outcome would give the information needed. For features with only two levels (binary), we tested the strength of the relationship using jaccard similarity. Using Jaccard similarity on features consisting of gender, whether a coupon was offered at a location within 15 minutes from the consumer, and whether a coupon was offered at a location within 25 minutes from the consumer. After running these tests, the features which have the strongest relationship with whether or not a coupon was accepted are “destination” (No Urgent Place, Home, Work), “passenger” (Alone, Friend(s), Kid(s), Partner), “weather” (Rainy, Sunny, Snowy), “time” (10AM, 10PM, 2PM, 6PM, 7AM), which have p-values at or very close to 0. More features include “income”, “age”, “gender”, how often the consumer visits a “coffee house” per month, and how often the consumer visits a restaurant (<$20) per month, which have p-values less than .05. The remaining variables included in this dataset either have either no relationship or very little association to whether or not a coupon is accepted (p-values are greater than or equal to .05). Therefore, some subset of these features will give the most accurate classification model.

Description of Methods Used & Analysis Results

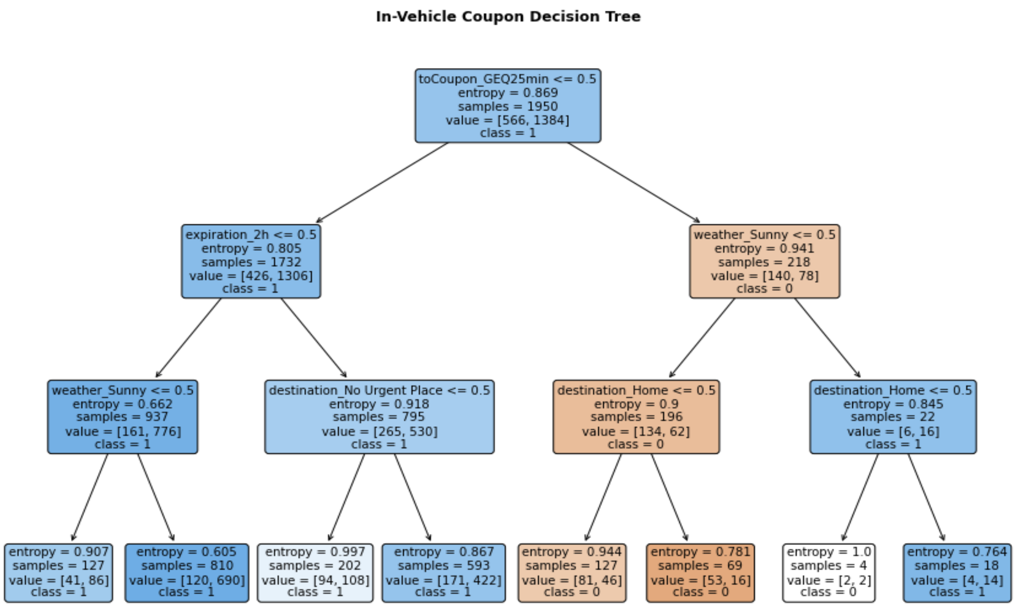
We solved the proposed problem using three different classification algorithms, with a combination of R and Python programming languages, to find the model(s) with the highest accuracy and F1-Score.

Baseline Model: A baseline model helps offer a comparison point. A good model must perform at least better than a baseline model. In this dataset, the coupon accepted class occurs 71% while the coupon denied class occurs 29%. Due to the imbalance in class frequency a good baseline model is the Zero Rate Classifier where the model predicts the largest group for all



instances. The baseline model has an accuracy of 71%. All the models described below have an accuracy above this, indicating that all three are acceptable models.

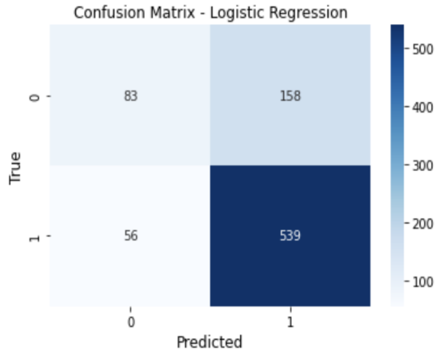
Decision Tree Method (Python): After cleaning and preprocessing the data, the categorical variables were converted to numeric variables for analysis. The most correlated features with each other(direction_same, temperature, and maritalStatus) were deleted from the dataset to prevent multicollinearity. Then, the most correlated features with the target variable (‘Y’) were ranked based on a threshold of r=0.1, which were: toCoupon_GEQ25min, weather, expiration, destination, toCoupon_GEQ15min, and passenger. These six features were used to create the first model and split into training (70%) and testing (30%) subsets. Using entropy as the criterion, the toCoupon_GEQ25min (driving distance to the restaurant for using the coupon is greater than 25 minutes) feature had the most information gain and became the root node. This first model yielded an accuracy of 76.7%. Adding additional features (e.g., gender) decreased the accuracy, leading to an issue of overfitting. In the second model, the passenger feature was removed since it had the lowest correlation of the six features mentioned above. Following the same methods as the first model, the second model yielded the same accuracy as the first model, and further removing features only decreased the accuracy. Therefore, the second model with five features (toCoupon_GEQ25min, weather, expiration, destination, and toCoupon_GEQ15min) and 76.7% accuracy is the best model because it is the simplest and most generalized model that will be able to best respond to new data. The F1-Score was also calculated and resulted in 85.6%. The max depth for the tree model is three, increasing the depth increased the accuracy but did not increase the F1-score and decreasing the depth below three decreased the F1-Score. Based on the decision tree below, if the distance to claim the coupon is greater than 25 minutes,the weather is sunny, and the destination is not home, then the driver will reject the coupon. However, out of the 22 samples in which the weather is sunny and the destination is home, four samples (in four instances) the coupon will be rejected; otherwise, the coupon will be accepted. In any other case, the driver will accept the coupon. With this in mind, there were 181 false positives and 13 false negatives.


In-Vehicle Coupon Decision Tree

Naive Bayes Method (R): Another classification method we decided to test was Bayes Classification because it is one of the most practical approaches to classification due to its calculations of probabilities for hypotheses and how each training example has an effect on those probabilities. Because we only have 2,786 rows of data, the computational cost of testing this method would not be too high. Before creating a classification model, there must be a split of the data into two sets, training and testing. By setting a seed for data reproducibility, assigning 2,000 random rows of data without replication of rows to a training dataset, and assigning the remaining 786 observations to a testing dataset, we can build a model on a subset of the data then test the model on a different subset of data to see how well the model performs in addition to observing any possible overfitting issues using R.

In order to perform Naive Bayes Classification, there must be an assumption that all attributes included are conditionally independent. After finding the independent variables by means of Chi-Squared Test of Independence using R, as stated earlier in this paper, our model takes into account relationship to the passenger (if one was present), weather, destination, time, how long until the coupon expires, and whether the restaurant to where the coupon was offered is within 25 minutes of the driver at the time of the offer. After creating a Naive Bayes model on the training data using a built-in function in R, we test the model on both the training data and the testing data. Then by using the predict function to see how our model would classify each individual transaction without the given output and creating a confusion matrix, we can calculate the approximate accuracy and F1-score of the model on both split datasets. When testing the Bayesian model on the training data, we get an accuracy of about 77% and an F-1 score of about 85%. When testing the model on our testing subset of data which the model hasn't become familiar with, we get an accuracy of about 76% and an F-1 score of about 84%. Getting similar accuracy and F-1 scores for both datasets we tested our model on indicates that the model is likely not overfitting to just the training data and the model could be applied to similar data and classify at an accuracy rate close to our testing data.

Logistic Regression (Python): Logistic Regression is another model used for categorical data. After determining the independent variables using the Chi-Square Test, the remaining features were mapped to numeric variables for an easier analysis process. The remaining features were tested for correlation in relation to the target variable 'Y'. Features with a correlation value below the absolute value of 0.18 were removed. The independent features used to fit the Linear Regression include destination, weather, temperature, expiration and toCoupon_GEQ25min. The data was then split and trained using a 3-Fold Cross-Validation. For replicability the split-train was done using a random seed of 2. The Logistic Regression model produces a mean Accuracy of 74.28% and an F1-score of 83.54%. To test for overfitting the model was trained on training data. The Mean Accuracy is 75.79% and the Mean F1-score is equal to 84.65% indicating that the model is not overfitting.



Observations and Conclusion

After testing three different methods to create a classification model, we observed a similar accuracy and F-1 scores for Decision Trees, Naive Bayes and Logistic Regression using a similar subset of features for each method. Features used in each method include destination, weather, expiration, and whether the restaurant to where the coupon was offered is within 25 minutes of the driver at the time of the offer. **When using the models created, we can predict whether a coupon would be accepted or declined at around 76% accuracy for each method.** While accuracy was around 76% for each model, F-1 scores were also similar at around 84%. F-1 score is included in this analysis because of the imbalance of the output variable, as around 71% of coupons offered were accepted whereas 29% of coupon offers were declined. We tested each of these methods to see if one model would be more effective than the others, but there was little to no difference. After seeing no difference, we wanted to choose the most simple model when looking at features used, however each method only used five or six variables. Based on these findings, we can conclude that any of these three methods could be used to predict whether a coupon would be accepted or declined.

Coupons for food establishments have gained popularity as a form of marketing and being able to target those who are more willing to accept coupons can be beneficial for restaurants using this service. By creating this classification model, we are enabling coupons to be pushed out to individuals who have a higher chance of accepting them based on our selected factors, ultimately bringing in more business to the establishment. The number of mobile device users is expected to reach 9.2 billion with a 5% compound annual growth rate by 2020; and in 2019, 1.05 billion shoppers are expected to use mobile coupons, up from just under 560 million in 2014 (Park). In this light, mobile promotions such as coupons can impact customer behavior because it is an enticing incentive to make a purchase they otherwise would not make. Bringing in more customers to restaurants also directly affects the coupon app's performance and with better performance brings more food establishments wanting to add their coupons to the app, benefitting all parties involved.

References

Ahmed, Kazi Afaq, and Zainab Sarwar. "Consumer Willingness to Use Digital Coupons: A Case of Karachi Market in Pakistan." International Journal of Experiential Learning & Case Studies 3.1 (2018): 33-42.

Johnson, Evelyn. “How AI Is Transforming Coupon Marketing Campaigns.” ClickZ, 2 Feb. 2021, https://www.clickz.com/how-ai-is-transforming-coupon-marketing-campaigns/264928/.

Park, C. H., Park, Y.-H., & Schweidel, D. A. (2018). “The effects of mobile promotions on customer purchase dynamics”. International Journal of Research in Marketing, 35(3), 453–470. https://doi.org/10.1016/j.ijresmar.2018.05.001

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.