**Final Report**

Likhitha Devineni, Nhan Le, Zoey Le

Data Mining: In vehicle coupon recommendation

1. **In vehicle coupon recommendation dataset:**
2. **Motivation:**

Customer Behavior analysis now is one of the key factors determining the success of a business. The younger generation, especially Gen-Z, prefers more rapid and direct responses with just the click of a mouse. More and more “brick-and-mortar" shopping or placing orders activities are now easily through a smartphone. Every touch on the screen triggers an action, and actions are, then, collected and gathered to become potential sets of data. Hence, this fast-moving circulation of dataflow has created a restless “virtual competition” for modern businesses where winners are those who get the most out of their data.

Data collected from customer’s past online purchasing behaviors helps businesses keep track of what kinds of products are purchased more often, why a product is more targeted than other products, what triggers the decision to buy, when the product is purchased the most, etc. By discovering these patterns, businesses understand more about their customers; thus, make more reasonable business decisions.

Many big tech companies, e.g., Netflix, or Amazon have been so successful in bringing the practice of behavioral analysis in learning the buying patterns of their users. Users like us get to see more and more product suggestions/ recommendations based on our past selections, for instance, we get recommendations for the next movies based on the genre of the movie we just finish; or after purchasing a book through a website, we get to see more book suggestions following the similar content.

The data our group chooses will be set as an example of how we as data analysts of a company discover the buying patterns, what factors impact their buying decisions, and how the marketing team can take on more customers based on our analysis.

1. **Problem definition:**

Customers tend to make their decisions to buy towards a product or service based on some common patterns. For example, if they are married couples, they are more likely to get attached with buying a new house, or long-term investments; or first-year students tend to join more on-campus activities than other group of students. The goal for this project is to find these target groups and the patterns in their actions through data mining.

A predicting model will be then developed to test if a customer will accept the offer. We as a team, will simulate the online dataset and gather newly answers as inputs to test out the model.

1. **Solution:**
2. Data mining tasks are done to discover the patterns from the dataset. These tasks include missing values detections, outlier detections, statistics, graphing and plotting.
3. A machine learning model, e.g., Decision Tree Classification, Random Forest and K- Nearest Neighbors, then used to predict future customers’ opinions.
4. Finally, new collected data will be tested on the model that yields a better predicting result.
5. **Data Mining Tasks:**

After collecting 752 surveys, 652 were accepted, the dataset was set to 12648 data cases with 25 categorical attributes:

* destination, - occupation,
* passenger, - income,
* weather, - Bar,
* temperature, - CoffeeHouse,
* time, - CarryAway,
* coupon, - RestaurantLessThan20,
* expiration, - Restaurant20to50,
* gender, - toCoupon\_GEQ15min,
* age, - toCoupone\_GEQ25min,
* car, - direction\_same,
* maritalStatus, - direction\_opp,
* has\_Children, - Y (customer’s decision)
* education,

1. Dataset Statistics:

Number of Variables: 25

Number of observations: 12684

Missing cells: 0

Duplicate rows: 74

Variable types:

Categorical: 21, Numeric: 4

Target variable Y:

1: 7210

0: 5474

There is a 56.84% of people accept the suggested coupons while 43.15% people say no with the coupons.

1. Data pre-processing:

**Missing data:** The ‘car’ attribute has 12576 missing values as customers either don’t want to share their properties’ information or the model of the cars are too old to be recognized by the system, so that we decided to exclude this attribute for our analysis. In addition to that, for the others, they include nearly 2% of the missing data, so we decided to replace it with Mode.

**Dummy coded:** Bar, CoffeeHouse, CarryAway, RestaurantLessThan20, Restaurant20to50, time, expiration, age, gender, weather, destination, passanger, coupon, maritualStatus, education, occupation, and income are encoded as 0, 1, 2, etc.

1. Data Exploratory:

*Destination*

People might have more time joining the survey and accept the coupon if they are heading to ‘No Urgent Place’. The number of accepts is higher than the other two groups.

*Chart, bar chart

Description automatically generated*

*Gender*

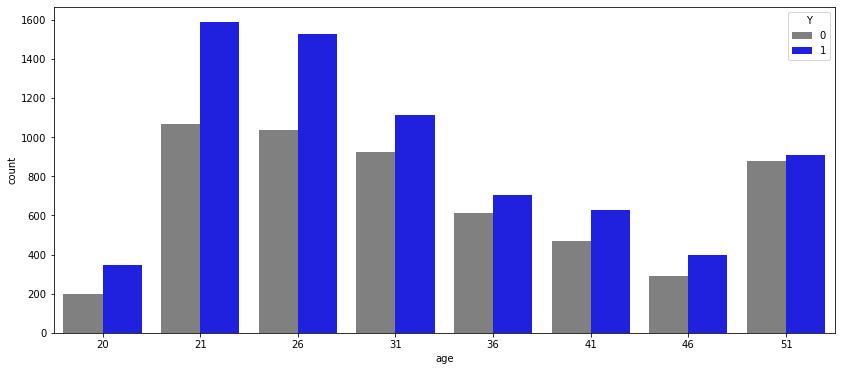
The accept rate (Y = 1) is approximately 56%, and both genders equally distributed throughout the dataset. We can learn that gender is not a factor that determine the decision of accepting or rejecting the recommendation.

Chart, bar chart

Description automatically generated

*Age*

Age ranges from over 20 to 51. The number of people accepting the coupon is mainly between the age of over 20 to 30.



*Occupation*

The group of people who are identified as unemployed and student yields the largest counts from the survey. In addition to that, a very interesting point we learn here is that the survey seems to interest loads of people who work in the Computer & Mathematical area; they are on the third place. :) Their group is encoded as 2 on the graph

Chart, bar chart, histogram

Description automatically generated

*Temperature:*

Temperature is also a factor that can determine customers’ decision. We can tell from the graph, the warmer the temperature, the more people join the survey.

*Chart, bar chart

Description automatically generated*

*Driving distance greater than 25 minutes:*

This graph shows that if the driving distance is too far, in this case, greater than 25 minutes, less people will accept the offer. We can see that there is a big gap between group 0 and group 1.

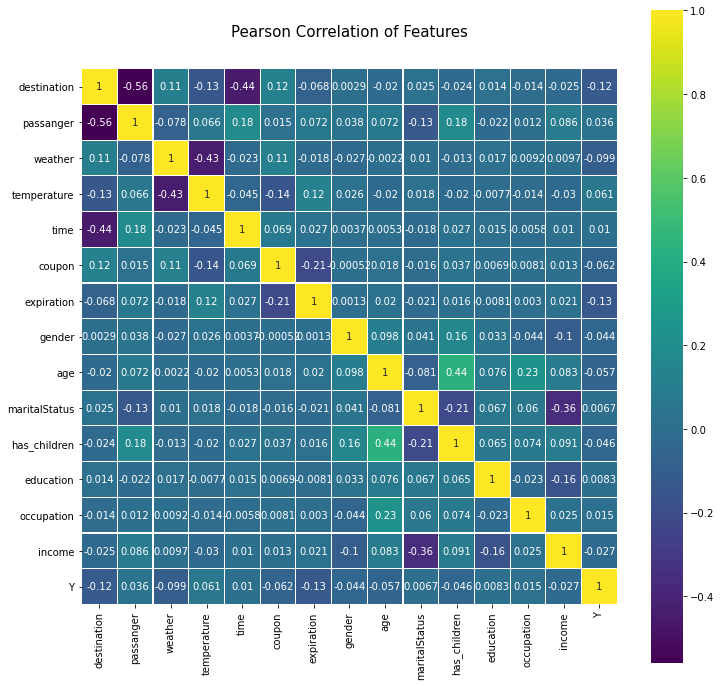
Chart, bar chart

Description automatically generated

*Pearson correlation of customers’ features:*

Explain: Pearson correlation measures the linear relationship between two variables. It has the value [-1, 1]. Negative values mean there is a negative relationship between two variables; positive values mean there is a positive relationship between two variables. On the graph showing below, the darker the color, the stronger the relationship.

In our data, the relationships are not significant as the possibly spotted widest range is from -0.56 to 0.44, between passenger and destination, and between has\_children and age.



1. **Machine learning models:**

Decision Tree Classification:

For the dataset, Decision Tree Classification algorithm is used to fit the actual data. The algorithm is the first option since most of the variables from the dataset are categorical.

Chart, treemap chart

Description automatically generated

Confusion Matrix:

[[ 534 576]

[ 187 1240]]

Accuracy: ~70%

K-Nearest Neighbors:

It is a data classification method for estimating likelihood that data point will become a member of one group, or another based on what group the data points nearest it belong to. For our data set the accuracy was low. To increase the performance of the model from the below graph we can observe that with increase of k value accuracy also increasing K value > 15 can give better results.

Chart, line chart

Description automatically generated

Confusion Matrix

[[ 628 474]

[ 342 1093]]

Accuracy: ~66.45%

Random Forests:

Random Forest selects the random samples from the dataset and create decision tree for selected samples and voting will be performed for predicted result. The below image is final tree after splitting the nodes. As random forest is ensemble learning one way to increase the accuracy would be opting K fold cross validation for splitting the test data and changing the parameters like increasing number of trees might give a good model, but this could lead to overfitting as our data will be well trained.

Diagram

Description automatically generated

Confusion Matrix:

[[ 547 527]

[ 218 1245]]

Accuracy: ~69.18%

1. **Microsoft Forms dataset:**

We created a survey through Microsoft Forms. Questions on the [survey](https://forms.office.com/Pages/ResponsePage.aspx?id=qIfso7ifWEG6j_EbrOHrqkOhIIBpr55ElMt1eJeCtKxUMFJZVTBNTjBZUFJTUVlVNktMSlFYQlZHWi4u) follow the attributes of the online dataset.

* 1. Target groups:

Based on the counts of each occupation group from the above analysis, we have many students and unemployed people joining the survey. Our focus group will be **students around NMSU campus.** We want to see if we can get a higher rate of accepting.

*(Note: Students will be guided through an imaginary scenario.)*

* 1. Scenario:

*“IMAGINE you were on the road waiting for the traffic light to go green. It had some incidents at the intersect, but the expected delay was guaranteed to be around 15 minutes.   
There was an invitation in your DM, and you would win a chance to get a food coupon at a local restaurant near the current location of your vehicle. And you decided to do it while waiting for the incident to be cleared out.  
Before taking this survey, think about the scenario a bit more detailed:****where were you heading to****?****who******were you driving with****?****what was the weather like (sunny, snowy)? was it cold? What was the estimated temperature, and what time was it?”***

* 1. Dataset Statistics:

Target variable Y:

﻿1: 24

0: 2

There is a 92.30% of participants accept the coupons; 7% participants reject it. However, this survey may result a strong bias as our collected data is significantly NOT big enough.

*(Note: More statistics about the dataset is included under our Python code.)*

* 1. Data pre-processing:

Clean the survey:

* Drop columns: The raw data set includes extra columns. These columns are to keep track each record from the survey, for example, ID, time taken for them to take the survey, etc. These columns, however, are not useful to extract insights from the data.
* Dummy coded: Convert categorical values into nominal values (1, 0, 2 etc.) following rules for the online dataset.
* Outlier detection: One attribute that includes outliers is age. Based on the boxplot showing below, age ranges from 15 to 30. We have a response with the age of 3, which doesn’t make sense in this case.

Chart, box and whisker chart

Description automatically generated

* 1. Data Exploratory Analysis:

*Occupation*

As mentioned earlier, our target group will be students around NMSU campus, hence, number of students is the majority, and the chance they accept the coupon is also higher in comparison with other groups.

Chart

Description automatically generated

*Distance greater than 25 minutes*

Different from the online dataset, we have a greater number of people chooses to take the coupon even the driving distance is more than 25 minutes. At NMSU, we seem to have many patient and diligent students. :)

*Chart, bar chart

Description automatically generated*

*(Note: more graphs will be displayed under our Python code. We only choose to mention attributes that stand out).*

* 1. Testing on Machine learning model:

After testing three models, we conclude that Decision Tree and Random Forest yield the approximately the same accuracy; but the running of the Decision Tree classification is faster.

*(Note: K-NN returns lowest rate of accuracy and takes longest time.)*

Decision Tree Classification:

﻿[[ 2 0]

[14 10]]

Accuracy: 46.15%

Run time: 0.0092s

Issue: Because of the time constraint, and the limitation to gather data from a large group of students, the number of the testing data is not large enough; hence, the results are **not reliable**.

1. **Conclusion:**

After the project, we learnt how to define a dataset’s issues, e.g., missing data, outliers, and learnt how to get a cleaner dataset that support our learning purposes. We found some interesting facts from the dataset and how people make their decision whether to get the coupons or not through both sets of data.

For predicting we have used 3 Machine Learning algorithms to find which model will predict whether customer accepts a coupon or not. Models we used are ‘Decision Tree’, ‘Random Forest’, ‘K nearest neighbor’. There is so much variance in terms of accuracies which might most likely depends on the dataset. With analysis we have, we have concluded that Decision tree predicts better. For test data we come up with an idea to collect data from fellow NMSU students, but we were not able to gather as much as we needed due to time constraints.

In future, we expect to get more people joining the survey so that the results from the machine learning models will be more accurate.

**References:**

Dataset:

<https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation>

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.