Student Research Project
Classification of Coupon Recommendations

(Computer Science Bachelor)

Maximilian Brückner

Summer semester 2021

Supervision by
Prof. Dr. rer. nat. Patrick Baier

Contents

1 Introduction

This project is based on the in-vehicle coupon recommendation data set.[Wan20] It contains information about whether a driver is interested in accepting a coupon for a restaurant or bar while driving. The data was collected via a survey on Amazon Mechanical Turk which presented participants with differing scenarios before asking them to make their decision. This poses a binary classification problem. Through examining different features like time of day, weather, the driver's destination and familial situation we ultimately need to come to a decision which can be either "yes, accept the coupon" or "no".

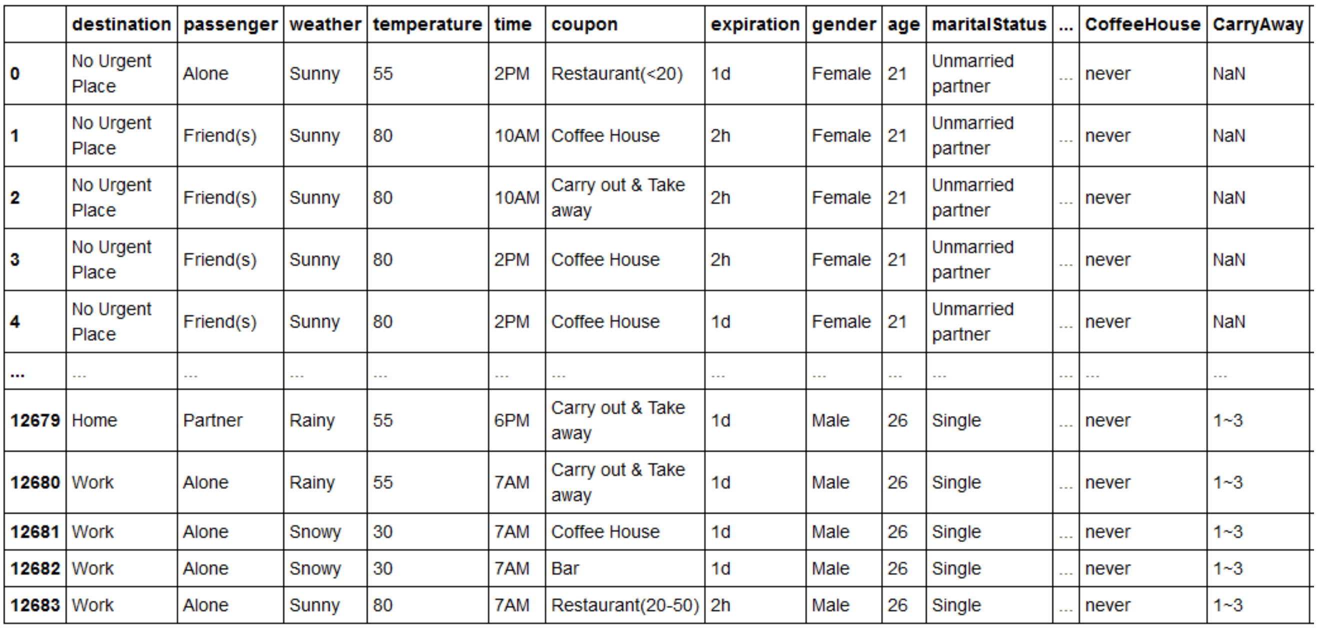
2 Topic

Firstly, we will be reviewing the data set in an exploratory data analysis. We will be looking at what kind of features are provided highlighting both their correlation to the label and amongst each other. Furthermore, the features will be graded based on their significance for making accurate predictions. We will also touch on the topic of feature engineering and the limitations of the provided data. Finally, our goal is to train varying models for classification as well as evaluate their performance and compare them against each other. In order to achieve best results we will thereby rely on cross-validation when determining hyper parameters. The implementation can be found in this repository. [Br1]

3 Implementation

3.1 Data Review

The data set contains 25 feature columns and one more for the label. This makes it difficult to display it in it's entirety, but here is an excerpt. For more details please refer to the source.[Wan20]



	destination	passenger	weather	temperature	time	coupon	expiration	gender	age	maritalStatus	...	CoffeeHouse	CarryAway
0	No Urgent Place	Alone	Sunny	55	2PM	Restaurant(<20)	1d	Female	21	Unmarried partner	...	never	NaN
1	No Urgent Place	Friend(s)	Sunny	80	10AM	Coffee House	2h	Female	21	Unmarried partner	...	never	NaN
2	No Urgent Place	Friend(s)	Sunny	80	10AM	Carry out & Take away	2h	Female	21	Unmarried partner	...	never	NaN
3	No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	2h	Female	21	Unmarried partner	...	never	NaN
4	No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	1d	Female	21	Unmarried partner	...	never	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
12679	Home	Partner	Rainy	55	6PM	Carry out & Take away	1d	Male	26	Single	...	never	1~3
12680	Work	Alone	Rainy	55	7AM	Carry out & Take away	1d	Male	26	Single	...	never	1~3
12681	Work	Alone	Snowy	30	7AM	Coffee House	1d	Male	26	Single	...	never	1~3
12682	Work	Alone	Snowy	30	7AM	Bar	1d	Male	26	Single	...	never	1~3
12683	Work	Alone	Sunny	80	7AM	Restaurant(20-50)	2h	Male	26	Single	...	never	1~3

Surprisingly all features are categorical, the only exception being "temperature". It could be considered numerical but there are only three different values in all 12684 data points. Even the drivers age is categorical because all drivers below the age of 21 and those above the age of 50 were grouped together respectively. Furthermore, the information provided is already quite expressive and does not leave any room for interpretation. This poses a significant challenge for Feature Engineering because we cannot derive new meaningful features. If there was a column containing something akin to a list of items in a shopping cart, we could generate a new feature like the number of items or their total value.

As often is the case not all entries are complete. We should therefore count the amount of missing values per feature.

- car 12576
- Bar 107
- CoffeeHouse 217
- CarryAway 151
- RestaurantLessThan20 130
- Restaurant20To50 189

Since for most features the proportion of NAs is small, we can simply fill the gaps using the mode. However, we must not include the "car" column where 99% of all entries are missing. In fact, it would be reasonable to drop the column altogether.

The next step is to prepare the data for model training through one-hot-encoding. Some features are already binary labeled, but most of them still need to be broken up into separate columns. Initially I wrote my own loop to adjust the data set. But you can also use pandas.get_dummies() which functions quite similarly.

```
dfOHE = df
featuresToBeOHE = featuresToBeOHE.drop(labels=[’has_children’, ’toCoupon_GEQ5min’, ’
    toCoupon_GEQ15min’, ’toCoupon_GEQ25min’, ’direction_same’, ’direction_opp’, ’Y’])
```

```
print(featuresToBeOHE)

for feature in featuresToBeOHE:
    print('Current feature: ' + feature)
    valueArray = df[feature].value_counts(dropna=False).index
    print(dfOHE[feature].value_counts())
    for value in valueArray:
        dfOHE[str(feature) + '_IS_' + str(value)] = df[feature].map(lambda x: 1 if x==
    value else 0)
        print('  Current value: ' + str(value))
        print(dfOHE[str(feature) + '_IS_' + str(value)].value_counts())
    print()

dfOHE = dfOHE.drop(columns=featuresToBeOHE)
```

Finally we perform the train-test-split and scale the features to facilitate convergence for gradient descent.

3.2 Exploratory Data Analysis

Before getting into training any models, we should examine the data more closely. Most importantly we need to ensure our data set is not imbalanced. Luckily, the label distribution appears to be quite balanced.

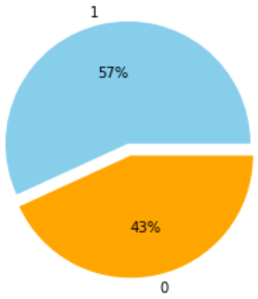


Figure 1: label distribution

3.2.1 Correlation

When examining correlations between features, I first attempted to generate a heatmap for all 115 columns. It is hard to read but can be viewed in notebook 02_LogReg_RF_GBT.ipynb.[ohe] There are patterns of boxes of negative correlation along the diagonal which can be attributed to one-hot-encoding. Obviously, a positive entry for one manifestation demands negative entries for all other values of the previously combined categorical feature. Apart from that I observed that there is a 1 to 1 correlation between people driving to work and the clock showing 7 am. Therefore we can drop one of the two because they effectively encode the same information. There are other unsurprising correlation like sunny weather indicating hot temperature or married drives being more likely to have kids as passengers. But no major new insights were gained.

3.2.2 Feature Importance

3.3 Model training

3.3.1 Basic models

3.3.2 AutoML

3.3.3 Neural Network

4 Evaluation

5 Summary

References

[Br1] Maximilian Brückner. https://github.com/Lumi1070/ML-Projektarbeit/, 2021.

[ohe] https://github.com/Lumi1070/ML-Projektarbeit/blob/master/02$_LogReg_RF_GBT.ipynb$.

[Wan20] Tong Wang. https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation, 2020.