Coupon Usage Prediction

# Introduction

In this paper and the follow documents presented along the way of this project, we will attempt to demonstrate if the online browsing history is correlated to final purchase.

In that case, we will be using four data files to learn from. The first file is composed with the list of 22.874 customers and some personal information that will be described in the dataset section. The second file to be used is a file that contains information related to users buying voucher. And the third file to be used contains the data related to users browsing in the site. The last file to be used is the coupon detail information that contains all the data specifically about the coupon.

# Literature Review

There are several researches and projects done related to online customer behavior and coupon usage.

• Blattberg et al. (1978) suggested that the coupon usage would be related to demographics characteristics where the consumers are assumed to minimize the sum of transaction costs, storage costs and the price of the item. He basically suggested that the upper income households, the more likely to redeem the coupon.

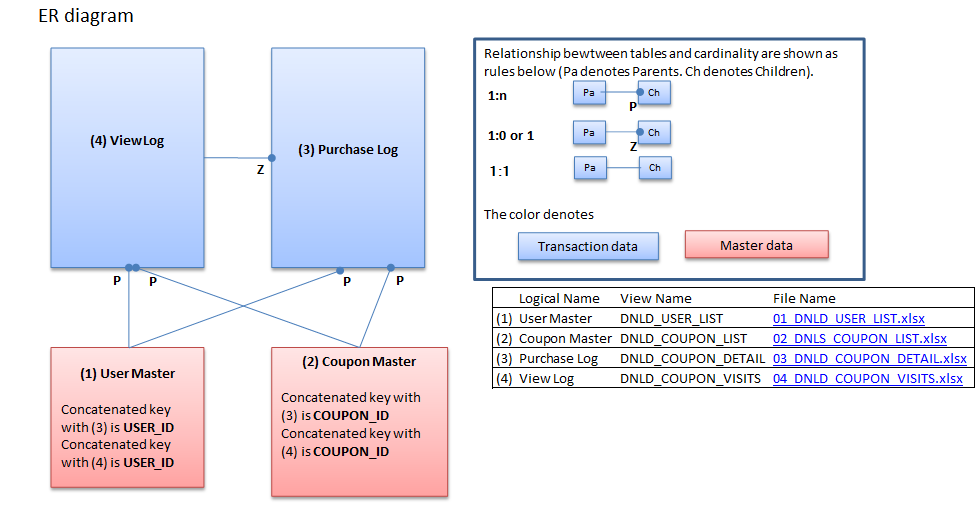
• Narasimhan (1984) proposed that intensity of coupon usage is related inversely to a household’s opportunity cost of time. Therefore it would be expected that in households that are more educated, have children under six and husband and wife are employed would have a lower prone to use coupons.

• Bawa and W. Shoemaker (1987) suggested that the intention of using the coupon (which in their project is called CPI - coupon proneness index) is a function of household characteristics and customer behavior.

-Kwon Jung (2010) suggested that the online usage of coupon is a funcion of the percentage of discount offered and demographics.

# Dataset

As mentioned before, the dataset used is composed by 4 files and they all have an entity in common.



Above you can see the ER diagram for the data.

• user\_list.csv: contains 6 features and 22,873 users. The features are related to (registered day, gender, age, date that unregistered, preferable name and user id).

* coupon\_detail.csv: contains 6 features and 168,997 entries. The columns consist in information about quantity bought, purchase date, geographic area that was bought, purchase identifier, user id and coupon id.
* coupon\_visits.csv: contains 8 features that refer manly to the browsing logs. The columns are purchase flag, purchase id in case it happened, log date, page serial, refer, coupon id, user id, session id.

coupon\_list.csv contains 24 features related to the coupon like category, expire date, what week days it’s available, discount value, and so on.

# Approach

In our process to model the data we will have 4 main steps:

## Step 1: Data

The first step that have been taken in this project is to transform the data into a more workable table with the features of the users and the coupons considering if the interaction was a purchase or not.

The code to transform the data can be found in this github [repository](https://github.com/dresenhista/StatisticsProjects/blob/master/Project/assignment%202/1-clean.R).

This final table has around 2 million of interactions, and 16 columns, therefore we selected randomly around 8 thousand interactions in order to run the model.

With the 8 thousand interactions we divided 80 % of the data to train and the rest to test. The training data was then used to work with the machine learning model using cross validation for 10 folds.

After the model selection the plan was to test in the test partition to re-validate the model avoiding overfitting.

The chart flow below exemplifies how the experiment was executed:

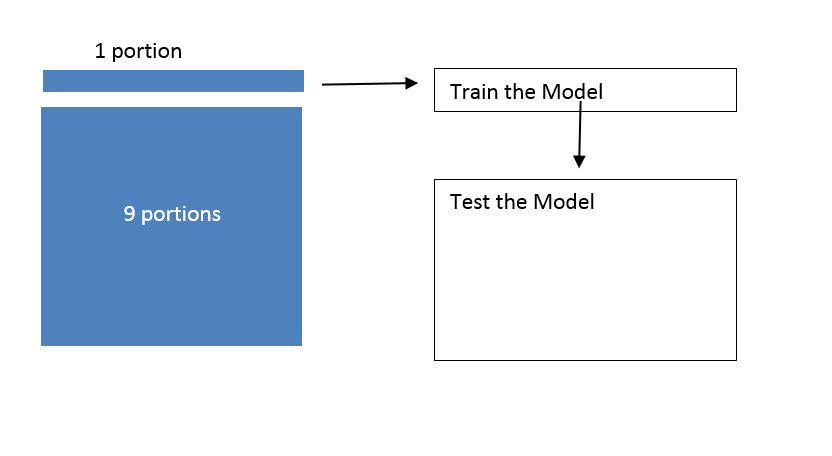
1 – Split the data in Training and Testing

20%: Test

80%: Train

The training data will be then used for the cross validation.

2 - Training the model using cross validation:



3 - As a result of the cross validation, we then had 10 models, and 10 predictions for the 9 partitions.

Select the model with the best performance during the training and the testing stages

Test in the 9 other portions

Prediction 1

Prediction 2

Prediction 3

Prediction 4

Prediction 5

Prediction 6

Prediction 7

Prediction 8

Prediction 9

Prediction 10

Model 1

Model 2

Model 3

Model 4

Model 5

Model 6

Model 7

Model 8

Model 9

Model 10

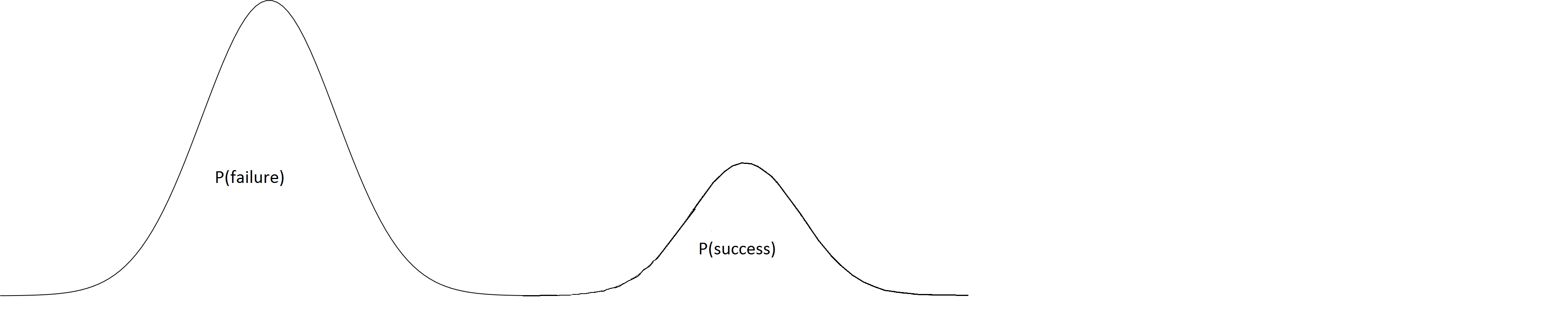
4 – Run the selected model for the test data set.

## Step 2: Cross Validation with Logistic Regression

The cross validation with model selection consists in splitting the data randomly in 10 pieces and then applying the logistic regression to each one. After the model is adjusted, compare what the prediction would be to the rest of the training data, i.e. 9 folds left. You can find the code to create the data [here](https://github.com/dresenhista/StatisticsProjects/blob/master/Project/assignment%202/3-rcode_final.R)

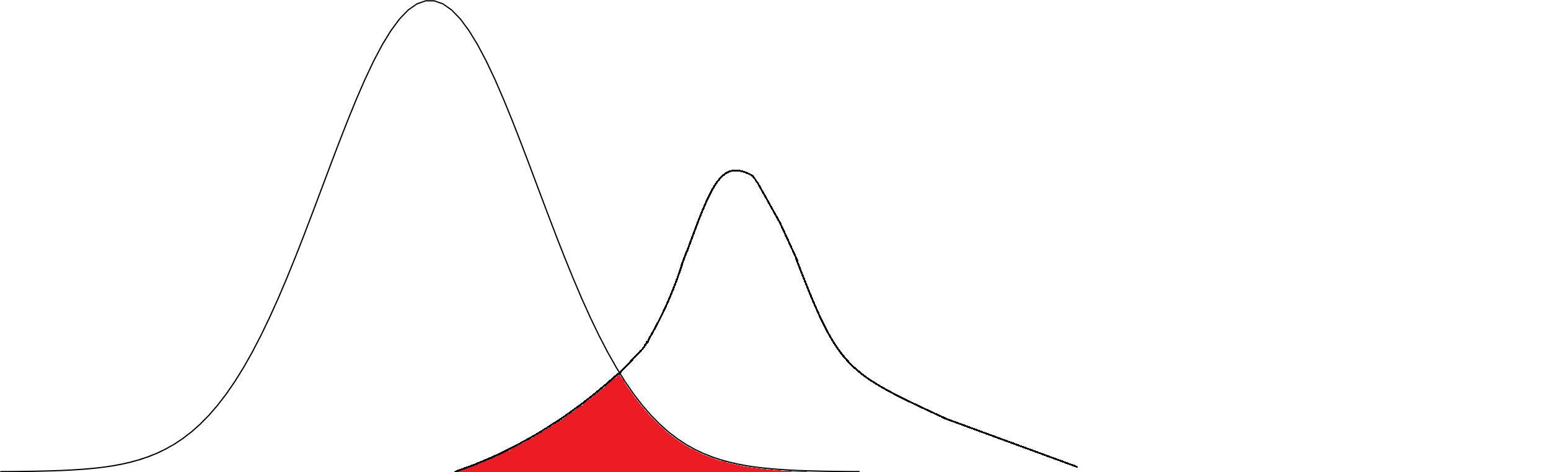
The model in this first stage consists in having the follow function: purchase = function (features). And because majority of the times we have more people not buying then buying, if the model predicts that everyone don’t buy we have a smaller error than the opposite. Therefore any model is biased to predict an unsuccessful purchase.

Ideally we would have something similar to the picture below:



Where there is no intersection between the success and the failures.

But in reality we have something that looks like:



So we need to minimize the intersection in order to reduce the error of predicting a purchase when it actually didn’t happen

Running the cross validation the model that minimizes the false positive and maximizes the true positive is the 5th interaction. See all the detailed data in this [link](https://github.com/dresenhista/StatisticsProjects/tree/master/Project/Cross%20Validation%20pictures).

So we ran the model for the

|  |
| --- |
| These are the parameters of the model:  And these is the qui-square test for the data.  We noticed that not all of the features are relevant to the mode, therefore based on p-value we decided to try the backward selection in order to reduce the number of features. |

This is the table of p-value

## Step N: Feature Selection

For the feature selection

Write details of the step N. If there is any source code that you’d like to share then provide the link of the Github.

# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.