Customer Purchase Predictions

Report for Assignment 2

Group ID: 6
Danas Dvilevičius (000000)
Rik Feunekes (000000)
Louise Giezman (000000)
Vid Tominec (560361)
Hasan Israeli (424293)

Learning from Big Data

Module 2 ”Purchase Prediction”

October 20, 2022

Link to github for code: https://github.com/Hashiraee/assignment2

1 Introduction

Retailers like to maximize the purchase of certain products (or product categories) by using personalized promotions/deals to stimulate sales, improve customer retention or introduce people to new products or new product categories. For example, using personalized discounts or offers to try to get customers to buy products in new or other segments. This raises the question whether purchase decisions can be influenced by using personalized/targeted discounts (by using discount coupons). Since using coupons is an effective way to provide personalized advertisement, we try to determine whether past coupon use have significant predictive power for future purchase decisions. Therefore, our central research question is: *Can past coupon usage and past purchase frequency predict future purchase decisions?*

Since people tend to stick to a particular product (or product category) either by brand loyalty (retention), habit, or do not like to try things outside of their comfort zone, using past purchase frequency can be a decent indicator of future purchasing decisions. Moreover, using personalized coupons gives customers an incentive to try out other products or other product categories since they might not be even familiar with a certain product (or category) or the discount might make them consider trying an alternative product. Therefore, we will analyze both the effects of past purchase habits (by means of past purchasing frequency) and the effects of using personalized coupons on the future purchase decisions of customers.

To answer our proposed research question, we use a dataset which consists of baskets: the week (of purchases), customers (basket of a customer), product (the product a customer purchases) and the price spent on a certain product in euro cents. Note that a ”basket” consist of a collection of multiple products in a week (multiple rows). Furthermore, we also have dataset which consists of the past coupons: which also has the week (of purchase), the customer (which used the personalized coupon), the product (which has a coupon for discount) and the discount in euro cents.

We find that indeed, that both past purchasing decisions and (past) coupon use have significant predictive power for the future purchasing decisions. These findings are important for retailers (especially if they collect purchase data and coupon usage data) because they can use both past purchase data and personalized advertising (by means of discount coupons) to either promote products or product categories to increase sales.

The remainder of this report is organized as follows. First, in Section 2, we give the problem formalization (methods, models and assumptions). Second, in Section 3, we describe our data and do some descriptive analysis. Third, in Section 4, we present our approach and baselines. Fourth, in Section 5, we present our findings/results. Finally, in Section 6, we conclude our analysis, answer the research question and give some recommendations.

2 Problem Formalization

In this section, we introduce our problem, give a mathematical formalization of the problem, introduce our models, and present and discuss our assumptions. Furthermore, we define our machine learning pipeline.

2.1 Problem Definition

To determine the effectiveness of using past purchase frequency and past coupon usage, we use two models which will be used to construct the predictions for future purchases. We will then also compare the predictive performance of the two models and choosing one to make our final purchase predictions for week 90. The before mentioned will help us answer our main research question: *Can past coupon usage and past purchase frequency predict future purchase decisions?*

2.2 Mathematical Formalization of the Learning Problem

In this (sub)section, we define the learning problem in a mathematical way. First, we define the target variable (the variable of interest). Note that the index i (i = 0, 1, ..., 1999) corresponds to a certain customer, the index j (j = 0, 1, ..., 249) corresponds to a certain product and the index k (k = 0, 1, ..., n = 1378720) corresponds to an observation of our customers dataset.

Let $y_{ij,k}$ be the probability of customer i purchasing product j for the k'th observation.

Next, we define our two selected features to predict future customer purchases, being past purchase frequency (past purchases) and past coupon usage.

Let $past\_purchases_{ij,k}$ be the past purchase frequency of customer i purchasing product j for the k'th observation.
Let $coupon\_use_{ij,k}$ be the (past) coupon usage of customer i purchasing product j for the k'th observation.

So our $y_{ij,k}$ is a function of our two selected features being past purchase frequency and (past) coupon usage:

$$y_{ij,k} = f(past\_purchases_{ij,k}, coupon\_use_{ij,k}; \theta). \tag{1}$$

Where θ is a parameter vector corresponding to our 2 selected features (variables). Note, that we can also include past purchasing data for different time horizons / time intervals (for example: weekly, monthly, quarterly and yearly).

2.3 Problem Approach

In this (sub)section we briefly introduce our approach to solve our research question defined in Subsection 2.1. We will tackle our problem by using two models: a Linear Regression model and a Random Forest model. Both of these models are Supervised Learning models, because our dataset(s) consist of data which already has labels (there are 250 product categories, and we need to classify

our test data and validation data into one of those 250 categories). Therefore, using Supervised Learning methods appears to be the most straightforward choice.

2.4 Assumptions

Here we state and discuss the assumptions needed/required for using our selected (supervised learning) models.

2.4.1 Selected Features: Assumptions

For our two selected features introduced in Subsection 2.2, we assume that there is no missing data and that there are no significant outliers. We will discuss/verify these assumptions in the next section (Section 3).

2.4.2 Linear Regression Model: Assumptions

To justify the use of linear regression, we need to state the seven underlying assumptions of Ordinary Least Squares **(OLS)**.
Assumption 1, independent (uncorrelated) explanatory variables.
Assumption 2, the expected values of error terms ($\epsilon_{ij,k}$) are zero.
Assumption 3, the error terms ($\epsilon_{ij,k}$) have equal variance (homoscedasticity).
Assumption 4, independence of the error terms.
Assumption 5, fixed and random values for the error terms.
Assumption 6, using a linear model.
Assumption 7, the error terms are normal independent identically distributed.

2.4.3 Random Forest Model: Assumptions

To justify the use of the Random Forest model/algorithm, we need to verify the following assumptions.
Assumption 1, at each step of building the individual (sub)tree we can find the best split of data.
Assumption 2, while building a (sub)tree we do not use the whole dataset, but a bootstrap sample.
Assumption 3, need to aggregate the individual tree outputs by averaging.
However, there is no formal distributional assumption, since random forest is a non-parametric model and can handle skewed and categorical data (like in our case).

3 Data and Descriptive Analysis

To perform the analysis to answer our research question raised in Section 2.1, we examine the dataset of the customers and coupons usage obtained from professor Sebastian Gabel during the Learning From Big Data **(LFBD)** (2022-2023) course. This section gives an insight to the datasets (included variables, descriptive statistics and key factors). These datasets enable us to extract (or

calculate) the features (past purchase frequency and past coupon usage) we need to determine whether future purchase decisions can be predicted.

3.1 Features and Target Variable Construction

First, we define the target variable (the variable of interest). Note that the index i (i = 0, 1, ..., 1999) corresponds to a certain customer, the index j (j = 0, 1, ..., 249) corresponds to a certain product and the index k (k = 0, 1, ..., n) where n = 1378720 corresponds to an observation of our customers dataset.

Let $y_{ij,k}$ be the probability that customer i purchases product j for the k'th observation.

Our target variable, the probability that customer i purchases product j for the k'th observation, (referred to as $y_{ij,k}$ from here on) is distributed between 0 and 1 since it is a probability. This is relevant because we are trying to predict the probability that a customer i will purchase a certain product j in the future.

Second, we define our first feature, the past purchase frequency.

Let $past\_purchases_{ij,k}$ be the past purchases (frequency) of a product j for customer i for the k'th observation.

The past purchases variable/feature is constructed by summing up the total number of times a certain customer i purchases a certain product j and dividing it by the total number of weeks. **Furthermore,** note that index k is not relevant for this variable, since the only things necessary for this feature is whether a customer i purchased a product j, how often a customer purchased that product, and what the total number of weeks is (90 weeks for our whole dataset).

Third, we define our second feature, the (past) coupon usage.

Let $coupon\_use_{ij,k}$ be the (past) coupon use of a product j for customer i for the k'th observation.

This feature is constructed by using the coupons dataset and appending the columns of the coupons dataset to the customers dataset for the corresponding customers, since the coupons dataset have a column indicating to which customer they were assigned/given.

3.2 Data Description

A total of 1378720 observations are included in the dataset provided by the LFBD course. In Table 1, we describe the customer/baskets dataset and its variables. In Table 2, we describe the coupons dataset and its variables. Note that the **price** and **discount** variables are given in euro cents.

Variable	Description
$Week$	The week of purchases
$Customer$	One of the 2000 customers
$Product$	One of the 250 products purchased
$Price$	The amount spent on a product

Table 1: Description of the Variables in the Customers/Baskets Dataset

Variable	Description
$Week$	The week of coupon use
$Customer$	The customer who used the coupon
$Product$	The product which is on the coupon
$Discount$	The discount a customer receives

Table 2: Description of the Variables in the Coupons Dataset

3.3 Data Characteristics

Next, we provide some data characteristics to provide an insight into the variables presented in Section 3.2.

To understand customers' future purchase decisions, we use the past purchases (frequency) and the past coupon usage to determine the future value of $y_{ij,k}$. The past purchases and past coupon usage, which we will use as key factors into understanding customers' future purchase decisions, provide a significant amount of information on customer purchases. Furthermore, the past coupon usage captures the impact of (personalized) advertisement which provides retailers an insight into the effectiveness of personalized advertising using coupons.

Next, we provide some data characteristics to provide an insight into some of the variables discussed in Table 1 and Table 2. We examine the mean, standard deviation (stdev), minimum and maximum. **Note**, we denote an "x" (a cross) for the cells that do not make sense (for example, the mean or variance of a customer).

	Week	Customer	Product	Price
$Mean$	x	x	x	584
$Stdev$	x	x	x	97
Min	0	0	0	234
Max	89	1999	249	837

Table 3: Data Characteristics of the Variables in the Customers/Baskets Dataset

	Week	Customer	Product	Discount
$Mean$	x	x	x	25
$Stdev$	x	x	x	10
Min	0	0	0	10
Max	89	1999	249	40

Table 4: Data Characteristics of the Variables in the Coupons Dataset

We observe that on average, customers spent 5.64 euros on products and got an average discount of 25 cents through coupons on qualifying products.

3.4 Data Preparation

The dataset is complete and does not contain any missing observations, or any (significant) outliers, therefore we use the dataset as is. Even if we find outliers, we will not exclude those observations since they provide valuable information. Furthermore, we will use the purchase data from the Customers/Baskets dataset to construct one of our features: the past purchase frequency. First,

this will serve as a baseline/benchmark model (more details in Section 4). Second, it will allow us to analyze the relationship between future purchase decisions and past purchases.

4 Approach and Baselines

In this section, we introduce our selected learning algorithms, present the definition of our model in mathematical terms (if possible), and discuss our approach and baselines. In addition, we will explain which model we selected to construct our predictions. With the predictions we will get using our selected model, we will be able to answer our research central question (defined in Section 2.1). The models we selected to use for our predictions are both supervised learning models, a linear regression model and a random forest model.

For our analysis, we do not plan to use multiple sequential model stages (because we did not select models which require that). We will also not stack models, but we will use two different supervised learning models in order to compare their performance against each other to determine which model produces the (significantly) more accurate predictions. Using multiple models for the same target variable ($y_{ij,k}$) gives us an idea about whether a model gives us more accurate predictions. Moreover, we will apply the linear regression and random forest separately, and not stack models (so we do not use multiple models to derive our target variable). Note, that for our random forest model, all the trees are built independently. Therefore, since there is no dependence between trees, all trees can be built in parallel.

The predictions we will make using one of our two selected models will be the predictions for the future purchase decisions customers make using past purchase frequency and coupon usage. This will help retailers who use personalized coupons as a way of personalized advertising understand whether the coupons indeed have significant effect on consumer purchasing behavior. And if retailers have access to both the coupons they provide and the past purchasing data of the consumers, they can target customers more effectively in order to increase brand awareness or increase/maximize sales. Past purchase data (or frequency) is a decent indicator for customer retention (or brand loyalty) and overall consumer behavior.

4.1 Data Preparation

After analyzing our data, we look for missing data and outliers to prepare the data for our analysis. The process for data preparation/processing is discussed in the data section (Section 3).

4.2 Data splits

Our data consist of 90 weeks. We use the first 88 weeks for training our models, to ensure we have enough data to make our predictions as accurate as possible. Next, we use the data of week 89 as our testing dataset. Finally, we will use the last week (week 90) as our validation dataset.

4.3 Model Training

We will use the first 88 weeks to train our models to provide sufficient data for our models. For the linear regression model, this will yield estimators which are as 'accurate' as possible. For our random forest model using the 88 weeks of data will provide enough data to make the classifications as accurate (low bias) as possible. We will evaluate whether our model generalizes well to unseen data (to the general population) by using the our 1 week test data (week 89) and our 1 week validation data (week 90) so 2 weeks in total. If we find that the performance metric we use shows us a significant improvement from our baseline model, we can conclude that our model generalizes adequately well to unseen data (more about the performance metric and baseline model in the following subsections).

4.4 Model Selection

4.5 Linear Regression Model

First, we use a linear regression. Here, we define our models in terms of our 2 selected features:

$$y_{ij,k} = \gamma_0 + \gamma_1 * past\_purchases_{ij,k}^{full} + \gamma_2 * past\_purchases_{ij,k}^{30\_weeks} + \gamma_3 * past\_purchases_{ij,k}^{4\_weeks} + \gamma_4 * past\_purchases_{ij,k}^{12\_weeks} + \gamma_5 * coupon\_use_{ij,k} + \epsilon_{ij,k}. \quad (2)$$

Where *past_purchases*$_{ij,k}^{t}$ corresponds to the past purchase frequency feature for a period of t weeks where $t = \{full, 30\_weeks, 4\_weeks, 12\_weeks\}$ corresponding to the all the weeks, 30 weeks, 4 weeks (a month) and 12 weeks (a quarter) respectively. And where *coupon_use*$_{ij,k}$ corresponds to the past coupon usage feature.

Our estimator for γ_0 corresponds to the intercept, $\gamma_1, \gamma_2, \gamma_3$ and γ_4 capture the effects of past purchases of customers for respectively all the weeks, 30 weeks, 4 weeks (a month) and 12 weeks (a quarter) and γ_5 captures the effect of (past) coupon usage.

These estimators can be used to construct our predictions and also give an insight to the significance of our selected features.

4.6 Random Forest Model

Second, we use a Random Forest model. Since models like the Random Forest Model are known as 'Black Box' models, we can not give an explicit mathematical definition. However, the probability of a certain customer (i) purchasing a certain product (j) for a certain observation (k) is still a function of the past purchases and past coupon usage, hence can be generally represented as:

$$y_{ij,k} = f(\sum_{m \in t} past\_purchases_{ij,k}^{m}, coupon\_use_{ij,k}; \beta). \quad (3)$$

Where β is a parameter vector corresponding to our selected features (variables) and $t = \{full, 30\_weeks, 4\_weeks, 12\_weeks\}$ corresponding to the all the weeks, 30 weeks, 4 weeks (a month) and 12 weeks (a quarter) respectively.

A random forest classification model builds on top of decision trees, another way of classifying information based on characteristics. Random forests use bagging (or bootstrap aggregating) to build a series of uncorrelated trees and average them for a better, less variant prediction. The way boosted trees are grown, is dependent on the concept of information gain, as measured by Shannon's entropy formula:

$$H(X) = \mathbb{E}\left[I(X)\right] = \mathbb{E}\left[-log(p(X))\right]. \tag{4}$$

Here, I(X) is defined as information contained by X and thus the expected information gain is defined as $-log(p(X))$, the probability of X occurring. Using this metric, the nodes of a decision tree nodes are grown until a certain level of information gain is reached.

Using bagging, which selects random samples with replacement, defined as $b = \{1, 2, 3, ..., B\}$ a decision tree T_b can be grown for each bootstrapped sample. The model first randomly selects m number of variables out of a total of n variables, picks the best split ratio between the m variables and splits the node into two daughter nodes. Repeating this for all b yields an ensemble of decision trees $\{T_b\}_1^B$. When it comes to making a prediction for say y, $\hat{C}_b(y)$, can be defined as a majority vote between the multiplicity of decision trees, or:

$$\hat{C}^B_{random\_forest}(y) = \text{ majority vote } \{\hat{C}_b(y)\}. \tag{5}$$

Reference: (https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf).
The assumptions which need to be satisfied for the Random Forest model, are stated in Section 2.4.

4.7 Baseline Model

For our baseline model, we use a simple model which only consists of one feature. The one feature it does contain is also part of our two selected features, namely the past purchase frequency ($past\_purchase\_frequency_{i,j,k}$). That serves as a decent baseline and does have a adequate cross entropy loss value.

4.7.1 Baseline Model: Implementation

Our baseline model is constructed by summing up all the purchases of a certain product j by a certain customer i and dividing the total sum by the total number of weeks. Which gives us a baseline for every customer and product combination.

4.7.2 Baseline Model: Relevance and Selection

Moreover, not only is it a decent baseline, but it is an adequate measure of customer purchase behavior. Since a significant amount of consumers tend to either stick to a certain product (or certain category), it is by itself a good indicator of future customer purchasing decisions. We selected this baseline because it not only uses one of our selected features (past purchase frequency), but it also has decent predictive power over future purchase decisions by itself.

These baselines are not only simple to implement, relevant and decently performing but can also be used as a benchmark for prediction performance between our models. If past purchase data itself is not available to retailers, it makes personalized advertising significantly harder because retailers have less information to make relevant coupons (or interesting coupons) for consumers. Also, the insight we can gain by comparing our baseline with our models is relevant. Since no improvement (or an insignificant improvement) against the baseline model implies that our model is not that useful.

4.8 Performance Monitoring

We monitor (or check) the performance by using (binary) cross entropy loss. The perfect prediction(s) have a loss of 0, a lower value indicates significant better performance in terms of (prediction) accuracy of a chosen model.

4.9 Findings / Insights

We will present our findings and insights in the next section (Results Section).

5 Results

In this section, we first present our findings, determine which model has the better performance and finally we discuss the results of our predictions. We evaluate model quality by using (binary) cross entropy (log-loss) on the predictions produced by one of our two models. A model with (significantly) more accurate predictions will yield a value closer to 0 (perfect predictions have a log-loss of 0). By using the log-loss on the predictions obtained from either models, we can get an insight into the accuracy/quality of the models. These metrics are meaningful because they show us which model has better predictive power on future customer purchase decisions.

5.1 Results: Linear Regression Model

We will consider the results (the estimates of the estimators) of our linear regression model of Equation (2) in Section 4.5. For the estimators and their interpretation you can refer to Equation (2) and explanation in Section 4.5.

	Estimate	Standard Error	P-value
$\gamma_0\ (constant)$	0.0003	0.000	0.219
γ_1	1.075	0.011	0.000
γ_2	0.024	0.013	0.061
γ_3	-0.066	0.004	0.000
γ_4	-0.062	0.008	0.000
$\gamma_5\ (coupon)$	0.129	0.006	0.000

Table 5: Value of Estimates

We find that the constant in not significant. Moreover, note that γ_1 (past purchase data of the whole sample), γ_3 (past purchase data of 4 weeks or a month) and γ_4 (past purchase data of 12 weeks or a quarter) are all significant even at a significance level of 1%. Also, the estimate for the (past) coupon use γ_5 is significant at a significance level of 1%. Hence, both the past purchasing data (for different time intervals) and the (past) coupon usage are (very) significant predictors of consumer purchasing decisions. This means that the predictions using this linear regression model will be decently accurate. These results are in line with our results, we indeed would expect that both past purchasing behavior and coupon usage have significant effects on (future) purchasing decisions.

5.2 Comparison of Models

In this (sub)section, we will consider which model gives us the more accurate predictions in terms of a lower log-loss values (closer to 0 is better).

	Baseline Model	Linear Regression Model	Random Forest Model
$log - losses$	0.10035	0.07883	0.06749

Table 6: Value of log-losses

After finding the log-losses for both our linear regression model and our random forest model, we can compare them with the log-loss of our baseline model (which is only based on the past purchasing data). We find that both our models are a significant improvement from our baseline model. This is not a surprise, since our models also included the feature we used for our baseline model. Therefore, our approach of choosing the past purchasing data as our 'simple' baseline model and including multiple time horizons and past coupon usage in our linear regression and random forest model is very decent.

Therefore, we can indeed say that there is significant evidence for the past purchasing data and past coupon usage being able to predict future purchasing decisions of customers. Moreover, the past purchasing data is the feature that is most influential for our results, since it has the 'stronger' effect on the probability of a customer purchasing a certain product.

Where our linear regression may fail is the fact that a linear regression assumes a linear relation between the dependent and independent variable, which might not be the case here whereas the random forest method does not have a formal distributional assumption which (may) explains the difference in performance between our two models. Hence contributing to the overall best performance among our models.

6 Conclusion

Retailers use personalized advertisement by using coupons to try to increase purchases, increase brand awareness or enter a new segment of a market. A insight into the future purchases of consumers gives retailers an edge in maximizing their profits by (for example) increasing sales. Accurate predictions into future consumer purchases are therefore a convenient way for retailers to determine the effects of both past purchasing habits and personalized advertising through coupons. In order to find predictions which are as accurate as possible, we compare a baseline model (with only past purchase frequency as feature) against a linear regression model which uses past purchasing data/frequency (with multiple time horizons such as monthly, weekly, quarterly etc.).

In this research, we therefore formulate an answer to our original research question: *Can (past) coupon usage and past purchasing frequency predict future purchase decisions?*

To answer the research question, we consider our baseline model (with only the past purchasing frequency feature), our linear regression model (with past purchasing frequency on multiple time horizons) and past coupon usage and our random forest model using the same features as the linear regression model mentioned before. We find that predictions using the Random Forest model significantly outperform the both the baseline model and the linear regression model in terms of cross entropy loss. Therefore, the best predictions are obtained using the Random Forest model with the past purchasing frequency (on multiple time horizons) and with past coupon use. The before mentioned implies that indeed the past purchasing frequency and coupon use are significantly import features into predicting the future purchase predictions of consumers.

Finding the significance of both using our Random Forest model, but more importantly the past purchasing frequency and coupon use, gives potential retailers an advantage for personalized advertising to consumers and also obtaining/collecting purchase data in order to optimally target customers for certain product (or categories). For retailers having access to both sales data and coupon usage data, the predictions of future purchases will be significantly more accurate than only having past coupon usage data. Therefore, retailers should either try to obtain past purchasing data or try to use their own platform for sales to be able to collect and analyze such data.

For future research, one may consider adding more data features (of customers) to be able more effectively target certain customers. Having additional data on consumers may result in more effective targeting. For example, the gender or age may also provide valuable information on future purchasing decisions (purchasing habits across different product categories). Moreover, we can also consider stacking models (combining regression with random forest by applying them sequentially) may result in even more accurate predictions than only using one model by itself. We may also consider using other performance metrics for the predictions, like mean absolute error or root mean squared error.