Machine learning Project

Federico Spatola

June 17, 2020

1 Introduction

Sales forecasting is important and beneficial for managers and companies as it helps them with the following (Furseth and Stumbaugh, 2017):

- Approximate the demand for their items
- Better manage inventory
- Financial and strategic planning for the company's growth

In this analysis, the goal was to predict which customers spend more than a fixed amount. Expecially, to build a classification model which gives a response taking in input some information about a customer.
If the predicted total amount for the purchases of a new customer is more than the fixed amount, the aforementioned model says yes, otherwise it says no. Having a similar model could be a real wish of a store, for example the store could have the intention of sending a 20 euro gift card as a present to the 30% customers who are expected to spend more.
The store interested to give gift cards to his customers could also decide of following a multi-class criterion. For instance, four classes of customers could be defined: the ones who spend less than a first amount represent the lower sales class, the ones who spend more than the first amount but less than a second amount represent the medium sales class, and so on a medium-high class and an high class. The store could give 20 euro gift card present to the customers who belong to the high class, 10 euro gift card to the medium-high class and 5 euro gift card to the medium sales class.
To help the store, in this project, there were built both the binary-classifier and the multiclass-classifier.

2 Data preprocessing and data visualizations

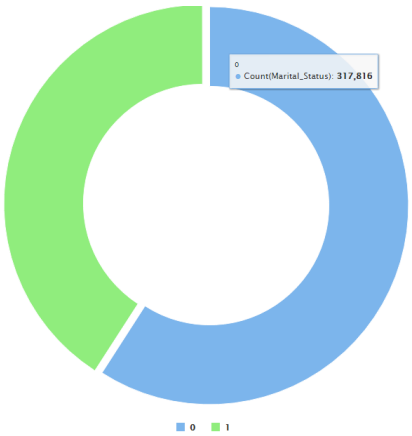
The Dataset, used in this study, contains eleven variables which give information about the sales details of a retail store, during the Black Friday. Before building a machine learning model, it is reasonable to get a better understanding about the data and thinking to transformations of the variables.
Some details concerning the data were not actually clear. In particular, the variable *Occupation* represents the ID number related to the occupation of each customer but for each ID is not specified the corresponding kind of job: for example the number 1 could mean that the customer is a doctor or a teacher or any other professional, the same happens for the indexes 2,3 and so on.
Similarly, there are not specified information about the meaning of the indexes related to the variables *Product-Category-1, Product-Category-2, Product-Category-3*.

	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
Occupation	1.000000	0.024691	-0.008114	0.006792	0.013452	0.021104
Marital_Status	0.024691	1.000000	0.020546	0.001146	0.019452	0.000129
Product_Category_1	-0.008114	0.020546	1.000000	-0.040730	0.229490	-0.314125
Product_Category_2	0.006792	0.001146	-0.040730	1.000000	0.543544	0.038395
Product_Category_3	0.013452	0.019452	0.229490	0.543544	1.000000	-0.022257
Purchase	0.021104	0.000129	-0.314125	0.038395	-0.022257	1.000000

Figure 1: Correlation matrix

With regard to the last two variables just mentioned, i.e. *Product-Category-2* and *Product-Category-3*, they have some missing values and also they are correlated. It was decided to ignore the variable *Product-Category-3*, from this analysis. Instead, the variable *Product-Category-2* was holded and the missing cells were filled with the value 0.
To improve the understanding of the data, it is relevant to provide some visualizations. There are more sales related to male customers and also it is possible to claim that a man spends more on average. It is interesting reasoning in the same way with the variable *marital-status*. There are more merried people's purchases but it is relevant that married people and not married people spend on average the same amount.

The cluster of 26-35 years old customers spent more than others (nearly 2 millions), but again it would be wrong claiming that 26-35 years old people spent on average more than younger or older customers.




```
number of  purchases from wedded and amount of purchases from wedded: 317816 , 2966282625
number of  purchases from not wedded and amount of purchases from not wedded: 219760 , 2051378878
rate not wedded purchase: 9334.632681106661
rate wedded purchase: 9333.333202230222
```

Figure 2: Number of rows per marital status

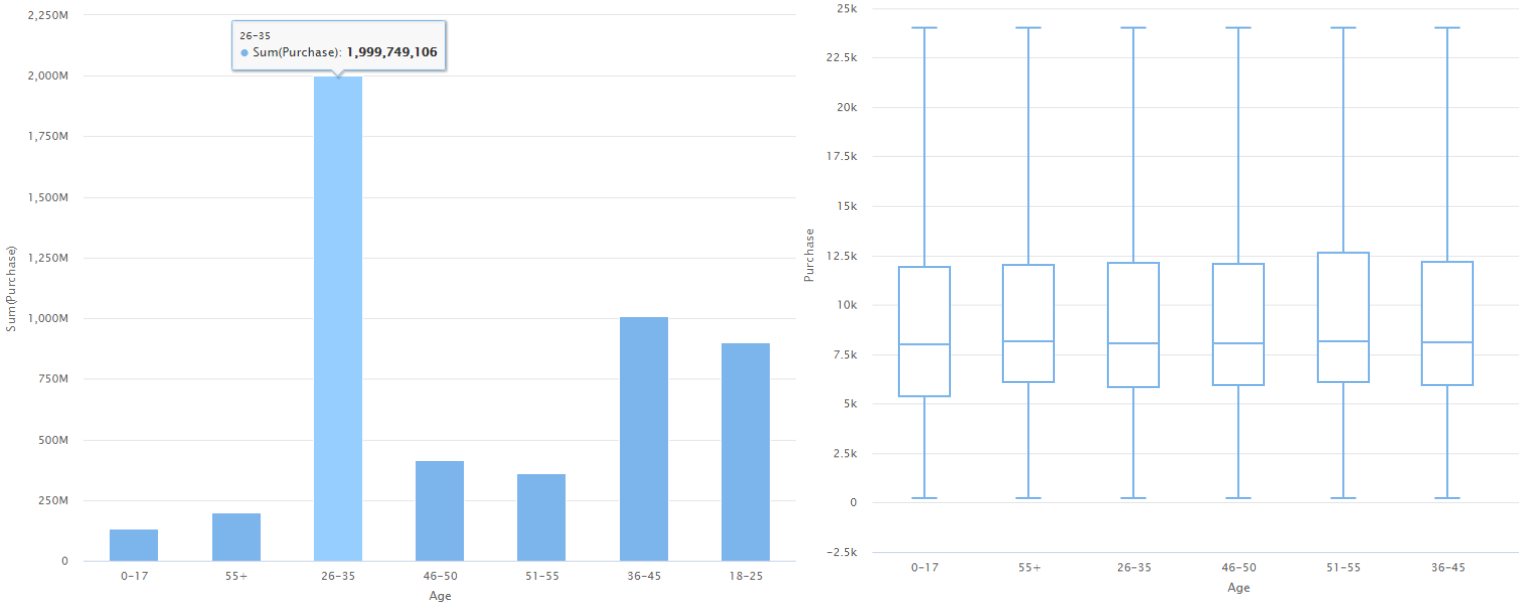


Figure 3: Number of rows per age

3 Data transformation

At the beginning, the idea was to use for the prediction the independent variables: Gender, Age, Occupation, City-Category, Stay-In-Current-City-Years, Marital-Status,Product-Category-1 ,Product-Category-2. In particular, to run the logistic regression model all the variables should be numerical. Therefore, it was necessary to do some data transformations. The most interesting transformations concern the variables *Occupation,Stay-in-current-city,Product-Category-1, Product-Category-2.* All of them were transformed in dummy variables. The alternative was to consider the indexes for these variables as numbers but it would be wrong. Indeed, these variables are ordered but there is no intrinsic ordering for them.

4 Binary-classification Model

In order of conducting the binary-classification analysis, it is possible to define a new variable, *boolean-purchase*. For every observation, it gets value 1, if the attribute *Purchase* is $>$ 11500, and 0 if *Purchase* is $<$ 11500. The value 11500 is the 7-th quantile of the sample distribution for the variable *Purchase*.
The dataset was splitted in a 75% training dataset and a 25% test dataset. The problem is a binary classification. There is a wide literature concerning this problem. Expecially, some suggested models to tackle this problem are logistic regression, discriminant analysis, support vector machine, classification tree and random forest. Looking at the accuracy, the logistic regression model resulted the best.

In the figure no.5 is rappresented the ROC curve, plotting the true positive rate (TPR) against the false positive rate (FPR). The models random forest and logistic regression are very close in outcomes and they both bette rperform than the Naive-Bayes classifier.

At the begining, the present analysis was conducted without doing great trasformations to the independent variables and the best model was random forest with an accuracy of 0.88. Using the dummy variables improved considerably the logistic

Model	Accuracy
Logistic regression	0.8983816362215856
Naive-Bayes classifier	0.8797425499460545
Decision Tree	0.888403586442948
Random Forests Classifier	0.89277874920942

Figure 4: Accuracy per different models

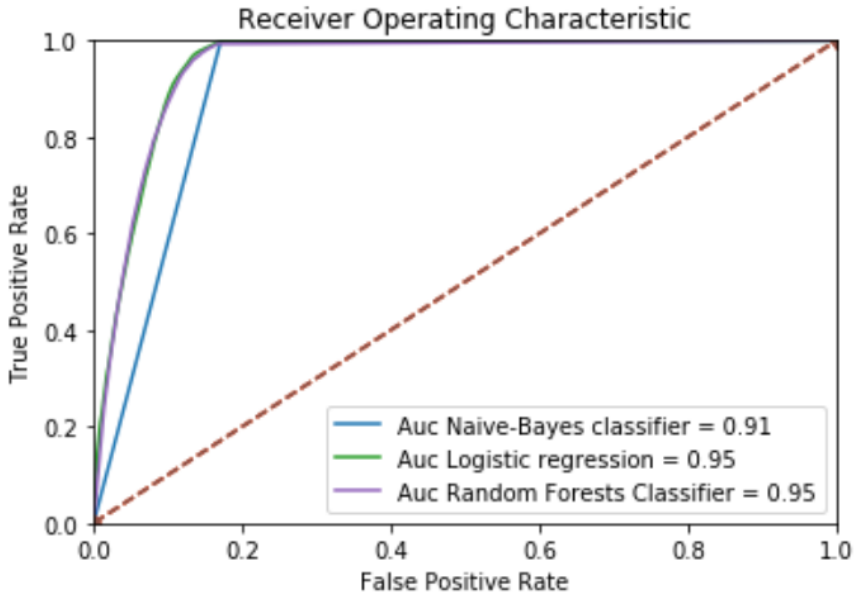


Figure 5: ROC curve

regression results, while the accuracy for the models decison tree and random forest remained constant.

5 Multiclass model

At this time, the target became to predict if the total amount of purchases related to a store would be lower than the 25th percentile (No-gift class) or between the 25th and the 50th percentiles (5-gift class) or between the 50th and 75th percentiles (10-gift class) or lastly, greater than the 75th percentile (20-gift class). Increasing the problem's complexity means getting a lower accuracy.
At the beginning, using this discretization and the same train-test datasets used before, the predictions were bad. Thus, this time, it will be considered a test dataset of only 10% of the full data. In order to have a better assessment the cross-validation technique will be used. The accuracy of tandom forest on the test dataset is 61.25%.

accuracy: 61.07% +/- 0.08% (micro average: 61.07%)

	true No-gift	true 5-gift	true 10-gift	true 20-gift	class precision
pred. No-gift	90763	53429	26471	1	53.18%
pred. 5-gift	13053	46419	37655	2	47.79%
pred. 10-gift	3002	5078	15266	4552	54.72%
pred. 20-gift	11078	11301	22750	142999	76.01%
class recall	76.99%	39.94%	14.95%	96.91%	

Figure 6: Confusion matrix related to the cross-validation

accuracy: 61.25%

	true 10-gift	true 20-gift	true No-gift	true 5-gift	class precision
pred. 10-gift	1746	522	358	570	54.63%
pred. 20-gift	2567	15978	1185	1315	75.92%
pred. No-gift	2966	2	10237	6025	53.23%
pred. 5-gift	4025	0	1294	4968	48.29%
class recall	15.45%	96.82%	78.30%	38.58%	

Figure 7: Confusion matrix related to the Test Dataset

6 Findings and Proposals

Using the binary-classification, for every ten predictions the model gives the right response nine times. Therefore, the results are enough satisfactory. It would be interesting setting a regression analysis without discretizing the variable *Purchase*. Browsing the internet, expecially on the website Kaggle.com, there are many regression analysis, that have been developed and published by some users. In these studies, one of the most widely used models was Gradient boosting Regression (Black Friday Regression Analysis 2020).

References

[1] Furseth, P., and Stumbaugh, J. (2017): *The Importance of Sales Forecasting.* https://www.orm-tech.com/app/uploads/2017/08/ORM-Technologies-The-Importance-of-Sales-Forecasting.pdf

[2] Black Friday Regression Analysis (2020): `https://www.kaggle.com/roshansharma/black-friday-regression-analysis#4.-Gradient-Boosting-Regression`