# Data-Driven Customer Classification for E-commerce Optimization

Kausik Chattapadhyay

Bellevue University

DSC680: Applied Data Science

Prof. Amirfarrokh Iranitalab

04/28/2024

**Title: Data-Driven Customer Classification for E-commerce Optimization**

**Business Problem:**

In the rapidly evolving landscape of e-commerce, understanding customer behavior is crucial for optimizing marketing strategies, enhancing user experience, and driving sales. The business problem addressed in this project is to develop a data-driven classification system that categorizes customers based on their purchase habits, allowing for targeted marketing and personalized recommendations.

This project aims at analyzing the content of an E-commerce database that lists purchases made by ~~4000 customers over a period of one year (from 2010/12/01 to 2011/12/09). Based on this analysis, I develop a model that allows to predict the purchases that will be made by a new customer, during the following year and this, from its first purchase.

**Datasets:**

The data for this project will be sourced from the E-commerce platform's database, which has details of purchases made by customers over a period. The dataset includes information such as customer demographics, purchase history, product categories, transaction amounts, and timestamps.

https://www.kaggle.com/datasets/carrie1/ecommerce-data

This data holds 8 variables that correspond to:

**InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it shows a cancellation.
**StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
**Description**: Product (item) name. Nominal.
**Quantity**: The quantities of each product (item) per transaction. Numeric.
**InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
**UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
**CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
**Country**: Country name. Nominally, the name of the country where each customer lives.

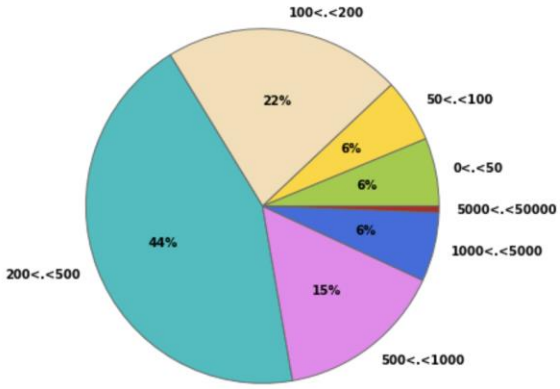
## Background/History:

E-commerce platforms generate vast amounts of transactional data, providing valuable insights into customer preferences and trends. Traditional marketing strategies often lack granularity and fail to capitalize on individual customer behavior. This project aims to leverage machine learning techniques to segment customers into distinct categories, enabling businesses to tailor their offerings and promotions effectively.

## Data Explanation:

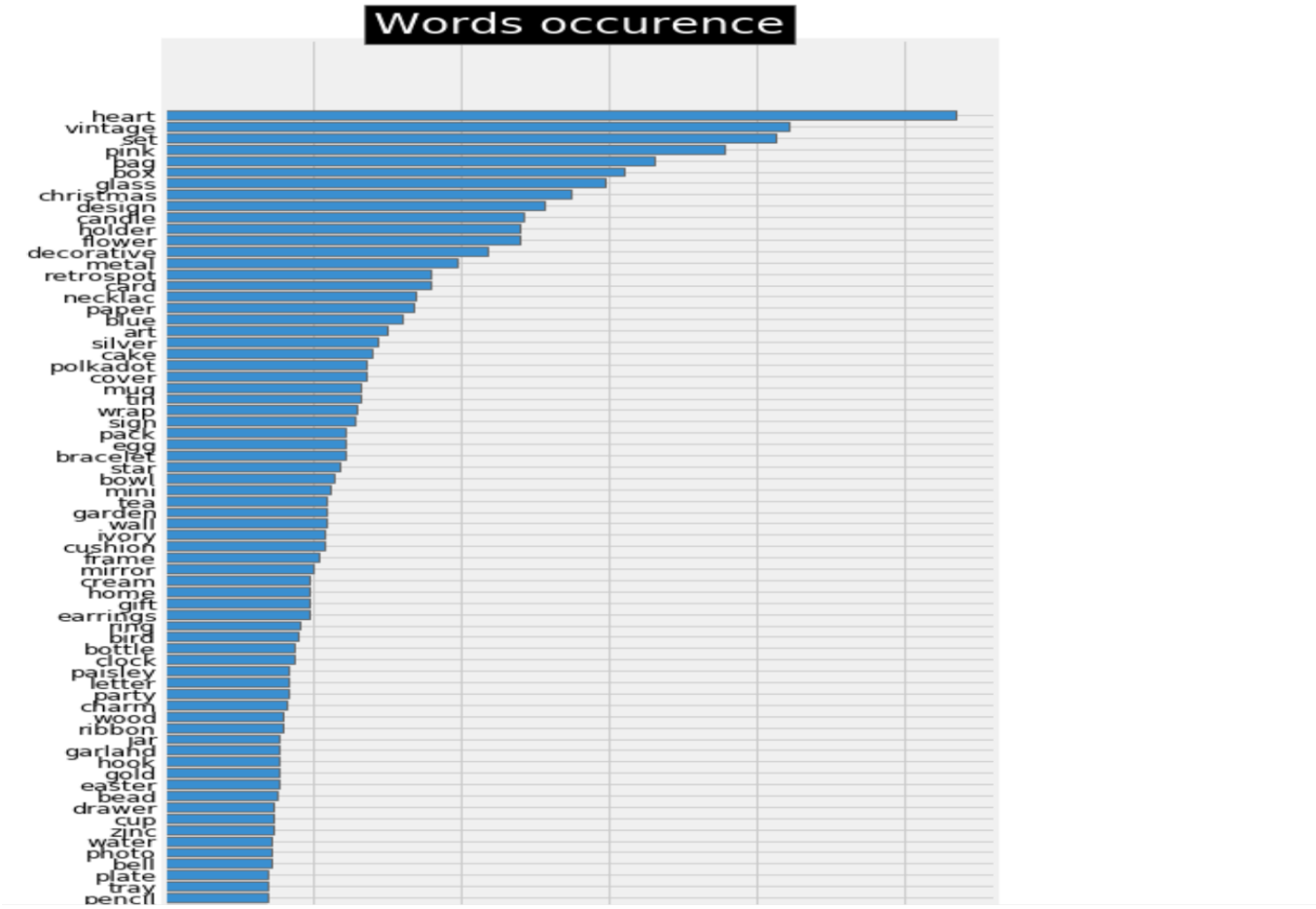
The dataset used in this project contains purchase data from an e-commerce platform over a period of one year. It includes information such as CustomerID, InvoiceNo, Basket Price, product categories, and InvoiceDate. Data preparation involved cleaning, feature engineering, and splitting into training and testing sets. A Data Dictionary was created to define variables and their meanings.

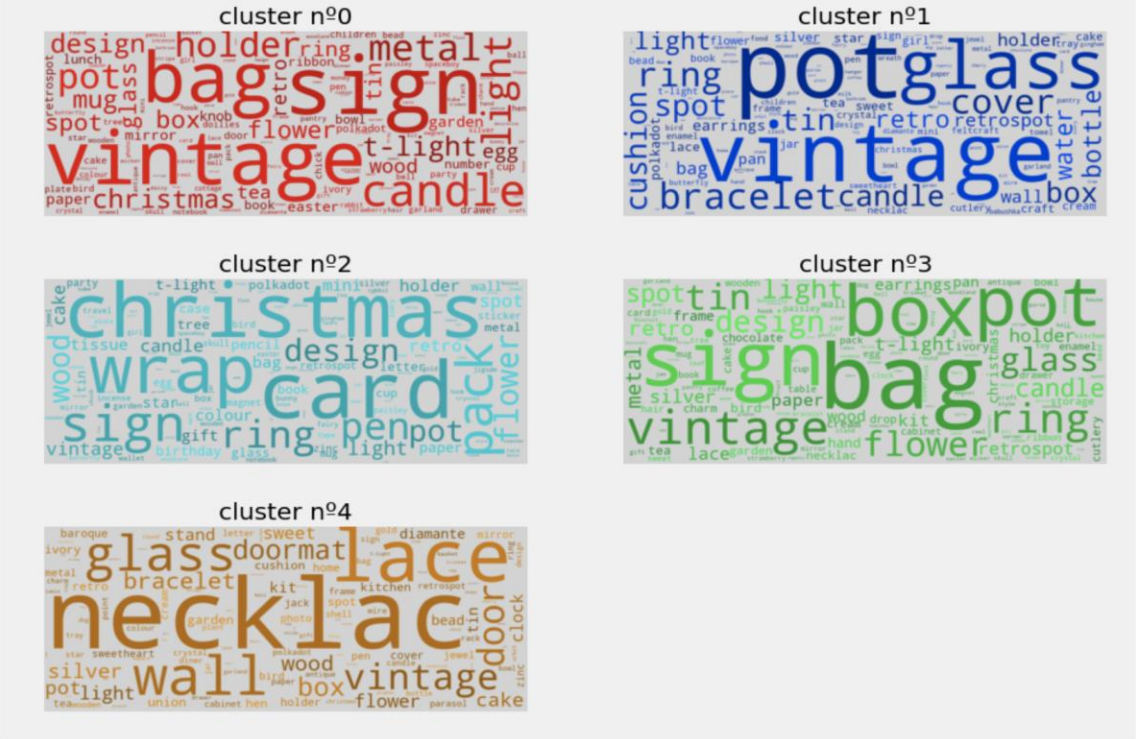

We see that the dataset is largely dominated by orders made from the UK.




It can be seen that the vast majority of orders concern relatively large purchases given that ∼65% of purchases give prizes in excess of £ 200.

How the purchases are divided according to total prizes.


Words occurence
heart
vintage
set
pink
bag
box
glass
christmas
design
candle
holder
flower
decorative
metal
retrospot
card
necklac
paper
blue
art
silver
cake
polkadot
cover
mug
tin
wrap
sign
pack
egg
bracelet
star
bowl
mini
tea
garden
wall
ivory
cushion
frame
mirror
cream
home
gift
earrings
ring
bird
bottle
clock
paisley
letter
party
charm
wood
ribbon
jar
garland
hook
gold
easter
bead
drawer
cup
zinc
water
photo
bell
plate
tray
pencil



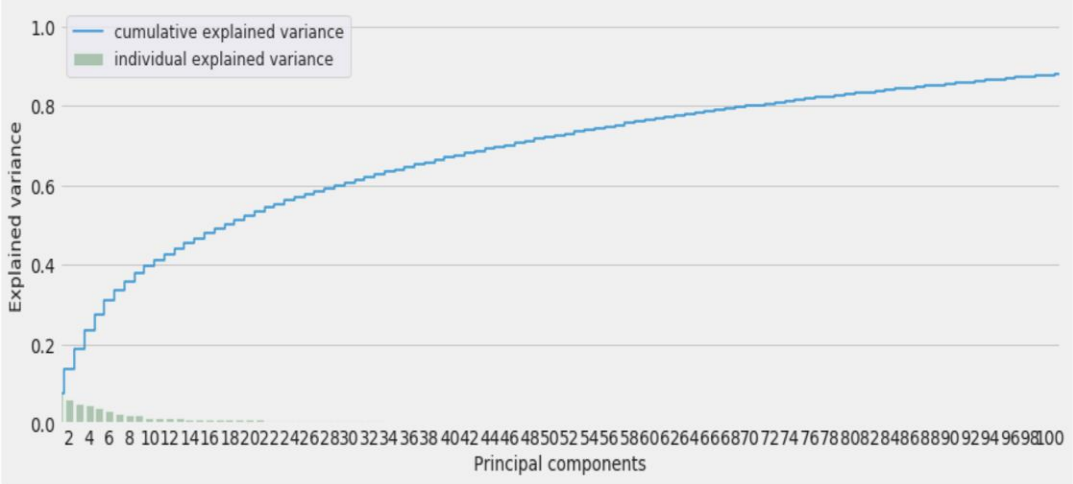

From this representation, we can see that for example, one of the clusters contains objects that could be associated with gifts (keywords: Christmas, packaging, card, ...). Another cluster would rather contain luxury items and jewelry (keywords: necklace, bracelet, lace, silver, ...). Nevertheless, it can also be observed that many words appear in various clusters and it is therefore difficult to clearly distinguish them.

**Methods:**

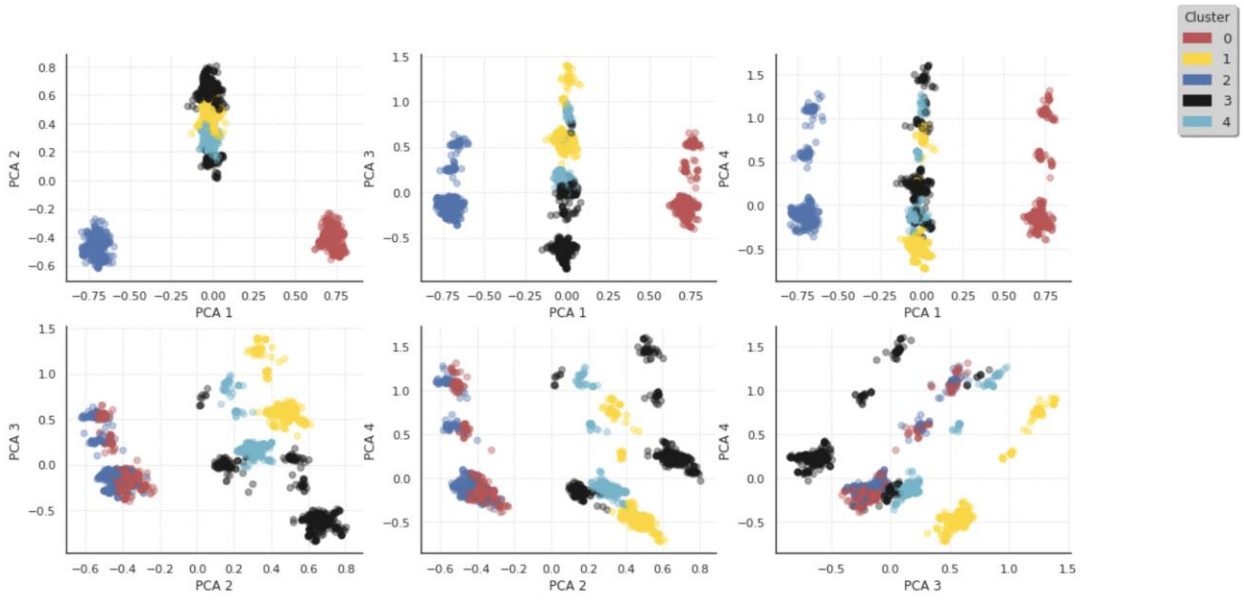
1. **Data Preparation:** Splitting the dataset into training and testing sets, grouping customer orders, and standardizing variables.
2. **Customer Categories Creation:** Utilizing PCA and k-means clustering to create customer clusters based on purchase behavior.
3. **Classification of Customers:** Testing multiple classifiers including Support Vector Machine, Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, AdaBoost Classifier, and Gradient Boosting Classifier.
4. **Ensemble Voting Classifier:** Combining predictions from multiple classifiers using a VotingClassifier approach to improve classification accuracy.

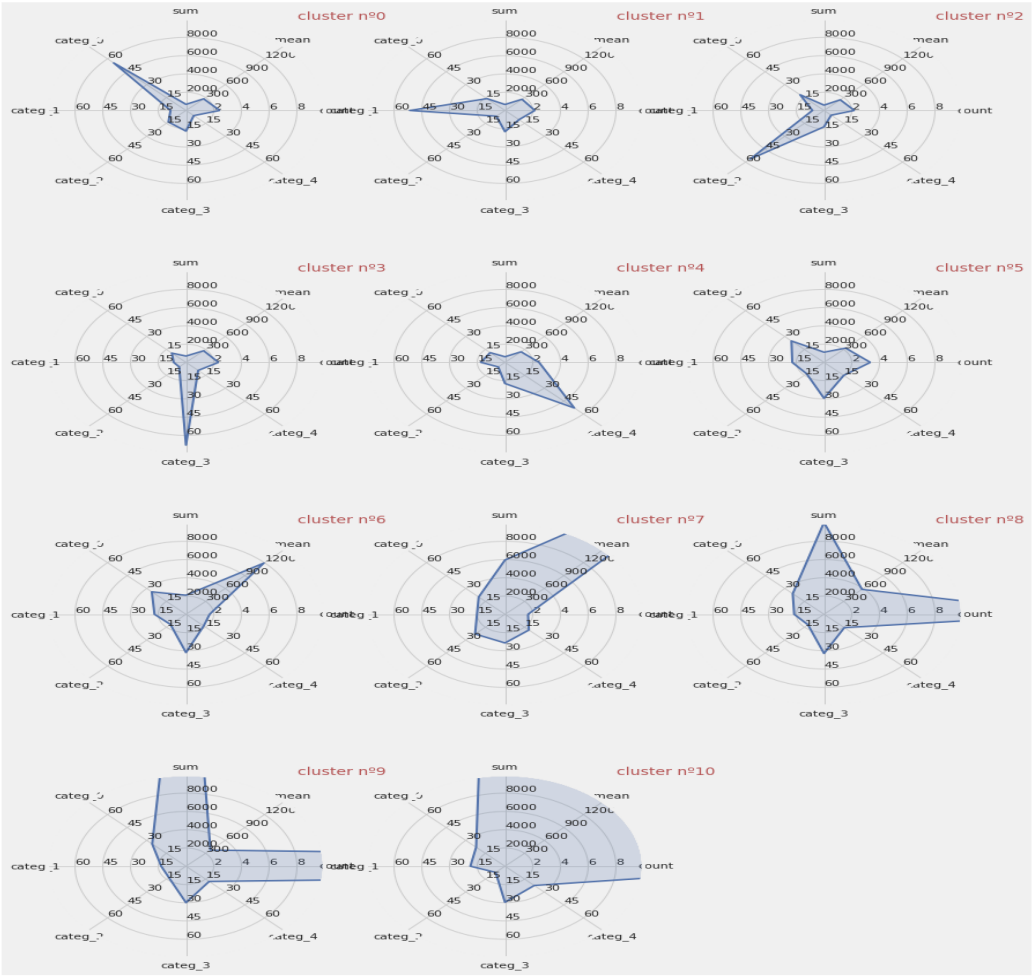
**Analysis:**

Evaluation metrics such as precision scores, learning curves, and confusion matrices were used to assess the performance of classifiers and ensemble methods. The analysis revealed promising results in accurately categorizing customers based on their purchase habits.



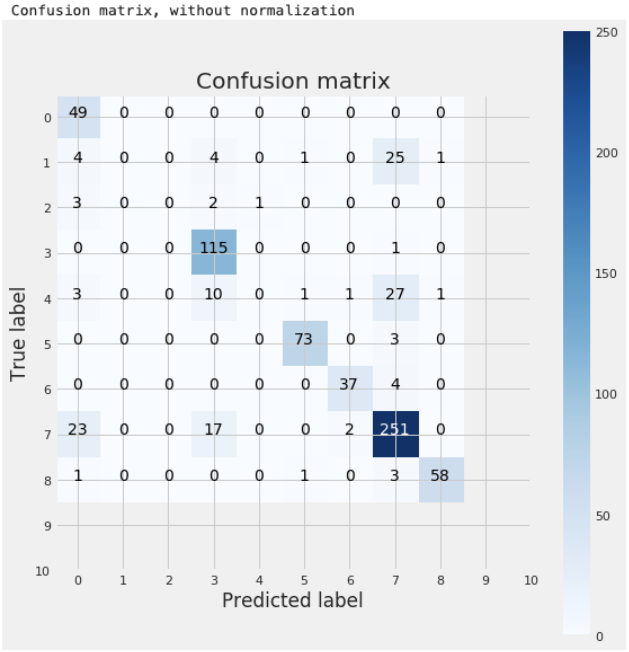

We see that the number of components required to explain the data is extremely important: we need more than 100 components to explain 90% of the variance of the data. In practice, I decide to keep only a limited number of components since this decomposition is only performed to visualize the data:



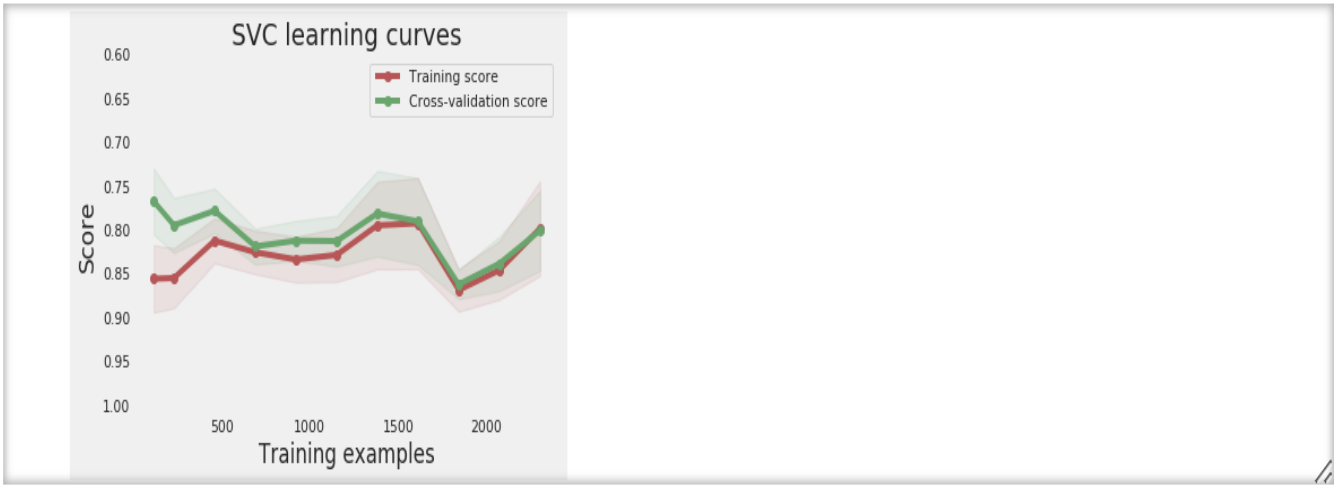




It can be seen, for example, that the first 5 clusters correspond to a strong preponderance of purchases in a particular category of products. Other clusters will differ from basket averages (** mean **), the total sum spent by the clients ( sum ) or the total number of visits made ( count **).

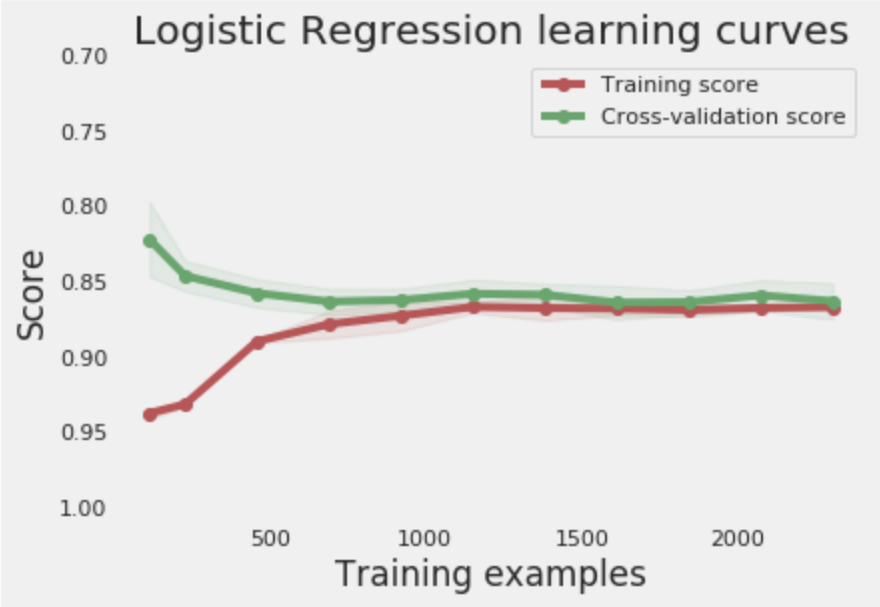
Confusion matrix, without normalization

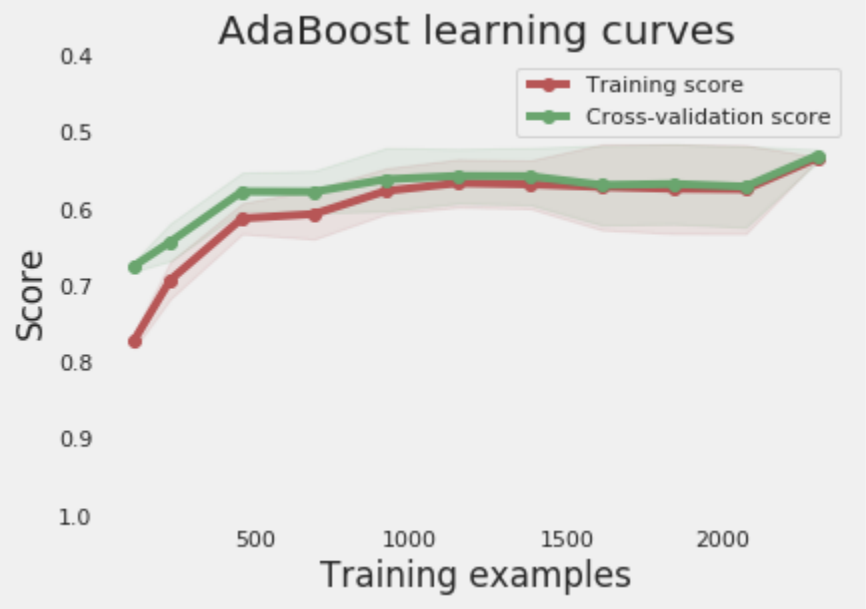


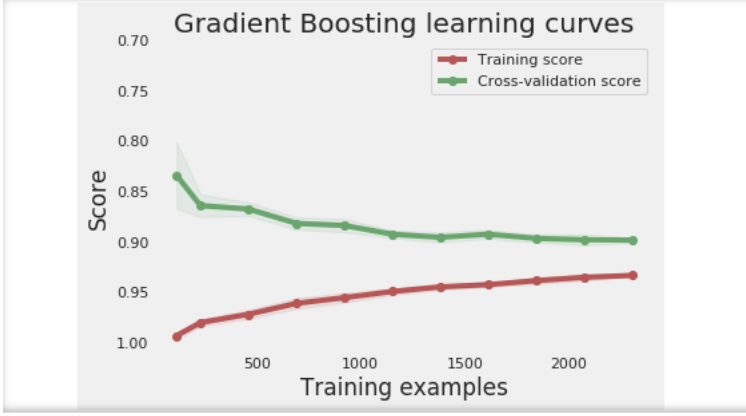

---




On this curve, we can see that the train and cross-validation curves converge towards the same limit when the sample size increases. This is typical of modeling with low variance and proves that the model does not suffer from overfitting. Also, we can see that the accuracy of the training curve is correct which is synonymous of a low bias. Hence the model does not underfit the data.


Logistic Regression learning curves
Training score
Cross-validation score
Score
0.70
0.75
0.80
0.85
0.90
0.95
1.00
500
1000
1500
2000
Training examples


AdaBoost learning curves
Training score
Cross-validation score
Score
0.4
0.5
0.6
0.7
0.8
0.9
1.0
500
1000
1500
2000
Training examples




[95]:
```
classifiers = [(svc, 'Support Vector Machine'),
               (lr, 'Logostic Regression'),
               (knn, 'k-Nearest Neighbors'),
               (tr, 'Decision Tree'),
               (rf, 'Random Forest'),
               (gb, 'Gradient Boosting')]
#_____________________________
for clf, label in classifiers:
    print(30*'_', '\n{}'.format(label))
    clf.grid_predict(X, Y)
```

```
_____________________________
Support Vector Machine
Precision: 65.93 %
_____________________________
Logostic Regression
Precision: 71.34 %
_____________________________
k-Nearest Neighbors
Precision: 67.58 %
_____________________________
Decision Tree
Precision: 71.38 %
_____________________________
Random Forest
Precision: 75.38 %
_____________________________
Gradient Boosting
Precision: 75.23 %
```

Finally, as anticipated in Section 5.8, it is possible to improve the quality of the classifier by combining their respective predictions. At this level, I chose to mix *Random Forest*, *Gradient Boosting* and *k-Nearest Neighbors* predictions because this leads to a slight improvement in predictions:

[96]:
```
predictions = votingC.predict(X)
print("Precision: {:.2f} % ".format(100*metrics.accuracy_score(Y, predictions)))
```

```
Precision: 75.46 %
```

## Conclusion:

The project demonstrated the effectiveness of machine learning techniques in classifying customers and providing valuable insights for targeted marketing. However, it also highlighted the need for addressing biases, seasonality effects, and longer-term data for more robust predictions.

The work described is based on a database providing details on purchases made on an E-commerce platform over a period of one year. Each entry in the dataset describes the purchase of a product, by a particular customer and at a given date. In total, approximately ~4000 clients appear in the database. Given the available information, I decided to develop a classifier that allows to anticipate the type of purchase that a customer will make, as well as the number of visits that he will make during a year, and this from its first visit to the E-commerce site.

The first stage of this work consisted in describing the different products sold by the site, which was the subject of a first classification. There, I grouped the different products into 5 main categories of goods. In a second step, I performed a classification of the customers by analyzing their consumption habits over a period of 10 months. I have classified clients into 11 major categories based on the type of products they usually buy, the number of visits they make and the amount they spent during the 10 months. Once these categories established, I finally trained several classifiers whose objective is to be able to classify consumers in one of these 11 categories and this from their first purchase.

Finally, the quality of the predictions of the different classifiers was tested over the last two months of the dataset. The data were then processed in two steps: first, all the data

was considered (ober the 2 months) to define the category to which each client belongs, and then, the classifier predictions were compared with this category assignment. I then found that 75% of clients are awarded the right classes. The performance of the classifier therefore seems correct given the potential shortcomings of the current model. In particular, a bias that has not been dealt with concerns the seasonality of purchases and the fact that purchasing habits will potentially depend on the time of year (for example, Christmas ). In practice, this seasonal effect may cause the categories defined over a 10-month period to be quite different from those extrapolated from the last two months. In order to correct such bias, it would be beneficial to have data that would cover a longer period of time.

**Assumptions:**

- Assumed that customer purchase behavior remains consistent over time.
- Assumed that the dataset is representative of the entire customer base.

**Limitations:**

- Seasonal variations in purchase behavior were not fully accounted for, potentially impacting the accuracy of long-term predictions.
- Limited historical data may restrict the model's ability to capture evolving customer behavior accurately.

**Challenges:**

- Handling imbalanced classes in customer categories.
- Ensuring model generalization to new customer data and unseen scenarios.
- Addressing potential biases in the dataset, such as selection bias or data quality issues.

**Future Uses/Additional Applications:**

- Integration with real-time data sources for dynamic customer segmentation and personalized recommendations.
- Incorporation of external data sources such as demographic information or social media activity for more comprehensive customer profiling.
- Application of deep learning models for more complex customer behavior analysis, including sentiment analysis or image recognition for product recommendations.

**Recommendations:**

- Collecting and integrating more historical data for improved model training and long-term predictions.
- Regularly updating and retraining the classification model to adapt to evolving customer behavior and market trends.
- Implementing A/B testing and conducting experiments to validate the effectiveness of targeted marketing strategies based on customer segments.

**Implementation Plan:**

5. Deploy the trained classification model to the e-commerce platform's backend, ensuring scalability and real-time performance.
6. Monitor model performance metrics such as accuracy, precision, and recall, and iterate on improvements based on feedback and data updates.
7. Collaborate with marketing teams to implement targeted campaigns, promotions, and product recommendations based on customer segments identified by the model.
8. Conduct regular evaluations and updates to ensure the model remains effective and aligned with business goals.

**Ethical Assessment:**

Considerations regarding data privacy, consent, and transparency in model usage were prioritized throughout the project. Steps were taken to anonymize sensitive customer information and comply with data protection regulations. Ethical implications of targeted marketing, customer profiling, and data-driven decision-making were also considered, emphasizing fairness, accountability, and transparency in model deployment and usage.

This detailed white paper draft provides a thorough exploration of the data science project, covering key aspects such as data preparation, methodology, analysis, limitations,

and future recommendations. It offers stakeholders and decision-makers a comprehensive understanding of the project's objectives, findings, and potential implications for business strategies.

## References

https://www.kaggle.com/datasets/carrie1/ecommerce-data