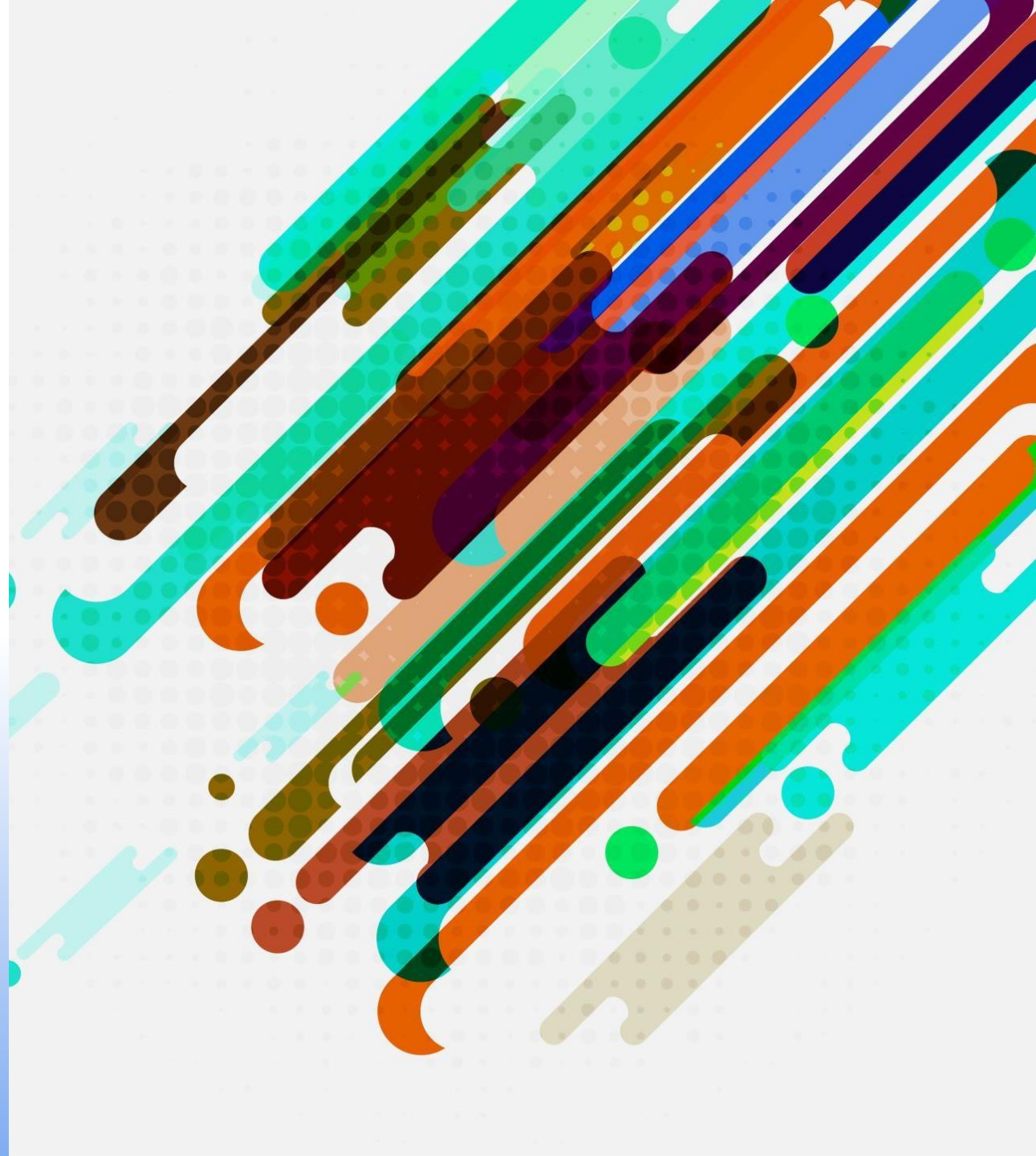
DATA-DRIVEN CUSTOMER CLASSIFICATION FOR E-COMMERCE OPTIMIZATION

BY KAUSIK CHATTAPADHYAY

04/28/2024



BUSINESS PROBLEM

- Understanding customer behavior is essential for optimizing marketing strategies and enhancing user experience in the dynamic e-commerce environment.
- The project addresses the business problem by developing a data-driven classification system to categorize customers based on their purchase habits.
- Analyzing an E-commerce database with records of purchases made by approximately 4000 customers over a year is central to this project's objectives.
- The goal is to create targeted marketing strategies and personalized recommendations through insightful customer categorization.

BACKGROUND/HISTORY

- E-commerce platforms offer valuable insights from vast transactional data about customer preferences and trends.
- Traditional marketing strategies often miss individual customer behavior nuances due to lack of granularity.
- This project utilizes machine learning to categorize customers, empowering businesses to personalize offerings and promotions effectively.

InvoiceNo: Invoice number.

StockCode: Product (item) code.

Description: Product (item) name.

Quantity: The quantities of each product (item) per transaction. Numeric.

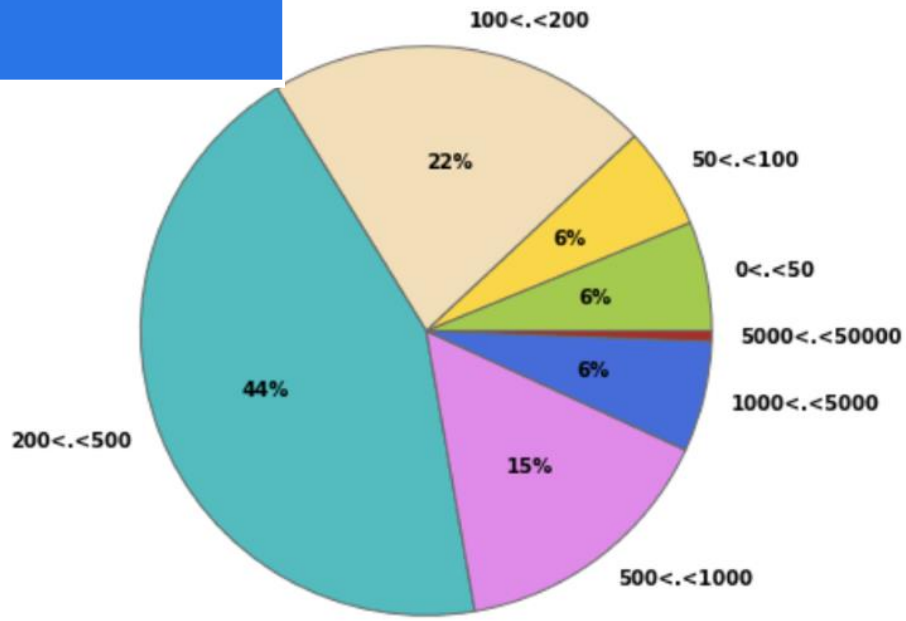
InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

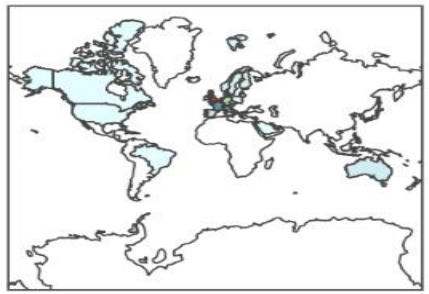
Country: Country name. Nominally, the name of the country where each customer lives.

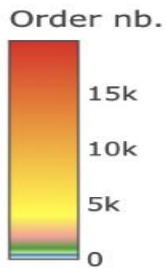
Data Explanation and Analysis



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.

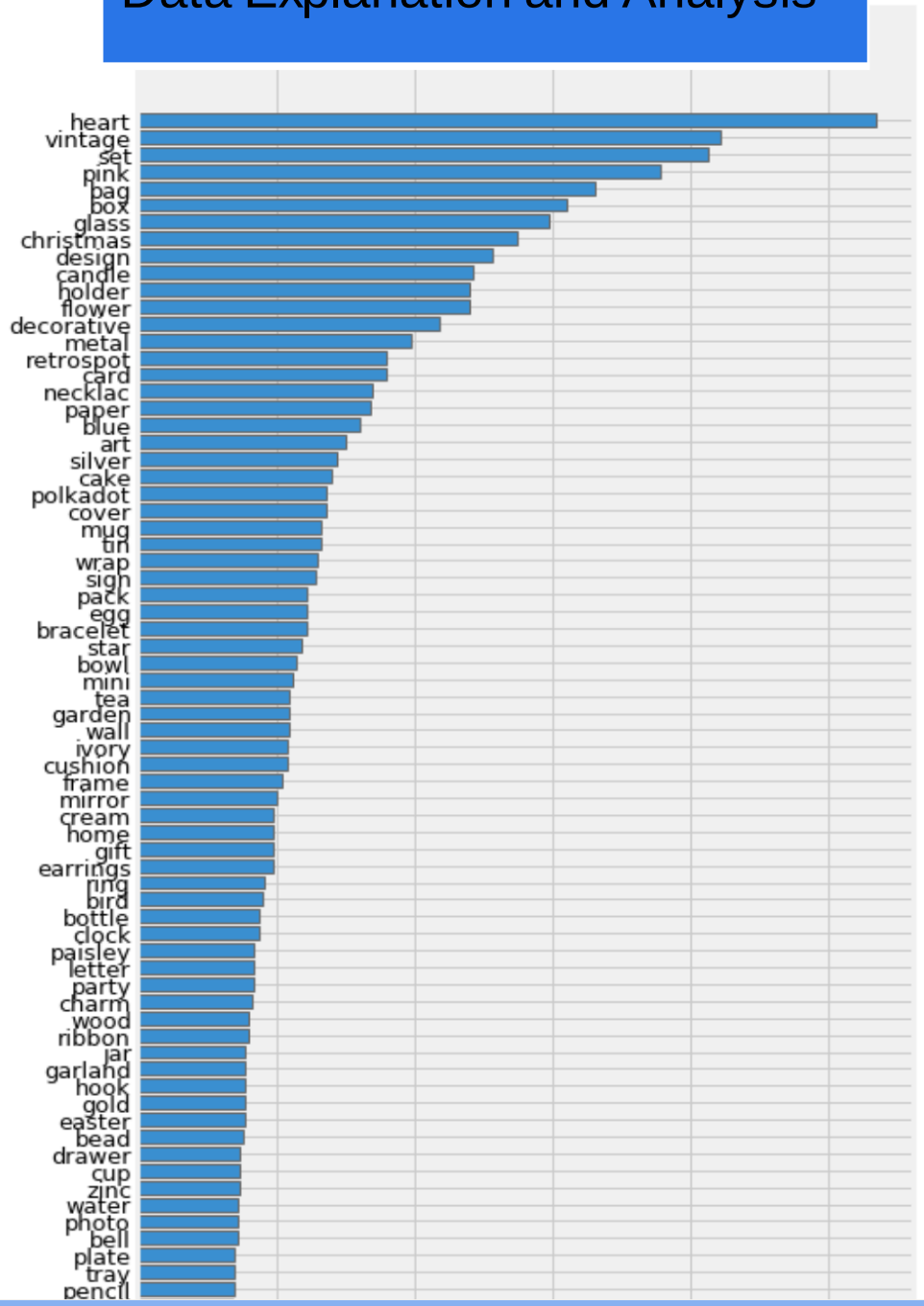
Number of orders per country





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

Data Explanation and Analysis



cluster nº0

cluster nº1

cluster nº2

cluster nº3

cluster nº4

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.

Method and Results

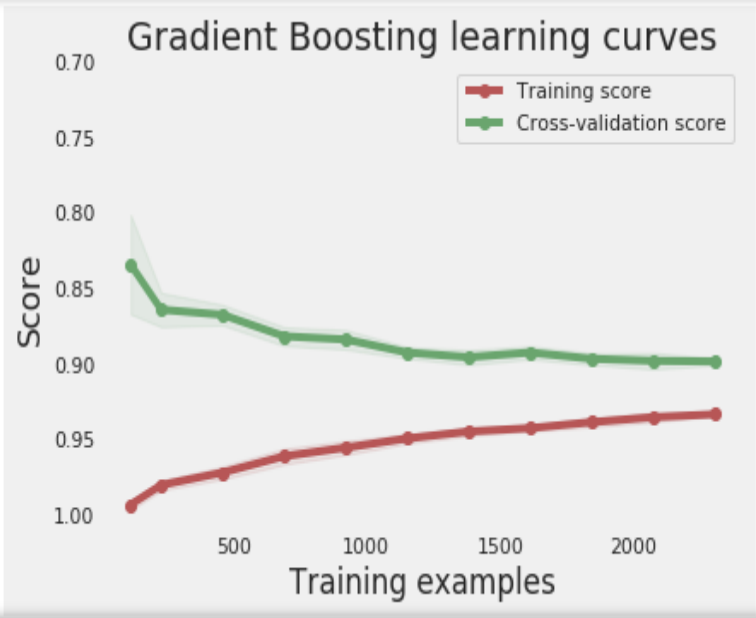
[95]:
```
classifiers = [(svc, 'Support Vector Machine'),
               (lr, 'Logostic Regression'),
               (knn, 'k-Nearest Neighbors'),
               (tr, 'Decision Tree'),
               (rf, 'Random Forest'),
               (gb, 'Gradient Boosting')]
#_____________________________
for clf,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

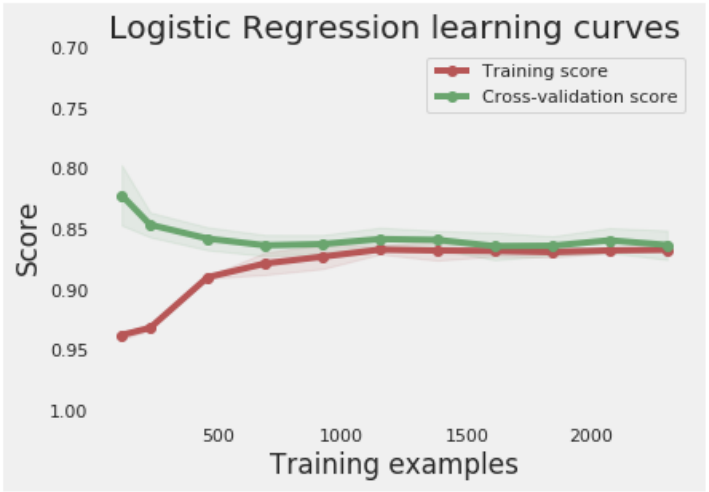
```
_____________________________
Support Vector Machine
Precision: 65.93 %
_____________________________
Logostic Regression
Precision: 71.34 %
_____________________________
k-Nearest Neighbors
Precision: 67.58 %
_____________________________
Decision Tree
Precision: 71.38 %
_____________________________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

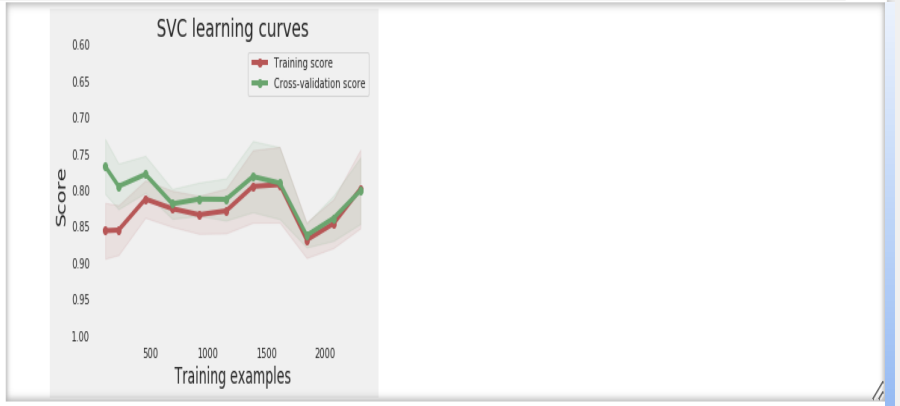
Finally, as anticipated in Section 5.8, it is possible to improve the quality of the classifier by combining their respective predictions. At this level, I chose to mix *Random Forest*, *Gradient Boosting* and *k-Nearest Neighbors* predictions because this leads to a slight improvement in predictions:

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

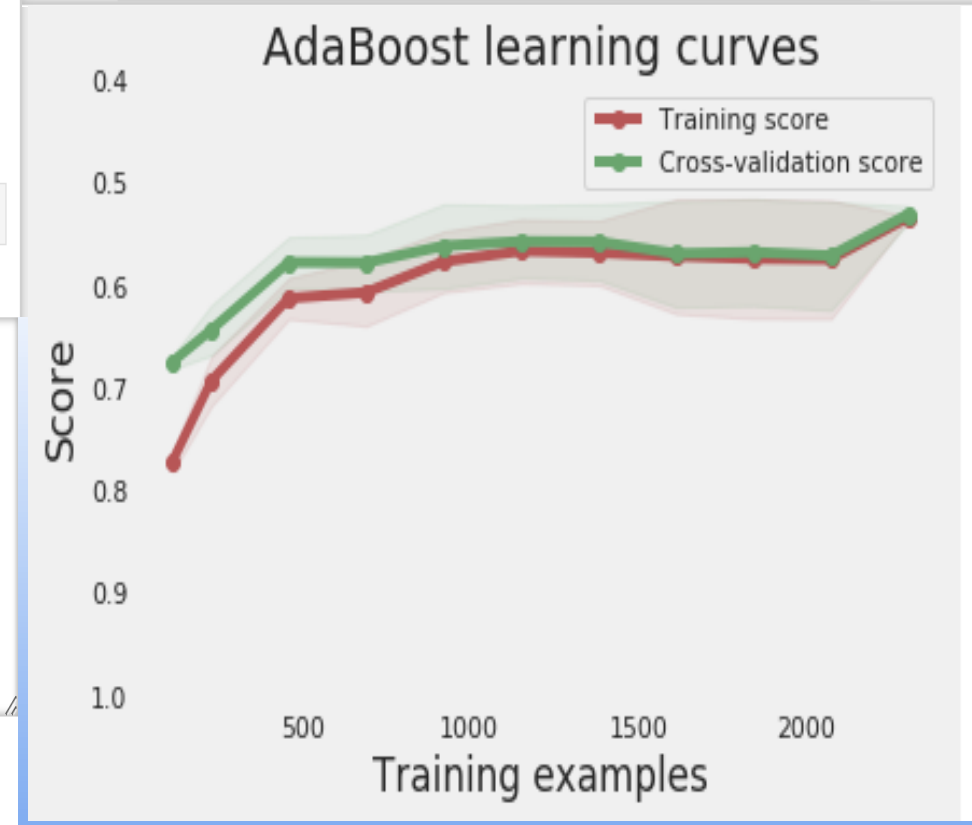
```
Precision: 75.46 %
```

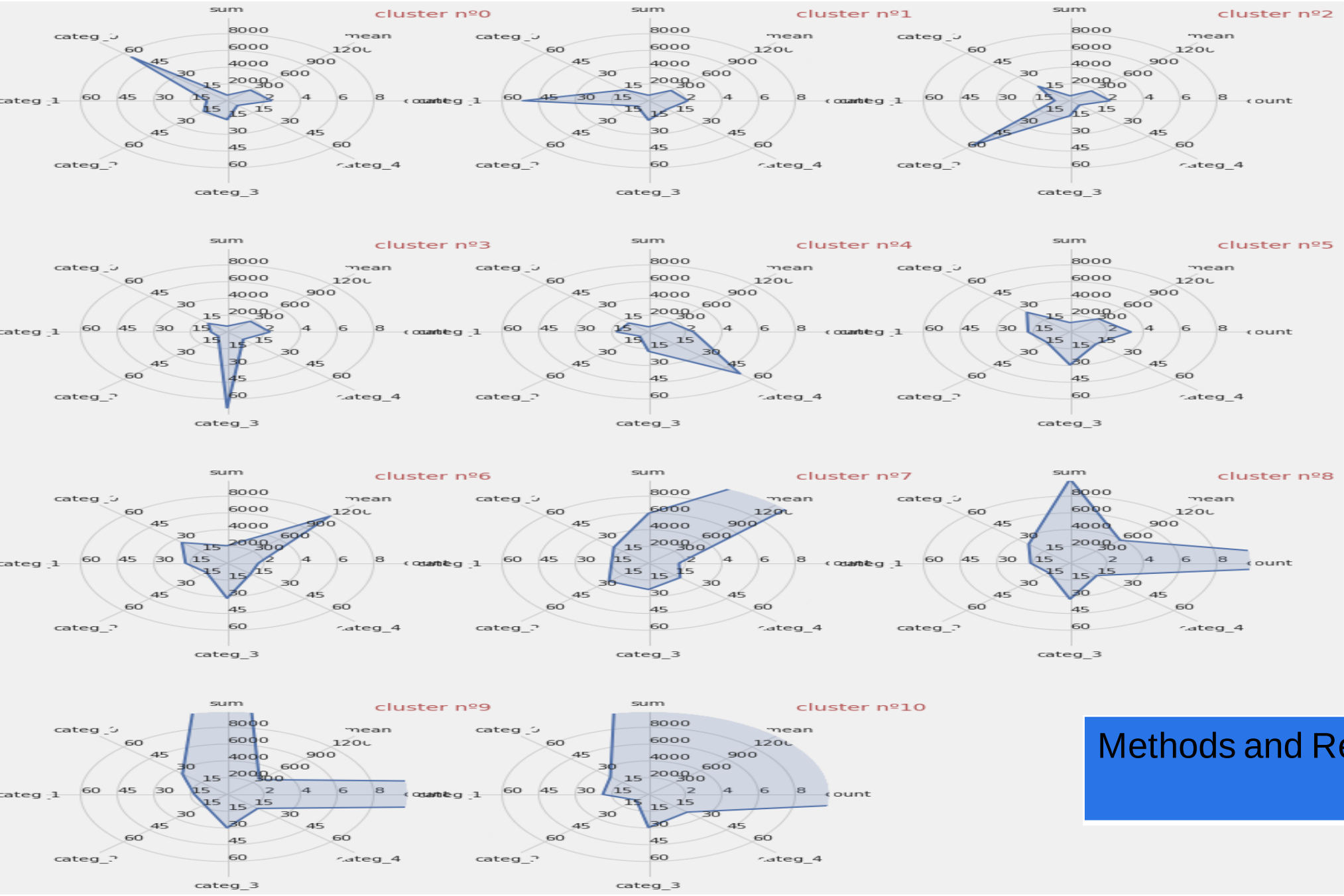






On this curve, we can see that the train and cross-validation curves converge towards the same limit when the sample size increases. This is typical of modeling with low variance and proves that the model does not suffer from overfitting. Also, we can see that the accuracy of the training curve is correct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

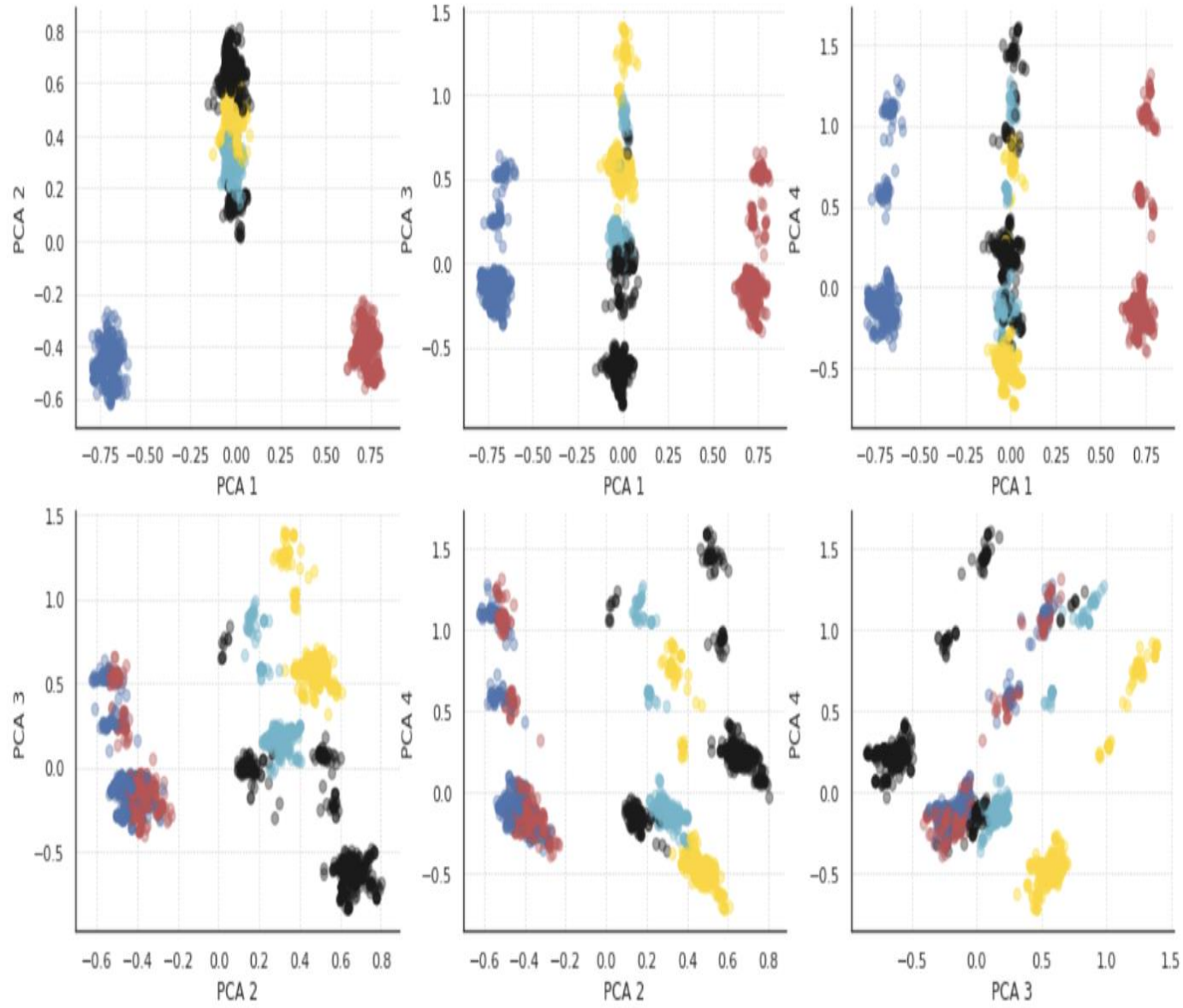


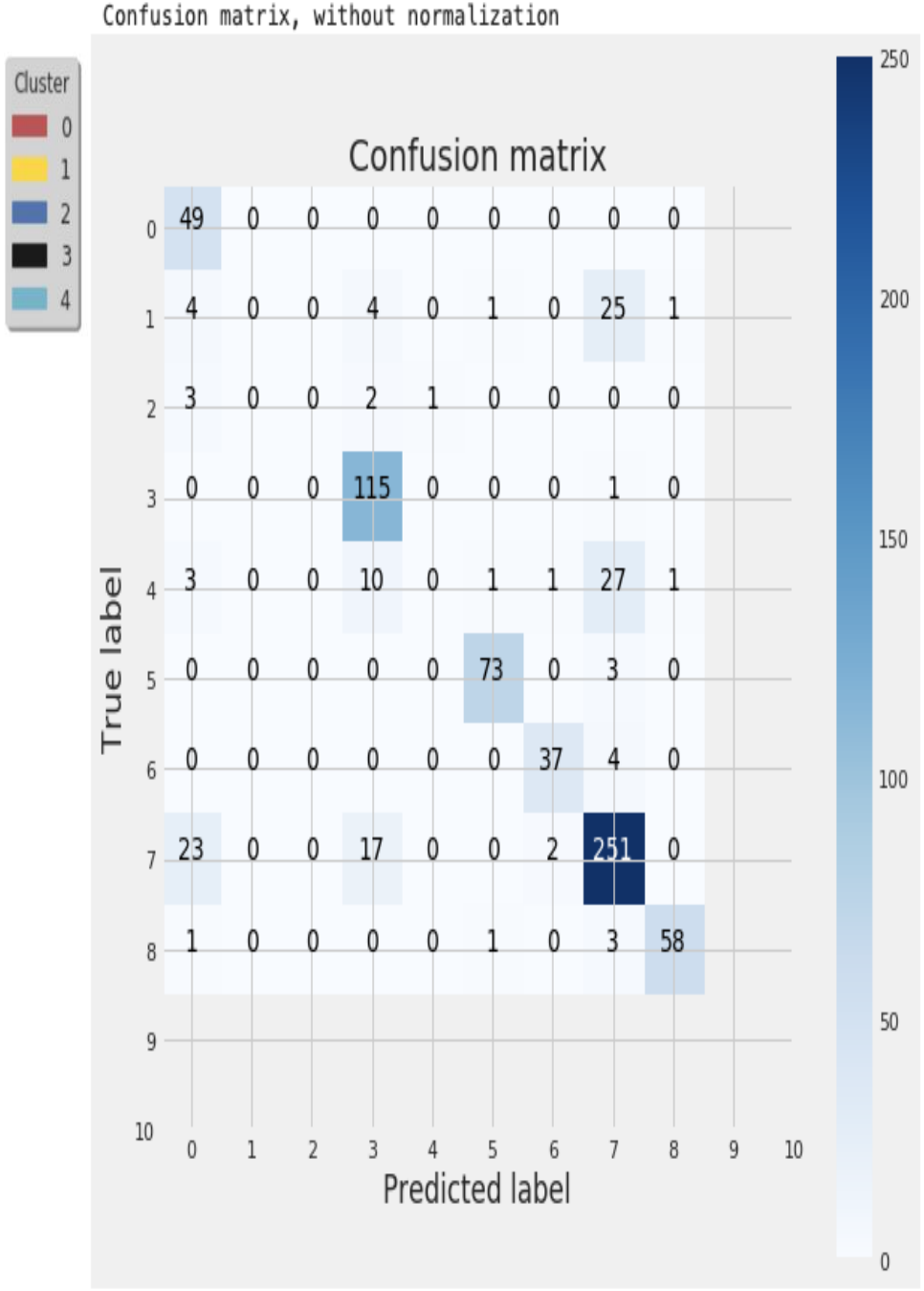


Methods and Results

t 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 rom basket averages (** mean **), the total sum spent by the clients (sum) or the total number of visits made (count **).

Methods & Results



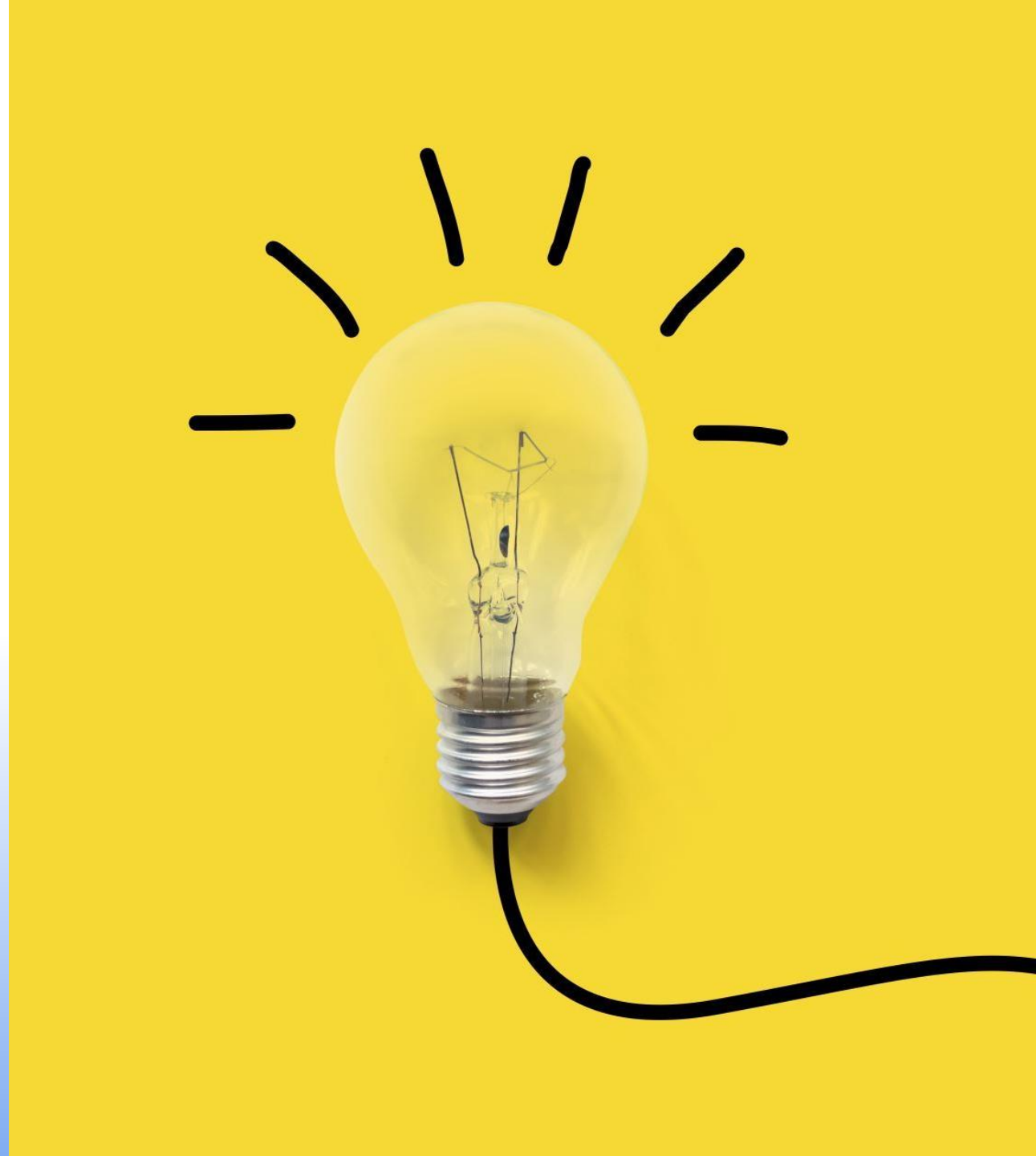


ANALYSIS

- 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

CONCLUSION

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



LIMITATIONS

- Address seasonal variations in purchase behavior to enhance accuracy of long-term predictions.
- Expand historical data collection to capture evolving customer behavior more accurately.
- Handle imbalanced customer categories effectively within the model.
- Ensure model generalization and mitigate biases through rigorous dataset evaluation and integration with real-time and external data sources.

LIMITATIONS

- Utilize deep learning models to analyze intricate customer behaviors, employing sentiment analysis and image recognition for enhanced product recommendations.
- Enhance model training and long-term predictions by gathering and integrating extensive historical data into the analysis.
- Ensure model adaptability to changing customer behaviors and market trends through regular updates and retraining, maintaining relevance and accuracy over time.

Assumptions:

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

Limitations:

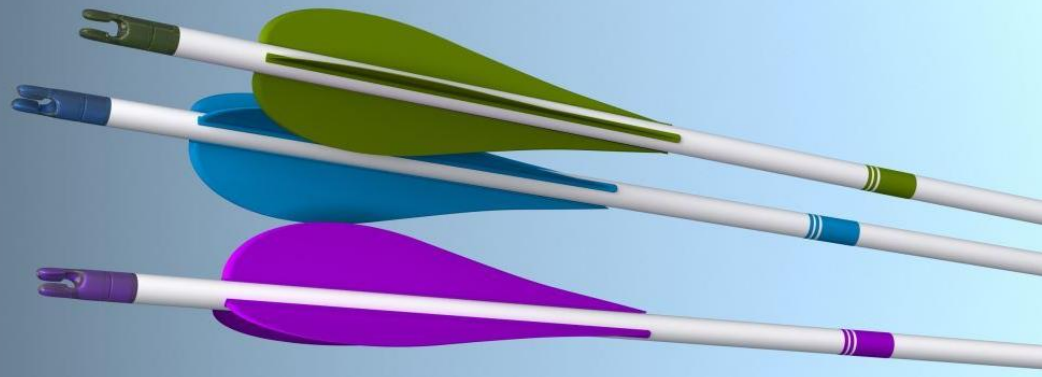
- Seasonal variations in purchase behavior were not fully accounted for, potentially impacting the accuracy of long-term predictions.
- Limited historical data may restrict the model's ability to capture evolving customer behavior accurately.

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.

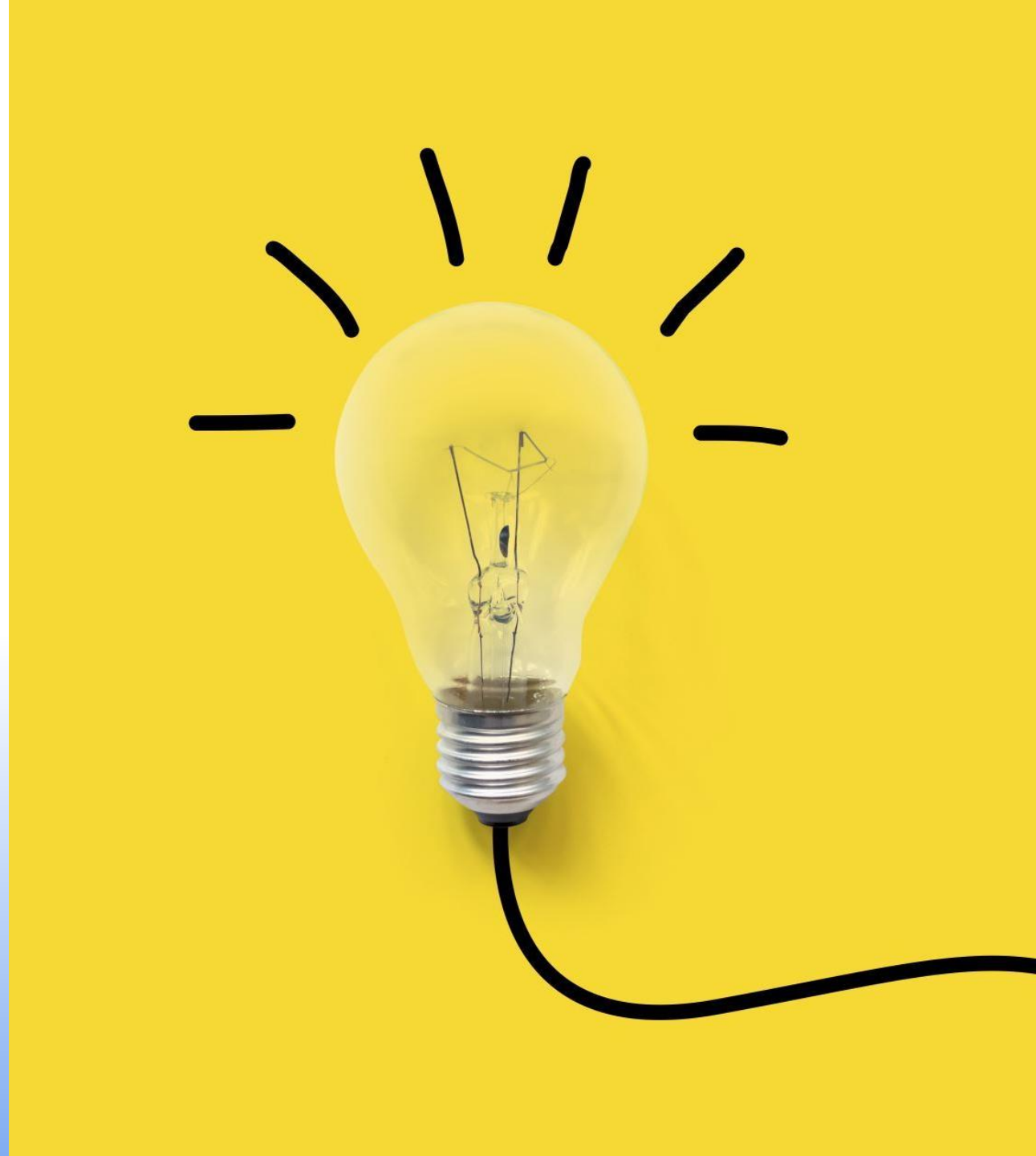
IMPLEMENTATION PLAN

- Seamlessly deploy the trained classification model to the e-commerce platform, prioritizing scalability and real-time performance.
- Monitor and refine model performance metrics like accuracy and recall, iterating improvements based on data updates and feedback.
- Collaborate with marketing teams to implement personalized campaigns and product recommendations driven by customer segments identified through the model.
- Conduct regular evaluations and updates to maintain the model's effectiveness and alignment with evolving business goals.



ETHICAL ASSESSMENT

- Data privacy, consent, and transparency were paramount, ensuring ethical usage of the model.
- The white paper draft meticulously explores data preparation, methodology, and future recommendations.
- Stakeholders gain a profound understanding of project objectives and strategic implications for business decisions.



REFERENCES

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