Capstone Project –DP1812 CP

Customer and Product Profiling

Presented by:
Nikita Dange, Sanjiv Sharan, Shriti Datta
11 July, 2020

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Agenda

- Overview
- Architecture and Workflow
- Data Exploration & Processing
- Model Implementation and Results
- Analysis of Results & Optimization
- Deliverables
- Next Steps
- Appendix

Overview

Objective

Perform Customer Analytics on e-commerce data to create solutions which help organization to increase sales by spending less money.

- Classifying customers into segments.
- Anticipate the purchases that will be made by a new customer.

Data Source

- Online Retail-ecommerce (UK Retailers)
- One Year data
- UCI Machine Learning Repository

Technology and Software

- Programming Language: Python
- Version Control: Git

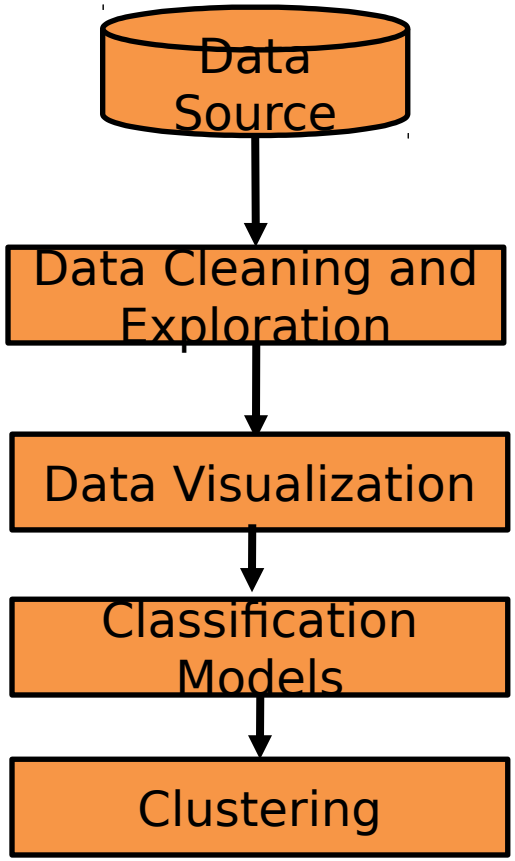
Learnings

Understood different classification techniques
Understood text analytics
Learned different plotting diagrams

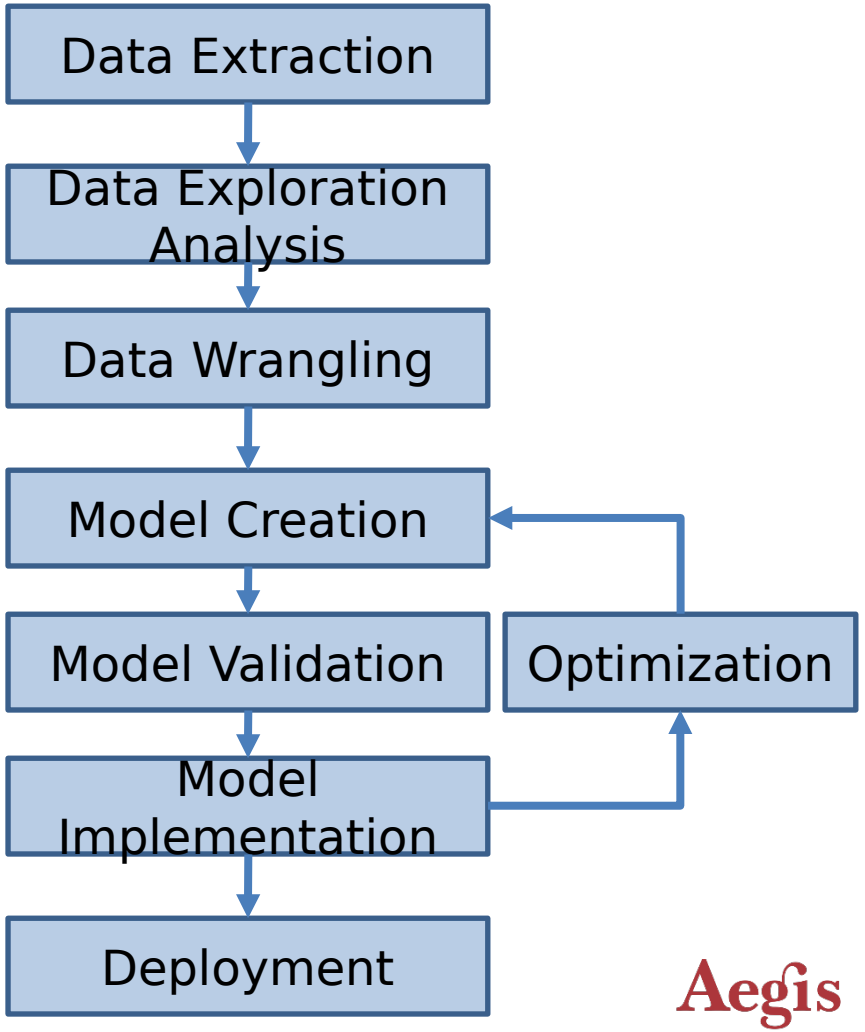
Output

- Customer Clustering
- Product Clustering

Architecture and Workflow



Architecture



Workflow

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Data Exploration & Processing

- 8 features present in the dataset. i.e. Invoice No., Stock Code, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country.
- Analyze significance of all data features and remove unwanted features
- Sanity check of the important data (missing values, NA, duplicates, Purchase records)
- Assessment of frequency dist
  extremes (especially extreme
  - Country
  - Customer and Products
  - Product Categories
  - Defining product categories

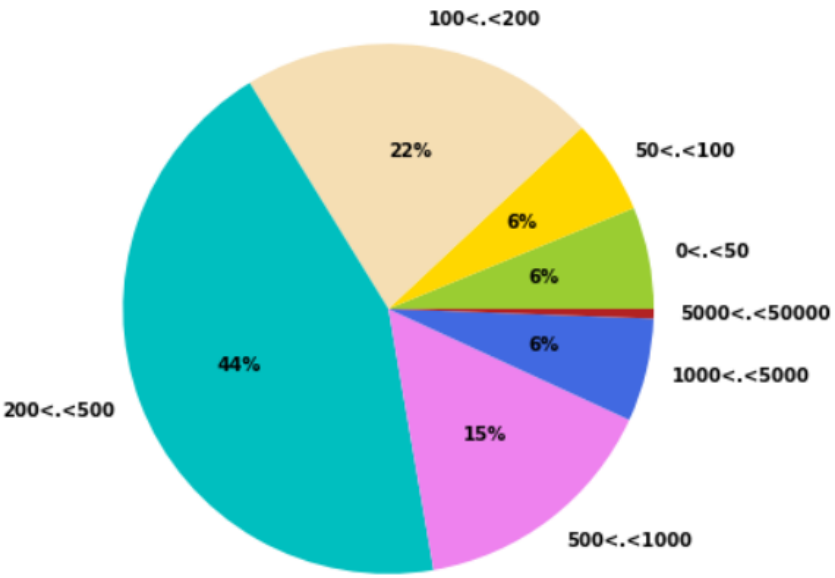
```
pd.DataFrame([{'products': len(df_initial['StockCode'].value_counts()),
               'transactions': len(df_initial['InvoiceNo'].value_counts()),
               'customers': len(df_initial['CustomerID'].value_counts()),
              }], columns = ['products', 'transactions', 'customers'],
              index = ['quantity'])
```

	products	transactions	customers
quantity	3684	22190	4372

- Identifying data for creating a prototype model (relatively uniform distribution across reviews preferred)
- Analysis of cancelled orders and reorders.
- Bulk orders

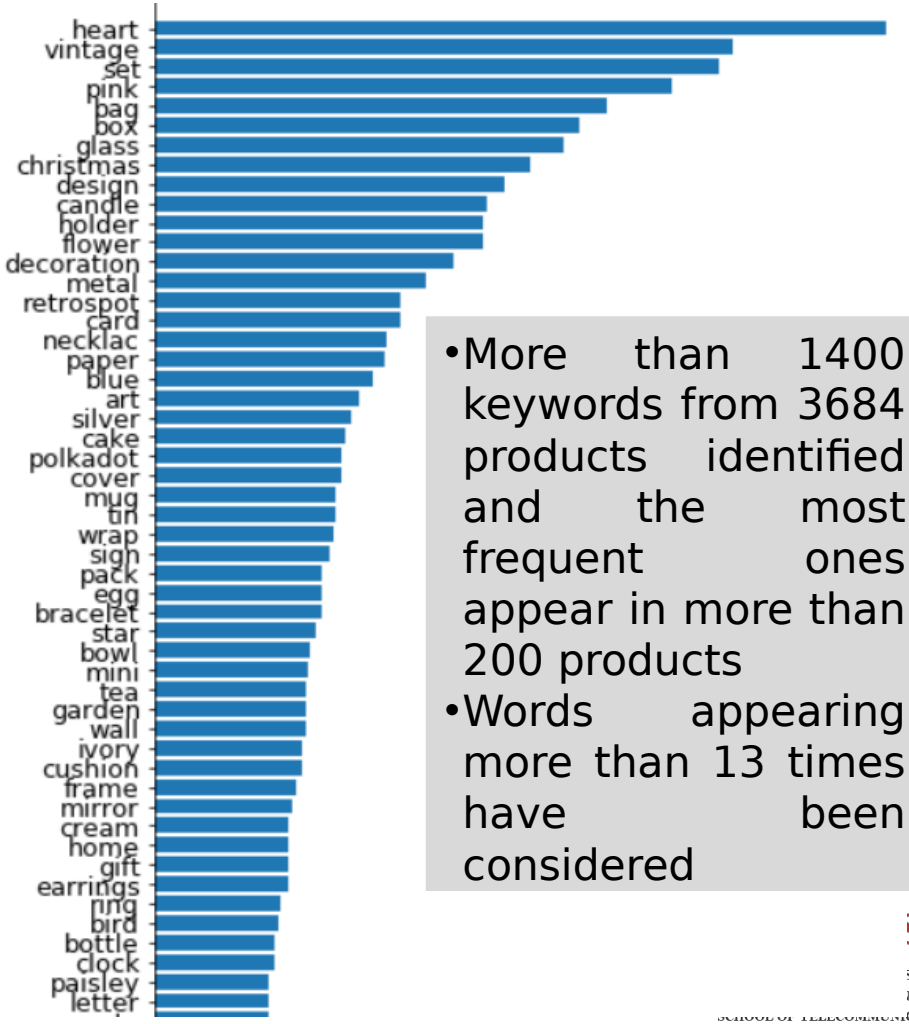
Data Exploration Charts

Distribution of order amounts



Majority of orders concern relatively large purchases given that ~65% of purchases give prizes in excess of £ 200.

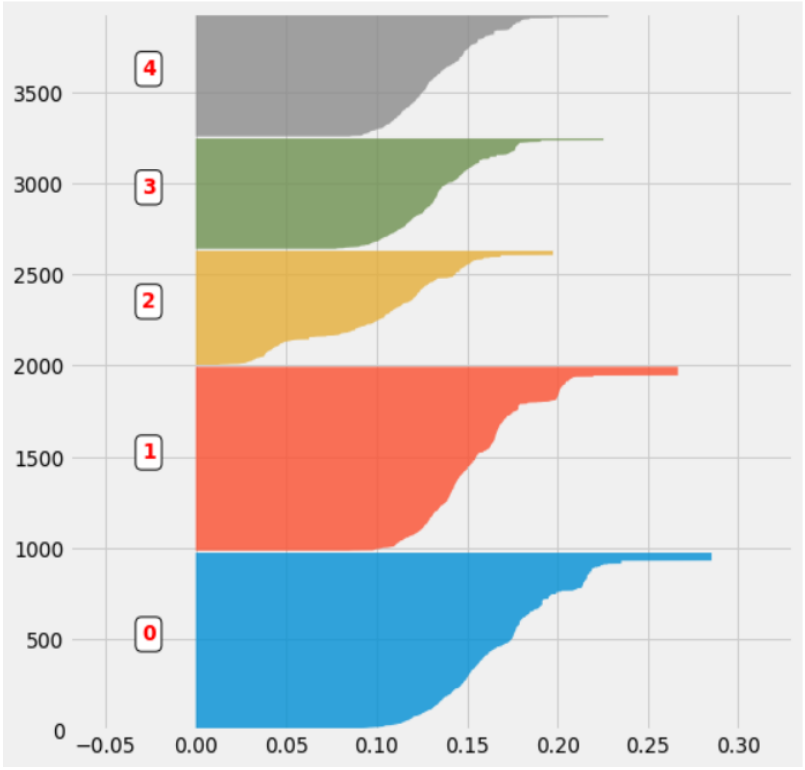
Distribution of product keywor



- More than 1400 keywords from 3684 products identified and the most frequent ones appear in more than 200 products
- Words appearing more than 13 times have been considered

Data Exploration Charts

Product Cluster Analysis



- K-means clustering
- Silhouette scores of each element of the different clusters

Word Cloud

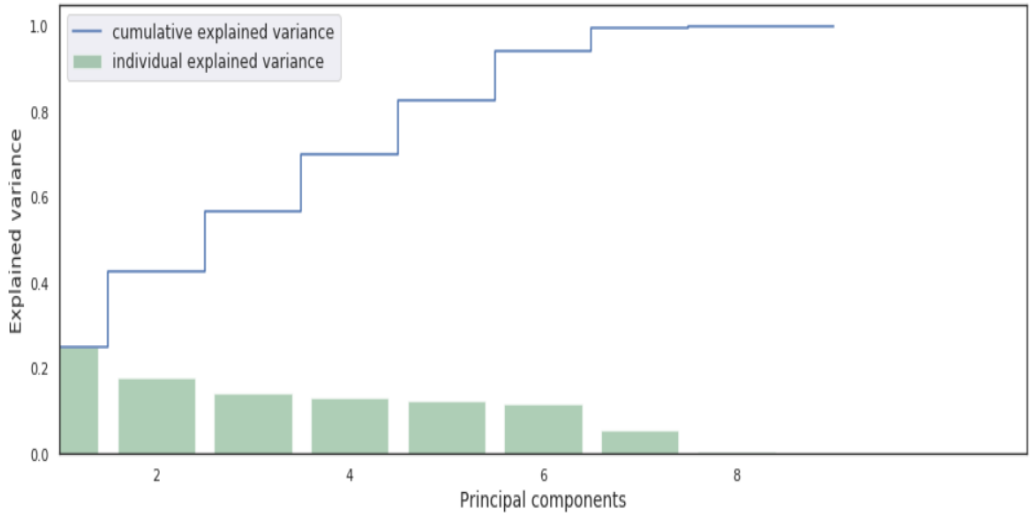
Word cloud based keywords from 5 clusters

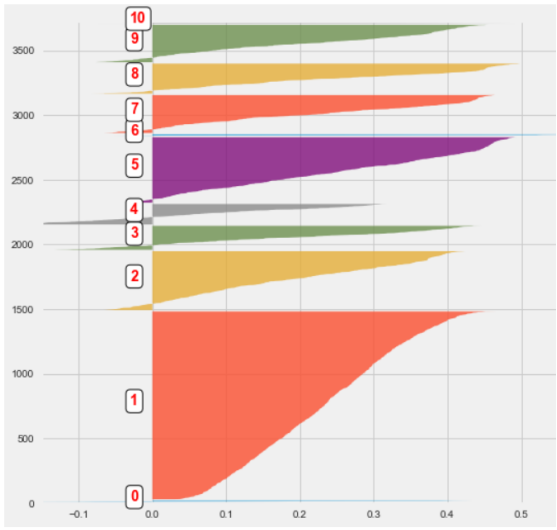
Data Exploration Charts

Customer categories

Formatting data

- Grouping products
- Splitting of the dataset
- Grouping orders
- Creating customer categories
- Data encoding
- Creating categories

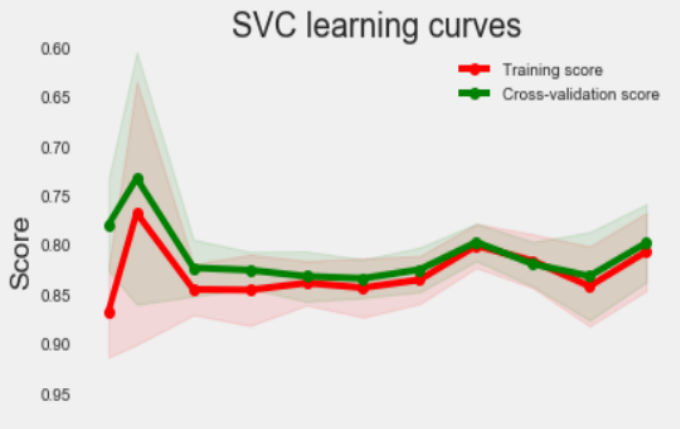


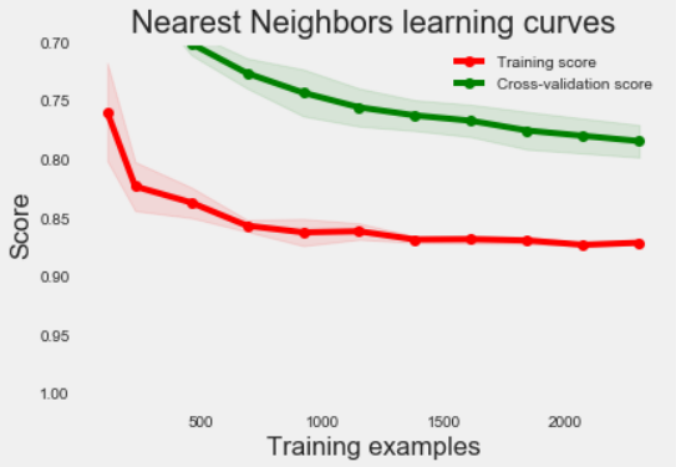


- Performed PCA to ensure that the clusters are truly distinct.
- Number of clusters based on the silhouette score
- Disparity in the sizes of different groups that have been created

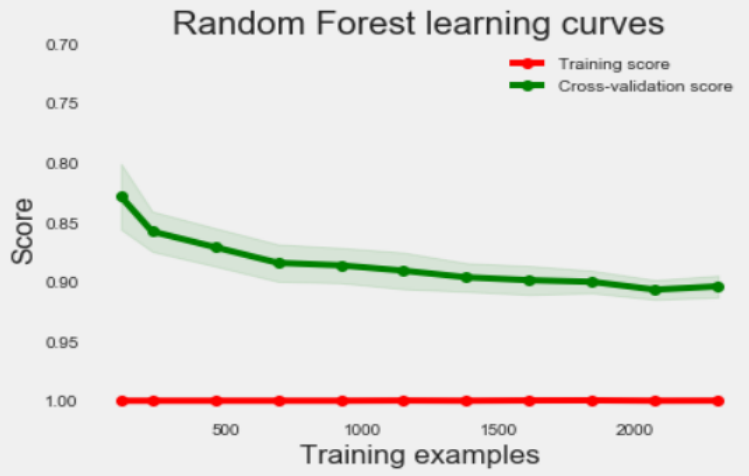
Different clusters are indeed disjoint

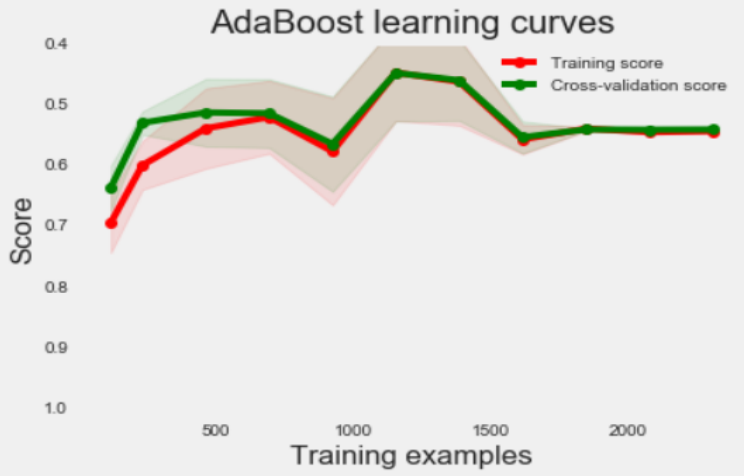
Model Creation
Classification of customers

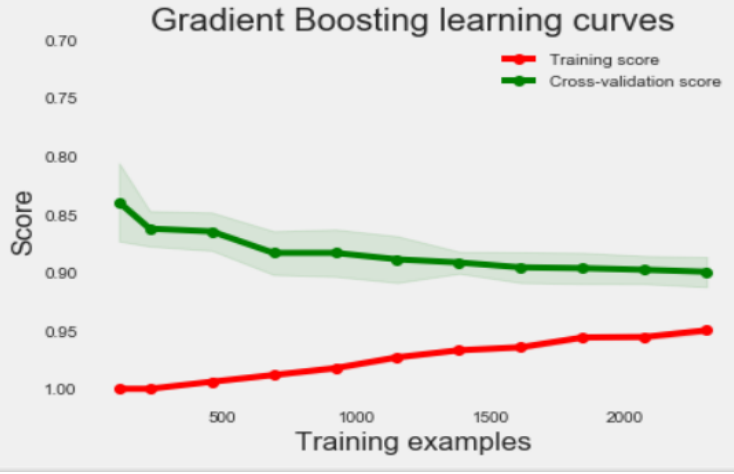




Model Creation
Classification of customers







Results of different classifiers can be combined to improve the classification model. This can be achieved by selecting the customer category as the one indicated by the majority of classifiers.
VotingClassifier method of the sklearn package was used.

Model Analysis

```
_____________________________________
Support Vector Machine
Precision: 71.97 %

_____________________________________
Logistic Regression
Precision: 75.81 %

_____________________________________
k-Nearest Neighbors
Precision: 67.50 %

_____________________________________
Decision Tree
Precision: 71.50 %

_____________________________________
Random Forest
Precision: 75.85 %

_____________________________________
Gradient Boosting
Precision: 75.38 %
```

- Quality of the classifier can be improved by combining different classifiers and their respective predictions.
- We chose *Random Forest*, *Gradient Boosting* and *k-Nearest Neighbors* predictions to improve predictions.
- Precision: 75.70 %

Further Optimization Areas

With regards to Machine Learning model

- Optimization in keyword generation
- Use other classification techniques
- Exploring other hyper-parameter tuning

With regards to overall functioning and performance

- Improve execution time and performance
- Comparison of results across R and Python
- Apply Deep Learning techniques for better results
- More visualization tool to be explored to better user friendly interface

Conclusion

- Combining multiple classifiers give better accuracy and prediction.
- Number of clusters created for product and customers plays an important role in achieving higher performance.

Thank You