NATIONALITY PREDICTION IN ONLINE RETAIL

Presented by: Group D
Yeji Choi, Yunji Kang

OVERVIEW

01	02	03
Research Question	Literature Review	Data set
04	05	06
ML Processing	Slected Approach	Limitation

RESEARCH QUESTION

Machine Learning Task

*Is it possible to **predict the country** (of an order) based on the stock code, invoice date, quantity, unit price and customer id?*

LITERATURE REVIEW

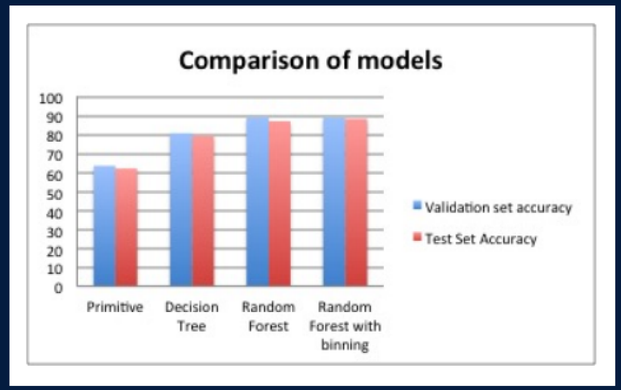
Predicting Airbnb User Destination Using User Demographic and Session Information

Predicting Airbnb User Destination Using User Demographic and Session Information

LITERATURE REVIEW

Predicting Airbnb User Destination Using User Demographic and Session Information

Predicting Airbnb User Destination Using User Demographic and Session Information



THE DATASET - ONLINE RETAIL

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

01

Instances : 541909
Features : 6

02

	InvoiceNo	StockCode	Description	Quantity	₩
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	
1	536365	71053	WHITE METAL LANTERN	6	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	

	InvoiceDate	UnitPrice	CustomerID	Country
0	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

03

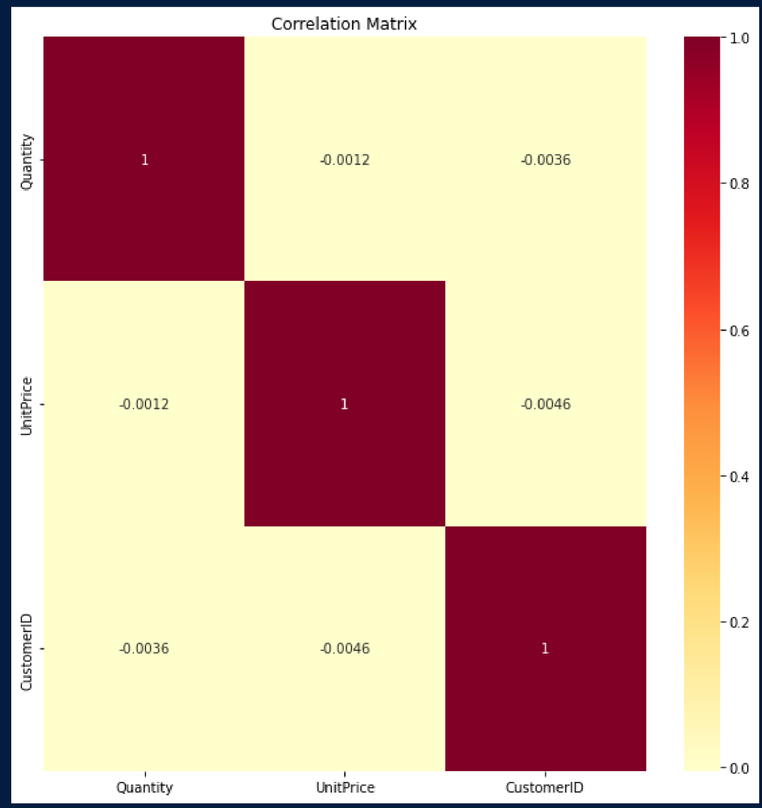
By Daqing Chen, Sai Laing Sain, Kun Guo. 2012
Published in Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates 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. Nominal, the name of the country where each customer resides.

Statistical Analysis

Index	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909	541909	541909	406829
mean	9.55225	2011-07-04 13:34:57.156386048	4.61111	15287.7
min	-80995	2010-12-01 08:26:00	-11062.1	12346
25%	1	2011-03-28 11:34:00	1.25	13953
50%	3	2011-07-19 17:17:00	2.08	15152
75%	10	2011-10-19 11:27:00	4.13	16791
max	80995	2011-12-09 12:50:00	38970	18287
std	218.081	nan	96.7599	1713.6

Description of the data set



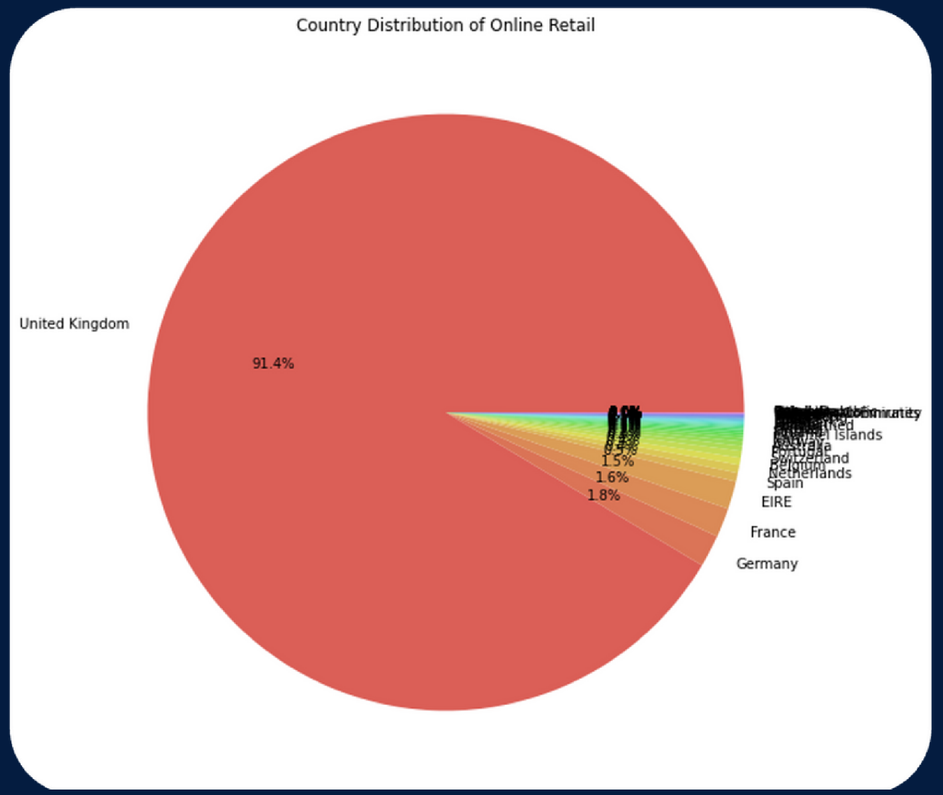
Correlation Matrix

*There are no highly interconnected variables

Statistical Analysis

United Kingdom has the highest proportion
=> We can predict that the online shopping mall is based in United Kingdom

df['Country'].value_counts()



Country	Count
United Kingdom	495478
Germany	9495
France	8557
EIRE 8196	8196
Spain	2533
Netherlands	2371
Belgium	2069
Switzerland	2002
Portugal	1519
Australia	1259
Norway	1086
Italy	803
Channel Islands	758
Finland	695
Cyprus	622
Sweden	462
Unspecified	446
Austria	401
Denmark	389
Japan	358
Poland	341
Israel	297
USA	291
Hong Kong	288
Singapore	229
Iceland	182
Canada	151
Greece	146
Malta	127
United Arab Emirates	68
European Community	61
RSA	58
Lebanon	45
Lithuania	35
Brazil	32
Czech Republic	30
Bahrain	19
Saudi Arabia	10

PREPROCESSING

Drop Null values

- NULL values in Description: 1454 rows
- NULL values in Customer ID: 135080 rows

-> Drop rows with NULL values using dropna() function

Divide the 'InvoiceDate' Column

- add 5 columns

```
df['Year'] = df['InvoiceDate'].dt.year
df['Month'] = df['InvoiceDate'].dt.month
df['Day'] = df['InvoiceDate'].dt.day
df['Hour'] = df['InvoiceDate'].dt.hour
df['Minute'] = df['InvoiceDate'].dt.minute
```

- Example

  2010.12.1 8:34 => 2010 | 12 | 01 | 08 | 34

Label Encoding

- 'Description' column contains the name of the stock (categorical values)
  => Encoding 'Description' column

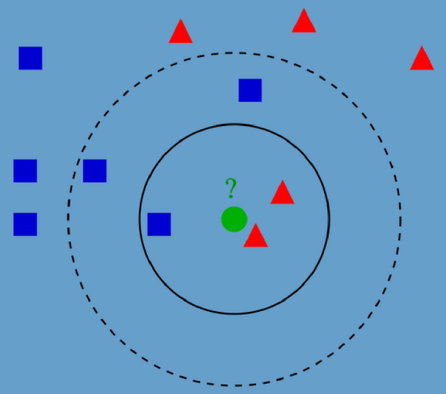
FEATURE SELECTION

- **Description**
  ‘Description’ column and ‘StockCode’ column imply the same information => the Stock name of Stocks
- **Quantity**
  Number of products ordered by the customer
- **UnitPrice**
  Price per product
- **CustomerID**
  Customer Identification number
- **Year / Month / Day / Hour / Minute**
  The date and time when the order was placed

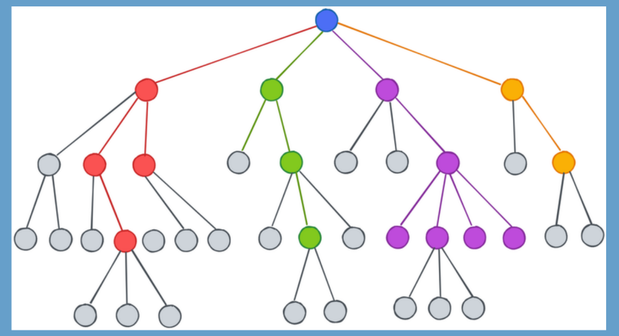
Target : Country

ML APPROACHES

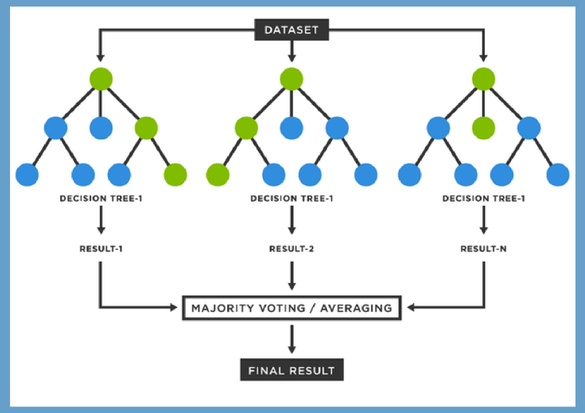
K-Nearest Neighbors



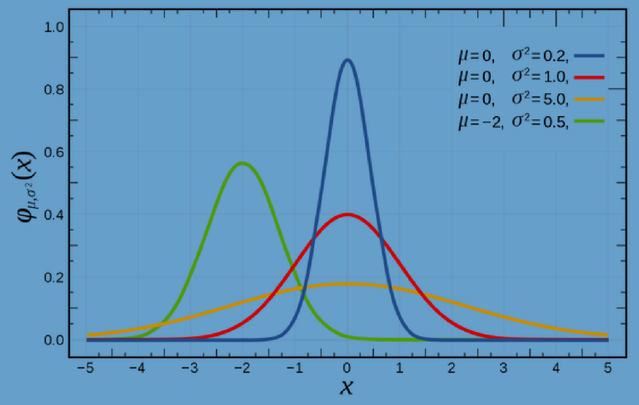
Decision Tree



Random Forest



Gaussian Naive Bayes



ML APPROACHES

K-Nearest Neighbors

KNeighbors Accuracy Score
0.989590246540324

KNeighbors Weighted average F1 score
0.9896155651026245

Decision Tree

DecisionTree Accuracy Score
0.9993732025661825

DecisionTree Weighted average F1 score
0.9993695795150729

Random Forest

RandomForest Accuracy Score
0.9970503650173291

RandomForest Weighted average F1 score
0.9970128638194027

Gaussian Naive Bayes

NaiveBayes Accuracy Score
0.08276184155544085

NaiveBayes Weighted average F1 score
0.1449078724518959

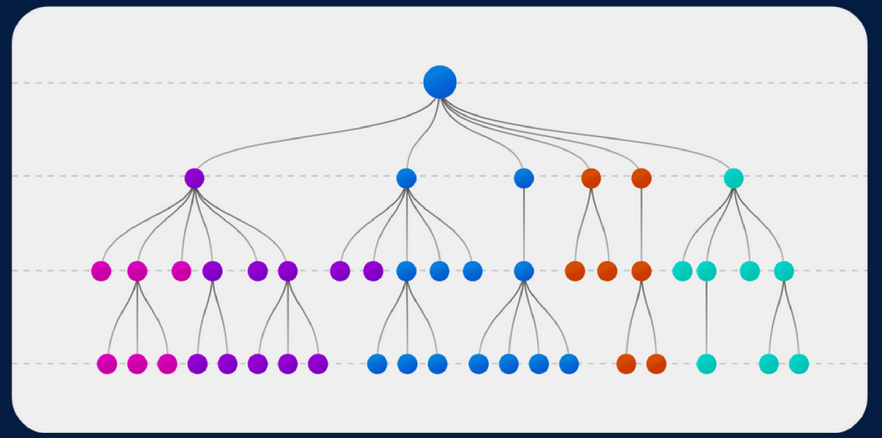
ed

SELECTED APPROACH

Decision Tree

Accuracy Score
0.999

F1 score
0.999



Features

- Automatically finds patterns in data and generates tree-based classification rules
- Simple and intuitive model
- Can handle both numerical and categorical data
- Can be prone to overfitting

Suitable Datasets

- When the relationships in the data are not too complex
- When interpretability of the model is important
- When the dataset is small

LIMITATION & FUTURE WORK

Profitability

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate nulla at ante rhoncus, vel efficitur felis condimentum. Proin odio odio.

Customer Value

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate nulla at ante rhoncus, vel efficitur felis condimentum. Proin odio odio.

Innovation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate nulla at ante rhoncus, vel efficitur felis condimentum. Proin odio odio.

Thank you For Watching