Topic Ideas

Roger Bukuru

2024-02-11

Business: Online Retail Data

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.

Research Question

1. **What items can be recommended with a given item ?**
    - What products are often purchased together ?
    - Are there cross-selling oportunities that exists ?

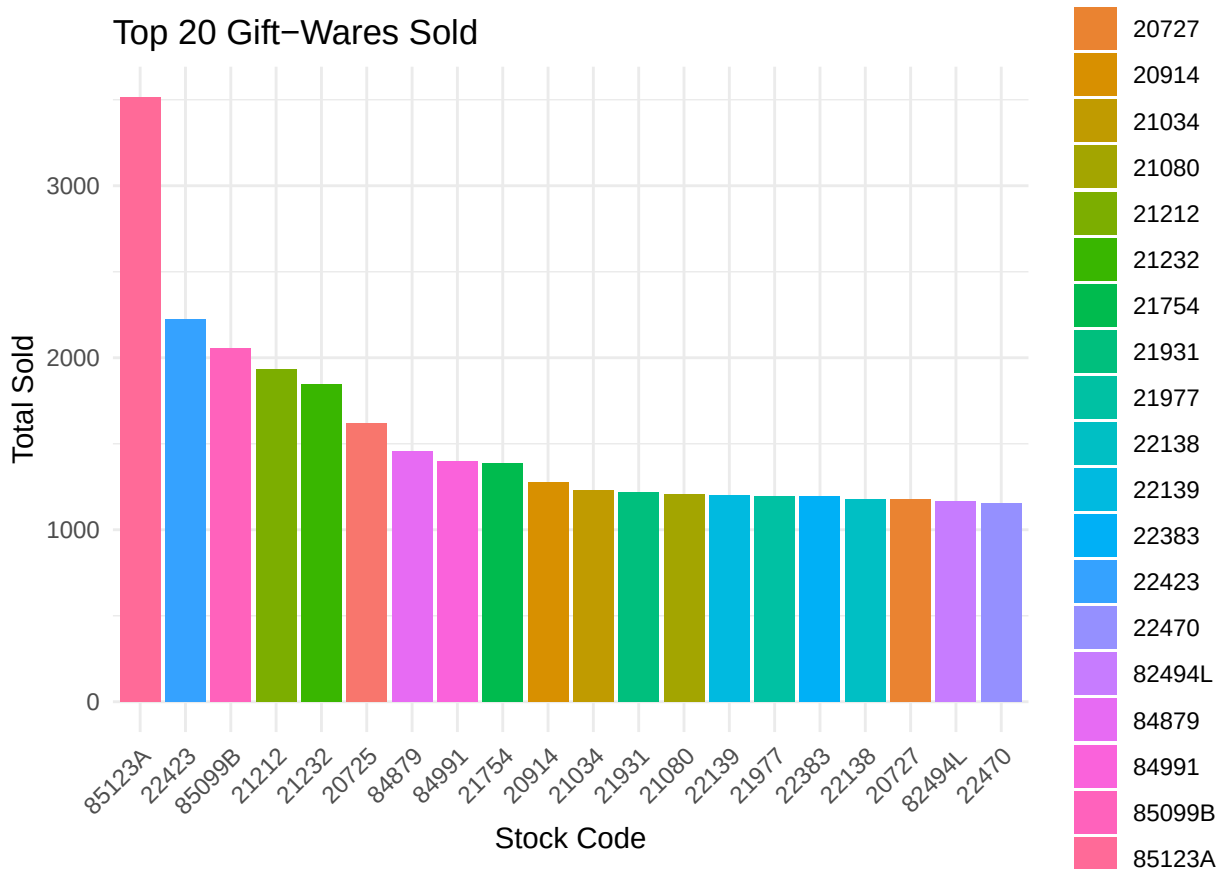
2. **Wholesaler customer retention over time ?**
    - Are the factors that contribute to possible customer churns ?

3. **Can wholesalers be grouped into various customer segements based on their purchasing patterns ?**
    - What are the trends of top wholesalers, how can they be targeted more effectively ?

4. **What gifts were are sold the most ?**
    - Top selling products ?
    - Probability of selling x product ?
    - Identify relationships between the sales of different products ?
    - What was the average basket size ?

Exploratory Data Analysis

There are 4631 unique gift-ware items, the top 20 selling items are as shown below. Gift ware with stock code 85123A was the highest selling item. Across the 2 year period, there was a total of 28816 sales across 4384 wholesalers, about 5229 sales(invoices) do not have associated customers.



The top wholesales

Customer ID	totalInvoices	totalItems	totalValue
NA	5229	107927	840616.41
14911	270	5710	40282.40
14063	13	44	39920.95
15760	5	5	33628.55
12918	3	3	32860.50
14156	138	2710	29888.03
17399	1	1	25111.09
17949	87	104	18585.93
17841	126	5114	15617.09
15202	8	8	14573.16
14606	135	3927	11247.37
12748	159	2665	10703.96
14527	108	1826	9625.22
15413	7	27	9416.82
17850	158	2515	7737.30
15311	158	2226	6436.08
18102	95	635	6408.17
15768	32	1213	6277.08
17017	18	181	6203.75
15480	6	172	5909.58

Business: Default of credit card clients

Additional Information

This research aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel Sorting Smoothing Method to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

Research Questions

1. Probability of customer defaulting
2. What other factors are more likely to effect a customer defaulting i.e do younger customers default more then older clients ?

Exploratory Data Analysis

Education: Student Performance Dataset

Student Performance Data was obtained in a survey of students' math course in secondary school. It consists of 33 Column Dataset Contains Features like

- school ID
- gender
- age
- size of family
- Father education
- Mother education
- Occupation of Father and Mother
- Family Relation
- Health
- Grades
- This dataset can be used for Regression (as target variable Grade) as well as Analysis tasks. it might contain imbalanced category features.

Research Question

1. What top factors contribute to students performing well ?
2. How do male students perform in relation to female students?

Exploratory Data Analysis