Topic Ideas

Roger Bukuru

2024-02-10

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 ?
   - Identify relationships between the sales of different products ?
   - What was the average basket size ?

Exploratory Data Analysis

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