Report

Introduction

The data given for this assignment is a banking dataset. The columns include 'id', 'age', 'sex', 'region', 'income', 'married', 'children', ‘car', ‘savings account', ‘current account', ‘mortgage' and ‘Personal Equity Plan’. The aim of the assignment is to find the strongest association rules in this dataset and look for the strongest rules for people who bought the personal equity plan.

Association rules are a data mining technique used to discover relationships between items in a dataset. It involves identifying frequent itemsets, which are groups of items that often appear together in the data, and then using these itemsets to generate rules that describe the relationships between them.

Data Cleaning

The data comes with no NA values and is fairly clean. The ID variables has no meaning in our analysis. Most variables namely, sex, region, married, car, save_act, current_act, mortgage, pep have only yes/no values. I have used the get_dummies function of pandas dataframe to perform one hot encoding for all variables.

The continuous variables like age and price have been discretized by passing the column to a discretizing function. Age has been separated into 10-25, 5-40,40-55,55-70,70+. Income has been separated into "< 17000”,”17000-25000","25000-35000","> 35000” roughly based on the quartiles. The end result is a fully one hot encoded dataframe.

Association Rules Mining

We use the mlxtend library in python to get the association rules. We modify the value of confidence and lift so we can to get the strongest rules.

Confidence is a measure of the reliability of a rule. It is calculated as the proportion of transactions in which the antecedent (the left-hand side of the rule) occurs and the consequent (the right-hand side of the rule) also occurs.

Lift is a measure of the strength of the relationship between the antecedent and consequent of a rule, taking into account how frequently they occur together compared to what would be expected by chance. It is calculated as the ratio of the observed frequency of both antecedent and consequent to the expected frequency under the assumption that they are independent. A lift value greater than 1 indicates that the antecedent and consequent are positively correlated and occur together more frequently than would be expected by chance. A lift value less than 1 indicates that the antecedent and consequent are negatively correlated and occur together less frequently than would be expected by chance.

Support is a measure that indicates the frequency with which a particular itemset occurs in a dataset. It is calculated as the proportion of transactions in the dataset that contain the itemset.

Experimenting with Lift and Confidence values

Our goal is to find 20 strong rules from the dataset. We start with confidence > 0.5 and lift >1.0. The rules have been sorted by list and confidence.

T1

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
177	(age_10-25)	(income_< 17000)	0.145000	0.240000	0.111667	0.770115	3.208812	0.076867	3.306000
1688	(current_act_YES, age_55-70, save_act_YES)	(income_> 35000)	0.150000	0.260000	0.123333	0.822222	3.162393	0.084333	4.162500
1707	(age_55-70, mortgage_NO, save_act_YES)	(income_> 35000)	0.128333	0.260000	0.103333	0.805195	3.096903	0.069967	3.798667
874	(age_55-70, save_act_YES)	(income_> 35000)	0.203333	0.260000	0.153333	0.754098	2.900378	0.100467	3.009333
1516	(married_YES, age_55-70, save_act_YES)	(income_> 35000)	0.141667	0.260000	0.106667	0.752941	2.895928	0.069833	2.995238
1678	(current_act_YES, children_1, save_act_YES)	(pep_YES)	0.121667	0.456667	0.105000	0.863014	1.889811	0.049439	3.966333
951	(mortgage_NO, children_1)	(pep_YES)	0.140000	0.456667	0.118333	0.845238	1.850886	0.054400	3.510769
865	(children_1, save_act_YES)	(pep_YES)	0.158333	0.456667	0.133333	0.842105	1.844026	0.061028	3.441111
928	(current_act_YES, children_1)	(pep_YES)	0.168333	0.456667	0.140000	0.831683	1.821204	0.063128	3.228039
678	(married_YES, children_1)	(pep_YES)	0.148333	0.456667	0.123333	0.831461	1.820717	0.055594	3.223778
174	(children_1)	(pep_YES)	0.225000	0.456667	0.183333	0.814815	1.784266	0.080583	2.934000
1908	(married_YES, children_0, current_act_YES, sav...	(pep_NO)	0.145000	0.543333	0.133333	0.919540	1.692405	0.054550	5.675714
1919	(married_YES, children_0, mortgage_NO, save_ac...	(pep_NO)	0.133333	0.543333	0.121667	0.912500	1.679448	0.049222	5.219048
1930	(married_YES, children_0, current_act_YES, mor...	(pep_NO)	0.146667	0.543333	0.133333	0.909091	1.673173	0.053644	5.023333
1767	(married_YES, children_0, sex_FEMALE, mortgage...	(pep_NO)	0.116667	0.543333	0.105000	0.900000	1.656442	0.041611	4.566667
1509	(married_YES, children_0, save_act_YES)	(pep_NO)	0.198333	0.543333	0.178333	0.899160	1.654895	0.070572	4.528611
1549	(married_YES, children_0, mortgage_NO)	(pep_NO)	0.193333	0.543333	0.173333	0.896552	1.650095	0.068289	4.414444
1845	(car_NO, children_0, married_YES, mortgage_NO)	(pep_NO)	0.111667	0.543333	0.100000	0.895522	1.648201	0.039328	4.370952
1759	(married_YES, children_0, current_act_YES, sex...	(pep_NO)	0.118333	0.543333	0.100000	0.845070	1.555344	0.035706	2.947576
1054	(married_YES, children_0, sex_FEMALE)	(pep_NO)	0.156667	0.543333	0.130000	0.829787	1.527216	0.044878	2.682917

We get a dataset of 1417 transactions. Slowly increasing the lift and confidence values we can see that even though we sorted the values by lift and confidence a lot of rows got eliminated.

Increasing the minimum criteria for lift (> 0.75) and confidence (>1.5) gives us the 20 strongest rules from the dataset. [Table T1]

The strongest rule suggests that anyone who is between the age of 10 and 25 is earning over 17000. This rule suggests that I could have made a different segregation between ages 10-18 and 18-25 as we know most people between the ages of 10 and 18 are still in school.

The next three rules have the consequents as income>35000. We see that there are several factors that can lead to this consequent but the most common ones are if someone is between the ages of 55-70 and they possess a savings account, we can be quite sure that they make over $35000.

All other rules are related to the Personal Equity Plan. One key insight from this data is that we can clearly see that most people who have 1 child opted for the personal equity plan while most people with 0 children did not opt for the plan.

Keeping the same threshold values for confidence and lift, we see what attributes lead to people buying the personal equity plan. We see here that most rules are combinations of people having a savings account, a checking account and a child.

T2

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1678	(current_act_YES, children_1, save_act_YES)	(pep_YES)	0.121667	0.456667	0.105000	0.863014	1.889811	0.049439	3.966333
951	(mortgage_NO, children_1)	(pep_YES)	0.140000	0.456667	0.118333	0.845238	1.850886	0.054400	3.510769
865	(children_1, save_act_YES)	(pep_YES)	0.158333	0.456667	0.133333	0.842105	1.844026	0.061028	3.441111
928	(current_act_YES, children_1)	(pep_YES)	0.168333	0.456667	0.140000	0.831683	1.821204	0.063128	3.228039
678	(married_YES, children_1)	(pep_YES)	0.148333	0.456667	0.123333	0.831461	1.820717	0.055594	3.223778
174	(children_1)	(pep_YES)	0.225000	0.456667	0.183333	0.814815	1.784266	0.080583	2.934000

The most interesting rules however are

1. If someone has a savings and checking account and a child they are most likely to but the personal equity plan. In this case, the support is 0.105000, meaning that 10.5% of the transactions in the dataset contain all three items in the antecedent and the item pep_YES. The confidence is 0.863014, meaning that 86.30% of the transactions containing the antecedent itemset also contain the item pep_YES. The lift is 1.889811, meaning that the presence of the antecedent itemset increases the likelihood of buying the personal equity plan being present by 1.89 times compared to the baseline likelihood.

2. The second rule means that there is a relationship between the antecedent itemset (mortgage_NO, children_1) and the consequent item (pep_YES) in the dataset. Specifically, the rule states that customers who do not have a mortgage (mortgage_NO) and have one child (children_1) are likely to purchase a personal equity plan (pep_YES). 11.83% of the transactions in the dataset contain both the antecedent itemset and the item pep_YES. 84.52% of the transactions containing the antecedent itemset also contain the item pep_YES. The lift is 1.850886, meaning that the presence of the antecedent itemset increases the likelihood of the item pep_YES being present by 1.85 times compared to the baseline likelihood.

3. The next rule (T2-174) states that customers who have one child (children_1) are likely to purchase a personal equity plan (pep_YES). 18.33% of the transactions in the dataset contain both the item children_1 and the item pep_YES. 81.48% of the transactions containing the item children_1 also contain the item pep_YES. the lift is 1.784266, meaning

that the presence of the item children_1 increases the likelihood of the item pep_YES being present by 1.78 times compared to the baseline likelihood.

4. This means that customers who have a savings account (save_act_YES), do not have a mortgage (mortgage_NO), and are not married (married_NO) are likely to purchase a personal equity plan (pep_YES). The support is 0.106667, meaning that 10.67% of the transactions in the dataset contain all of the antecedent items and the item pep_YES. 74.42% of the transactions containing all of the antecedent items also contain the item pep_YES. In this case, the lift is 1.629604, meaning that the presence of the antecedent items increases the likelihood of the item pep_YES being present by 1.63 times compared to the baseline likelihood.

   We can see however this rule did not meet our threshold values. It is an interesting rule nonetheless as it goes in the opposite direction from the trend we saw where having 1 child usually means that the person will but the PEP

5. This means that female customers who are not married are also likely to purchase a personal equity plan (pep_YES). 10.67% of the transactions in the dataset contain all of the antecedent items and the item pep_YES. 74.42% of the transactions containing all of the antecedent items also contain the item pep_YES. the presence of the antecedent items increases the likelihood of the item pep_YES being present by 1.63 times compared to the baseline likelihood.

   Just as the rule 4 this one is also slightly lower than threshold value for confidence but a high value for lift.