Data stream classification of red and white wines marketing essay

Introduction

Well we got 2 data sets to analysis using SPSS PASW 1) Wine Quality Data Set and 2) The Poker Hand Data Set. We can do this using CRISP methodology. Let us look what is CRISP by wikipedia " CRISP-DM stands for Cross Industry Standard Process for Data Mining It is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems." PASW Modeler is a data mining workbench that enables you to quickly develop predictive models using business expertise and deploy them into business operations to improve decision making. Designed around the industry-standard CRISP-DM model, IBM SPSS PASW Modeler supports the entire data mining process, from data to better business results.

CRISP DM, Clementine's own " lightweight" methodology of 5 stages

Business Understanding, Data Understanding, Data Preparation

Modelling, Evaluation and Deployment.

CRISP Methodology

Business Understanding: Understanding the project requirements & objectives from a business perspective, and then converting this knowledge into a data mining problem definition

Data understanding

In this step following activities are going on, Data understanding, Collecting Initial Data then describing Data, Exploring Data and lastly verifying Data Quality

The data preparation phase

Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools. Cleaning Data using appropriate cleaning and cleansing strategies then Integrating Data into a single point.

Modeling:

Selection and application of various modeling techniques done in this phase, and their parameters are adjusted to optimal values. Basically, there are more than one technique for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. Steps consist of Generating a Test Design, Building the Models assessing the Model

Evaluation

Building of model (or models) takes place in this phase. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model.

Deployment

In the final stage Knowledge gained is organized presented so that an end user can easily use it. As per the requirements this can be a report or a complex data mining process. Normally Customers carry out the deployment step

Wine quality data set

Wine quality is modeled under classification and regression approaches, which preserves the order of the grades. Explanatory knowledge is given in terms of a sensitivity analysis, which measures the response changes when a given input variable is varied through its domain

The red wine data set contains 1600 samples out of which I have selected 200 random samples and doing the analysis(" Data mining cannot discover patterns that may be present in the larger body of data if those patterns are not present in the sample being " mined" ") . So I selected the data set bearing in mind. The data set I have selected has high confidence. With measurements of 13 chemical constituents (e. g. alcohol, Mg) and the goal is to find the quality of red and white wine.

Input variables

1 – fixed acidity

2 – volatile acidity

3 – citric acid

4 – residual sugar

5 – chlorides

6 – free sulphur dioxide

7 – total sulfur dioxide

8 – density

9 – pH

10 – sulphates

11 – alcohol

Output variable is quality (score between 0 and 10)

CRISP methodology has been followed through out the phase . By checking the web site and resources learned about the wine domain . the next step was to check whether incorrect, missing or " abnormal" values in the data set end ensure the data quality. Data quality of the data set is very good.

PASW Data stream classification of red and white wines

Classification for Red and White wine

2 data sets red wine and white wine have been imported using variable file nodes Use of type node here is to describe the characteristics of data. . The Classification and Regression (C&R) Tree node is a tree-based classification and prediction method. Similar to C5. 0, this method uses recursive partitioning to split the training records into segments with similar output field values. The C&R Tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is

subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary (only two subgroups)

Red Wine’s variable importance White wine variable importance

From variable importance diagram we can say that important attribute to determine Red wine quality is pH. The variable importance is in the order pH, citric acid, chloride as shown in the figure1. But for determining White wine’s quality the most contributing attribute is chloride and 2nd attribute is Alcohol.

Decision Tree Model of white wine (only a portion)

Analysis and conclusion

The above generated tree consists of nodes and its children. The top node represent the total number of wine samples and how many number belongs to different categories(1 to 9). The first split is on chloride. This implies that most of the wine belongs to chloride level <0. 041 and some belongs to chloride level> 0. 041. We see that good quality wine has chloride level <0. 041. If we stop after one split we would say cl <0. 041 are good wine. Second split from LHS node is on the basis of fixed acidity varies most between good quality and bad quality wine. In RHS wine quality difference in citric acid.

It has been found from count Vs Quality graph that how many belongs to good quality categories. Alcoholic concentration of white wine samples is more than that of red wine sample. Good wines normally have high concentration. So we can conclude that White wine samples are good. In the white wine chloride level is normally high that implies it has got good Aroma. Where as in red wine the citric level is between particular levels that shows the red wine is very tasty!!

PASW has got a number of 2-D and 3-D charts like bar, pie, histogram, scatter etc for time being I am using linear graph and 3-d scatter graph. You can use any of the graph as per the requirements. Some graphs are easy to interpret . Let us consider a 2-D graph between most contributing variable pH and quality from the graph it is clear that the relation ship between pH and quality is in such a way that if pH is in between 3. 23 and 3. 27 quality is good. Quality is very low for 3. 38 and 3. 50. We can plot similar graph between quality and citric acid or towards what ever contributing variable then find out the relation ship between them

2-D graph represent the relation ship between quality and pH of Red wine

Let us plot a graph between chloride and Quality for the white wine. In the below figure it shows the quality is very good when chloride level below 0. 036. And quality in the range 5 to 6 when chloride level is above . 048.

Like this if plot a graph between quality and alcohol we will see the quality is too good if alcoholic concentration in between 12. 5 and 13(as per the sample I have analyzed)

3D graph which shows the relation ship between alcohol, quality and chloride level of white wine from the 2d analysis it was shown how the quality is being affected by single variable. If the one variable does not tell about how quality being related we can check relation ship between 3 variables using a 3d graph. It is having 3 axes.

How Regression is useful

In this multiple regression , Predictors such as (Constant), alcohol, fixed acidity, residual sugar, chlorides, volatile acidity, free sulfur dioxide, sulphates, pH, total sulfur dioxide, citric acid, density determine the value of quality. Below gave a Pasw stream for regression.

As per the variable importance graph volatile acidity, total SO2 and alcohol are most important variables in Regression analysis.

Model

R

R Square

Adjusted R Square

Std. Error of the Estimate

1

. 792(a)

. 626

. 474

. 542

Each by changing the independent variable’s value we can get value of dependent variable quality. With the help of a hypothesis we need to understand and build a relation ship among the variables. To predict the mean quality value for a given independent variable (say volatile acidity) we need a line which passes between the mean value of both quality and volatile acidity and which minimize the sum of distance between each of the points and predictive line. This fits into a line.

The Poker Hand Data Set

Each record is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is described using two attributes (suit and rank), for a total of 10 predictive attributes. There is one Class attribute that describes the “ Poker Hand”. The order of cards is important and there are 480 possible Royal Flush hands. Below discussing about how to determine poker hands using data mining. I am considering classification only. If we consider clustering/Regression it does not make any sense

PASW MODEL CLASSIFICATION USING CRT ALGORITHAM

We got training and testing data set . First applying a model on training data set. Source file is a

Comma separated file (CSV) with 1 million rows. It is difficult to do analyse on this input data set so selected sample data set and doing the analysis.

Problem faced

The given source data was not in a meaning full format so I have given meaningful attribute name and Values by using Vlookup function in MS excel, now the data has become more meaning full and it looks like below. Data cleansing is very important and comes under data preparation phase of the methodology

Accuracy of predictive model

The accuracy of predictive model is checked by analysis node. It has been found that accuracy is 90%.

Using the Algorithm need to predict any of these:

0: Nothing in hand; 1: One pair; 2: Two pairs; 3: Three of a kind; 4: Straight; 5: Flush;

6: Full house; 7: Four of a kind; 8: Straight flush; 9: Royal flush;

Pocker hands variable importance diagram

Let me say what did I understood from the diagram. Rank2 (rank of card2) is most contributing variable to predict poker hands. It is clear that Rank of 1st, 4th and 2nd cards are more contributing than suit of those cards.

The different section of pie chart represents number of cards in a particular poker category. Blue represents No Poker; Red represents ONE PAIR, Green represent Royal flesh

How Pasw helps to do classification

Pasw has got number tree constructing algorithms(C&R, c5. 0) to do classification. I considered Classification and Regression (C&R) though this is not a time efficient algorithm time complexity is more when compared to c5. 0)I selected C&R. The data set I have got is simple one and I am not considering the deep analysis all I need to do is to predict poker hands so C&R can do it. Below shows the constructed tree using C&R (Ashort description of tree already given above)

Analysis

Data has been classified into Training set and Testing set . Here most of the data set into a training set and small portion of data is used for testing. After a model has been processed by using the Training set, we can test the model by making predictions against the Test set. Since the data in the training set already contains known values for the attribute that you want to predict. Below giving the portion of training set being used.

Integrating classification and association rule mining can produce more efficient and accurate classifiers. Here each row is an instance

Trial: pair of 5 attributes (SUIT and RANK) + classification class. So this can be used to predict the classification of other unclassified instances. consider the training set given below

suit1

rank1

suit2

rank2

suit3

rank3

suit4

rank4

suit5

Rank5

poker

Heart

ASS

heart

KING

spades

4

Spades

3

heart

QUEEN

Nothing in hand

Diamonds

QUEEN

diamonds

2

diamonds

JACK

Clubs

5

spades

5

ONE PAIR

Hearts

10

hearts

jack

hearts

king

Hearts

queen

hearts

1

royal flesh

Spades

Jack

spades

king

spades

10

Spades

queen

spades

1

royal flesh

Diamond

queen

diamond

jack

diamond

king

diamond

10

diamond

1

royal flesh

Hearts

5

diamond

king

spades

king

spades

7

clubs

5

two pairs

Hearts

4

hearts

1

hearts

3

diamond

5

diamond

2

straight

Suppose want to predict below hand is what type of Poker hand?

suit1

rank1

suit2

rank2

suit3

rank3

suit4

rank4

suit5

rank5

poker

Club

10

club

jack

club

A

club

king

club

queen

???

From the training set data the testing set is predicted, answer is Royal Flesh

Conclusion

Two data sets the wine and poker have been analysed using CRISP methodology and using the tool IBM SPSS PASW, used different modelling techniques which suits. Analysed the knowledge elicited by each model

DATA MINING & KNOWLEDGE DISCOVERY IN MARKETING (PART 2)

Abstract

Now-a-days Using the high power computing and information technology enables to collect store and process complex Marketing data. Data mining is used to extract knowledge from this marketing data. This report discuss about Data mining process, short discussion about different mining techniques such as classification tree, neural network, Regression and their application in marketing domain. My report Also cover different type of analyzes and tasks being used

Introduction

From the given topics I have selected the topic Data mining and Knowledge discovery for marketing since my cup of tea is Business and computing. I would always like to do research in Business analytics . Well let us look at what is data mining

Data mining is the process of discovery of interesting, meaningful and actionable patterns hidden in large amounts of data . This is one of the tools to transform data into information. It is widely used in almost all fields of science and business' profiling practice such as marketing, fraud detection, and scientific discovery.

The technique to uncover pattern on data can also apply on sample data . so the sample data should be so the sample should be a good representative of larger data set. data mining can not find out the pattern which may be present in larger body of data and not contains in the small sub set of data. So this is very useful when sufficiently represented data are collected

Most well known branches of data mining is knowledge discovery or KDD

It derives knowledge from input data . This knowledge which have got from the process will become additional data and can be used for further discovery in related field normally an analyst can analysis and predict it. DM can generate thousands of pattern but all these patterns are not interested and useful.

In this I am considering Data mining in a marketing field prospective. The data coming from different sources like transactions, loyalty cards, and discount coupons; customer complaint calls public life style studies using this data we can make Target marketing like

to identify appropriate customer segments for new marketing initiatives

determine customer purchasing pattern over time

associations/co-relations between product sales, & predict based on such association I mean cross market analysis

what type of customer buys what type of product that is customer profiling

Predict likelihood of customer churn and target those likely to leave with retention campaigns

Customer requirement analysis like Identify the best products for different groups of customers and Predict what factors will attract new customers

Provision of summary information such as Multidimensional summary reports and Statistical summary information (data central tendency and variation)

Another question is why can not we go for a traditional data analysis instead of data mining? Answer is the field like marketing has tremendous

Amount of data and it has multi dimension and complexity. A Marketing firm would likely to segment their customers into similar groups or clusters in order to better understand consumer behavior and more effectively market their products. In the past for a small business initiatives did not have trouble to understand their customers. They knew what they have to do once a customer approach them . Today’s business is more competitive, more customer oriented, more products oriented so it is very difficult to understand the customer behavior, wants, needs the hidden relation ship between the data and preferences. With the help of data mining an analyst can deliver timely, personalized promotional offers.

1

Knowledge Discovery (KDD) Process

S2

S1

S3

Data Cleaning

Data Integration

Databases

Data Warehouse

Knowledge

Task-relevant Data

Selection

Data Mining

Pattern Evaluation

Normally in the huge DWH & data mining environment data coming from various sources integrated and put it in data warehousing. Various data mining soft wares like teradata intelligent miners are used to mine Tera bytes of data and find market prediction.

As I mentioned the DM is a Tools for developing predictive and descriptive models. Some are statistical method such as regression. Other use non statistical method like neural networks, classification trees. Here I considered some important tools then their

How Classification trees are being used in marketing data mining

Classification tree partition the data to maximize the difference in the dependent variable. it is also called a decision tree. Aim of classification tree is to classify the data into distinct groups or branches that create the strongest separation in the values of the dependent variables. The tree can identify segments. This can be helpful when a company is trying to understand what is driving market behavior. It detects nonlinear relationship.

Mailed 10000

2. 6%

Male 4677

3. 2%

Female

2. 15

2. 1 %

<£30 1. 7%

1. 7%

£30-45

3. 6%

>£45

4. 1%

Age> 40

4. 3%

Age <40

0. 7%

Box shows resp rate in percentage

The tree growth is through series of steps and rules . say for example sales pieces were mailed to 100000 names and yielded a response rate of 2. 6%. the first split is on gender. This indicates that greatest difference between responders and non responders is gender. We see that males are much more responsive than females. We would consider males the better target group If we stop after one split. Our goal is to find out group with in both genders that discriminates between responders and non responders. In the next level split male and female groups are considered separately

The second level split from the male node is on income, this implies that the income level varies in most between responders and non responders among the males. For female greatest difference is among the age group . It is very easy to identify the group with the highest response rate. Lets say that management decides to mail only to groups where the response rate is more than 3. 5%. the offers would be directed to males who makes more than £30000 a year and female over age 40

Some typical Classification tree Algorithms are

“ 1) C4. 5: Quinlan, J. R. C4. 5: Programs for Machine Learning. Morgan Kaufmann., 1993. & 2) CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984”

Linear regression and its applicability in marketing

Knowledge of deviation from normal is very important for a marketer. In the past such deviations were very difficult to detect. Now-a-days data mining tools give great flexibility to detect and classify these changes.

It is a statistical technique that quantifies the relationship between dependent variable and the independent variable, these are continuous. Consider the below equation, it shows a relation ship between sales and advertising along the regression equation . Our goal is to predict the sales based on the amount spend on advt. Plot a graph sales vs. advt that would be linear. A key measure of the strength of the relationship is the R-square. It measures the amount of overall variation in data that explained by the model. More than 70% Of the variation in sales can be explained by variation in advertising. Some times the relationship between sales and Advt is non linear (may be curvilinear) . By using the square root of advertising we are able to find better fit for the data.

Sales= 17. 813+. 0897*Advertising

£120

£1, 503

£160

£1, 755

£205

£2, 971

£210

£1, 682

£225

£3, 497

£230

£1, 998

£290

£4, 598

£315

£2, 937

£375

£3, 622

£390

£4, 402

£440

£3, 844

£475

£4, 470

£490

£5, 492MINIMIZE SQUARED ERROR Advt sales

ADVEGRTISIN ——-ƒ x axis

When building targeting models for marketing, risk and CRM, it is common to have much predictive variable. Using multiple predictive or independent continuous variables to predict a single continuous variable is called multiple linear regression . Targeting model created using linear regression is generally very robust. In marketing they can be used alone or in combination with other model.

Neural Networks and its applicability in marketing

Neural network does not follow any statistical distribution (Neural network is very vast topic a complete discussion is beyond the scope of this report) . it is modeled after the function of the human brain. The process is one of pattern recognition and error minimization.

we can say it as nodes that are arranged in layers. The figure tells simple neural network with one hidden layer. Data has been classified into training and testing set (before the process). Then weight or " input" is assigned to each of the nodes in the first layer. During each iteration , the input are processed through the system and compared to the actual value . the error is measured and fed back through the system to adjust the weights. The weights get better at predicting the actual results. A error limit is defined and it check with the error limit the process finishes when the minimum error limit reached

" One specific type of neural network commonly used in marketing uses sigmoidal functions to fit each node. This technique is very powerful in fitting a binary or twoilevel outcome such as response to an offer or a default on a loan"

Neural network not only pick linear data but also do a good pick up with non linear relation ship in the data. So this allows fitting data which is not possible to fit using regression. One disadvantage we can say that the result of neural net work is some what difficult to interpret

A brief description on how Clustering can applicable in data mining

Cluster analysis Cluster analysis group respondents with similar behaviors, preferences, or characteristics into segments. By doing so we can understand important similarities and differences between the respondents. Analyst can use this information to develop targeted marketing strategies, or to provide subgroups for analysis. In market survey data, clustering enables market researchers to group respondents who provide similar responses on several questions. In Clustering we use more than one variable that analyzes responses to several questions in order to find similar respondents. Clustering is based on the concept of creating groups based on their proximity to, or distance from, each other. Respondents within a cluster, therefore, are relatively homogenous.

Most widely used Algorithms are

“ 1)K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967

2) BIRCH: Zhang, T., Ramakrishna, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD ’96”

Let us look at some more major areas of application of data mining in the marketing like Customer profiling, Deviation analysis and Trend analysis. The pattern which formed after mining the data helps in analytics.

Customer profiling

This help to predict several marketing decision. A customer profile is a model of customer based on this marketer decides on the right strategies and tactics to meet the needs of that customer . The data mining task used in customer profiling can be dependency analysis, class identification and concept description. Below giving set of transaction that can help marketer to construct useful customer profiles.

Frequency of purchases

Marketing firm can build targeted promotion offer such as “ frequent buyer programs” by looking how often their customer purchases product from their shop.

Rcency of purchases

The meaning of term is How long has it been since this customer last placed an order? Suppose a customer frequently visit the shop. It has been found that the specific customer or customer group not visiting the firm over long period of time . Market investigate the reason. By knowing this they can take appropriate offer or action.

Size of purchases

It tells, on a particular transaction how much he or she spends. This information helps to give resources to those customer groups.

Identifying typical customer groups

It gives characteristics of each group . For example a profile indicating that the customer has purchased a WINDOWS 7 SOFTWARE CD may hold to the marketer offering a special deal for MICROSOFT OFFICE SOFTWARE CD.

Prospecting

Customer profiles like buying patterns, give clues to the marketer on prospective customers. Say for example, consider the pattern Purchase of Norton Anti Virus package with one year validity is followed by purchase of Norton Up gradation version /or new version within 11 months about 85% of the time by high income customers discovered by data mining. Analyst who analysis pattern can identify the prospective customers for Upgraded/new version based on first time purchase details and tailor the mail catalog accordingly, thus, increasing the prospect of sales.

2 Deviation analysis

Deviation analysis is one of the important analysis for example a higher than normal credit purchase on a credit card can be a fraud anomaly or a genuine purchase by the customer changes. Once a deviation has been discovered as a fraud, the marketer takes appropriate steps to prevent such frauds and initiates corrective action. If the deviation has been discovered as a change, further information collection is necessary. For example, a change can be that a customer got a new job and moved to a new house. In this case, the marketer has to update the knowledge about the customer.

3) Trend analysis

Trends are patterns that persist over a period of time. Trends could be short-term trends like the immediate increase and subsequent slow decrease of sales following a sales campaign. Or, trends could be long-term, like the slow flattening of sales of a product over a few years. Data mining tools, such as visualization, help us detect trends, sometimes very subtle and hidden in the database, which would have been missed using traditional analysis tools like scatter plots. In marketing decisions, trends can be used for evaluating marketing programs or to forecast future sales.

Data mining task in marketing data mining domain

These tasks present in all data mining process we are just looking it into marketing prospective

Dependency analysis

Data Visualization

Class identification

Deviation Detection

Concept Description

Dependency analysis

The market basket analysis gives the relationship between different product purchased by a customer . Using this techniques we can develop marketing strategy for promoting product that have dependency relationship in customers mind.

Class identification

It groups customers into classes which are defined in advance. Mathematical taxonomy and clustering are being used for class identification task. What the first one does is it maximizes the similarity with in classes but minimize similarity between classes. In clustering approach it determine the clustering according to attribute similarity as well as conceptual cohesiveness as defined by domain knowledge (describe above). A company doing business over the net, based on the session log data of internet users, the firm can classify the web users into “ email only users” “ Surfers” or “ Just for fun Surfer” etc

Concept description

Comparison analysis will be done using statistical techniques. Using this we can compare marketing and customer knowledge.

Deviation detection

This helps us to determine the anomaly and changes. We can find the anomaly from various statistical techniques. This is already being explained above.

Data visualization

This kind of software’s allows the market research team or concerned people to view complex 3-D and 2-D patterns. They also provide drill down drill up &slice facilities. In the KDD (knowledge discovery from data base) process, data visualization is used in association with other tasks such as dependency analysis, class identification, deviation detection and clustering. IBM SPSS

PASW has got good data visualization techniques. Some of them are explained in Part 1 of the report.

Conclusion

Report discussed about Data mining process,