DATA MINING AND ANALYTICS FOR MANAGERS-MGMT63582

A PROJECT REPORT ON ANALYSIS OF GERMAN CREDIT DATA SET

BY (GROUP -9)
SOWMYA NELLORE (sn478)
SAINATH DUTKAR (sgd23)
AQSA SHEIKH (ars92)

INDEX

1.ABSTRACT:

Data mining provides us with various methods and very useful tools to extract knowledge from a complex data set. In this project we aim to use these tools to analyse and gain knowledge from the “German Credit data set”, which contains valuable data that helps bank to make decisions regarding loan approvals.In this paper Information theory and domain knowledge are used to choose minimum attributes to reduce complexity. These attributes are used analyse the credit data using predictive modelling techniques-neural networks and decision trees implemented using MATLAB. Prior to building the model, the dataset is pre-processed, reduced and made ready to provide efficient predictions. The final model is used for prediction with the test dataset and the experimental results prove the efficiency of the built model.

2.INTRODUCTION:

Data mining is a way of finding patterns in expansive data sets including techniques at the convergence of machine learning, database and statistics. It is a basic procedure where keen strategies are connected to extract data patterns. In brief the general objective of the data mining process is to extract data from a data set and change it into a reasonable structure for additional use. Finding an accurate outcomes, however, requires an insightful determination of information mining tools, strategies as well as procedures that are suitable for the business, organization to get the desired result. Data mining is a new and demanding field for the present age as a result of its extensive applications. For the most part of saying, it has an pulled in a lot of consideration in the data business and in the public eye, due to the wide accessibility of large amount of data and the unavoidable requirement for transforming such data into valuable data and learning.

This paper describes the techniques we use to analyze “German Credit Data Set” which contains valuable data that help bank to make decision regarding the approval of loan. This paper further discusses the components of Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which we have used for the mining of “German Credit Data Set”, including Business Understanding, Data Understanding, Data Preparation, Model Building Based on Neural Network and decision trees, Testing and Evaluation and, Deployment.

3.DATA MINING METHODS:

3.1 CLASSIFICATION ANALYSIS:

Classification is a data mining technique that assigns categories to a collection of data in order to aid in more accurate predictions and analysis. This method is utilized to recover critical and important information about a data set. It is utilized to organize information in various classes.

Very large databases are becoming the norm in today's world of "big data." The primary challenge of big data is how to make sense of it.
And sheer volume is not the only problem: big data also tends to be diverse, unstructured and fast-changing. Consider audio and video data, social media posts, 3D data or geospatial data. This kind of data is not easily categorized or organized. To meet this challenge, a range of automatic methods for extracting useful information has been developed, among them *classification*.

Applications:
Classification, and other data mining techniques, is behind much of our day-to-day experience as consumers.
Weather predictions might make use of classification to report whether the day will be rainy, sunny or cloudy. The medical profession might analyze health conditions to predict medical outcomes. From fraud detection to product offers, classification is behind the scenes every day analyzing data and producing predictions.

3.2 CLUSTERING ANALYSIS:
The cluster is a collection of data objects; those objects are comparable inside a similar cluster. That implies the items are like each other inside a similar group also, but they are fairly unique or they are different or random to the objects in different groups or in different clusters.
Clustering analysis is the way of finding clusters and groups in the data such that the level of relationship between two objects is maximum if they belong to the similar group and least generally.

Applications:
Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
Clustering also helps in classifying documents on the web for information discovery.

3.3 ASSOCIATION RULE TECHNIQUE:
It alludes to the technique that can enable you to distinguish some intriguing relations (reliance demonstrating) between various factors in extensive databases. This procedure can enable you to unload some shrouded designs in the data that can be utilized to recognize variables inside the data and the simultaneousness of distinctive variables that seem much of the time in the dataset. Association rules are valuable for looking at and determining client conduct. It is profoundly suggested in the retail business investigation. This strategy is utilized to decide shopping crate data analysis, item clustering, designing catalog and store design. In IT, developers utilize association standards to build programs capable of doing machine learning.

Applications:

Association rules are widely used in marketing, medical or credit card fields, to name just a few. A classic example of association rules that comes in the literature quiet often is the market basket analysis, which involves data collection by scanning bar codes in the supermarket to identify what items are being bought together in a single transaction. Based on that information, the store managers can adequately arrange the store layout in hopes of increasing the sales or cross selling.

3.4 INFORMATION THEORY:

Information theory studies the quantification, storage, and communication of information**.**. Information Theory is nearly connected with a gathering of unadulterated and connected orders.A key factor in information theory is "entropy". Entropy gives the measure of the uncertainty in the information. The entropy is calculated for each variable or attribute to measure the measure the uncertainty in the attribute.The entropy is calculated using the below formula

$$entropy=-p1\ log(p1)-p2\ log(p2)-\ldots.$$

Where p is the probability of the attribute within the particular set. To analyse the contribution of the attribute towards the output information gain is used.Information gain is a function of parent attribute and its child.In supervised learning we use information gain to calculate how much information an attribute has provided by partitioning ,i.e, how much purer the child is than the attribute.The information gain is calculated using the below formula

$$Information\ gain= entropy(parent\ attribute)-[-p1\ log(p1)-p2\ log(p2)-\ldots.$$

Applications:
Information theoretic concepts apply to cryptography and cryptanalysis. Concepts from information theory such as redundancy and code control have been used by semioticians such as Umberto Eco and Ferruccio Rossi-Landi to explain ideology as a form of message transmission whereby a dominant social class emits its message by using signs that exhibit a high degree of redundancy such that only one message is decoded among a selection of competing ones

3.5 DECISION TREE:

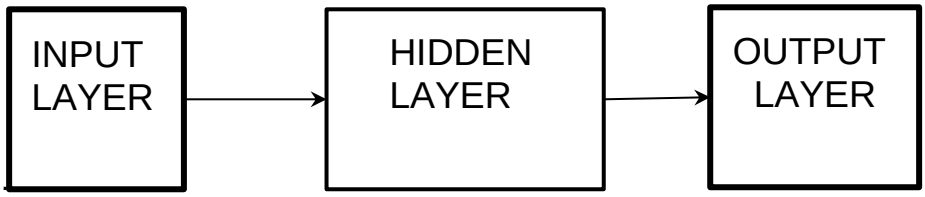
A decision tree is a guide of the conceivable results of a progression of related choices. Decision tree is one of the most used techniques in data mining because of its simplicity to explain the results. Besides, there are decision tree algorithms that work with parallel and incremental techniques, which help to process large databases for classifying new objects faster than traditional algorithms.
A decision tree ordinarily begins with a single node, which branches into conceivable results.

Applications:
Decision trees are very “user-friendly” because they are easy to understand by practically everyone and provide reasonably accurate results. They provide an easy to analyze breakdown

of the data and can be used practically in every business area that requires decision-making, including, but not limited to marketing, pharmacology, financial analysis, manufacturing, production, etc.

3.6 NEURAL NETWORKS:

A Neural Network, frequently, is a mathematical model propelled by biological neural network. A neural system comprises of an interconnected gathering of counterfeit neurons, and it forms data utilizing a connecting way to deal with calculation. As a rule a neural network is a versatile framework that progressions its structure amid a learning stage. Neural systems are utilized to display complex connections amongst information sources and yields or to discover designs information.



Applications:

The utility of counterfeit neural system models lies in the way that they can be utilized to induce a capacity from perceptions. This is especially helpful in applications where the many-sided quality of the information or undertaking makes the plan of such a capacity by hand illogical. Application regions incorporate quantum chemistry, pattern recognition, sequence recognition, game-playing and decision making, medical diagnosis, visualization and e-mail spam filtering, financial applications and data mining.

4.TECHNIQUES AND APPROACHES:

4.1 BUSINESS UNDERSTANDING-

The aim of the project is to analyse the German Credit data set and create a predictive model based on supervised learning. The model is used to help the bankers find out whether a customer is good or bad so that, the decision on giving a loan or not can be taken.There are two types of risks associated with wrong decisions-

-Predicting a good customer bad and not approving a loan could cause business loss to the bank.

-Predicting a bad customer good and approving a loan could cause financial loss to the bank.

4.2 DATA UNDERSTANDING-

The data set contains 1000 rows, where each row provides information about one customer.It contains 20 attributes that represents different features.And also a target attribute which shows

the actual outputs, showing whether the customer is good(1) or bad(2).All the values in the data set are numerical.The data set uses a cost matrix of-

Actual\Predicted	1(good customer)	2(bad customer)
1(good customer)	0	1
2(bad customer)	5	0

4.3 DATA PREPARATION:

By manual analysis and using microsoft excel, the data is preprocessed making sure that there are no null or missing values.The analysis of the data is based on supervised learning .The target variable given in the data is well defined and uses 1 to represent good customer and 2 to represent bad customer.The cost of the target is also well defined and the cost matrix is defined as above.Out of the 1000 tuples given the 980 are considered as the training set and 20 are taken for testing the models accuracy. The number of attributes are reduced to reduce the complexity of analysis. The selection of the attributes is done based on information theory and domain knowledge.The following attributes are selected:

1. Status
2. Duration
3. Credit History
4. Purpose
5. Credit Amount
6. Savings

4.4 MODEL BUILDING BASED ON NEURAL NETWORK:

The Basic aim is to select a Model that will help us solve the Problem. The main aim is to understand the problem Classify the data, Find out class probability estimation, its ranking, regression.

The Model we are using here is the Neural Networks in MATLAB. This provides algorithms, pretrained models to create, train, visualize, and simulate both shallow and deep neural networks. We will use this to perform classification, regression to get results.

By selecting Attributes according to our need and using Neural Network we can find appropriate result.

4.5 TESTING AND EVALUATION-

We ran number of test for Decision Tree and Neural Network with the aim to identify the best solution to fit the given data**.** The aim is to identify the best Neural Network with appropriate

attributes, neurons and layers. We have run number of test cases by building networks with different Attribute, Neurons and Layers.

Neurons,Layers	Attributes	Precision	Accuracy	Cost points
(10,2)	20	59%	65%	25
(10,2)	11	51%	60%	24
(10,2)	8	57%	70%	22
(10,2)	6	47%	80%	16
(86,7)	8	58%	75%	20
(75,9)	11	58%	70%	25
(90,6)	6	59%	70%	20
(10,1)	5	48%	65%	28
(80,5)	5	56%	65%	30
(210,10)	20	75%	80%	16
(210,10)	6	55%	70%	18
(720,14)	6	64%	75%	13

Please find the attached document containing Testing and Evaluation for the Neural Network

NeuralNetwork_Testing.docx

Please find the attached document containing Testing and Evaluation for the Decision Tree

DecisionTree_Testing.docx

4.6 DEPLOYMENT-

Neural Networks and decision tree models are used to obtain results. We would like to validate our test observations with the Domain experts and gets feedbacks to improve the models. During this process benefits and the risks are taken into account and cost matrix is used to

calculate business cost. Different models are compared and tested to find the correct model for the data. Holdout data with 5K cross validation is used to test the model.The Output we get is used to compare with the target output provided to know the actual performance of the Model.

5. RESULTS:

When compared between the decision tree and Neural network on the basis of simulation output, Neural Network seems to perform better with an accuracy of 75%.

DECISION TREE VS NEURAL NETWORK

	PASS	FAIL	ACCURACY	GOOD customer Predicted BAD	BAD customer Predicted GOOD	COST POINT
NEURAL NETWORK	15	5	**75%**	3	2	**13**
DECISSION TREEE	14	6	**70%**	4	2	**14**

NEURAL NETWORK

PREDICTION			
PASS	FAIL	TOTAL	ACCURACY
15	5	20	75%

	PREDICTIONS	COSTPOINTS	TOTAL
BAD TO GOOD	2	5	10
BAD TO BAD	2	0	0
GOOD TO BAD	3	1	3
GOOD TO GOOD	13	0	0
TOTAL	**20**		**13**

NeuralNetworkReults.xlsx

Please refer the attached results document

DECISION TREE

PREDICTION

PASS	FAIL	TOTAL	ACCURACY
14	6	20	70%

	PREDICTIONS	COSTPOINTS	TOTAL
BAD TO GOOD	2	5	10
BAD TO BAD	2	0	0
GOOD TO BAD	4	1	4
GOOD TO GOOD	12	0	0
TOTAL	**20**		**14**

DecisionTree_Results.xlsx

Please refer the attached results document

6.CONCLUSION:

Model was built in neural network for the data on trial and error based analysis, used to select the best model with maximum accuracy. The model's accuracy is compared to another classification model that is decision trees. It's found that the number of higher the number of neuron and layers the maximum accuracy of the output. And the attributes Status, Duration, Credit History, Purpose, Credit Amount and Savings contribute more towards the target.