

CSE 471 Project Writeup

Team Name

David Ganey

Kerry Martin

Evan Stoll

Michael Theut

Ben Roos

April 6, 2015

Contents

1	Part 1	2
2	Part 2	2
2.1	Problem Formulation	2
2.2	Solution Proposals	3
3	Part 3	4
4	References	4
5	Appendices	4

1 Part 1

2 Part 2

Part 2 of the project describes another task given by the CEO of the CactusCard Credit Company, wherein a program should be developed to provide “a *machine learning based approach* to identify individuals who should/should not be given the CactusCardPlus.” This section will formulate this problem and describe the steps taken to solve it.

2.1 Problem Formulation

The problem itself is a machine-learning problem, which at its core simply means a program designed to recognize patterns. Machine learning can encompass a wide variety of problems, which can be divided into “regression” problems and “classification” problems. In this case, the consumer research division has given the team a dataset which contains information about customers *for whom the decision to award or not to award the CactusCard has already been made*. This is important information – because customers can either be given the card or not (a binary *classification*), this falls under the realm of “classification” machine learning

problems. Additionally, the fact that the classification program will be trained with data means this is a “supervised” learning problem.

As a supervised classification machine, the program must be able to read the data set and, with that information, draw conclusions about which credit variables have the largest impact on whether or not the customer was given the card. To be an effective tool for the CactusCard corporation, the dataset supplied by the consumer research division should only include proper decisions – cases where the card was awarded to the correct individual and denied to individuals who would have abused the line of credit. Otherwise, the machine will learn “bad habits” and will draw conclusions from the data which are not useful to the company.

The dataset itself is comprised of 15 credit attributes for the individuals, as well as an indication of whether the individual received a card. The credit attributes are not specified, though the problem specification does include the datatypes for each attribute. (For example, attribute 1 can be either “b” or “a”, and therefore represents some binary information about the customer such as whether they are male or female).

The goal of this component of the project is to demonstrate a program which can use the dataset to learn how the credit attributes affect the decision to award the card, and from there use that knowledge to predict card decisions within a certain degree of accuracy.

2.2 Solution Proposals

Weka, a software tool developed at the University of Waikato, is a software package with implementations of a large number of machine learning algorithms. Weka can be used in multiple ways. One option is to simply use the Weka API to access the algorithms from Java code. This is advantageous when one wishes to write a full machine learning system. Another option is to use the Weka Explorer, which is a wrapper around the algorithms. This GUI tool allows the user to load a data set and run the algorithms on it, then view the results in various formats.

It is this method which is proposed to solve the task given by the CEO. Using the Weka Explorer, we have a straightforward method for gaining insight into the effect of the 15 credit attributes. We can use the Explorer to test various machine learning algorithms, and based

on their success rates, determine which should be used in the future to make the actual issue decision.

3 Part 3

4 References

5 Appendices