Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 158 (2019) 254–259

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd World Conference on Technology, Innovation and Entrepreneurship (WOCTINE)

Feature Selection in Credibility Study For Finance Sector

Ali Tunç[1]

[1] *Kuveyt Türk Katılım Bankası AR-GE Merkezi, KONYA-TURKEY*

Abstract

With the advances in the Information Technology field, banks can evaluate the credit requests of the customers via effective analytical methods and risk analysis. The software products, named Credit Scoring Systems, consist of collecting customer data based on pre-determined credit factors, processing the data with various statistical or machine learning methods, and conducting a credit risk analysis to make the final credit decision. The effects of the quantitative values of the properties formed in the data set vary according to the results. Determination a subset set of columns with a high impact on the outcome and meaningful and removal of irrelevant columns without effect according to the values in the features of a dataset is called the property selection. It is generally used for accuracy and scaling. In this study, it was studied to determine the areas that most impacted the credit result on the sample dataset to meet the need of the structure that can apply the credit to the consumers who apply for the loan and manage the assessment in the consumers. Information Gain and Gain Ratio algorithms were used to determine the most useful features. As a result of the study conducted on the data set, the valuable values of function were committed by using the Gain Ratio, and Information Gain algorithms, and the characteristics were listed according to their magnitude.

Peer-review under responsibility of the scientific committee of the 3rd World Conference on Technology, Innovation and Entrepreneurship

Keywords: Information Gain; Gain Ratio; Feature Selection; Credit Score;

1. Introduction

With the developments in information technologies, bank companies can evaluate their customers' credit demands through effective analytical methods and risk analysis. The software products, which are called credit-scoring systems, generally consist of collecting the data of the customer according to the credit factors previously determined, processing the obtained data with various statistical or machine learning techniques and learning the credit decision by making a credit risk analysis.

* Corresponding author. Tel.: +90-505-529-8223
E-mail address: ali.tunc@kuveytturk.com

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd World Conference on Technology, Innovation and Entrepreneurship
10.1016/j.procs.2019.09.049

The purpose of these systems is to develop models that help prevent errors in the credit decision stages and offer a standard solution for the evaluation of different decision factors. Instead of the statistical analysis methodologies used in credit scoring systems, a solution has been used in which artificial learning techniques are used which can be adapted according to the credit criteria of each bank.

Various scoring models are widely used in the evaluation process of loan applications. In these models, the past bank transactions of the customers can be processed, and a credit decision can be made. This study is based on a systematic methodology developed to determine the factors affecting credit risk analysis and to make a logical analysis.

2. Literature and Methods

Credit scoring is a numerical expression of an individual's creditworthiness. The overall goal is to determine the individual's credit score. The amount to be given to an individual and the term of repayment is established in the credit scoring process. Criteria such as credit history determine credit scoring. Subsets are created using feature selection algorithms to create the most impacting areas from historical information. Some credit scoring studies were conducted in this area. Credit scoring models for the microfinance industry have developed a scoring system using neural networks and have proven these improvements in Peru [1]. A psychological approach to scoring micro-finance credit was conducted through the classification and regression tree [2]. They developed a credit risk assessment model for commercial banks and called it the nerve scoring approach [3].

One of the essential topics for machine learning studies is the classification process of the dataset. Data classification studies are among the leading issues in the field of data mining. Data mining is finding a meaningful data subset by finding and extracting important ones from data sets containing large amounts of data. Data mining aims to extract useful summary data from large data groups. The concept of data mining is to create helpful information by evaluating the existing data[4]. The primary method is to convert secret information and relations in the raw data into predictive information. Utilizing these developed methods, the connections between the data were determined, and the results based on these relations were tried to be revealed. To establish these relationships correctly, the data must be passed through the pre-processing techniques and the necessary statistics and learning algorithms [5]. If the feature selection is to be defined as the process; it can be said that various methods analyse a lot of data on the data set and it is the process of trying to uncover undiscovered information and data.

Data pre-processing techniques of the data mining process can be detailed, as in the following [6].

- Data Clearing
- Data Integration
- Data Reduction
- Data Convertion
- Implementation of Data Mining Algorithms
- Results & Reviews

One of the most critical steps of the study is the feature selection part. Some studies have been conducted in this area as follows. In 2011, Manimala and his friends proposed the hybrid soft calculation technique for feature selection and parameter optimization to classify [7]. Garcia and her friends, have studied the optimization of indexes obtained for clustering of feature selection methods in gene expression microarrays [8]. In some studies, he has developed a new feature selection method based on two-stage logic function [9]. Ghamisi and Benediktsson performed a feature selection with the use of genetic algorithms and particle swarm optimization in their study on salinas hyperspectral data set [10]. It has been observed that the method automatically selects the most instructional features within an appropriate processor time.