A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree

ABSTRACT

This paper predicted a model that indicates whether to buy a car based on primary health insurance customer data. Currently, automobiles are being used to land transportation and living, and the scope of use and equipment is expanding. This rapid increase in automobiles has caused automobile insurance to emerge as an essential business target for insurance companies. Therefore, if the car insurance sales are predicted and sold using the information of existing health insurance customers, it can generate continuous profits in the insurance company's operating performance. Therefore, this paper aims to analyze existing customer characteristics and implement a predictive model to activate advertisements for customers interested in such auto insurance. The goal of this study is to maximize the profits of insurance companies by devising communication strategies that can optimize business models and profits for customers. This study was conducted through the Microsoft Azure program, and an automobile insurance purchase prediction model was implemented using Health Insurance Cross-sell Prediction data. The program algorithm uses Two-Class Logistic Regression and Two-Class Boosted Decision Tree at the same time to compare two models and predict and compare the results. According to the results of this study, when the Threshold is 0.3, the AUC is 0.837, and the accuracy is 0.833, which has high accuracy. Therefore, the result was that customers with health insurance could induce a positive reaction to auto insurance purchases.

**EXISTING SYSTEM**

Azure is Microsoft's cloud computing platform that has been in service since 2010. In 2011, Platform as a Service (PaaS) was followed by Infrastructure as a Service (IaaS) service in 2013. The Azure platform provides over 600 services. You can collect and manage data in the cloud using the Azure platform, create models through Azure Machine Learning Studio, and easily build web services and apply them to various devices. At Microsoft Ignite, It announced Azure Machine Learning designer's general availability, the drag-and-drop workflow capability in Azure Machine Learning studio, which simplifies and accelerates building, testing, and deploying machine learning models for the entire data science team, from beginners to professionals (Nam et al, 2019). In addition, unlike existing cloud platforms and machine learning libraries and tools, it provides an easy-to-access GUI environment in consideration of user convenience (Kang et al., 2018).

Machine Learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying instead on patterns and inference. It is seen as a subset of artificial intelligence. It also focuses on representation and generalization. Representation is the evaluation of data, and generalization is the processing of data that is not yet known. The term Machine Learning was first used by Arthur Samuel, an IBM researcher in the field of artificial intelligence, in his paper "Studies in Machine Learning Using the Game of Checkers." Machine learning is a part of artificial intelligence, focusing on representation and generalization as a field for developing algorithms and technologies that enable computers to learn. In addition, data can be analyzed using algorithms, learning through analysis, and make decisions or predictions based on learning. Machine learning is largely classified into supervised and unsupervised learning, and supervised learning includes classification algorithms, regression algorithms, and deep learning (Kim et al., 2018).

According to Yu-Jin Nam, The methodology consists of three main steps: data set selection. The second step includes preprocessing, in which the original data is prepared for classification. The last step contains training models according to the accuracy of each algorithm (Nam at el., 2019). In machine learning, Support-Vector Machine (SVM) is supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, and an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the separate categories that are divided by a clear fap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. The reason for using a two-class support vector machine is that technically it can be used in both classification and forecasting problems. For its experimental usage, the performance of two-class logistic regression and two-class neural networks were lower than the two-class support vector machine (You et al., 2020).

Disadvantages

1) .The system doesn’t have technique to analyze large number of datasets.

2). There is no technique for Prediction Using Two-Class Logistic Regression.

**PROPOSED SYSTEM**

Logistic regression is a kind of simple machine learning that performs regression estimation on proportional, proportional, or categorical data. Compared to more advanced classification and regression techniques, it is based on a very simple theory, but probabilistic analysis is possible for categorical data. In other words, it is possible to predict the occurrence of a specific event by using an independent variable that has a direct influence on the dependent variable (Kim et al., 2020). Logistic regression analysis can explain the types of associations and interactions caused by the model structure and can evaluate the influence of explanatory variables on response values through parameter inference. In addition, since it is possible to perform discrimination and classification based on predicted probability, various industries such as medicine, telecommunications, and finance are performing tasks to predict the probability of an event occurring using logistic regression analysis.

**Advantages**

1. The proposed system is more effective due to presence of Prediction Using Two-Class Logistic Regression.
2. The system is more effective due to analyzing of large number of datasets.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).