A PROFIT MEASURE WITH DETERMINISTIC AND STOCHASTIC EFFECTS FOR DATA MINING MODELS

FRANCO EDUARDO GARRIDO CÉSPEDES
MAGÍSTER EN GESTIÓN DE OPERACIONES

RESUMEN

En la actualidad los entornos de negocios se están volviendo más competitivos, solo aquellos que toman decisiones informadas se mantienen exitosos, por lo tanto es imperativo tomar siempre decisiones bien informadas con el objetivo de lograr el fin último de los negocios, alcanzar la utilidad más alta posible. Business Analytics es un área que incluye el uso de modelos de Data Mining para tomar decisiones de negocios de manera informada, pero a pesar de que existe una amplia variedad de modelos, existen pocas medidas para estos modelos que consideren los costos y beneficios asociados a las decisiones conducidas por éstos. En este contexto nosotros tratamos de enriquecer el estado del arte de la toma de decisiones a través de una mejora a una medida de utilidad, para lo cual identificamos una medida de utilidad y extendimos su dominio esperando modelar la variabilidad de los costos y beneficios en los ambientes de negocios con alta variabilidad, proporcionando una medida capaz de modelar un número más amplio de contextos de negocios. Este nuevo enfoque consiste en la adición de choques aleatorios a las distribuciones de costo y beneficio de una medida conocida como Expected Maximum Profit, la nueva medida es llamada R-EMP. La investigación puso en marcha un experimento sintético y otro empírico en un contexto de credit scoring. El experimento sintético fue desarrollado simulando un conjunto de datos, y el caso empírico fue basado en un conjunto de datos de una empresa financiera de Chile. Para ambos experimentos se utilizó Regresión Logística para clasificar si un cliente cometió default o no, utilizando como criterio de selección nuestra medida (R-EMP), una medida frecuentemente usada conocida como AUC, una media de perdidas conocida como H-measure y nuestra medida primaria EMP. Luego seleccionamos nuestros choques aleatorios como información aleatoria normalmente distribuida, replicamos 5.000 veces y simultáneamente muestreando de forma estratificada las dos clases desbalanceadas, teniendo como resultado de ambos experimentos un acuerdo sobre la mejora en la utilidad cuando la medida R-EMP fue utilizada como criterio de selección. Como conclusión validamos la

incorporación de choques aleatorios para mejorar una herramienta para la toma de decisiones como lo es la medida EMP, entonces nosotros recomendamos el uso de la medida R-EMP como criterio de selección en ambientes de negocios muy variables. Para investigaciones futuras sería interesante incorporar choques aleatorios en otras aplicaciones de negocios como la fuga de clientes, además de probar otras variedades de choques y además capturar choques asociados a contextos reales para incorporarlos como entrada a la medida R-EMP.

ABSTRACT

Nowadays business environments are becoming more competitive, just those who take informed decisions remain successful, so it is imperative to take informed decisions to reach the businesses ultimate goal, achieve a profit as high as possible. Business Analytics is an area, which includes the use of Data Mining models to take informed business decisions. There is a wide variety of models, but there are few measures for those models that takes in consideration the costs and benefits associated to the decisions driven by them. In this context, we aim to enrich the state-of-art on decision making tools by generating a new profit-driven measure. We identify a profit-driven measure and extent its domain aiming to model the variability of costs and benefits for highly-variable business environments, providing a measure able to model a wider number of business contexts. This new approach consist into adding random shocks to the cost-benefit distributions of a measure known as Expected Maximum Profit, the new measure is called R-EMP. The research established a synthetic and empirical experiment in a context of credit scoring. The synthetic case was developed simulating a credit scoring data set, and the empirical case was based on Chilean financial institution dataset. For both experiments we used the Logistic Regression model to classify if a client fall in default or not, using as selecting criterion of attributes our measure (R-EMP), a commonly used measure known as AUC, a measure of losses known as H-measure and our primary measure known as EMP. Then, we selected our random shocks as random normally distributed information, replicating 5,000 times and simultaneously applied stratified sampling accordingly to the two imbalanced classes. The results of both experiments agrees that using R-EMP measure as selection criterion drives to the improvement of the total profit for the company.

As conclusion we validate the incorporation of random shocks to improve a decision making tool EMP measure. We recommended the use of R-EMP measure as selection criteria on highly-variable business environments. For future research it would be interesting to incorporate random shocks to another business applications like churn prediction, also trying to test another variety of shocks, and also to capture a real business shock and incorporate this as input for the R-EMP measure.