Tittle

Methods Boostrap and Cross-Validation, in the Construction of the Model of Vector Machine of Support for the Prediction of the Defaulting of the Banwest Bank's Clients

Participant

REYES RAMIREZ GIANCARLOS

Teacher

RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ

Index

Abstract

This research has as its main aim to determine the best method of resampling, in the construction model vector machine of support for the prediction of the defaulting of the Banwest Bank's clients, in order to analyze different scenarios that allow us to analyze

the efficiency of the Model of Vector Machine of support, through historical information of Banwest Bank corresponding to the period of July and September of 2016.

Given that it is important for financial institutions count with predictive models for defauls, we seek to raise an optional way to predict the already stated variable. Therefore, through this application, it will allow us to show that the Cross-Validation method can be more powerful than the Bootstrap method, and this is due to the fact that it has higher GINI and sensitivity indicators.

Key Words

Support Vector Machine, Cross Validation, Bootstrap, Defaulting

CONCLUSIONS

The best resampling method, for the construction of the vector machine of support, for the prediction of the defaulting of the Banwest Bank's clients turn out to be the Bootstrap's model, it shows us an indicator of sensibility of the 86.78%, for which reason is best in 8.9% in comparison to the Cross-Validation (77.80%), which means, through the Vector Machine of Support, with the Bootstrap's model we was achieved to classify the clients of the delinquent portfolio.

With the Boruta's algorithm, it was determinated that the factors most influence in the defaulting of the delinquent portfolio's clients are the loan state, Category of the Superintendence of Bank and Insurances, Internal Segmentation, Total Debt, Type of Credit, Disbursement Amount, Interest Rate, Instalment Rates, Balance Amount, Days past due, Overdue Balance.

The scenario, which under the model of the vector machine of support, for the prediction of the defaulting of the Banwest Bank's clients, through the Bootstrap's model, showed better efficiency, considering 40 Bootstrap's samples, for which it was presented a Gini indicator of 77.78%. Whereas the capacity to predict the defaulting event was 86.78% with an estimation error of 891.123.

The scenario, which under the model of the vector machine of support, for the prediction of the defaulting of the Banwest Bank's clients, through the Cross-Validation method, presented a better efficiency, considering 10 partitions, for which it was presented a Gini indicator of 77.26%. Whereas the capacity to predict the defaulting event was 77.80% with an estimation error of 1547.45.

Tittle

Predicting of Success of the Telemarketing Campaing through Deposits to Term to the Clients of the Company "Bankuni", Through the Comparison of the Random Forest's Technique and the Naive Bayes' Technique

Participant

CHERO CAJUSOL CARLOS ENRIQUE

Teacher

AMELIDA PINEDO

Index

Abstract

The BankUNI Company is a financial institution whose main activities are to provide credit services to customers, receiving savings deposits, etc. In a context of customer loyalty it is necessary to implement relational marketing tools that help improve the company's position in the financial sector in the city of Lima.

It is for this reason that this research has focused on an analysis of the success of the telemarketing campaign offering term deposits to its customers in order to increase customer loyalty.

Thus the study of the problem: What data mining technique, Naive Bayes or Random Forests has better predictive ability to determine the success of the telemarketing campaign term deposits?, This predictability will be measured by indicators such as Specificity, sensitivity and the ROC curve.

KEYWORDS: Random Forest, Naive Bayes, specificity, sensitivity, Gini Index.

Conclusions

- The Random Forest Model turned out to be appropriate that showed positive indicators to classify if a customer will access to the subscription of a term deposit after being managed through the telemarketing's campaign.
- The most significant variables to the Random Forest model was the duration of the last call in seconds and the contacts of the last month to have success the telemarketing's campaign.
- In the Naive Bayes's model, the most significant variables was the Numbers of days that has passed after the customer have been contacted by a previous company and the duration of the last call in seconds.

Tittle

Identification of the Main Factors than Influence in the Mortal Car Accidents In Metropolitan Lima, by te Outcomes of the Importance of Variables of the Models Of Data Mining, Random Forest, Boosting and Cart's Decisions Tree - National Census of Police Stations 2014

Participant

Tarazona Tocto Sherly Sindia

Teacher

RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ

Index

Abstract

Every year, more than 1.2 million people die in the world in traffic accidents according to the World Health Report 2015; Likewise, a report from the Andean Community of Nations (CAN) published in 2013 placed Peru as the country with the highest traffic accident rate per 100,000 vehicles, followed by Bolivia, Colombia and Ecuador.

In 2014, one of the last National Census of Commissaries was held, which contains a section on Traffic Accidents. As a result of the census it was possible

to know that 45% of the total traffic accidents in the country occur in Lima Metropolitan.

Addressing the problem of fatal traffic accidents is a complex study because there are many factors involved, but at the same time interesting. The present investigation is an application in the field of the study on road accidents. The study started from the problem What are the main factors that influence the fatal traffic accidents in Metropolitan Lima, through the importance of the variables of the Data Mining models: Random Forest, Boosting, and CART Decision Tree?

To answer the research problem we used the use of Data Mining Models such as Boosting, Random Forest and CART Decision Tree, whose results showed a table of importance of variables from which it was possible to identify that the type of vehicle used (Road or avenue), Invasion of the lane and disrespect to the sign of transit by the driver, are the main factors that influence in the fatal result of the accidents of Transit in Metropolitan Lima.

Keywords

Fatal Traffic Accident, Sensitivity, Gini, Cross Validation, Boosting, Random Forest, Decision Tree CART, Smooth

Conclusions

- ✓ The main factors that influence in the fatality of the car accidents in Metropolitan Lima, which they were obtained from the outcomes of data mining techniques: Random Forest, Boosting and Decisions Tree, were: That the car accident occurred in the type of highway or avenue, it was because of the contempt for the transit signal by the driver, the type of the vehicle was a rural van (combi), that it was caused for invade a track, that the type of vehicle was a mototaxi and/or three-wheeled motorcycle

- ✓ The values of the performance indicators of the models used are: Gini higher to 50% for all models and higher to 70% to the Boosting Model, Sensibility higher to 55% for all the models, and Specificity higher to 60% for all the models.

- ✓ The error rates of the build models, obtained through the croos-valid model k-fold, with k=10, are fewer to 9% for the Boosting Model, fewer to the 15% for the Random Forest model and fewer to 25% for the CART Decisions Tree Model.

- ✓ The variables (top 3) that listed in the importance table of variables of the Boosting Model are: Type of path of occurrence of the car accident (highway), type major vehicle involved (truck-rural bus) and disrespect to the transit signal by the driver; for the Random Forest Model are: Type of major vehicle involved (van - rural van), Type of transport (public) and the car accident by shove; finally to the Decisions Tree are: Type of route of accident occurrence (highway), type of minor vehicle involved (three-wheels motorcycle taxi) and for invasion in the opposite track.

Tittle

Comparison Between Neuronal Network Techniques and the Decision Tree for the Relapse Prediction of Young Offenders using Information of the Census in 2016

Participant

Adama Espirilla Jhonny Ivan

Teacher

RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ

Index

Abstract

Peru currently presents problems with juvenile delinquency, for this study is aimed at determining the factors that are part of the young offenders using decision tree techniques and neural networks, the theoretical information and the database received from the first census National to youth centers held on August 14 this year, which had 5 modules. The treatment of the variables was performed using data imputation methods, the sampling methodology was used for the treatment of asymmetry, and to verify that the parameters were not overestimated the cross-validation was performed, the results of the tree of the Which has been proved the best stability of the errors worked in the validation of the cross, in addition to its approximation of the indicators that had with the neural networks, finally the indicators for the techniques and the selection of the most adequate technique were determined It was the tree-making technique for its indicators.

Conclusions

- ✓ For the obtaining of the statistic technique was obtained that the most appropriate is the decision tree for his parameters of stability in the cross validation and his indicators.
- ✓ The indicators found in the neuronal network were:

INDICATORS	NEURONALNETWORK
sensitivity	66.2%
specificity	68.1%
accuracy	66.7%
VP	52.3%
VN	47.7%
GINI	38.2%

✓ The indicators found in the decision tree were:

INDICATORS	DECISION TREE
sensitivity	68.1%
specificity	68.0%
accuracy	67.2%
VP	53.1%
VN	46.9%
GINI	54.5%

Title

Effects of truncation and censoring of the dependent variable in 2-stage multilevel regression models

Participant

Huanca Huamaní Milagros Luz

Teacher

RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ

Index

Abstract

It has demonstrated that the information censured and truncated affects the estimation of the parameters by the MCO in the models of linear regression (Zuehlke & Kassekert, 2008). Moreover, we know that the multilevel models constitute the methodology of analysis more appropriate for treat information "Hierarchical" or "Nested" (Murillo, 2008). The multilevel models are considered extensions of the models of classics linear regression.

What was done in this investigation is study the marginal effect of the dependent variable censured and truncated in the models of multilevel regression of 2 periods in the next criterions: Akaike information criterion (AIC), Akaike information criterion corrected (AICC) and conditional Akaike information criterion (CAIC) in the models of multilevel regression of 2 periods. This under the diagram of 2 stages: in the first stage worked with a sample of

200 searches and different proportions of truncation and censoring as 10, 20 and 30%. Keeping in mind that is consider a high censoring since 20%. In the second stage was fixed a proportion of truncation and censoring of 10% for different sample sides as 50, 200 and 600 registers.

It was concluded that the fact of the dependent variable is censored and truncated, it has effects (in different percentage) about Akaike information criterion (AIC), Akaike information criterion corrected (AICC) and conditional Akaike information criterion (CAIC), these values tend to decrease up to 12% in the case of truncation and increase up to 190% in the case of censuring.

Conclusions

- ✓ The fact that the dependent variable is censured or truncated has effects (In different percentages) about Akaike information criterion (AIC), Akaike information criterion corrected (AICC) and conditional Akaike information criterion (CAIC), these values tend to decrease up to 12% in the case of truncation and increase up to 190% in the case of censuring.

- ✓ As much for the stage 1 where the sample was fixed with 200 registers and was made different proportions of truncation as 10, 20 and 30%, as for the stage 2 where was fixed the proportion of truncation in 10% for different samples sides

as 50, 200 and 600 registers , were obtained values of Akaike information criterion (AIC), Akaike information criterion corrected (AICC) and conditional Akaike information criterion

(CAIC) below the obtained values in the block control, where it doesn't exist any type of modification of information. The difference in percentage values become being up to 12% (When the sample is 50).

- ✓ The effect of the censure of the dependent variable in the values of Akaike information criterion (AIC), Akaike information criterion corrected (AICC) and conditional Akaike information criterion (CAIC) as much for the stage 1 where the sample was fixed with 200 registers and was made in different proportions of censoring of 10, 20 and 30%, as the stage 2 where was fixed the proportions of truncation in 10% for different samples sides as 50, 200 and 600 registers, is too negative so these values of information criterions with respect to the values in the control block where wasn't made censure have a percentage difference from up to 190% (For a sample of 600).

Tittle

Classification of scam with the model of Neuronal Networks using a response surface to select the independent variables that optimize the malign connections in simulated datums of the United States Armed Forces in 2016

Participant

Huamani Gonzales L. Dayana

Teacher

Richard Fernando Fernández Vásquez

Abstract

Based on the fundamentals and techniques of data mining, models can be designed and elaborated that allow finding clandestine behaviors that are easily detected at first sight. In particular the usefulness of data mining in this area lies in a series of techniques, algorithms and methods that imitate the human characteristic of learning to be able to extract new knowledge from experiences. These characteristics can be of vital importance to be applied in information security through the detection of intruders. This paper aims to show the contribution to the information security using the Response Surface as a technique that with the experience of the investigator can prevent the new modalities of information theft.

Keywords

Denial of services, ANN, intrusions, data mining, model, prediction.

Index

Conclusions

The comparison of the Neuronal Networks, using or not using the surface response previously it wasn't clear and conclusive; in determining that the model of Neuronal Networks without the use of the Surface response got a Gini=95% whereas that the model that used the protocol variables, flag, logged, count range as part to optimize the malign connection got a Gini=98%.

It was determined the model of second order is the most appropriate, since it was compared the R-adjust in both models y got a 62.03% of prediction of the variable malign connection, that is precisely the variable that we want to maximize.

It was obtained that the protocolo variables, flag, logged and count range are the variables that optimize the maligne connections in the model of first order, subsequently with the ascend method more pronounced we make the calculus of the model of second order in which we got as significant variables a protocol, flag, logged, count range, protocol*flag, protocolo*loged, loged*rango_count y flag_loged