Hasilian Research Snapshot

by Mr. Raja Azhan Syah bin Raja Wahab

Digital Economy Tax Compliance Model in Malaysia using Machine Learning Approach

Abstract

- Digital economy income tax compliance is still in its infancy.
- The ability to diagnose the taxpayer's compliance will ensure the IRBM effectively collects the income tax and gives revenues to the country.
- This paper proposes the descriptive and predictive analytics models for predicting the digital economic income tax compliance in Malaysia.
- Through a brief description of the descriptive model, the data distribution in a histogram shows that the information extracted can give a clear picture of influencing the results to classify digital economic tax compliance.
- In predictive modelling, single and ensemble approaches are employed to find the best model and factors contributing to the incompliance of tax payment among digital economic retailers.
- Based on the validation of training data with the presence of seven single classifier algorithms, three performance improvements have been established through ensemble classification, namely wrapper, boosting, and voting methods, and two techniques involving grid search and evolution parameters.
- The results show that the ensemble method can improve the single classification model's accuracy with the highest classification accuracy of 87.94% compared to the best single classification model.
- The knowledge analysis phase learns meaningful features and hidden knowledge that could classify the contexts of taxpayers that could potentially influence the degree of tax compliance in the digital economy are categorized.

Problem Statement

- A sudden increase in business and digital services, including web advertising, social media, e-commerce, and online blogs, need to be taxed accordingly. The limited collection of government income taxes has forced the Inland Revenue Board of Malaysia (IRBM) to develop a solution to improve the tax compliance of the digital economy sector. Companies operating in the digital economy and domiciled in Malaysia are not able to distinguish conventional economics and are often considered as income to seal business operations that making it more difficult for the IRB to collect income tax.

Research Questions

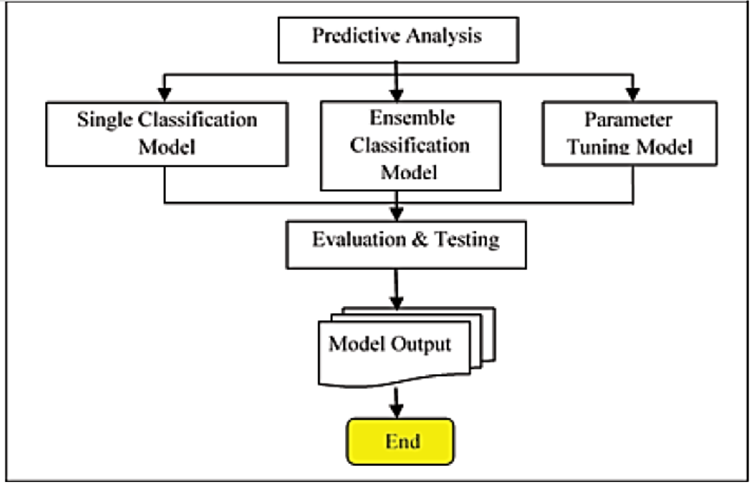
- How well can predictive analysis help to define target class categories once complex data from external and internal resource matching is provided and various patterns and knowledge rules can be generated after the development of a data modeling?
- Are there connections between these important features and hidden information sufficient to generate valuable knowledge?

Objectives

1. To track down and predict the tax compliance and non-compliance among digital businesses:
   - To describe data features and values that affect non-taxation criteria for taxpayers through perspective and inferential interpretation so that it can be simplified toward a better understanding
   - To establish predictive models such as single classification, improvement of performance through ensemble classification, and tuning of parameters to obtain the best predictive model.
2. To identify factors that influence the tax compliance and non-compliance of digital business

Framework

Model development methodology (Han & Kamber 2002)



Predictive Analysis

Single Classification Model

Ensemble Classification Model

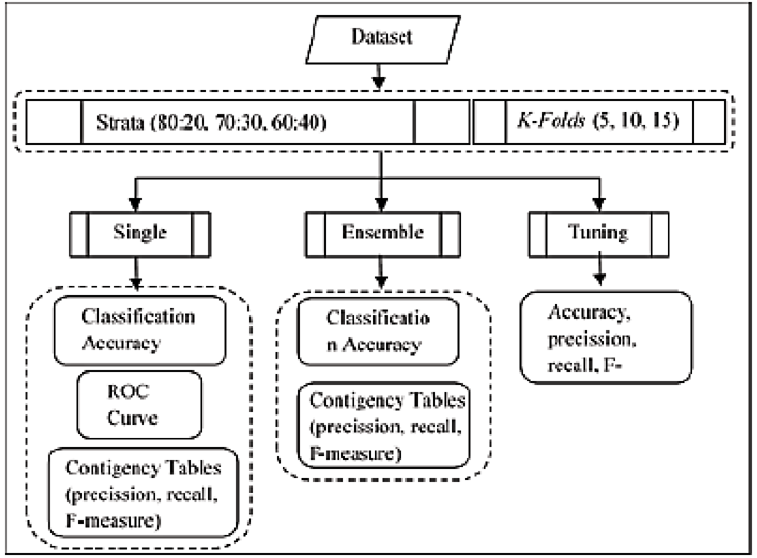
Parameter Tuning Model

Evaluation & Testing

Model Output

End

The overall experimental design of predictive analysis (Hamsagayathri & Sampath 2017)



Dataset

Strata (80:20, 70:30, 60:40)

K-Folds (5, 10, 15)

Single

Ensemble

Tuning

Classification Accuracy

ROC Curve

Contigency Tables (precission, recall, F-measure)

Classificatio n Accuracy

Contigency Tables (precission, recall, F-measure)

Accuracy, precission, recall, F-

Methodology

- This study adopted the Cross Industry Standard Process for Data Mining (CRISP-DM) standard data analytics research methodology introduced by Crisp (1999).
- The data for this study is obtained from the IBRM Department of Tax Operation (tax assessment years 2015, 2016, and 2017). Raw data of digital economy external sources were retrieved using website crawler software named *Kapow* and internal source data were obtained from an in-house database
- Total of 11,706 business taxpayer's data with 29 conditionals attribute and a class attribute involved in the modelling. The class attributes are the status of tax compliance namely Compliance (Comp), and Non-Compliance (Non-Comp).
- Development of a predictive model involves classification algorithms i.e. Classification and Regression Tree (CART), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN). Seven (7) widely used classification models including Naive Bayes (NB), functions (SVM, LR), meta (ANN, KNN), rules, and tree (CART, RF) are used, and the k-fold cross-validation method (k=5,10,15) is used for percentage split of training: testing data (80:20, 70:30, 60:40).
- Two modeling schemes are proposed in this study i.e. Single Model and Ensemble Model.
- As for the experimental design and evaluation of predictive analysis, the **three metrics of assessment** are **classification accuracy, contingency table** (True Positive - TP, True Negative - TN, False Positive - FP, and False Negative - FN), and **classification reports** which include accuracy measurement, retrieval, and F-measure.
- The TP parameter shows the correct classification prediction for the target class 'Non-Comp'. The FP parameter shows incorrect predictions for the target class 'Non-Comp'. The TN parameter shows the correct classification prediction for the target class 'Comp'. The FN parameter shows the incorrect classification prediction for the target class 'Comp'.

Findings

- For **Single Classification, CART algorithm** models have the highest classification accuracy. Based on the results of this experiment, four models of selected algorithms namely CART, RF, ANN, and LR will be used in performance improvement through ensemble classification and parameter tuning.
- For **Ensemble Classification, RF (Wrapper)** models achieve higher classification accuracy.
- For **the expert evaluation of non-compliance (non-comp) rules**, under the RF rules (2nd tree), there are 1487 non-compliance taxpayers correctly classified with the digital economy sector, where they do submit their first business tax and stamp duty return information with a specific value of vehicle assets in possession).
- Under the RF rules (4th tree), there are 1486 non-compliance taxpayer cases that are correctly classified when they are registered yearly based on the receipt of IRBM's tax return. They also have several vehicle assets in their possession and submitted their stamp duty return information without hesitation. This should indicate that the taxpayer was committed to performing their responsibilities. However, the voluntary aspect of tax reporting and taxpayers may seek to claim tax relief for many reasons, in order to avoid the real loss of the digital economy combined with overall taxable revenues.
- For **the expert evaluation of compliance (comp) rules**, under the CART rules (1st branch tree), there are 1561 compliance taxpayers who have no track record of owning a vehicle, no property assets, and no stamp duty amount which signifies no purchase of real estate assets, and further enables the taxpayer's potential not to hide revenue generated from the digital economy sector as no additional revenue is reported.
- Under the CART rules (2nd branch tree), there are 1338 taxpayer records that show a presence of bank account number information proving that banking transactions can occur for online income tax repayment payments or used in making financial loans such as real estate/housing and vehicle loans. This in turn provides an overview of the availability of banking status information available to the IRBM to resolve previously taxpayer cases.

Conclusion

- This study proposed machine learning algorithms for classification modelling of tax compliance and non-compliance cases.
- Overall, this study has three important research findings to the IRBM.
  - It supports the initiative of the big data analytics project in the IRBM
  - To determine the non-compliant taxpayers' category and vice versa for future use.
  - The experimental results can be used as a reference and guide for future research in improving the classification model related to determining the digital economy sector's level of tax compliance.

Research Gap

- The use of massive tax data lakes can further enhance the digital economy tax compliance model, and more discovered knowledge help the IRBM in making a strategic decision.
- It will also help the government manage the revenue and plan for development programs that benefit the nation.