AN ANALYSIS OF COMPUTATIONAL LEARNING MODELS FOR QUIT RENT REVENUE ESTIMATION

MUHAMAD HAMIZA BIN HAMDAN

MASTER OF COMPUTER SCIENCE
(DATABASE TECHNOLOGY)

2017

© Universiti Teknikal Malaysia Melaka

MUHAMAD HAMIZA BIN HAMDAN MCS. (Database Technology) 2017

© Universiti Teknikal Malaysia Melaka

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

AN ANALYSIS OF COMPUTATIONAL LEARNING MODELS FOR QUIT RENT REVENUE ESTIMATION

Muhamad Hamiza Bin Hamdan

Master of Computer Science

2017

© Universiti Teknikal Malaysia Melaka

AN ANALYSIS OF COMPUTATIONAL LEARNING MODELS FOR QUIT RENT REVENUE ESTIMATION

MUHAMAD HAMIZA BIN HAMDAN

A thesis submitted
in fulfillment of the requirements for the degree of Master of Computer Science
in Database Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

© Universiti Teknikal Malaysia Melaka

DECLARATION

I declare that this thesis entitled "An Analysis of Computational Learning Models for Quit Rent Revenue Estimation" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ……………………………

Name : ……………………………

Date : ……………………………

© Universiti Teknikal Malaysia Melaka

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Computer Science in Database Technology.

Signature : ………………………

Supervisor Name : ………………………

Date : ………………………

© Universiti Teknikal Malaysia Melaka

DEDICATION

Thank you Allah SWT for this gift.

© Universiti Teknikal Malaysia Melaka

ABSTRACT

Quit rent is a major income for states in Malaysia. Hence, quit rent revenue projection is a crucial component for yearly state budget presentation to ensure the sustainable physical and development in the particular state, as well as throughout the whole country. Identifying predictors is essential to accurately predict the revenue for the next coming year. Current practice of quit rent revenue projection in the state of Negeri Sembilan, Malaysia is to increase the past year revenue by a certain percentage according to the performance indicators in the state. This manual prediction has posted an overrated projection every year. Hence, a more intelligent quit rent revenue estimation technique is needed to automate and improve the projection. This project aims to analyse three benchmarking techniques, namely the Neural Networks (NN), Support Vector Machine (SVM) and Logistic Regression (LR) techniques in quit rent revenue estimation. The studies follows the data science methodology using experimental approach starting from problem formulation to data preparation, model building, results analysis, and concluding on the research findings. The experiment data was built on the quit rent payment transaction in the year of 2015 from the state of Negeri Sembilan, Malaysia. The indicators of account active duration, category of land use, arrears, and late payment charges were used as conditional features. The learning models were built to first predict the payment status before estimating the total quit rent revenue for the year of 2016. The estimation results were compared with actual results and further analysis in details using the performance measures of classification accuracy, precision, weighted mean precision, recall, weighted mean recall, and root mean square error (RMSE). The analysis showed that all the three estimation models have demonstrated good performance. The LR model has achieved the best payment status accuracy of 94.78% followed by the NN model at 94.72% and the SVM model at 91.57%. However, the measurement of RMSE has showed a slight different. The NN model has the closest estimation to the actual total amount of quit rent revenue in the coming year with only 2.07% estimation error in Ringgit Malaysia, while the LR and SVM models were recorded with 2.10% and 4.69% difference respectively. In summary, both the LR and NN models are good to be used as the quit rent revenue estimators. However, all the three models are not able to predict the minority class of payment done after yearly quit rent estimation in October because of imbalance data problem. Further research should focus on treating the imbalance data problem before feeding the data into the learning model. Besides, improving the prediction strategy by taking into account the payment trend and behaviors of land owners is another research direction worth to follow.

© Universiti Teknikal Malaysia Melaka

ABSTRAK

Cukai tanah merupakan salah satu pendapatan utama negeri-negeri di Malaysia. Oleh itu, jangkaan hasil cukai tanah adalah komponen yang sangat penting dalam pembentangan bajet tahunan negeri bagi memastikan pembangunan fizikal yang mampan dan berterusan di negeri tertentu, serta di seluruh negara. Pengenalpastian ke atas predictors adalah sangat penting untuk menjangkakan dengan tepat hasil pendapatan cukai pada tahun akan datang. Pada masa kini, amalan semasa bagi unjuran pendapatan cukai tanah di Negeri Sembilan, Malaysia adalah dengan menetapkan peratusan tertentu ke atas pencapaian hasil cukai tanah tahun sebelum sebagai peningkatan hasil pendapatan cukai tahun tahun berikutnya selaras dengan petunjuk prestasi untuk negeri tersebut. Amalam manual ini, telah menyebabkan unjuran yang dibuat adalah berlebihan berbanding dengan pencapaian sebenar pada setiap tahun. Justeru itu, teknik pengiraan cukai tanah yang lebih pintar diperlukan untuk mengautomasikan dan menambahbaik ketepatan unjuran. Kajian ini bertujuan untuk menganalisa tiga teknik sebagai penanda aras iaitu Neural Networls (NN), Support Vector Machine (SVM) dan Logistic Regression (LR) dalam membuat anggaran hasil cukai tanah. Kajian ini dijalankan mengikut metodologi data sains menggunakan pendekatan eksperimental bermula dengan formulasi masalah hingga penyediaan data, pembinaan model, analisa hasil dan merumuskan hasil penemuan yang dihasilkan dalam kajian. Data yang digunakan merupakan transaksi bayaran cukai tanah pada tahun 2015 di Negeri Sembilan, Malaysia. Petunjuk seperti tempoh akaun yang aktif, kategori kegunaan tanah, tunggakan cukai tanah/denda dan denda lewat yang dikenakan pada tahun semasa digunakan sebagai "conditional features". Model pembelajaran dibangunkan pada awalnya akan membuat ramalan ke atas status bayaran sebelum membuat anggaran ke atas jumlah hasil cukai tanah untuk tahun 2016. Anggaran hasil cukai tanah yang diperolehi akan dibandingkan dengan hasil sebenar dan seterusnya penganalisaan secara terperinci menggunakan pengukuran prestasi seperti ketepatan klasifikasi, precision, weighted mean precision, recall, weighted mean recall, dan root mean square error (RMSE). Analisa yang dibuat menunjukkan ketiga-tiga model anggaran menghasilkan prestasi yang baik. Model LR mencapai prestasi paling baik untuk ketepatan status bayaran dengan 94.78% diikuti dengan model NN 94.72% dan model SVM 91.57%. Walaubagaimanapun, pengukuran ke atas RMSE menunjukkan sedikit perbezaan. Model NN telah membuat anggaran hasil cukai tanah yang paling hampir dengan hasil sebenar untuk tahun berikutnya dengan 2.07% ralat dalam Ringgit Malaysia, manakala model LR dan SVM merekodkan 2.10% dan 4.69% ralat daripada hasil sebenar. Secara kesimpulannya, kedua-dua model LR dan NN adalah model yang baik untuk digunakan dalam membuat anggaran hasil cukai tanah. Namun, ketiga-tiga model ini tidak dapat meramalkan kelas minoriti untuk bayaran yang dilakukan selepas anggaran cukai tanah tahunan pada bulan oktober disebabkan masalah ketidakseimbangan data. Kajian selanjutnya perlu memberi fokus untuk menyelesaikan masalah ketidakseimbangan data sebelum memasukkan data ke dalam model pembelajaran. Selain

© Universiti Teknikal Malaysia Melaka

itu, penambahbaikan terhadap strategi peramalan dengan mengambil kira tren pembayaran dan sikap pemilik tanah dalam membuat bayaran cukai tanah merupakan halatuju kajian lain yang patut dibuat.

© Universiti Teknikal Malaysia Melaka

ACKNOWLEDGEMENTS

I am very thankful to Allah SWT that he granted me the chance to finish this thesis. A lot of courage, sacrifices, time and strength needed to fully complete this thesis. First and foremost, I must acknowledge my limitless gratitude to my supervisor, Associate Professor Dr Choo Yun Huoy for her expertise, patient, encouragement and guidance throughout my candidature. Thank you for the countless hours of effort spent on me.

I also owe a deep debt of gratitude to my beloved wife, Enda Diana Yahya for her sacrifice and constant support on me throughout the journey. Finally, I want to give my appreciation to all my friends and lecturers for giving me strength to complete this master degree.

© Universiti Teknikal Malaysia Melaka

TABLE OF CONTENTS

PAGE

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

LIST OF TABLES

© Universiti Teknikal Malaysia Melaka

LIST OF FIGURES

© Universiti Teknikal Malaysia Melaka

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Every country has land policy and land planning systems to enhance social, physical, spatial and financial awkwardness. Land policies give the structure, bearing and congruity of choices made for the capacity of land in the execution of national advancement arranges which include nearby, state and local arrangements. In Malaysia, land policies are executed inside a more extensive system, which is supervised by the government. Some land policies were executed in view of the National Land Code (NLC). The National Land Code, 1965 came into constrain on first January 1966 to control all land matters in Peninsular Malaysia and the Federal Territory of Labuan while Sabah and Sarawak kept on utilizing the Sabah Land Ordinance 1930 and Sarawak Land Code 1958 separately. The reason for the NLC is to guarantee the consistency of land approach and land law concerning land tenure, transfer, registration, leases, easements, chargesand different interests and rights to land.

One important element in the NLC is a quit rent. It is a tax imposed on the owner of the land and if the land owner fails to pay, late payment penalty will be imposed and the state government has the right to confiscate the land if the owner continues to pay taxes not follow the procedure laid down in NLC. In Negeri Sembilan, the main contribution in the state revenue income is a quit rent apart from other income such as land premium, non-tax revenue, non-revenue receipts and others.

© Universiti Teknikal Malaysia Melaka

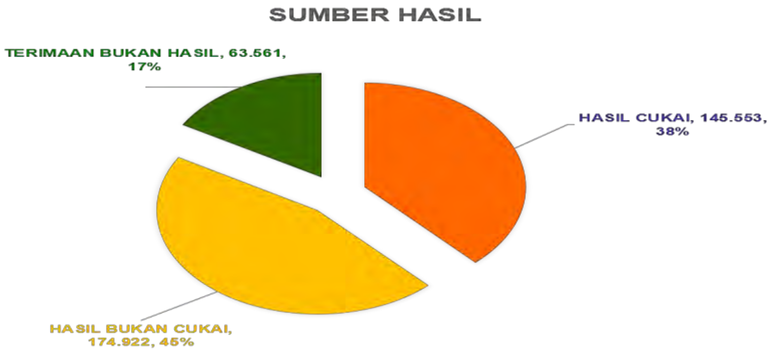

Figure 1.1: Forecast State Revenue Income in 2016

Through the state budget in 2016 laid, the quit rent will be contributed 38% (RM145,553,000.00) of state revenue. The increase on the income of the country is important to facilitate the state government plans for development of the state of various aspects which will be presented during the state budget. Normally, the state budget will be around October or November for the following year's budget. In the state budget, expected total revenue for the coming year will be presented in addition to the details of the expenditure will be allocated according to the main focus of the state government as a welfare program, the program of physical development, education, tourism and so on.

There are several methods which can be used in making revenue forecasting such as time series method, causal/econometric forecasting method, judgemental method, artificial intelligence method and many others. Tax revenue time series forecasting was investigated deeply previously. They have indicated many characteristics which made them difficult to gauge because of the requirement for traditional statistical method to take care of the parameter estimation issue. Most common forecasting model used is econometric model which based on Statistical approached which popular with linear regression, Auto regression and Auto Regression Moving Average (ARMA). However, when using this model there are assumption must be considered as linearity and stationary of the revenue time series data. For example, non-realistic assumption can reduce the quality of prediction

© Universiti Teknikal Malaysia Melaka

result. Another one is soft computing based model which include Artificial Neural Network (ANN), Fuzzy Logic, Support Vector Machine (SVM), Particle Swarm Optimization (PSO) and so on.

Thus, there is an interest to develop a prediction model by comparing between statistical base model and computational base model to estimate accurately the quit rent collection will be available next year. This will improve the efficiency of state administration for the revenue management and planning of the state budget with efficient and optimal.

1.2 Problem Statement

Revenue management is an essential element in the administration of a country. It can give the image as a good government and transparent in carrying out the trust and confidence of the people. In addition, the increase in revenue can also give an opportunity to the state government to plan better for them to spend the most important sectors in the development of a state successfully. At the present, a large portion of specialist utilize statistical techniques to forecast tax revenue, which depends on the full examination of changes in tax revenue factors and the related tax history information, utilization of forecast hypothesis, strategy and model, to make a judgment on the assessment income without bounds.But the statistical modeling approach is limited to some specific function, and the actual tax problems in the application of variables and dependent variables have complex non-linear relations, these statistical methods in the practical application does not have a good predictive effect (Yong Zhang, 2014). With the development of the technology, the researcher using another prediction model such as computational intelligent model which has features likes classification, regression and clustering. Artificial Neural Network (NN), Support Vector Machine (SVM), NaiveBaiyes, K-Nearest

© Universiti Teknikal Malaysia Melaka

Neighbour (KNN) among the technique used in computational intelligent technique to enhance the precision of the tax revenue for the future.

Sheng Lu et al (2009) using Genetic Algorithm and SVM to forecast tax gross. However, it only uses regression arithmetic of SVM to forecast tax with genetic algorithm to find the optimal parameters. Tax gross data in China from 1990 till 2001, which is a time series of data with single value, was used in this study without mention of other predictors that influence the predicted result. Despite SVM can achieve greater forecasting accuracy than ANN but the main question is there are no predictor models can be used for forecasting tax revenue. Han Liu et al (2008), states an attribute reduction SVM in the light of its good performance in classifying high dimensional nonlinear data. The outcomes show that model created by them performs well both in predictive accuracy and data classification accuracy. 13 indicator of tax assesment were utilized to see its adequacy to assess the duty installment of the taxpayers. Based on the result, they finish up the ROC (Rate of change) of Prime working income increments with the ascent of ROC of Account receivable, Current proportion and Quick proportion can't be both beneath the typical level for the endeavor wage charge evaluation. But in this study comparison was made using only the kernels in SVM as linear, polynomial, RBF and sigmoid kernel.

A study by (Alaa F. Sheeta in etl, 2015), conducted a research to predict the stock market index by comparing the statistical method (Regression) with a computational method (ANN and SVM). In a comparison of these three techniques, it was found SVM make a more accurate prediction than the other technique. However, predictors that are used in the stock market index are very different from that used in the predictor of quit rent revenue. In predicting stock market index, potential variables used were very dynamic, taking into account various factors that may influence on the direction of stock index. Exchange rate between country, economic indicators and financial such as gold and

© Universiti Teknikal Malaysia Melaka

olilprice and also the return of the biggest company in a stock index is a factor to be taken into account in this field. In addition, it is also the period is very short as daily and weekly as events of shares is based on market fluctuations. In the quit rent scenario, predictors used is an internal predictor centered on land development activities in a country. Predictors chosen as the new titles, cancellation of title due to foreclosure, payment transaction records in the current year, and the updating of the title due to changes in quit rent amount, can be an impact on the achievement of quit rent revenue for the upcoming years.

Therefore, there is a need to propose the model predictors for Negeri Sembilan State Government in forecasting revenue of quit rent collection for upcoming years, by comparing two different method or approach which are statistical based and computational intelligent based. This comparison will show us which forecasting method are most accurate and will be used as a tool in Negeri Sembilan Land Administration.

1.3 Objective of The Research

Objective of this research are:

i. To propose the model predictors for Negeri Sembilan State Government in estimating revenue of quit rent collection for coming years;

ii. To analyze forecasting model between statistical method which is regression technique and computational intelligence method which are Artificial Neural Network (ANN) and Support Vector Machine (SVM) in order to estimate revenue of quit rent with accurately.

© Universiti Teknikal Malaysia Melaka

1.4 Scope of study

The study was conducted using data obtained from the database of SistemHasil Tanah Negeri Sembilan (SIHAT NS) for each district land office. Quit rent payment history data in year 2015 was used in this study and involved three area in district of Port Dickson. The data will be processed and then used for training and testing data using 3 prediction model mentioned earlier. In order to predict and verify the accuracy of the model, Classification Accuracy Performance, Class Recall, Class Precision, Root Mean Square Errors (RMSE) are employed to evaluate the model. The best model with higher classification accuracy percentage will be choose as toolsin order to make prediction of quit rent revenue for future and at the same time facilitate budget planning affairs.

1.5 Significant of Study

The significant of this study is to design a good prediction model for Negeri Sembilan State Government especially land district office to predict quit rent revenue will be collected for the coming years. This is very important factor because for the budgeting purpose, each land district office must report a forecast value of quit rent will be collected for the next years. Therefore, this prediction model will be beneficial to them in order to get the most accurate value.

1.6 Conclusion

In summary, this chapter describes the research background related to the prediction of quit rent revenue for future years in Negeri Sembilan State Government. The problem statement describes in this research is to propose a suitable prediction model which generate the most accurate by comparing 3 prediction model of Regression,

© Universiti Teknikal Malaysia Melaka

Artificial Neural Network and Support Vector Machine. This chapter also has stated objective of the research, scope of the study and the significant of this study.

© Universiti Teknikal Malaysia Melaka

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will focus on literature review related to quit rent revenue prediction. This literature review is essential for assessing prediction models to be compared in this study, namely Regression, Artificial Neural Network and Support Vector Machine in prediction of quit rent revenues for the year ahead. This chapter will also provide a clear explanation on previous studies that have been made using a prediction model involved. This is very important in ensuring that the prediction model will be develop soon, have to take into account lessons learned from previous studies and indirectly will generate a high level of accuracy of the quit rent revenue collection.

2.2 Predictive Model

Prediction is required in many situations. As an example, stocking an inventory requires predict of stock requirement, predict for future sales performance, predict for future tax revenue, closing stock market and others. Whatever the conditions or time horizons included, prediction is an essential guide to successful and proficient arranging. The consistency of an occasion or an amount relies on upon a few elements including:

i. how well understanding the factors that contribute to it;

ii. how much information are accessible;

ii. is it the figures can influence the value we are attempting to predict.

© Universiti Teknikal Malaysia Melaka

Frequently in prediction, a key stride knows when something can be anticipated precisely, and when forecast will be no superior to flipping a coin. Great forecast catch the genuine patterns and connections which exist in the historical data, however don't reproduce past occasions that won't happen once more.Prediction situations vary widely in their time horizons, factors determining actual outcomes, types of data patterns, and many other aspects (Rob J. Hyndman,2014).

2.3 Time Series Model

A time series is a sequence of real-valued signals that are measured at successive time intervals. Autoregressive (AR), moving normal (MA), and autoregressive moving normal (ARMA) models are regularly utilized with the end goal of time-seriesmodelling, analysis and forecasting. These models have been successfully used in a wide range of applications such as speech analysis, noise cancelation, and stock market analysis (Hamilton (1994); Box et al. (1994); Shumway and Stoffer (2005); Brockwell and Davis (2009)). A standout amongst the most prevalent models in time series model is Moving Average (MA). Technical analysis has been around for a considerable length of time and as the years progressed, dealers have seen the innovation of many pointers. Moving averages come in different structures, however their basic reason continues as before is to enable specialized brokers to track the patterns of financial assets by smoothing out the everyday value vacillations or commotion. By distinguishing patterns, moving average enable us to make those patterns work to support them and increment the precision of the expectation.

Time series model includes techniques for analyzing time series data to concentrate significant measurements and different attributes of the information. Time series forecasting is the utilization of a model to forecast future values based on previously

© Universiti Teknikal Malaysia Melaka