# Prediction of Accrual Expenses in Balance Sheet Using Decision Trees and Linear Regression

Chih-Yu Wang
Department of Information Engineering
and Computer Science
Feng Chia University
Taichung, Taiwan
M0217588@mail.fcu.edu.tw

Ming-Yen Lin*
Department of Information Engineering
and Computer Science
Feng Chia University
Taichung, Taiwan
linmy@mail.fcu.edu.tw

***Abstract*—In response to globalization, International Financial Reporting Standards (IFRS) has become the norm of the global capital markets. Companies preparing financial statements using IFRS may make the financial situation fully disclosed. Nevertheless, an overestimated accrual expense of a balance sheet may not only underestimate the earnings data, but also increase the cash outflows of the statement of cash flows. When the accrual expense is underestimated, corporate earnings will inflate earnings statistics. In addition, the problem of funds shortage may occur upon actual payment because the cash outflows of the statement of cash flows is underestimated. In this paper, we adopt the prediction mechanism in data mining to predict the unused vacation time of employees, which in turn becomes a part of the accrual expenses in the balance sheet. The prediction target is the bonus of unused annual leave in terms of unused hours so that the estimated amount of fees payable accuracy in the balance sheet can be improved. Both decision-tree models and regression analysis are used. Comprehensive experiments show that the decision-tree method outperforms the regression analysis method, with MAE of -23.1 and RMSE of 43.1.**

***Keywords—balance sheet, accrual expenses, bonus of unused annual leave, data mining, decision tree, regression analysis***

## I. Introduction

The government of Taiwan aims to enhance the global competitiveness of domestic enterprises, and further promote the integration of international capital markets. On May 14 2009, FSC members of the Executive Yuan announced that the International Financial Reporting Standards (IFRSs) was to be adopted comprehensively by domestic enterprises for work schedule reporting.

A full disclosure of the enterprises' financial circumstances were required for corporations using financial statements prepared by the IFRSs. Inaccurate estimation of accrual expenses in balance sheets could underestimate the corporate earnings and profit, or result in shortage of funds.

Therefore, the main purpose of this study is to apply data mining technology in order to establish a prediction model for the unused annual leaves of employees. This research aims at improving upon the previous estimations of unused annual leaves to accurately forecast the amount of bonus payable to employees. Previous estimations normally guess a certain amount without using any methodology. Therefore, we present two estimation methods for improvements in this paper. We have conducted experiments using real world data from a middle-size company. The results show that the estimations using decision trees are quite satisfactory. Both estimations have a large amount of improvements over traditional estimations.

The rest of the paper is as follows. Section II describes the annual leave entitlement and relevant data mining methods. Section III details the research methods used in this paper. Section IV presents the experimental results. Section V concludes this study.

## II. Employee Annual Leave Entitlement and Data Mining Methods

### A. *Employee annual leave entitlement*

In accordance with the provisions outlined in the Labor Standards Act [1], employees employed by the same employer or corporation, who continue to work over a certain period of time are entitled to annual leave according to the provisions as follows.

Employees who have been with the company for (1) Over 1 year but less than 3 years are entitled to 7 days of annual leave. (2) Over 3 years but less than 5 years are entitled to 10 days of annual leave. (3) Over 5 years but less than 10 years are entitled to 14 days of annual leave. (4) Over 10 years will get a day applied for each subsequent year of employment. A maximum entitlement of 30 days is applicable.

For employees who did not use their annual leaves due to the end of the year or termination of contracts, the employer should compensate for that accordingly [2].

### B. *Data mining methods*

Data mining [3] is an approach adopted to discover specific rules or knowledge that exist in large databases. Alternatively, it is considered as the science of extracting useful information from large datasets or data warehouse [4]. Constructing a prediction model is one of the major functionalities in data mining. Many methods have been presented to achieve

978-1-5090-5732-0/16/$31.00 ©2016 IEEE

successive predictions. Although there are many techniques Among the techniques, predictions using decision trees and linear regression have attracted many researchers in recent years [6]. The objective of the study is to predict the unused annual leaves of employees, in terms of numeric hours [7]. Therefore, we utilized the C4.5 algorithm to build a decision tree and used the tree to justify our predictions. We also build a linear regression model to predict the numerical value for the unused annual leave of an employer, after selecting appropriate vaiables from our data. In addition, some attributes of the data are converted to fit into the construction of the decision tree model, as described in Section III.

## III. METHODOLOGY

Fig. 1 outlines the research steps in this study. First, the required data was interpreted and defined. Second, real data from a company is collected. The data is cleansed in the third step. Four, we convert the data for proper modeling and construct relavant prediction models. Finally, the models are tested and evaluated.

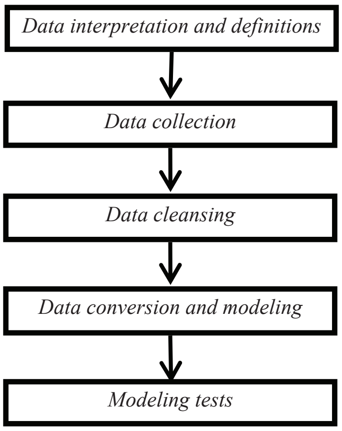
Data interpretation and definitions
Data collection
Data cleansing
Data conversion and modeling
Modeling tests

Fig. 1. Research steps

### A. Data interpretation and definitions

At first, the applicable independent variables and dependent variables were obtained from the personnel system of the case study company. To better capture the semantics of the unused annual leaves of employers, we have discussed with the personnel experts of the company in this case study. The result of attributes and definitions is shown in Table 1.

### B. Data collection

Data collection was considered as an annual concept as each employee would generate a set of data each year. This data included the independent variables for the year, i.e., the age of the employee, the position held, whether the employee was in a management position, and the available annual leave hours. When combined with the dependent variable, the "unused annual leave hours," these variables could be used as a basis to develop a prediction model for the estimation of unused annual leave. Additionally, the employee's date of birth was converted to its horoscope code based on Table 2.

The counties/cities of the employee's registered residence and mailing address were converted to English codes based on Table 3.

TABLE 1. DATA ITEMS IN THE CASE STUDY

| Data Item | Comments |
|---|---|
| Gender | Female (F) / Male(M) |
| Blood Type | A/B/AB/O |
| Date of Birth | Convert Age & Horoscopes |
| Employment Date | Convert Group seniority |
| Position | D/I/K/L/M/N/P/Q/R/S/T/U/W/Z |
| Management level | Y/N |
| Education | Doctor of Philosophy / Masters/ Bachelor/ … |
| Registered Residence | Taipei/ Taichung/ Kaohsiung/ … |
| Mailing Address | Taipei/ Taichung/ Kaohsiung/ … |
| Performance | A/ B/ C |
| Available Annual Leave Hours | 0~240 hours |
| Unused Annual Leave Hours | 0~240 hours |

TABLE 2. ZODIAC TABLE AND THE CONVERSION CODE

| Date of Birth | Horoscope | Code | Date of Birth | Horoscope | Code |
|---|---|---|---|---|---|
| 12/21~1/20 | Capricorn | A | 6/22~7/22 | Cancer | G |
| 1/21~2/19 | Aquarius | B | 7/23~8/22 | Leo | H |
| 2/20~3/20 | Pisces | C | 8/23~9/22 | Virgo | I |
| 3/21~4/19 | Aries | D | 9/23~10/24 | Libra | J |
| 4/20~5/20 | Taurus | E | 10/24~11/21 | Scorpio | K |
| 5/21~6/21 | Gemini | F | 11/22~12/20 | Sagittarius | L |

### C. Data cleansing

A total of 15,261 data items were obtained from the company personnel database for the period between 2004 and 2014. Next, a total of 3755 data items for the years 2007, 2010, and 2011 were processed by the performance review procedure, which excluded the data items "unused annual leave hours" as they were not applicable to the purpose of this study. Therefore, a total of 11,506 data items were kept for further review after this step.

TABLE 3. CODE CONVERSION TABLE FOR THE COUNTIES/CITIES OF THE EMPLOYEE'S ADDRESS

| County/City | Conversion Code | County/City | Conversion Code |
|---|---|---|---|
| Taipei | A | Miaoli | K |
| Taichung | B | Nantou | M |
| Keelung | C | Changhua | N |
| Tainan | D | Hsinchu | O |
| Kaohsiung | E | Yunlin | P |
| New Taipei | F | Chiayi County | Q |
| Yilan | G | Pingtung | T |
| Taoyuan | H | Hualien | U |
| Chiayi City | I | Taitung | V |
| Hsinchu | J | | |

TABLE 4. CROSS REFERENCE TABLE OF DIFFERENT BRACKETS OF UNUSED ANNUAL LEAVE HOURS AND THE QUANTITY OF DATA (25 LEVELS)

| Number of Hours | Level | Number of Hours | Level |
|---|---|---|---|
| 0 | A | 120.5~130 | N |
| 0.5~10 | B | 130.5~140 | O |
| 10.5~20 | C | 140.5~150 | P |
| 20.5~30 | D | 150.5~160 | Q |
| 30.5~40 | E | 160.5~170 | R |
| 40.5~50 | F | 170.5~180 | S |
| 50.5~60 | G | 180.5~190 | T |
| 60.5~70 | H | 190.5~200 | U |
| 70.5~80 | I | 200.5~210 | V |
| 80.5~90 | J | 210.5~220 | W |
| 90.5~100 | K | 220.5~230 | X |
| 100.5~110 | L | 230.5~240 | Y |
| 110.5~120 | M | | |

A subsequent review of the annual data identified a number of anomalies, such as specialists whose number of available annual leave hours were inconsistent with their seniority, employee records containing incorrect job information, and other similar data discrepancies. In order to avoid prediction inaccuracies, 278 such records were excluded. Hence, a total of 11,228 data items were kept for further analysis.

Subsequently, a total of 8896 data items pertaining to the years 2004, 2005, 2006, 2008, 2012, 2014 were selected as a training dataset. A total of 2332 data items were selected as a test dataset for model testing and verification. The selection builds a ratio about 80% for the training set and 20% for the test set.

### D. *Data conversion and modeling*

Decision trees are commonly applied and functioned as a classification model. Hence, the dependent variable, "unused annual leave hours," could not be used as a numerical value in this model. Thus, "unused annual leave hours" needed to be converted into codes representing the different levels that were based on the magnitude of differences between the hour brackets. The rule of conversion is outlined in Table 4. An interval of about 10 hour was applied to build the label (named level) of the leaves in the decision tree.

### E. *Modeling tests*

After the application of the established model to the test dataset, it was possible to obtain the number of correctly classified instances and the results of the classification accuracy for the decision tree model. True-positive ratio (TP Rate), which marks the number of correctly predicted labels with respect to the total number of instances, and false-positive ratio (FP Rate), which marks the incorrected number of predictions were recorded. In fact, the precision was calculated.

Once the model was applied to the test dataset, the following values could be obtained, Correlation Coefficient, Mean Absolute Error, and Root Mean Squared Error for the linear regression model.

## IV. EXPERIMENTAL RESULTS

### A. *Decision Tree*

The open data mining tool Weka [5] was used to establish a decision tree model based on the C4.5 algorithm. An example has been used to demonstrate the application of the decision tree model as shown below.

Case: A female employee had 192 hours of the available annual leave, with 20 years of seniority. She was 48 years old and did not hold a management position. She held the Code I. work position, with a performance record of B. Her horoscope was E, blood type was A, education level was N, registered residence was K, and mailing address was K.

On the basis of the decision tree model, it was predicted that the "unused annual leave hours" belonged to the N bracket, which was approximately 120.5-130 hours.

The test dataset was used to test the model. In total, there were 404 Correctly Classified Instances, with an accuracy rate of 17.3%. On the basis of the same approach, adjustments to the different independent variables were made and the rules for setting the dependent variable "unused annual leave" brackets were formed. Subsequently, the decision tree models were established and the test dataset was applied to test the models. The weighted averages of accuracies of the sub-models were obtained. The results are summarized in Table 5.

TABLE 5. SUMMARY TABLE OF THE ACCURACY OF THE DECISION TREE MODELS

| Applied Variables | Unused Annual Leave Hours | TP Rate | FP Rate | Precision | ROC Area |
|---|---|---|---|---|---|
| **12** | **25 Levels** | **0.173** | 0.062 | 0.183 | 0.648 |
| 7 | 25 Levels | 0.163 | 0.064 | 0.175 | 0.647 |
| 12 | 6 Levels | 0.325 | 0.149 | 0.329 | 0.635 |
| **7** | **6 Levels** | **0.53** | 0.269 | 0.477 | 0.747 |

TABLE 6. CROSS REFERENCE TABLE OF DIFFERENT BRACKETS OF UNUSED ANNUAL LEAVE HOURS AND THE QUANTITY OF DATA (SIX LEVELS)

| Number of Hours | Level |
|---|---|
| 0~56 | A |
| 56.5~80 | B |
| 80.5~120 | C |
| 120.5~168 | D |
| 168.5~216 | E |
| 216.5~240 | F |

TABLE 7. SUMMARY TABLE OF REGRESSION MODEL TESTS AND VERIFICATION DATA

| Applied Variables | CC | MAE | RMSE |
|---|---|---|---|
| 12 | 0.7171 | 33.7807 | 46.5819 |
| 11 | 0.7155 | 33.9401 | 46.8738 |
| 11 | 0.7222 | 33.1687 | 45.4415 |
| 11 | 0.7143 | 33.9215 | 46.759 |
| 9 | 0.7186 | 33.2874 | 45.5737 |
| 7 | 0.7172 | 33.7517 | 46.3613 |

The unused annual leave hours were divided into six levels. The conversion rules are detailed in Table 6.

### B. *Linear Regression*

The Weka software was used to establish a model based on the least-squares algorithm. The regression model was applied to the aforementioned case study. The unused annual leave hours for this case was predicted to be 66.16 hours.

The test data set was applied again to obtain the following test results, CC = 0.7171, MAE = 33.7807, and RMSE = 46.5819.

On the basis of the same approach, adjustments to the different independent variables were made and the rules for setting the "unused annual leave" brackets were formed in order to build different regression models. The test data was used to test the sub-models. The verification results are listed in Table 7.

TABLE 8. CROSS-REFERENCE TABLE OF THE CLASSIFICATION RESULTS AND THE CONVERTED HOURS

| Hours/Brackets | | Converted Hours | Hours/Brackets | | Converted Hours |
|---|---|---|---|---|---|
| 0 | A | 0 | 120.5~130 | N | 125 |
| 0.5~10 | B | 5 | 130.5~140 | O | 135 |
| 10.5~20 | C | 15 | 140.5~150 | P | 145 |
| 20.5~30 | D | 25 | 150.5~160 | Q | 155 |
| 30.5~40 | E | 35 | 160.5~170 | R | 165 |
| 40.5~50 | F | 45 | 170.5~180 | S | 175 |
| 50.5~60 | G | 55 | 180.5~190 | T | 185 |
| 60.5~70 | H | 65 | 190.5~200 | U | 195 |
| 70.5~80 | I | 75 | 200.5~210 | V | 205 |
| 80.5~90 | J | 85 | 210.5~220 | W | 215 |
| 90.5~100 | K | 95 | 220.5~230 | X | 225 |
| 100.5~110 | L | 105 | 230.5~240 | Y | 235 |
| 110.5~120 | M | 115 | | | |

### C. *Overall descriptions*

In terms of the decision tree model, when the dependent variable, "unused annual leave hours," was set at 25 brackets, regardless of whether properties of 12 or 7 independent variables were assumed, the accuracy of the model was lower than 20%. If the dependent variable, "unused annual leave hours," was set at 6, the model established on the assumption of properties of 12 independent variables, had an increased accuracy of 32.5%. If the model assumed only 7 independent variables, the accuracy of the model was increased to 53%.

In contrast, for most of the linear regression models, regardless of the assumed independent variable properties, the relevance of the training dataset and the test dataset was approximately 0.71.

The same dataset was used to assess the accuracies of the different modeling methods, i.e., decision tree model and regression model. On the basis of "12 independent variables with the unused annual leave hours set at 25 brackets," the classification results of the decision tree was obtained. The classification codes of the decision tree were converted to the corresponding number of hours in order to facilitate the calculation of the model's MAE and RMSE. The conversion rules are shown in Table 8.

When the classification results of the decision tree were converted to hours, the calculated MAE = -23.14, and RMSE = 43.1.

### *D. Model implementation*

The decision tree model and the linear regression model was applied to the employee data of year 2015 to predict the unused annual leave time. The obtained results were compared with the actual "unused annual leave time" in order to assess the accuracy of the implementation. The implementation results are shown in Table 9.

TABLE 9. MODEL IMPLEMENTATION RESULTS

| | |
|---|---|
| The Total Available Annual Leave Hours | 199084 |
| The Total Unused Annual Leave Hours | **90722** |
| The case study company estimated that 20% of the "Total Available Annual Leave Hours" will be "unused annual leave hours". | 39817 |
| The difference between the actual "unused annual leave hours" | **-50905** |
| The "unused annual leave hours" predicted by decision tree model | 85355 |
| The difference between the actual "unused annual leave hours" | **-5367** |
| The "unused annual leave hours" predicted by linear regression model | 76044 |
| The difference between the actual "unused annual leave hours" | **-14678** |

According to the data shown in Table 9, if the prediction of the case study company was used, i.e., 20% of the total available annual leave as the unused annual leave, the difference between the predicted and the actual unused annual leave was -50905 hours. The error rate was 56.1%. If the decision tree model was applied, the difference in unused annual leave was -5367 hours. In this case, the error rate was 5.9%. If the linear regression model was adopted, the difference in unused annual leave was -14678 hours, with an error rate of 16.2%. In this case study, we conclude that the decision tree model outperforms the regression model.

## V. CONCLUSIONS

This study applied data mining technology to establish a prediction model for the unused annual leave hours of employees in a company. This research improved upon the estimation method that was being used by the case study company. The decision tree and linear regression models have predicted higher number of unused annual leave hours in comparison to the method currently used by the company. Consequently, the estimated bonus of unused annual leave was relatively high. However, the predicted results were closer to the actual unused annual leave hours. In addition, these prediction methods could accurately forecast the cash flow of the enterprise which helped avoid inaccurate fund allocations caused by cost underestimation.

In accordance with the results of this study, the decision tree model produced more accurate estimation of unused annual leave hours. However, the linear regression models could easily identify the extent of influence that each dependent variable could have on the independent variables. Thus, the Human Resources department of the company could apply different forecasting models based on their need. For example, at the beginning of the year, the decision tree model could be adopted to estimate the number of unused annual leave hours, whereas the linear regression model could be used to explore the utilization of annual leave based on different job roles.

## REFERENCES

[1] R.O.C. Labor Standards Act. 38.

[2] R.O.C. Enforcement Rules of the Labor Standards Act. 24.

[3] D. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.

[4] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus, "Knowledge Discovery in Databases: An Overview," AI Magazine, Vol. 13 No. 3, pp. 57-70, 1992.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and Ian H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, Vol. 11, Issue 1, 2009.

[6] S. S. Rathore and S. Kumar, "A Decision Tree Regression based Approach for the Number of Software Faults Prediction, " ACM SIGSOFT Software Engineering Notes, Vol. 41, No. 1, pp. 1-6, 2016.

[7] M. Y. Chen, "Comparing Traditional Statistics, Decision Tree Classification And Support Vector Machine Techniques For Financial Bankruptcy Prediction," Intelligent Automation & Soft Computing, Vol. 18, No. 1, pp. 65-73, 2012.