English Learning Platform Student Behavior Analysis in Buying Courses by Using Data Mining Techniques

Yohanssen Pratama[a]
[a] Faculty Informatics and Electrical Engineering, Institut Teknologi Del, Indonesia
Corresponding author's e-mail:[a] yohanssen.pratama@del.ac.id

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: English Learning Platform is an online English learning platform which utilizes subtitled YouTube video as learning platform. This learning platform is said to be personalized, by recommending next material based on how well one performs for each exam given per video. If students are able to learn successfully, they may have a tendency to buy more courses. Our goal is to use data mining techniques to research on the relationship about number of courses that students purchase with their learning behavior. After comparing 5 methods by using 4 machine learning algorithm we found that student score attributes has stronger impact to determine the number of packages that bought by students and the login frequency doesn't affect the number of purchased courses. The neural network algorithm gives the best accuracy which around 90% compare to another algorithm there are decision tree, random forest, and k-means.

Keywords: Data Mining, Decision Tree, K-Means, Neural Network, Random Forest

1. Introduction

English Learning Platform is an online English learning platform which utilizes subtitled YouTube video as learning platform. English Learning Platform uses big data and machine learning technology to provide adaptive learning system which offers everyone a tailor-made curriculum based on their level, interest, and learning curves [1]. In this platform, the users are initially given customized sequence of videos based on their performance on the initial introduction video. This learning platform is said to be personalized, by recommending next material based on how well one performs for each exam given per video. This platform aims to improve essential skills in learning a language which are listening, reading, vocabulary, and understanding.

In this research, we have the chance to make implementation of data mining technique that we studied by using data that was provided by English Learning Platform. We were given three kind of possible direction (but not limited to) to apply our data mining technique. The possible direction is to do research about students learning behavior and try to search method that can identify hard working student (beginner level), the number of courses they buy (medium level), and the optimization of choosing suitable video for student to learn (advanced level). For this research, we choose to analyze the relation of student's learning behavior with the number of course they buy.

Quoting Alex Schultz, VP of Growth of Facebook, "Retention is the single most important thing for growth", we would like to figure out what are the factors which have impacts on user retention and drives them to buy more packages [2]. We did several behavior analyses using data mining techniques to figure out what factors which may affect student's decision to buy more packages. The data mining techniques that we used in here is Decision Tree, K-Means, Random Forest, and Neural Networks.

Students learning behavior in here are includes a survey, students testing score, how often they come into the system, how many vocabularies they save during the learning process, and also how often they ask teacher question. From several behaviors above we want to know if certain type of students is more willing to buy more courses.

2.Research Method

There are 2 stages to analyze the student behaviors relationship with the number of course they buy, data preprocessing stage and data analysis stage (fig 1.).

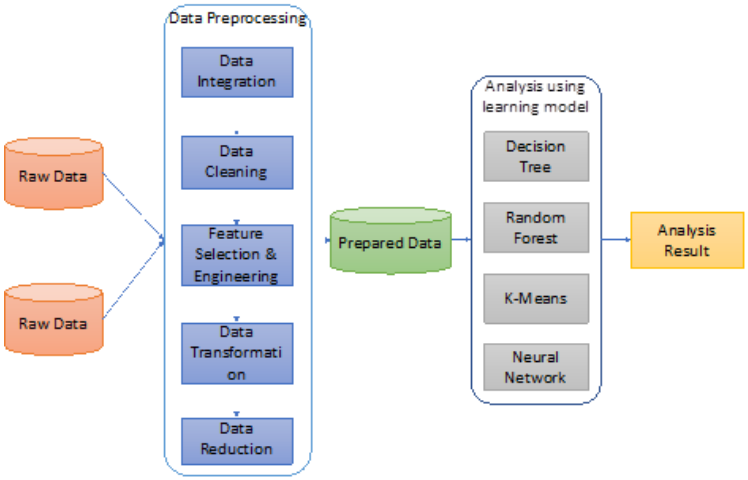


Figure 1. Data Flow Preprocessing and Analysis Stage

Fig 2. Show the student learning behavior data and course content:

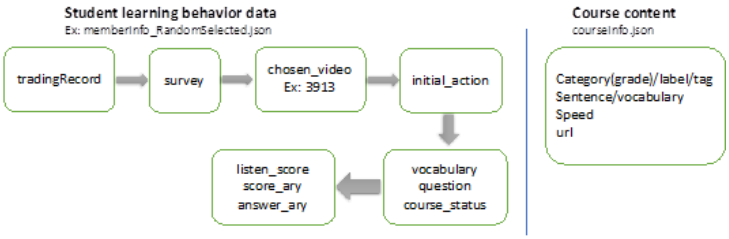


Figure 2. English Learning Platform Learning System User Flow

Course content description

Classification: Category, Label, Tag

Content: Sentence

You can define the level of video

- Speed
- Vocabulary (how many words of vocabulary are within gept1)
- Sentence difficulty (by length) <- design by yourself

2.1.Preprocessing

2.1.1 Data Integration

English Learning Platform provided us with 3 different sets of data:

1. courseInfo.json, which consists all information about the courses (298)
2. memberInfo.json, which consists information about the users (20000)
3. testAnswer.json, which consists user's exam information (20000)

Each set of data contains important information which depends on other data set. For example, if we would like to find out the information about a video content which was chosen by user from "memberInfo.json", we need to find that information from "courseInfo.json". For this purpose, we need to do data integration.

2.1.2 Data Cleaning

Before we started the data mining process, one of the most important things to do is to make sure that we have a good quality of data [3]. Usually, data that we are about to process cannot be used directly, there always are inaccurate or even invalid data among them which need to be corrected or removed.

Same case with the data provided by English Learning Platform, we tried our best to find any irregularity inside the dataset before we start the mining process. Since the data inside courses and test answers dataset are all being handled by the system without user interference, there is nothing much to clean here. On the other hand, the member info data have a lot of input which came from the user. The data became dirty that we need to clean up first. Therefore, we focused more on this dataset.

In this process, we identified and removed outliers and noises that can affect our result in data mining.

a. Trading Record Package (Total 142 out of 20000)[0.71%]

The first finding was about the user's trading record's package which is used to determine the actual package that user have bought. According to the English Learning Platform official website (https://www.English Learning Platform.com/course/products) and some extra information which we got from the email they sent to us, they only have 4 kinds of package to sell as their products which are package 30, 60, 120 and 200. But after we spent some time analyzing the dataset, we found out that there are actually a lot of other kinds of package beside those mentioned above.

Some example for that irregular package is 10, 0, 80, 240, 45, 400, 375, 200 堂, 90, 170, 70, 6, 150, 250, 600, 300, 215, 420, 570, 480, 540, 660, 870, 180, 360, 780, 690, 420, 250, 160 and so on.

Our first thought about the problem was that it might be related to the promotion gift/campaign. But, after we did some verification with the extra information that they sent to us via email, we are sure that it is not related to the promotion gift. Therefore, we decided to label those data records as invalid and remove them from the dataset.

b. Corrupted Timestamp Data (Total 154 out of 19858)[0.77%]

The second irregularity which we found out is about the record's timestamp. There were a small number of records which don't have their last_login_time or verifyDate logged correctly from the system such as last_login_time being equal to '0000-00-00 00:00:00' or even "None".

For verifyDate being equals to "None", we have checked all the records and are sure that those accounts are the accounts which was being used by developers upon developing or testing the system. There are around 7 records for this kind of irregularity

c. Malformed score_ary format (Total 1053 out of 19704)[5.34%]

score_ary is the attribute which are used to log the user examination scores. The correct format of this attribute is as following "score_ary": {"0": 78, "1": 83, "2": 92}. The key represents the i-th exam the user took. We also are able to determine the which video's section it exactly is by tracing it using chosen_video attribute from memberInfo dataset and section attribute from courseInfo dataset.

The irregularity which we found is the data type. Instead of being a dictionary or array list as shown above, there were another type such as a regular array like [0, '82997', '82997', '82997', '82997'] which consist of integer or string value inside which we do not have any idea what it is about.

d. Abnormal Score Average (Total 38 out of 18651)[0.2%]

Score Average is an attribute which is used to record user's average of exam's score. There is nothing to worry about for all of those scores beside the zero one. Since the zero average score have two kind of meanings; the first one is that it represents those users have not taken any exam yet, and the other meaning is that it represents that those users have taken some exam and actually get zero score.

We further analyzed those users who actually got zero score and discovered that almost all of them have irregularity phenomenon. One of the irregularities is that half of them are took the exam only twice and all of them got 0 score for their first time and got 1 score for their second time or vice versa. Another irregularity is about around 5 users who have taken more than 15 exams and got 0 for all of it.

e. Corrupted Timestamp Data No. 2 (Total 44 out of 18613)[0.2%]

We found out that there are a small number of records which have their last_login_time earlier than their verifyDate, which means that have some activity recorded even before they have register. After we spent some time investigating them, we found out that most of them are passive users with almost no activity at all on the system. Therefore, we removed them from the dataset rather than risking for it to be later become an outlier.

f. User Refund Problem (Total 878 out of 18146)[4.7%]

After we done some deep observation on user behavior when they are buying the package, we found out that a lot of them only login on the same date when they register and never come back again to the system. The same phenomenon also happened for several days after a user registered.

Later, we found out that the company have some refund policy that can explain this problem. On their official website, they said that as long as the user have not spent the package they bought, they are still eligible to make refund request and get full refund as long as the period of time is not over 7 days. They also still can get 90% refund when they only have finished 1 course only.

To deal with this issue, we removed all of those users, who never come back to the system again in the period of 7 days after they registered. Since most of them do not have any activity recorded by the system at all and even for small number of them who have activity recorded, those activities are not accurate for representing their actual behavior since most of those activity are generated because most likely, the users are testing the platform and are trying to get used to it.

g. No Activity User (Total 142 out of 17268)[0.8%]

We also found out that there are some users who did not have any activity recorded at all, for example: not taking exam at all, not saving any vocabulary, not asking any question, have not chosen any video, and also not even a single last traced activity by the attribute course_status.

h. No Trading Record User (Total 15 out of 17253)[0.8%]

This is probably one of the most interesting irregularity that we found out. There are a small number of users who have some activity being recorded by the system. Otherwise, those users did not have any trading record.

i. Unreasonable Number of Package (Total 15 out of 17253)[0.0%]

Upon analyzing the dataset, we found out that there are users who bought a lot of packages that if we combined the total course credit they have, it is over 500 thousand. But after we have applied a lot cleaning task above, all of those outliers have been taken care indirectly.

2.1.3 Feature Selection & Engineering

Since our goal is to analyze the relation of student's learning behavior with the number of course they buy, not all of the attributes provided within the dataset are important for us [4]. We need to select some of them, which we think that might be important for us to determine how likely the user will buy the course.

We also needed to create some new feature based on the original one by applying aggregation or calculating a new value from some other features (feature creation).

2.1.4 Data Transformation

In this project, we used data normalization on several attributes.

"scoreAvg", "questionCnt", "vocabularyCnt", "monthlyActivity", "examCnt", "vocabExtraCnt" and "remainderCredit" was normalized in User Behavior Analysis on Buying Packages using Decision Tree to map the data so the data will have the same range of values before fed into the decision tree algorithm.

"login_interval" was normalized in User Behavior Analysis using K-Means on Login Interval. The purpose of this normalization is to make the data has the same range of values before fed into K-Means clustering algorithm.

General formula for the normalization that we use is: $\frac{att_n - \underline{att}}{att_{max} - att_{min}}$

att_n= attribute value that we want to normalize

$\underline{att}$ = mean of attribute value

att_{max}= maximum value of the attribute

att_{min} = minimum value of the attribute

2.1.5 Data Reduction

Several data reduction process was done in this final project. First, we only selected relevant features that related to the user behavior to buy more packages. We also did binning on total packages (grouping the total

packages bought) in User Behavior Analysis using K-Means on Login Interval to simplify our calculation without affecting our calculation accuracy [5].

For "surveyBool" and "firstPkg" attribute we do data discretization. In "surveyBool" we reduce the data that contain information or no into "0" or "1". In "firstPkg" we replace the value 30, 60, 120, and 200 into 0, 1, 2, and 3.

We also used downsample in User Behavior Analysis on Buying Packages using Decision Tree to have balanced train data on both categories.

2.2 Analysis

2.2.1 User behavior analysis on buying Packages using Decision Tree, Random Forest, and Neural Network

Retention is important in business [6]. We would like to figure out whether some behaviors or factors may affect student's decision to buy more package. The key attribute which we would to obtain from this analysis is the "transaction count". "Transaction count" is the number of transactions which a user did. Based on "transaction count", we can classify students into 2 categories [7]:

1. "SingleTx": Students who bought only one package. This group of students may be these who bought 30, 60, 120, or 200 packages. They may be the students who already finished the course or has reached the timing constraint but never buy the course again, and also the students who just bought the course and haven't finished all the course.

2. "MultipleTx": Student who has bought the packages more than once.

For this analysis, we extracted and created few features from the "memberInfo.json" as we can see in the Table 1.

Table 1. Feature Selection and Feature Creation

New Attribute	Description
surveyBool	Represent whether the user had done the survey [0,1]
scoreAvg	Represent user average exam's score
tradingCnt	Represent total course transaction count
questionCnt	Represent total question user had asked
vocabularyCnt	Represent total vocabulary stored
monthlyActivity	Represent total question+exam activity count for duration of 3 months before last login
examCnt	Represent total exam taken
vocabExtraCnt	Represent extra saved vocabulary due to course change event happened
firstPkg	Represent the first packageID that user bought
remainingCredit	Represent the current course credit that user have

All of this data along with memberID then merged together into a dataframe. The next step is to do data normalization on "scoreAvg", "questionCnt", "vocabularyCnt", "monthlyActivity", "examCnt", "vocabExtraCnt". The sample of this normalized attribute is displayed on Table 2.

Table 2. Normalized Attribute

	memberID	surveyBool	scoreAvg	tradingCnt	questionCnt	vocabularyCnt	monthlyActivity	examCnt	vocabExtraCnt	firstPkg	remainingCredit
0	51458	0	-0.613192	1	-0.006954	-0.028492	-0.032749	-0.030836	-0.004038	0	-0.041015
1	80795	0	0.236808	1	-0.003439	0.004042	-0.032749	-0.007858	-0.004038	0	-0.065432
2	91558	0	-0.263192	1	-0.005196	-0.027236	-0.032749	0.028906	-0.004038	2	-0.017707
3	99944	0	-0.063192	1	0.007106	-0.012665	-0.032749	-0.023483	-0.004038	2	0.027798
4	51110	0	0.106808	1	-0.006954	-0.023593	-0.032749	-0.018888	-0.002658	0	-0.056553

We then examined our data distribution by user transaction (who have been active for 180 days). We found out that the number of users with single transaction 6831 is users, and multiple transactions 4090 is users. Because the group of no transaction user is very small compared to the other two groups, we decided to only observe the behavior from the single and multiple transaction group [8].

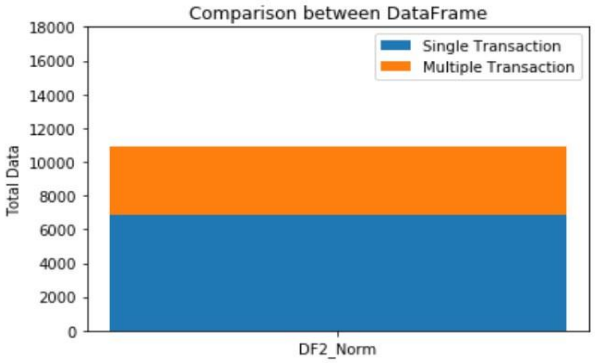


Figure 3. Data Distribution of Single and Multiple Transaction

We used only 4090 data on both categories to make sure that our learning algorithm has a balanced input data, and divide our data with shuffled distribution 80:20 for train data (3272 data) and test data (818 data). Then, we input our data into 3 kinds of learning algorithm: decision tree, random forest, and neural network [9].

2.2.2 User behavior analysis using K-Means on Login Interval

In this part, we tried to figure out whether students' login frequency may affect his or her decision in buying more courses.

a. Data Selection

For data selection, we used interval measurement scale to group students based on the total package they bought [10].

The Package they bought ("package") are divided into four groups: 30, 60, 120, and 200. If the student bought more than one packages, we use their total packages and floor down the total packages into the nearest category [11]. For example, user A who bought 30 and 120 courses will be categorized as 120-group. User B who bought 30, 60, and 120 courses will be classified into 200-group.

Other attributes which we used for this clustering method are "memberid", "package_num", and "login_interval". "memberid" is the member id of the student, we use this as the primary key in our dataframe. "package_num is the total transaction of each student.

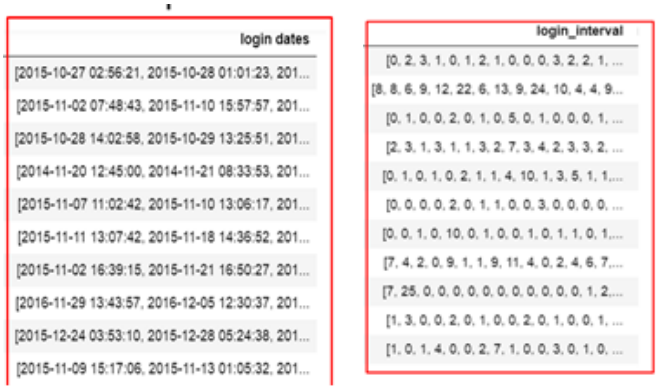


Figure 4. Login dates transformation into login_interval

We also select our data from a window of number of login (Fig. 5

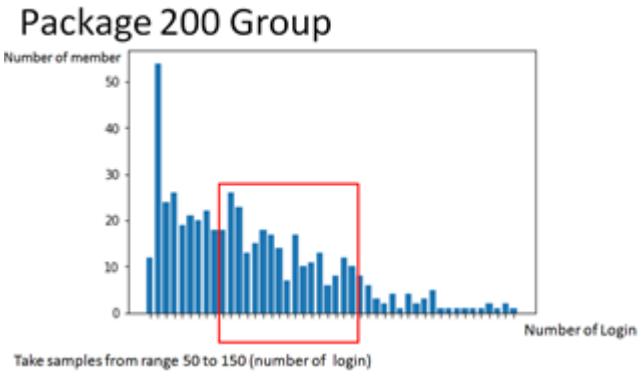
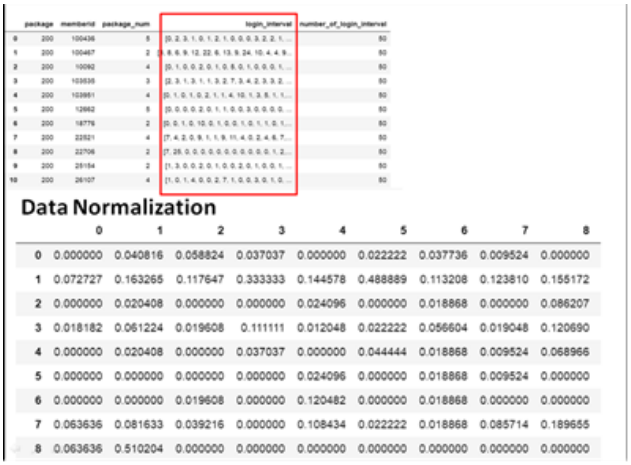


Figure 5. Data sampling

We took sample range 50 to 150 number of login so that our target member is the member who already made some progress and have the enough experience using this learning platform. The next step is to remove the outlier for "login_interval" [12].

Outliers (m=2) = Absolute value (data – mean value of data) < m * standard deviation (data)]

Then we do data normalization on "login_interval" (Fig. 6), and do K-mean clustering with k=3.



Data Normalization

	0	1	2	3	4	5	6	7	8
0	0.000000	0.040816	0.058824	0.037037	0.000000	0.022222	0.037736	0.009524	0.000000
1	0.072727	0.163265	0.117647	0.333333	0.144578	0.488889	0.113208	0.123810	0.155172
2	0.000000	0.020408	0.000000	0.000000	0.024096	0.000000	0.018868	0.000000	0.086207
3	0.018182	0.061224	0.019608	0.111111	0.012048	0.022222	0.056604	0.019048	0.120690
4	0.000000	0.020408	0.000000	0.037037	0.000000	0.044444	0.018868	0.009524	0.068966
5	0.000000	0.000000	0.000000	0.000000	0.024096	0.000000	0.018868	0.009524	0.000000
6	0.000000	0.000000	0.019608	0.000000	0.120482	0.000000	0.018868	0.000000	0.000000
7	0.063636	0.081633	0.039216	0.000000	0.108434	0.022222	0.018868	0.085714	0.189655
8	0.063636	0.510204	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 6. Data Normalization of login_interval

3.Results and discussion

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily. The discussion can be made in several sub-chapters.

3.1 User behavior analysis on buying Packages using Decision Tree, Random Forest, and Neural Network

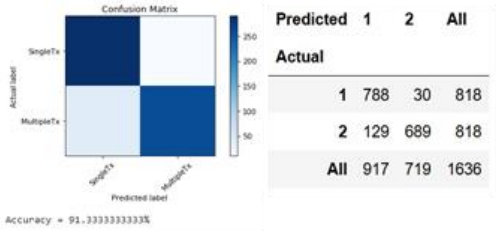
3.1.1 Decision Tree

After we obtained our decision tree model, we tested our model on the test data set [13]. Its accuracy is around 90%. The result of the test is illustrated in the Fig 3. From Fig. 3, we can see that the most important feature that affect the user to do single or multiple transaction is the "remainingCredit". If the remaining credit within 6 months of becoming active user is below -0.04 (25.5 credit), user can be categorized as single transaction user, while user who has more 25.5 credit tends to be multiple transaction user.

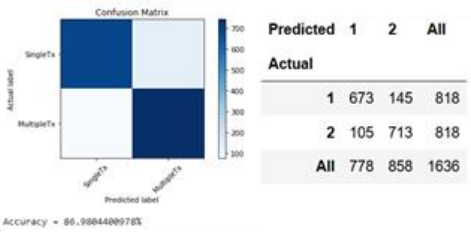
Figure 7. Decision Tree Graph

3.1.1 Random Forest

For our analysis using Random Forest algorithm, we generated 10 trees. We picked the one that has the highest information gain on the first split [14]. Surprisingly, the main attribute in our random tree is also "remainingCredit" (Fig 7.). The accuracy of our model is around 86%, which can be seen on Fig. 8.



Predicted	1	2	All
Actual			
1	788	30	818
2	129	689	818
All	917	719	1636



Predicted	1	2	All
Actual			
1	673	145	818
2	105	713	818
All	778	858	1636

(a) (b)

Figure 8. Confusion Matrix from Decision Tree Analysis (a) and Random Forest Analysis **(b)**

3.1.2 Neural Network

Lastly, we try to create neural network model [15] with 200 hidden layers, 500 max iteration, and alpha=1e-5 and for this model we get around 98% accuracy.

Figure 9. Confusion Matrix from Neural Network Analysis

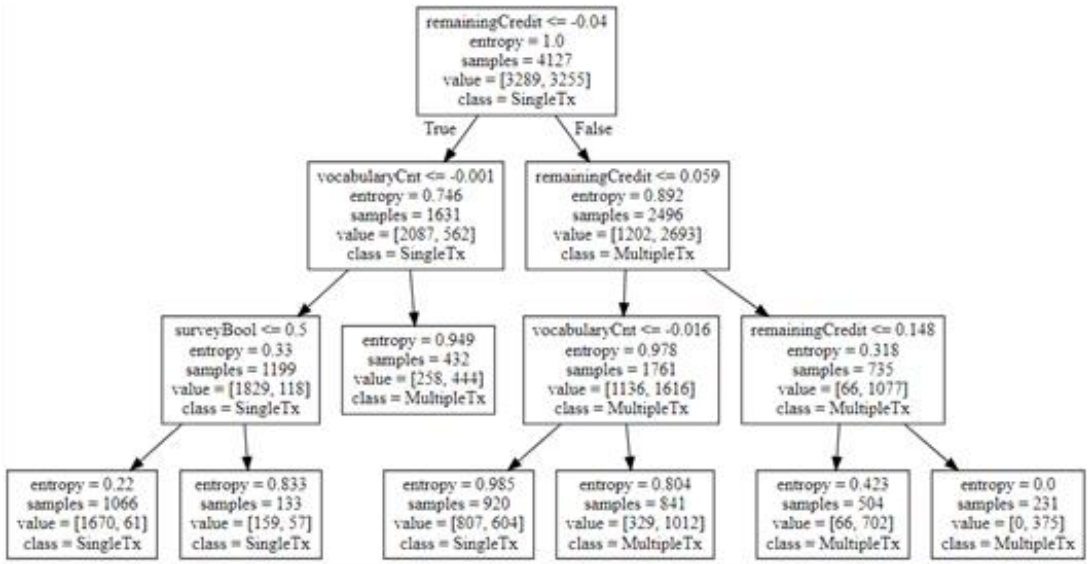


Figure 10. Random Forest Graph

3.1 User behavior analysis using K-Means on Login Interval

Table 3. interval measurement scale to group students based on total package

K-Mean: Package 200 Group

Group Number:	Number of Members:
0	305
2	77
1	37

K-Mean: Package 120 Group

Group Number:	Number of Members:
0	356
2	85
1	31

K-Mean: Package 60 Group

Group Number:	Number of Members:
1	379
2	84
0	29

K-Mean: Package 30 Group

Group Number:	Number of Members:
1	382
2	21
0	10

From Fig 11, we can find out the plot of login interval and number of logins from each package group. We can see that for all of the group, the majority group has the same login behavior. They are the group of people who actively login and has shorter login interval compared to other people group.

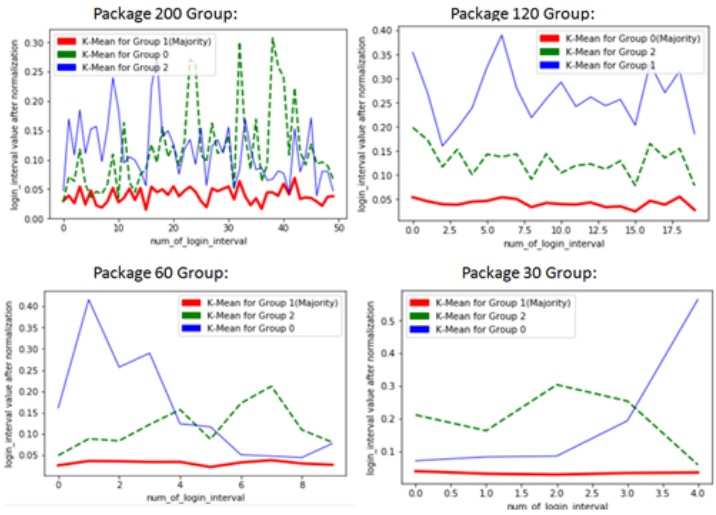


Figure 11. User Behavior Graph by Packages

4.Conclusions

Student learning behavior is something that we need to analyze for most learning platforms nowadays. If students are able to learn successfully, they may have a tendency to buy more courses. Our goal is to use data mining techniques to research on the relationship about number of courses that students purchase with their learning behavior.

1. We found out that the number of course left for user that has been validated for at least 6 months can determine whether they will buy more course or not.

2. Majority students in each package group tend to login to the system actively and frequently. Therefore, it may imply that how often the student's login to system does not affect the number of courses they purchase (since members in package 200 group have the same behavior as other groups).

3. There are two models trained here. The first one is linear regression with penalty, we found that there are five attributes influencing the prediction most seriously. All of these five attributes belong to LEVEL. It is reasonable that those with bad level want to buy courses due to satisfying some temporal criteria. The other one is Random Forest; in this model we realize that most of features in the trained model weight equally. As a result, since there are no any attributes that have a stronger impact toward prediction classes, the prediction become less accurate than the linear regression.

4. Students score is one of the attributes that has an impact to determine the number of packages that bought by students.

5. Many people are taking their training in serious since they have saved more than 3000 vocabulary and scores tend to exceed the mean.

5.Acknowledgment

This work was supported in part by Institut Teknologi Del and English Learning Platform.

References

1. Y.L. Chen, K. Tung, RJ Shen, and Y.H. Hu, "Market basket Analysis in Multiple Store Environment", Decision Support system., 2005, Elsevier
2. S. Vijaylakshmi, V. Mohan, S. Suresh Raja, "Mining of users access behavior for frequent sequential pattern from web logs", International Journal of Database Management Systems (IJDM), vol.2, August 2010
3. J. Watada, K. Yamashiro, "A Data Mining Approach to consumer behaviour", Proceedings of the first International Conference on Innovative Computing Information, 2006.
4. P.S. Sandhu, D.S. Dhaliwal and S.N. Panda, "An efficient approach based on profit and quantity", International Journal of the Physical Sciences, vol.6(2), pp. 301-307, 2011
5. B. Yildiz and B. Ergenc, "Comparison of Two Association Rule Mining Algorithms without Candidate Generation", International Journal of Computing and ICT Research 2010.

6. P. Peter, W. Waiswa, and V. Baryamureeba, " Extraction of Interesting Association Rules Using Genetic Algorithms", International Journal of Computing and ICT Research, vol.2, no.1, 2008.
7. J. Pillai, " User centric approach to itemset utility mining in market basket analysis", International Journal on Computer Science and Engineering, 2011.
8. Shrivastava, R. Sahu, "Efficient association rule mining for market basket analysis", Journal of e-Business & Knowledge Management, vol.3, 2007.
9. R. Agrawal and L. Srikat, " Fast Algorithms for mining association rules", Proceedings of the 20th Database Management Systems, vol.3, no.3, 2011.
10. R. Agrawal, T. Imilienski, and A. Swami, "Mining Associations Rules between Sets of Items in large databases", International conference on management of data, pp. 207-216, 1993.
11. S. Brin, R. Motwani, and C. Silvertrin, "Beyound Market Baskets: Generalizing association rules for correlation", SIGMOD Record (ACM Special Interest Group on Management of Data), vol.26(2):265, 1997.
12. S. Brin, R. Motwani, J.D. Ullman, and S. Tsur, " Dynamic interest counting and implication rules for market basket data", SIGMOD Record (ACM Special Interest Group on Management of Data), vol 26(2): 255, 1997
13. G. Paulo, "Applied Data Mining: Statistical Methods for Business and Industry"
14. J. Han and M. Kamber, " Data Mining Concepts and Techniques", Simon Fraser University
15. S.Moro, P. Cortez, and R.M.S Laureano, "A Data Mining Approach for Bank Telemarketing Using the Rminer Package and R Tool", 2013.