Mindanao Journal of Science and Technology Vol. 19 (1) (2021) 224-235

Ensemble Predictive Model for Academic Churn Risk Using Plurality Voting

Jovelyn C. Cuizon
School of Computer Studies
University of San Jose - Recoletos
Cebu City, 6000 Philippines
jcc@usjr.edu.ph

Date received: July 1, 2020
Revision accepted: March 27, 2021

Abstract

Academic churn analysis involves identifying students who are most likely to discontinue schooling. Although churn is an unavoidable phenomenon, timely detection and early intervention have been proven to be effective retention mechanisms. The study aimed to develop a model that predicts the likelihood of students to churn to provide insights to school administrators to initiate activities to prevent student attrition. This study examined academic, demographic and psychological data of students admitted as freshmen from 2005 to 2010 in two programs (Bachelor of Science in Information Technology [BSIT] and Bachelor of Science in Computer Systems [BSCS]) of the University of San Jose-Recoletos, Cebu City, Philippines. The psychological data representing the personality traits of students were gathered through Manchester Personality Questionnaire (MPQ). This study applied the ensemble method in machine learning to create a predictive model to define profiles for churners and non-churners. The predictive model was created by bagging three different classification models, namely support vector machine (SVM), random forest (RF) and k-Nearest Neighbor (k-NN) via plurality voting. The performance of the model was verified through 10-fold cross-validation with an overall accuracy of 78%. The model will be integrated into the college student advising system to provide notifications to administrators on students who need intervention on their subsequent enrollment.

***Keywords**: predictive modeling, data mining, academic churn, churn analysis, student attrition*

1. Introduction

Voluminous student psychological and academic data have been accumulated over the years. Lying hidden in all these data is potentially useful information that might be of significance to the organization. It has been a recurring dilemma in higher educational institutions to ascertain the root causes of student attrition. The major challenge for academic institutions is to uncover the evidence of potential abandonment (churn) among students. In the educational setting, churners discontinue a program to enroll in another program or leave the university without completing an academic degree. Involuntary churn occurs when the school terminates a student's enrolment due to violations of school policies by expulsion or dismissal. Voluntary churn occurs when the student intentionally leaves the university to join another university (transferee) or completely abandon schooling (dropout). Deliberate discontinuance of a program to join another university program (shiftee) is also considered voluntary churn.

Student attrition is an essential issue for any academic institution. To manage this problem, schools need to understand students' behavior and classify performance indicators to look for churn and non-churn students so that the necessary decisions can be made before churn happens (Klepac *et al.*, 2015). Numerous literature has been written on empirical methods to determine various indicators of academic churn. A report by Ramist (1981) identified academic matters as relatively the most frequent reason for dropping out, which might have stemmed from having poor grades, boredom with courses, change in career goals and inability to take desired courses. Financial difficulties, motivational problems, personal considerations, dissatisfaction with the college and having jobs were also part of the list (Ramist, 1981). The instrument developed by McRoberts and Miller (2015) revealed that the factors influencing students' decision to leave a college program were related to wellness, finances and overall college experience. Chai and Gibson (2015) evaluated demographic, academic history, enrolment, course, resource use, and engagement. They studied features to build student attrition models at four different periods – pre-enrolment, enrolment, in-semester and end-semester.

Predictive modeling enables researchers to model potentially relevant student attributes to make inferences or identify meaningful relationships. The use of these relationships to better predict future events significantly provides an overall assessment of the students and meaningful insights into churn risk analysis (Sharma and Panigrahi, 2011). Figini *et al.* (2009) used survival analysis models to define profiles of students with a high churn risk, evaluating the dependence of risks based on academic and demographic factors. Luna (2000) used classification models such as logistic regression, discriminant analysis and classification and regression trees (CART) to predict academic standing. Wei and Chiu (2002) incorporated a multiclassifier class-combiner approach to identify potential churners in telecommunications using subscriber contractual information and call details. Raju (2012) found that the decision tree model is best for predicting student persistence. Chai and Gibson (2015) recommended logistic regression over a tree-based model for ease in interpretation.

Numerous studies have been conducted investigating this phenomenon. However, most of them were more focused on academic, demographic and social attributes than personality attributes as causal factors for college persistence.

This study aimed to develop a software model to predict at-risk students at an early stage based on the students' academic, demographic and personality attributes. It applied the ensemble method in machine learning to create a predictive model that combines probabilistic (support vector machine [SVM]), decision tree induction (random forest [RF]), and instance-based (k-Nearest Neighbor [k-NN]) classification models via plurality voting. The study evaluated three models using academic-demographic and psychological features set. Churn prediction allows administrators to design strategies to increase student retention. As early as enrolment, school administrators may assess and identify at-risk students and provide early intervention schemes through personal tutoring, academic counseling, and follow-ups. On the other hand, the students will be aware and are guided to achieve academic success.

2. Methodology

2.1 Dataset Description

A dataset of 1,338 students admitted as freshmen from 2005-2010 in Bachelor of Science in Information Technology (BSIT) and Bachelor of Science in Computer Systems (BSCS) programs of the University of San Jose-Recoletos, Cebu City, Philippines, with 778 churners and 560 non-churners was utilized.

The dataset was acquired as a comma-separated values (CSV) file, which is a dump file saved as a result of a query from the Electronic Data Processing (EDP) Center – the office that serves as this study's data source. The data integrity depended on the correctness of the query used to extract them from the data source.

The conversion of dataset into a representation suitable for the model took a considerable amount of time. Variable-by-variable data cleaning, a straightforward filter approach (Kotsiantis *et al*., 2006), was done to detect possible data quality problems such as duplicate instance identification, spelling inconsistencies and illegal values. Parents' educational attainment and occupation were discarded due to a large number of missing values.

New attributes were derived from existing attributes. Age upon enrolment was obtained by calculating the year difference of the term enrolled with the student's birth date. Global positioning system (GPS) coordinates in longitude and latitude of the students' hometown location were obtained through Google Distance Matrix API to compute its proximity from the school location. Numerical (continuous) features such as age, intelligence quotient (IQ), college grade average and distance were discretized. Discretization is the process of putting values into buckets so that there are a limited number of possible states. Table 1 shows the academic and demographic features considered in the study.

Table 1. Demographic and academic data set attributes

Attribute	Description
stud_no	Uniquely identifies the student
age_upon_enrolment	Age of the student at the time of enrolment
gender	Either male or female
religion	Religion
marital_status	Marital status
proximity	Hometown location: to determine the proximity of hometown from school
highschool_type	Private or public high school
average_grade	Accumulated average grade

The features with no predictive information were eliminated using a variable importance estimate derived by building a learning vector quantization (LVQ) model. This is done to select the most relevant input variables. Top 3 predictors identified by LVQ included average_grade, highschool_type and terms_enrolled (Table 2). After data cleaning, only 1100 records remain usable for further processing.

Table 2. Attribute importance

Feature	Importance
average_grade	0.8089
highschool_type	0.6455
terms_enrolled	0.6154
age_upon_enrolment	0.5806
proximity	0.5344
gender	0.5254
religion	0.5075
marital_status	0.5023

On the other hand, the psychological data representing the personality traits of 560 students were gathered through Manchester Personality Questionnaire (MPQ). The MPQ factor version 14 is a 90-item standardized questionnaire where items are short statements that require a response to describe the way the respondents tend to think, feel and act (Tandoc and Tandoc-Juan, 2014). This examination was taken by students in their second year of college through the Student Development and Placement Center (SDPC). The factors measured by the full-length test included personality traits in the primary dimension and the significant five factors. The same questionnaire was also used by Lopez and Santelices (2012) to investigate the personality characteristics of table tennis athletes that contributed to their athletic performance. Table 3 shows personality traits in the primary dimension with its corresponding descriptors as measured by the MPQ.

Table 3. Primary dimension factors (Center for Creative Leadership, 2000)

Factors	Low score description	High score description
Originality	Implements ideas and changes	Originates action, invents
Rule Consciousness	Keeps to rules, reduces risk	Challenges assumptions, take a risk
Openness to Change	Practical, grounded, task-oriented	Imaginative, change-oriented, experimental
Assertiveness	Holds back ideas, gives way to others	Assertive, persuasive, convincing
Social Confidence	Less comfortable in a social situation	Fits in quickly interacts
Empathy	Individualistic, self-reliant	Supportive, sensitive, considerate
Communicativeness	Reserved, quiet, distant from people	Communicative, open, expressive
Independence	Sociable, group-oriented	Self-contained, works well alone
Rationality	Intuitive, spontaneous	Logical, reflective, systematic
Competitiveness	Accommodating, less committed to a career	Committed to career, contesting
Conscientiousness	Radical, challenging, expedient	Conscientious, preserving
Perfectionism	Less methodical, less detail-oriented	Quality-driven, detail-oriented, methodical
Decisiveness	Cautious, slower to take initiative	Decisive, controlling
Apprehension	Calm, relaxed	Apprehensive, worried

Table 4 shows personality traits regarded as big five factors with their corresponding descriptors as measured by the MPQ.

Table 4. Big five factors (Center for Creative Leadership, 2000)

Factors	Low score descriptor	High score descriptor
Creativity	Adaptive, pragmatic, implementation focused	Innovative, change-oriented, non-conformist
Agreeableness	Individualistic, self-reliant, independent	Participative, rational, team player
Achievement	Accommodating, expedient	Quality-driven, achieving, conscientious
Extroversion	Reserved, socially inhibited, introvert	Communicative, outgoing, extrovert
Resilience	Apprehensive, worried, anxious	Calm, stable, decisive

Feature importance estimate was derived through recursive feature elimination (RFE) to reduce feature dimensionality. Top seven predictors included independence, extroversion, assertiveness, resilience, creativity, conscientiousness and perfectionism. Table 5 shows the values of the mean attributes of the top predictors of students measured in this questionnaire.

Table 5. Personality profile of students on top 7 predictors

Personality Trait	Churn = Yes	Churn = No
Independence	5.852941	6.167702
Extroversion	5.445378	5.515528
Assertiveness	4.647059	4.664596
Resilience	4.294118	4.133540
Creativity	4.617647	4.822981
Conscientiousness	6.785714	6.975155
Perfectionism	4.756303	4.993789

Non-churners showed higher mean scores in the top seven predictors' personality traits except for resilience. The two-tailed T-*t*est was used to determine if there is a significant difference in the personality traits of churners and non-churners. The result indicated evidence that churners and non-churners significantly differed in personality traits such as independence, creativity, rationality and originality (Table 6).

Table 6. Significant personality traits

Personality Trait	Churn = Yes	Churn = No	P-value
Independence	5.852941	6.167702	0.02418
Creativity	4.617647	4.822981	0.03074
Rationality	4.995798	5.298137	0.04507
Originality	4.899160	5.096273	0.06599

Level of significance $\alpha = 0.10$

Taking into consideration the top predictors and the significant personality traits, the features selected for evaluation and model building encompassed rationality, originality, extroversion, independence, resilience, creativity and assertiveness.

2.2 Ensemble Model Building

Empirical evaluation of classification models on psychological data and academic-demographic features was conducted to ascertain which yields better predictive performance. The 10-fold cross-validation on the selected features (Table 5) was used to determine the accuracy of each model. This consists of splitting the original dataset into 10 complementary subsets of size n/10. The classifier is fed with the nine subsets for training, and the remaining subset is used for validation testing. This process is repeated for ten rounds. Random forest (RF), k-Nearest Neighbor (k-NN), and support vector machine (SVM) showed acceptable predictive performance on separate tests on psychological and academic-demographic features, as shown in Table 7.

Table 7. Accuracy of experimental models

Algorithm	Accuracy	
	Psychological	Academic-demographic
RF	62%	83%
k-NN	57%	83%
SVM	56%	82%

A predictive ensemble model was developed from three identified experimental models on two datasets through plurality voting (Figure 1), a method used to increase classification accuracy by training a set of classifiers and aggregating their output by voting (Leon *et al.*, 2017).

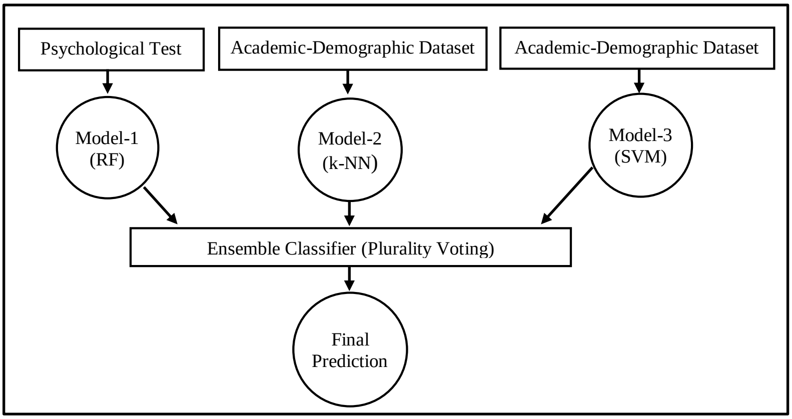

Figure 1. Ensemble predictive model

The base classification algorithms are RF, k-NN, and SVM. Based on the result of the experimental models, RF returned the highest accuracy when trained with the psychological dataset. The academic-demographic training dataset showed better predicting performance in all three algorithms. Thus, two of the models, k-NN and SVM, were trained with it.

2.3 Parameter Settings

The base algorithm for model-1 is RF. An RF is a multitude of tree-based classifiers which were grown from a random subset of the MPQ factor scales sampled independently; each tree casts a vote for the most popular class (Breiman, 2001). Model-1 used 439 samples with seven predictors and two classes ('NO,' 'YES'). The value of *mtry*, the number of variables for splitting at each tree node, was selected using the largest accuracy value in the resampling results after five-fold cross-validation (Table 8). The final value used for the model was 7.

Table 8. Resampling results across tuning parameters

mtry	Accuracy	Kappa
2	0.5648642	0.1216168
4	0.5648119	0.1206069
7	0.5808516	0.1550451

The base algorithm for model-2 is k-Nearest Neighbor. K-NN is an instance-based learning (IBL) algorithm that identifies *k* student instances with similar churn behavior based on the academic-demographic data and uses the popular class to predict the class of the instance to be predicted (Hardoon and Shmueli, 2015). The algorithm utilized Euclidean distance as similarity metric, which is computed by summing up the squared difference between the pair of attribute values. Model-2 used 979 samples with eight predictors and two classes ('NO,' 'YES'). Accuracy was employed to select the optimal model using the largest value after five-fold cross-validation. The final value used for the model was $k = 5$.

Table 9. Resampling results across tuning parameters

k	Accuracy	Kappa
5	0.8293964	0.6590938
7	0.8181509	0.6367633
9	0.8140745	0.6287139

On the other hand, SVM separates instances into churners and non-churners by looking for the optimal hyperplane maximizing the margin on student academic-demographic features closest point. The radial kernel is used for model-3 since the data is not linearly separable. Figure 2 shows R Script of SVM model creation with the following parameter settings: SVM-type – C-classification; SVM-kernel – radial; and cost – 1.

```
svm(formula = CHURN ~ .,data = train,type = 'C-classification', kernel = 'radial')
```

Figure 2. R Script for SVM model creation

2.4 Plurality Voting

Prediction was made through plurality voting, a method used to increase the accuracy of classification by training a set of classifiers and to aggregate their output by voting (Leon *et al.*, 2017). Each model returns individual prediction (CHURN, NOT CHURN) on the test dataset. Votes were tallied, and the majority vote was considered the ensemble prediction. Figure 3 shows the R script to perform plurality voting.

```
pred <- list()
pred$knn<-predict(object = en_model$knn,a_test[,-1],type='raw')
pred$rf<-predict(object = en_model$rf,p_test[,-1],type='raw')
pred$svm = predict(en_model$svm, newdata = a_test[,-1])

votes <- data.frame(pred$rf, pred$knn, pred$svm)
votes$YES <- rowSums(votes=='YES')
votes$NO <- rowSums(votes=='NO')

pred$maj_vote<-as.factor(ifelse(votes$YES>votes$NO,'YES','NO'))
pred$accuracy = Accuracy(p_test$CHURN,pred$maj_vote)
pred$precision = Precision(p_test$CHURN,pred$maj_vote)
pred$recall = Recall(p_test$CHURN,pred$maj_vote)
pred$f1 =  2*(pred$recall * pred$precision) / (pred$recall + pred$precision)
```

Figure 3. R Script for plurality voting

3. Results and Discussion

Considering that the two datasets come from different unrelated data providers, there were student records with academic-demographic attributes but no personality test results and vice versa – instances of academic-demographic and psychological features were aggregated. An inner join of the datasets was performed to determine the common instances based on student numbers. One hundred twenty-one student instances with complete attributes were extracted to test the ensemble model. The test set was excluded from the training of the models to prevent bias in the result. k-NN and SVM were trained using 979 academic-demographic instances, while 439 instances of student psychological attributes for RF.

Table 10. Accuracy of individual models in the ensemble

Algorithm	Dataset	Accuracy
SVM	Academic-demographic	78%
k-NN	Academic-demographic	77%
RF	Psychological	56%

The performance of individual models in the ensemble was evaluated to determine if there is a need to allocate more votes to one or more models. Models vote on each instance in the test set whether the student is likely to churn. As shown in Table 10, SVM and k-NN yield better accuracy compared with RF. Academic-demographic attributes provide better predictive performance compared to psychological attributes. All three models are allocated one vote each. Figure 4 shows the performance of the ensemble model. The confusion matrix presents the number of correct and incorrect predictions in each category. The overall accuracy of the ensemble model was at 78%. The model was able to predict better on students who are likely to retain (Churn = No) with an F1-score of 84% than those who are likely to stop or churn (Churn = Yes) with an F1-score of 64%.

```
Confusion Matrix and Statistics

          Reference
Prediction NO YES
       NO  70  17
       YES 10  24

               Accuracy : 0.7769
                 95% CI : (0.6922, 0.8475)
    No Information Rate : 0.6612
    P-Value [Acc > NIR] : 0.003799

                  Kappa : 0.4804

 Mcnemar's Test P-Value : 0.248213

              Precision : 0.8046
                 Recall : 0.8750
                     F1 : 0.8383
             Prevalence : 0.6612
         Detection Rate : 0.5785
   Detection Prevalence : 0.7190
      Balanced Accuracy : 0.7302

       'Positive' Class : NO
```

Figure 4. Confusion matrix and statistics

4. Conclusion

The study aimed to employ data mining techniques to build a predictive model to identify student academic churn risk. While there are other qualitative factors affecting students' decision to abandon schooling, the academic, demographic and personality attributes play an important role in determining the probability of academic churn. The proposed ensemble classifier obtained an overall accuracy of 78%, which could be very helpful in identifying at-risk students for targeted intervention schemes. Early intervention of at-risk students may lead to better student retention.

5. References

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-16.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Center for Creative Leadership. (2000). Leadership resources: A guide to training and development tools (8th ed.). Greensboro, NC: Center for Creative Leadership.

Chai, K., & Gibson, D. (2015). Predicting the risk of attrition for undergraduate students with time based modelling. Proceedings of the International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA), Dublin, Ireland, 109-116.

Figini, S., De Quarti, E., & Giudici, P. (2009). Churn risk mitigation models for student behavior. Electronic Journal of Applied Statistical Analysis, 2(1), 37-57.

Hardoon, D.R., & Shmueli, G. (2015). Getting started with business analytics: Insightful decision-making. Florida, USA: CRC Press.

Klepac, G., Kopal, R., & Mršić, L. (2015). Developing churn models using data mining techniques and social network analysis. Pennsylvania, USA: IGI Global.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), 4104-4109. https://doi.org/10.5281/zenodo.1082415

Leon, F., Floria, S.-A., & Badica, C. (2017). Evaluating the effect of voting methods on ensemble-based classification. Proceedings of the 2017 IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA), Gdynia, Poland.

Lopez, A., & Santelices, O. (2012). Personality characteristics of elite table tennis athletes of the Philippines: Basis for a proposed recruitment program. International Journal of Table Tennis Sciences, 7, 1-4.

Luna, J. (2000). Predicting student retention and academic success at New Mexico Tech. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.4720&rep=rep1&type=pdf

McRoberts, T., & Miller, T. (2015). Instrument development for examining student attrition. Journal of Academic Administration in Higher Education, 11(2), 43-53.

Raju, D. (2012). Predicting student graduation in higher education using data mining models: A comparison (Dissertation). University of Alabama, Alabama, USA.

Ramist, L. (1981). College student attrition and retention. Retrieved from https://eric.ed.gov/?id=ED200170

Sharma, A., & Panigrahi, P. (2011). A neural network-based approach for predicting customer churn in cellular network services. International Journal of Computer Applications, 27(11), 26-31

Tandoc, J., Jr., & Tandoc-Juan, M. (2014). Students' personality traits and language learning strategies in English. International Refereed Research Journal, 5(3), 1-10.

Wei, C.P., & Chiu, I.T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. Data Expert Systems with Applications, 23(2), 103-112. https://doi.org/10.1016/S0957-4174(02)00030-1