Predicting Student Performance to Forecast University Enrollment

Fernando A. Zambrano, M.S.

Nima Zahadat, Ph.D.

The George Washington University

ABSTRACT

Not all universities have dedicated registrar departments, nor access to specific software to keep track of student enrollment to forecast future enrollment. The ability to develop reliable student forecast brings multiple benefits in regard to both university finances and student investment. This study explores prior student data from the Institution Tecnologico de Chihuahua (ITCH), with the intent to develop such forecast models. This is achieved through an ensemble of different predictive models such as a Support Vector Machine (SVM) binary classification, ordinary least squares linear regression (OLS), and Holt-Winter Exponential Smoothing.

Keywords: data mining, time-series, linear regression, binary classification, Holt-Winters, Random Forest Regression, Support Vector Machine Classifier

INTRODUCTION

Universities benefit knowing their future revenue and operating costs, as it allows for reliable budget projections and resource allocation as tuition payments make up a significant portion of revenue (Salley, 2016). This is the current challenge that Institutio Tecnologico de Chihuahua (ITCH) faces. ITCH is the primary engineering and business institution in the state of Chihuahua, Mexico. However, it lacks any software or dedicated registration department that can help calculate the number of sections necessary for each course for the following semester. Their primary focus is on the essential math classes that all students, regardless of their major, are required to take, which are Differential Calculus (DifCal), Integral Calculus (IntCal), Vectoral Calculus (VecCal), and Differential Equations (DifEqs). The number of students who have to take DifCal is affected by the number of new students, plus those who have to retake it if they failed the previous semester. Likewise, the number of students who have to take IntCal are equally affected by the number of students who failed DifCal. The reason this task is important is because the university needs to know how many instructors are needed to maintain an adequate and cost-effective student to teacher ratio for the required number of sections each semester.

Since ITCH is a public university, the federal government of Mexico only pays for the electrical bills and salaries of the faculty. However, the number of faculty positions are limited and insufficient to support the student body. In order to accommodate for the lack of faculty, the university contracts temporary and adjunct professors each semester for single semester contracts. These contractors are paid directly from the university's budget which is funded solely from student tuition. Yet, from this budget the university also has to pay for the upkeep of its own facilities, technology maintenance, and various other costs. Therefore, ITCH needs a better way to predict how many students will fail the essential math courses each semester to determine the number of class sections required for the following semester, which will them allow them to determine how many contracted instructors will be needed. This will allow them to ensure that there are enough instructors the following semester to avoid students having to wait another semester to take the course, which adds extra costs to the budget for the proceeding semester. Likewise, the university wants to avoid paying for more instructors than are needed to reduce its out of pocket expenses that could instead be used to pay for maintenance, new equipment, or in reserve for more instructors.

ITCH would significantly benefit from an in-depth exploratory analysis to develop a predictive model that would estimate the number of students expected to enroll in each math course the following semester. Not only would this help the financial health of the institution, but it also allows the performance of students and instructors to be tracked.

LITERATURE REVIEW

Predicting student performance is widely popular amongst academics and researchers, and many have studied the issue for various applications. One prevalent reason is to discover high-risk and slow learning students. For example, several classifiers such as random forests, decision trees, neural networks (NN) and SVMs for binary classification to predict student performance on Portuguese and Mathematics courses of secondary students to asses high risk students in Portugal. The data consisted of several academic variables such as past grades and failures. However, the bulk of the data features were behavioral and demographic. These included if the student paid for extra classes, attendance, eager to attend university, and if they consumed alcohol or not. Furthermore, the demographic data included sex, age, family size, and guardian's level of education. The study achieved high accuracy rates of 94% using tree-based algorithms (Cortez & Silva, 2008).

A similar study was also undertaken to discover slow learners in secondary school using classification techniques. The primary classifiers included Multilayer Perceptron (MLP), Naïve Bayes, and tree-based classifiers such as Decision trees and Random Forests. The predictive features included more descriptive variables rather than behavioral, as seen in Cortez & Silva. These features included geographic location of urban vs rural, access to internet via computer and mobile phone, and whether education was at a public or private institution. With such features a 75% accuracy score was attained using MLP with an F-score of 82% (Kaur, Singh, & Josan, 2015). Instead of trying to predict overall course performance, classification algorithms such as Decision Trees, K-Nearest neighbors, and Naïve Bayes classifiers can be used to predict student performance on final exams in an effort to find weak academic students. The predictive variables were a mix of both academic performance and demographic data in regard to living location and family education (Yadav & Pal, 2012). One predictive performance study aimed at using various student in-class participation and online participation features to determine performance.

Furthermore, the study aimed to test the accuracy scores between traditional classifiers such as MLP, Logistic Regression, and Random Forest vs clustering algorithms. The study demonstrated that clustering algorithms can in fact compete with traditional classifiers (Lopez, Luna, Romero, Ventura, 2013). Beyond supervised classification, there are cases of using unsupervised binary classification models such as k-means to predict student performances. Similar to other studies, the study focuses on using behavioral and family demographic features (Bhardwaj, Pal, 2011).

There are some challenges that come with educational performance classification. There is a lot of variance in the educational system that can distort data interpretation, such as changes in course offerings across years and semesters. Furthermore, there are variances in academic evaluations and exposures across times such as topics, exercises, and resources. These issues can become more challenging to overcome as data is scaled up, which is a current problem given the trend in data collection and analysis (Merceron & Yacef, 2008).

In the realm of forecasting student enrollment there are several methods that can be used depending on the data being analyzed. Discussed in Lavilles and Arcilla (2012), universities use a wide variety of forecasting methods, such as Naïve forecasting, Markov chain, regression models, and either single or double exponential smoothing. Moreover, the system at the University of Hawaii outlined that it makes individual forecasts for different classes of its student populations. Students are classified as either first-time, transfer, returning, or continuing. The key prediction feature uses previous continuation rates with trends taken into account. Similarly, the University of California planned its 10-year forecast by creating four different forecast models for first-time students, transfer students, projected continuing students, and overall student enrollment.

RESEARCH METHODOLOGY

The research methodology follows a standard process, in which the first step is to preprocess and clean the original raw data from ITCH so that it can be prepared for machine learning algorithms and understood during Exploratory Data Analysis (EDA). Python 3.7 and popular packages such as pandas and numpy were used throughout the process. The original data was spread across 16 Microsoft Excel sheets, in which many contained duplicate records or records from courses and majors outside the scope of study. Once the non-essential records were filtered out, the data features needed to be translated form Spanish into English where necessary, such as

course name, for example. The final step in this process was to recode some features from text into numerical values and fill in any null values with a zero. The preprocessed data was then exported to a single CSV file. With clean data, EDA is used to inspect the data and underlying relationships between the independent variables and the target variable (pass or fail). Afterwards, the most important variables in the dataset needed to be determined. Not using the right ones, or using all of them, can lead to overfitting (Brieman, 2001). Once the right features had been selected, they are applied to multiple classification models. The models were compared and evaluated on several metrics to determine the best one for this study. Next, the number of students that were eligible to take the classes the following semester was predicted based on the classification results using ordinary least squares (OLS) package form the Stats.Model API. This process was necessary, since there are policies that ITCH has for its students that are outside the scope of this study that leak in.

One such policy the university has, is that if a student fails any class, that student must take that same class again the following semester and is restricted to the number of classes they can take. Therefore, if students fail courses not included within the scope of this study, but pass one of the four courses, the data for those students will not be accounted for in the following semester. This forced the creation of the "dropout" feature, which captures students who passed one the four math courses in one semester but did not appear in the following semester. There are other facets which this feature captures such as urgent leave of absence and actual university dropouts. The inverse of the dropout feature then left the "retention" feature. These are the students that, regardless of whether they passed or failed, continue their enrollment at the university for the following semester. The last feature engineered is student "re-admission." These are students that have been enrolled before but took a leave of absence or dropped out the semester prior. Additionally, new students enter each semester and fall under the category of re-admission. These students need to be considered in order to fully complete the student enrollment forecast. The final component is to then forecast how many new students are expected to enroll the following semester as well as returning students from previous semesters using Holt-Winters Exponential Smoothing, also found within Stats.Model API. The results from the three models are aggregated to form the final forecast of next semester students. Model results and research methods are then reviewed, and further recommendations are made in the conclusion.

DATA

The historical data from ITCH was given directly by the university and not publicly available. The original dataset included 14 features and totaled over 140,000 records over a span of 10 years from 2010 – 2019. Within each year the data is classified into either spring, summer, or fall semesters. After preprocessing, the remaining dataset had 17 features and 14,949 records. This data only contained information relevant to the scope of the study which are the required math courses, Differential Calculus, Integral Calculus, Vectoral Calculus, and Differential Equations. Each record represents a student taking a particular course and can be identified by the 'studentID' field. The rest of the features describe that specific student, including their major, how many times they have repeated the course, their instructor, assigned section, their numerical grade, and if they passed or failed the course (grades below 70 are considered failing). Furthermore, students can only fail a course twice before they are ultimately suspended from the university. A more detailed description of the variables can be found in the Appendix section of this report. Of the 17 resulting features, the majority are categorical or nominal, and of the few numeric features that exist most are tied to the target variable itself and cannot be used for regression or classification. The final dataset was skewed in regard to the years in which data was available. The majority of the data is from after 2014, with the prior semesters only having a handful of samples and only for some the courses. These samples were kept for classification purposes, but not utilized for forecasting.

DATA ANALYSIS

The first part of the data analysis focused on EDA. In this part the data is highly inspected looking for patterns and trends within each feature and among the data as a whole. The tools used to inspect and visualize the data were bar charts to show differences in frequencies, line graphs to demonstrate change over time, violin plots, similar to box plots, to show distributions between features in relation to another, and correlation matrices to illustrate feature dependency. Python packages using matplotlib and seaborn provided the necessary tools to visualize the data. Four methods were implemented for feature selection. From the sklearn library, The Chi-Square test, F-test, mutual-information classifier, and Random Forest Regressor were used to select the best features. This created four extra datasets to train with the different classification models. Six classification models were used, Logistic Regression, Ridge Classification, Random Forest

Classifier, Decision Tree Classifier, Support Vector Machine (SVC), and Gaussian Naïve Bayes (NGB). This allowed for 24 different models to be compared and evaluated. Before the data is piped into the classifiers, it was split into training and test sets. The test set consisted of one semester. The models were first implemented with default settings from sklearn to serve as a baseline. Afterwards, the hyper-parameters were tuned with a grid-search with cross-validation to optimize each classifier for prediction. The optimized models were compared among themselves and cross referenced with the baseline models, then evaluated on several measurements, including accuracy score, precision, recall, specificity, F-1 score, Cohen-Kappa (kappa), and Area Under the Receiver Operator Curve (AUC).

The results from the best model were then fed into an evaluation function to verify how well the model performed to predict pass and fail for each major and math course. The classification predictions were then used to determine the retention rate for passes and fails for each math course using the OLS linear regression model. The expected number of new students, re-admission, and summer fails (since the test set is a spring semester) were predicted using the Holt-Winters seasonal exponential smoothing method. These features were predicted separately from OLS since they are independent from the classification results. The predicted retention rates and forecasts from the Holt-Winters model were then aggregated to provide the final prediction for each math course for the upcoming semester.

KEY FINDINGS

The EDA illustrated how the data could be used for modeling and which features may work well. These patterns and distributions were highlighted primarily with bar charts, violin plots, and categorical plots or, catplots. Basic descriptive statistics show that there are more samples of students passing than failing. As seen in (Figure 1), the overall passing rate (total passes/ total samples) for the university is 67.16% with an average grade of 77.37 with failure set at 69. Moreover, majors and courses are not evenly distributed in terms of enrollment.

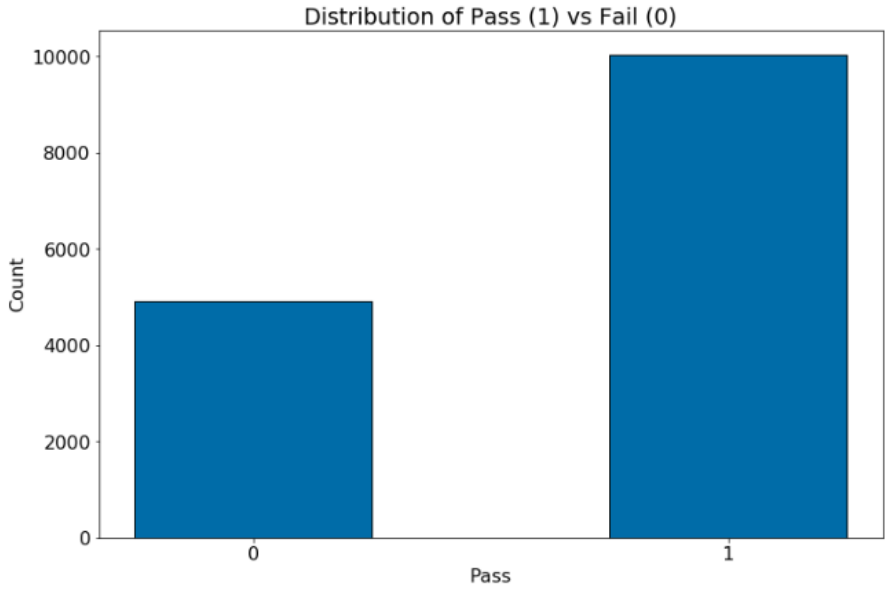

Figure 1

Features that clearly show separability are easier to classify, however this dataset lacked those kinds of variables. Since pass and fail are not continous variables, the students numerical grade was used instead for visual purposes. The best features that showed some form of separability were 'RepEn', the number of times a course was repeated, instructor, and the section the student was assigned. However, as the catplot shows below (Figure 2), this separability is seen in students who passed and not between students who failed. In the case of the assigned section, some groups only have students from certain courses. Feature weakness is further supported by the correlation plots (Figure 3) that show that there is a weak relationship between the predicting features and passing a course. The highest correlated features reached (±) 0.15 with a mean of 0.01.

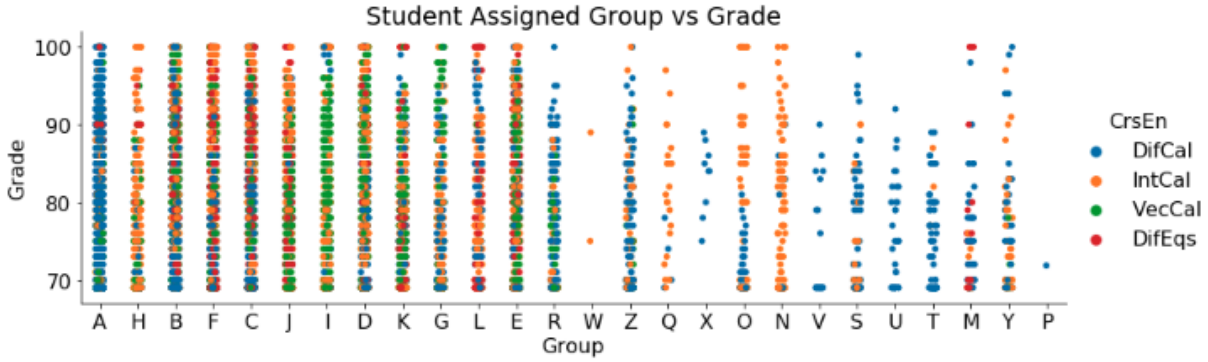

Figure 2

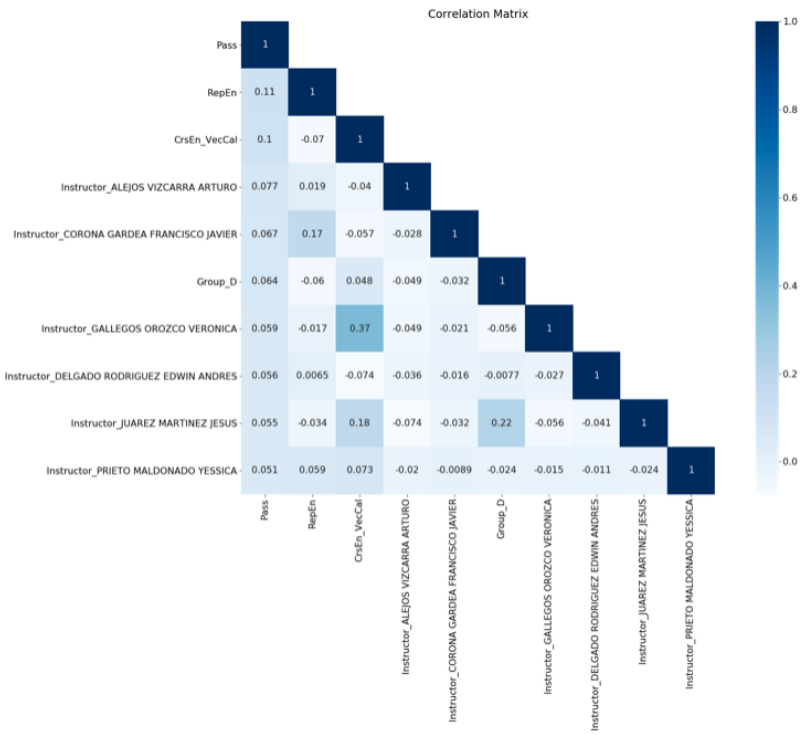

Figure 3

Aside from showing what features may work well for classificaiton, the EDA demonstrated the distribution of key features such as student enrollement in each course, major, and semester. The choice of visualization was the violin plot, as it captured the median grade, distribution, inner quartile range (IQR), as well as the frequency for each category. (Figure 4) shows the violin plots of grade distribution by course and (Figure 5) by major. These plots demonstrate that there are subtle differences in student performance among most majors. However, Industrial (IND) and Industrial Online (EAD) both stand out from the rest, as they have clearly different distributions, with a higher frequency around the fail threshold. In regard to the courses, differential calculus, or 'DifCal', has the most obvious difference in distribution, with a much higher frequency around failure. Furthermore, the median increased with each higher level course with differential calculus close to 70 and differential equations closer to 80.

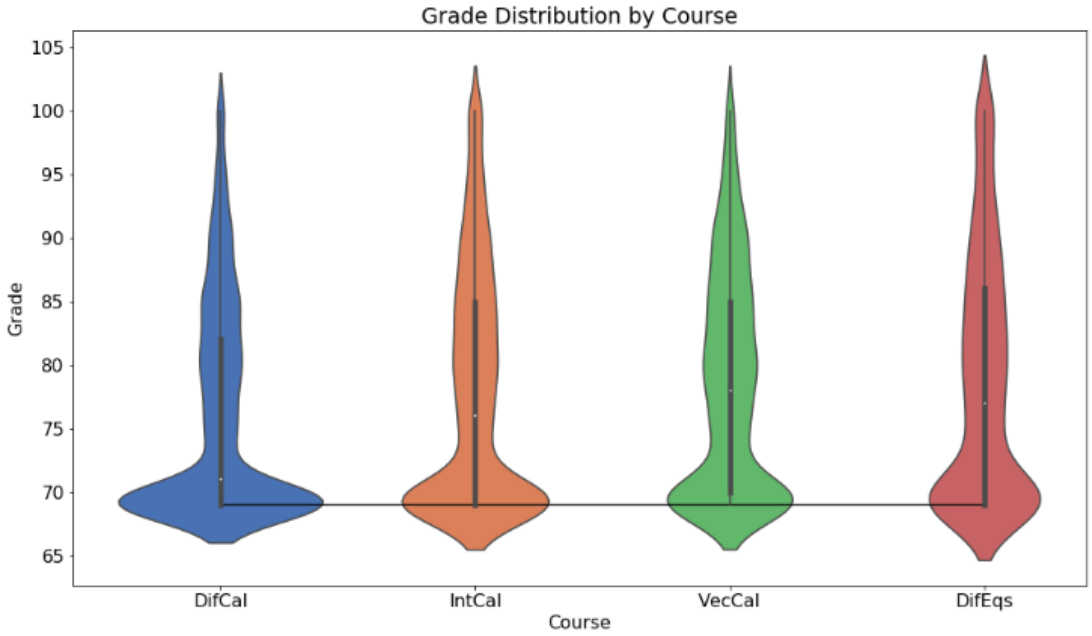

Figure 4

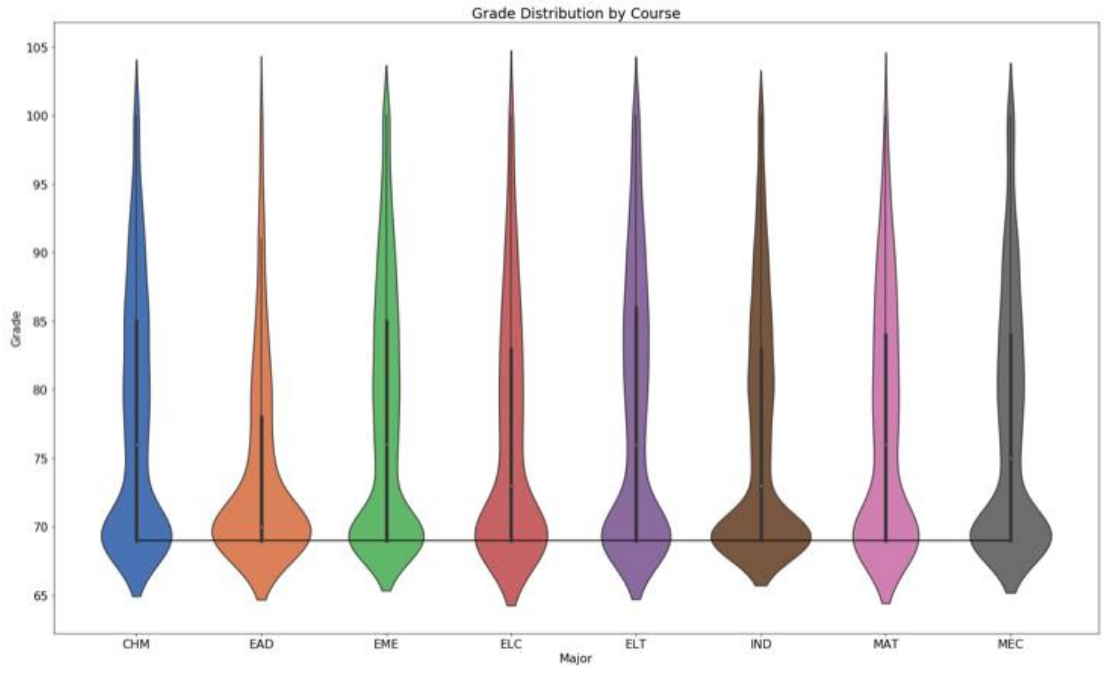

Figure 5

Despite showing which features may work well, the EDA also demonstrated that the classification would not be easy, nor to expect exceptional results beyond a baseline test. The use of bar charts demonstrated signs of seasonality and trends in the overall number of students and courses at ITCH. There are more students in semester three (Fall) than in one. Semester two is the

summer semester with has significantly less student and no new students enrolled. Understanding if there are trends and signs of seasonality in enrollment is key for time-series forecasting.

Since most of the data was categorical or nominal, those features were one-hot encoded in order to gain quantitative metrics. This changed the original dataset of 17 features to 91, after dropping features that were tied to passing or irrelevant such as “Grade”, “studentID”, and “Period.” This made it so that each class within each categorical feature became its own feature. Of the four feature selection methods, the top 30 features selected by the Random Forest Regressor had the best classification results. The top feature, based on Gini feature importance (in which higher more important variables are given higher numbers), was course repetition, ‘RepEn’, at 0.07. This was followed by the semester in which students enrolled, ‘Sem’, at 0.06. The rest of the features were a mixture of major, course, assigned section, and instructor. Overall, the average feature importance was 0.026. (Figure 6) shows the top five features.

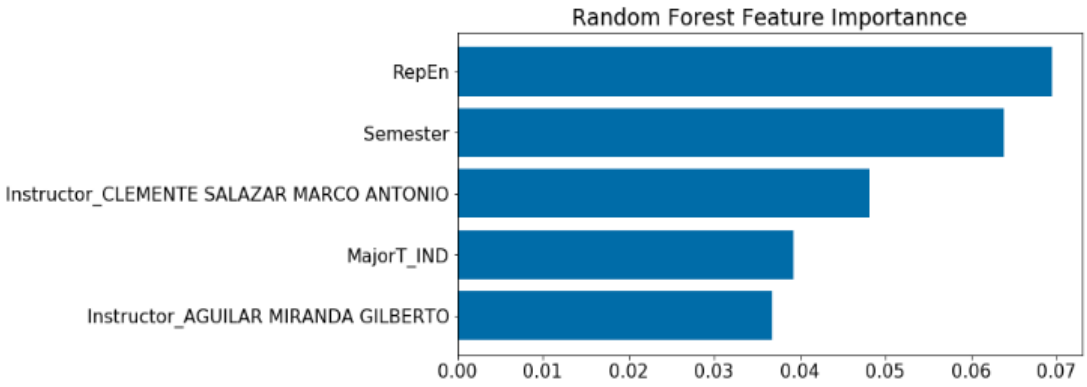

Figure 6

Before the data was fed into the classification algorithm it needed to be split into a training and test set. The training set included 13,051 samples which were all data points, except those from the first semester of 2019, which served as the test set with 1,898 samples. As found in the EDA, the target features are highly imbalanced with there being more than twice as many passes than fails. This can easily throw off the classifier during training to only predict the mean of the target feature, which then leads to biased testing scores. To overcome Target imbalance, a random over-sampler from the imblearn library was used to increase the number of samples of the minority target from, 4,096 fails, to match the same number of passes, almost 8,955. This resulted in the models being trained on almost 18,000 samples.

As expected, the optimized classification models outperformed their baseline results. Of the six different classifiers, SVC outperformed the rest in terms of accuracy, kappa, and AUC (Figure 7). Ridge classifier had the best performance in regard to recall, 84.58%, and F1-score 72.69% (Figure 8). GNB had the highest scores for specificity, 82.57%, and precision, 74.68% (Figure 9). Since future enrollment predictions depend equally on the number of students that pass and fail, the best classifier for this study needs to be able to correctly classify true positives (passes) and true negatives (fails) equally. The ridge classifier will predict true positives well but has an extremely high false positive rate, because it does not classify samples as negatives. On the other hand, the GNB will classify true failures well, but at the expense of a high false failure rate. Hence the SVC is the most suitable classifier for this scenario. SVC scored 69.75% on accuracy, 70.45% on precision, 81.20 % on recall, 54.4% specificity, 75.43% F-1 score, 36.52% Kappa, and 67.84 AUC. Therefore, SVC was the best classifier for this situation as it was consistently either the best or runner-up in the evaluation metrics. Furthermore, its Cohen-Kappa score of 36.6%, is significantly higher than the rest which demonstrates that its classifications were fairly correct based upon a stronger model rather than random chance (Figure 10).

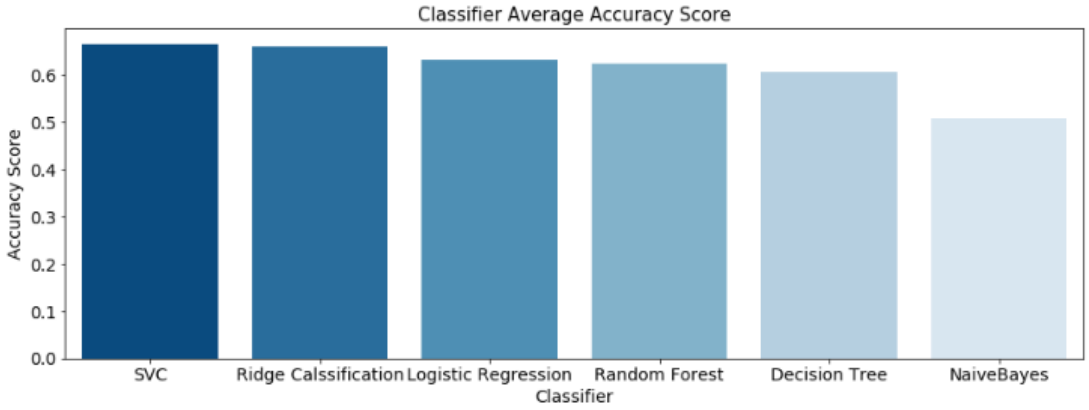

Figure 7

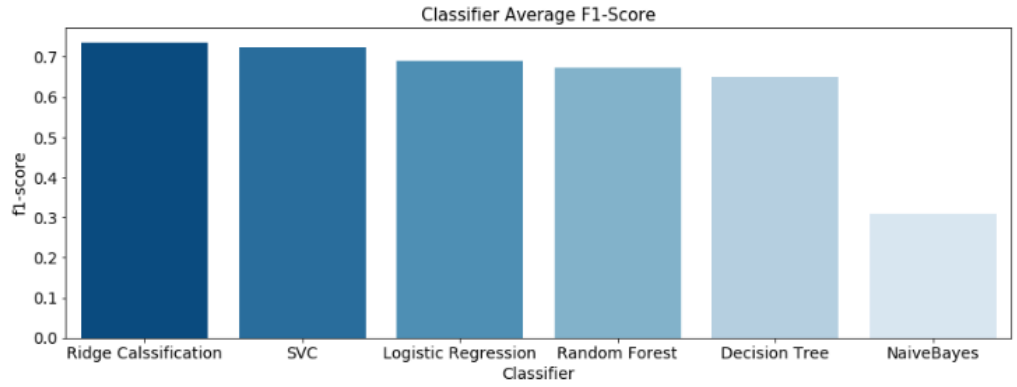

Figure 8

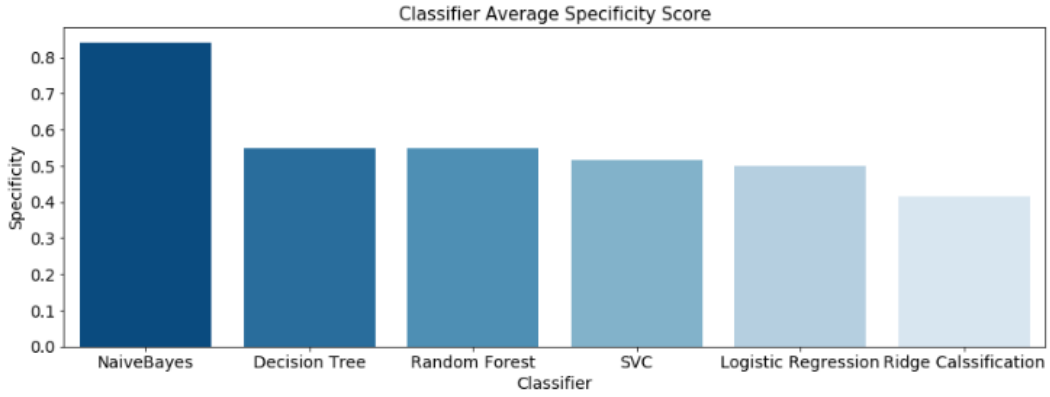

Figure 9

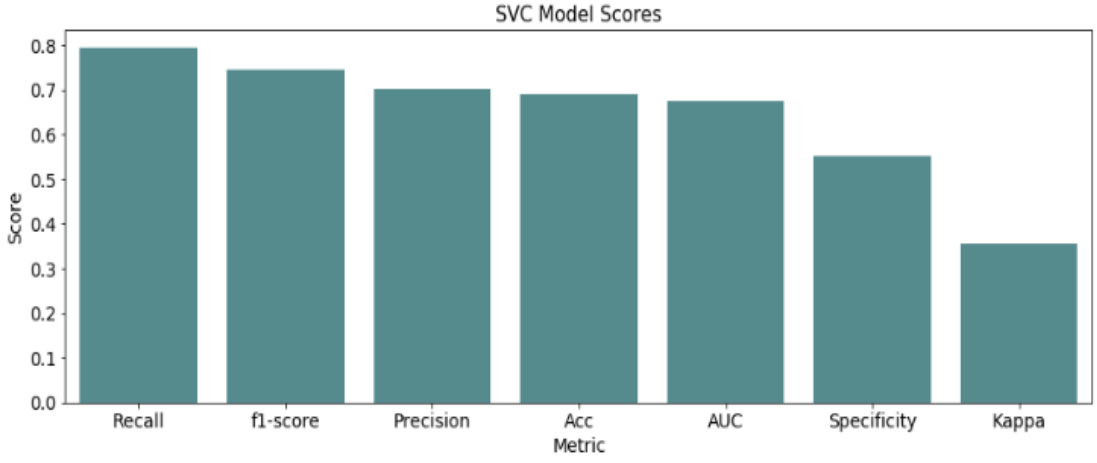

Figure 10

The results from the SVC estimated that 302 (39.58%) students will pass from differential calculus into integral calculus, while 423 (55.43%) students failed and would repeat the following semester. The remaining 38 (04.99%) students failed more than twice and were kick out, and 464 (83.91%) students will advance from integral calculus to vectoral calculus. Meanwhile, 89 (16.09%) students failed integral calculus. 266 (81.10%) students will advance from vectoral calculus into differential equations and 61 (18.60%) failed. 219 (86.22%) students passed out of differential equations and 35 (13.78%) failed. The overall passing rate for this semester (Fall of 2019) was 59%, however the classifier predicted 66%. Unfortunately, this is almost exactly the same passing rate as the overall dataset as discovered in the EDA. Hence, the model classified more false positives and overestimated the number of passes.

The pass and fail data from the first semester of 2013 to the first semester of 2018 were used to train OLS models using the Stats.Model library to predict the retention rate for the next semester for each math course. The t-test with a 95% confidence interval was used to verify that pass and fail as predictive features were significant.

The last components of the overall student enrollment forecast were the number of expected new students and expected re-enrollment. These features were predicted by using the Holt-Winters Exponential Smoothing method. Unlike the OLS model, Holt-Winters can forecast data that is both seasonal and has a trend. This makes the method even more useful as it can apply to seasonal characteristics of student enrollment seen in the EDA (Figure 11). The Holt-Winters method forecasts through a combination of four different equations: the forecast equation and three smoothing equations. The smoothing equations are for three components of the data, level(lt), trend(bt), and seasonality(st). Through optimization, in the Stats.Model package the parameters for each equation were optimized using a training set.

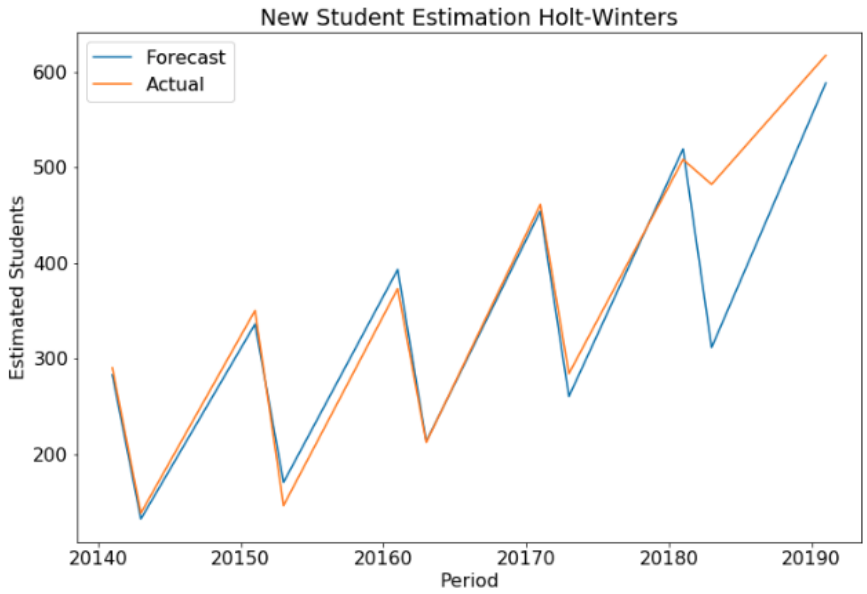

Figure 11

The overall forecast was partially accurate in terms of the number of students forecasted for the next semester, but with several caveats. The forecast was fairly accurate for differential calculus, integral calculus, and differential equations (Figure 12). However, this accuracy is the result of balancing out of overestimates and underestimates, rather than true accuracy. The majority of the error occurred in vectoral calculus forecast. However, this is not the true source of the error, but rather from the overt false positive misclassification of integral calculus since it is the primary component of the following semester's vectoral calculus enrollment. In regard to the results ensemble model of SVC pass classification + OLS retention prediction + Holt-Winters forecast; the model performed poorly. The classification was a result of over half the error in the model, followed by the OLS model with almost 45% of the total error. These two models accounted for hundreds of students worth of error. The Holt-Winters forecast did well by comparison, only accounting for 5% of the overall error, with only a few tens worth of student error (Figure 13).

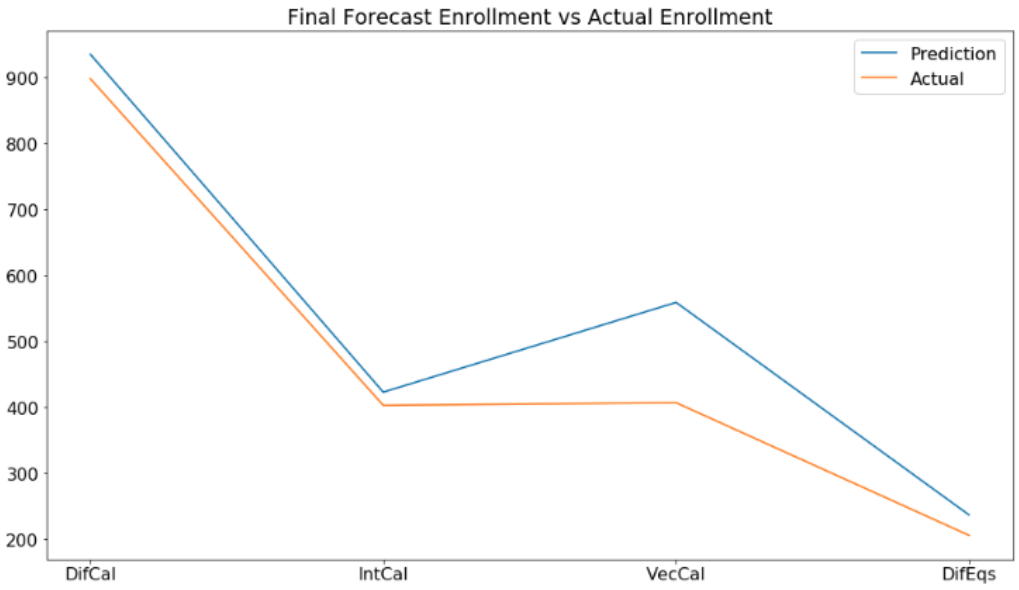

Figure 12

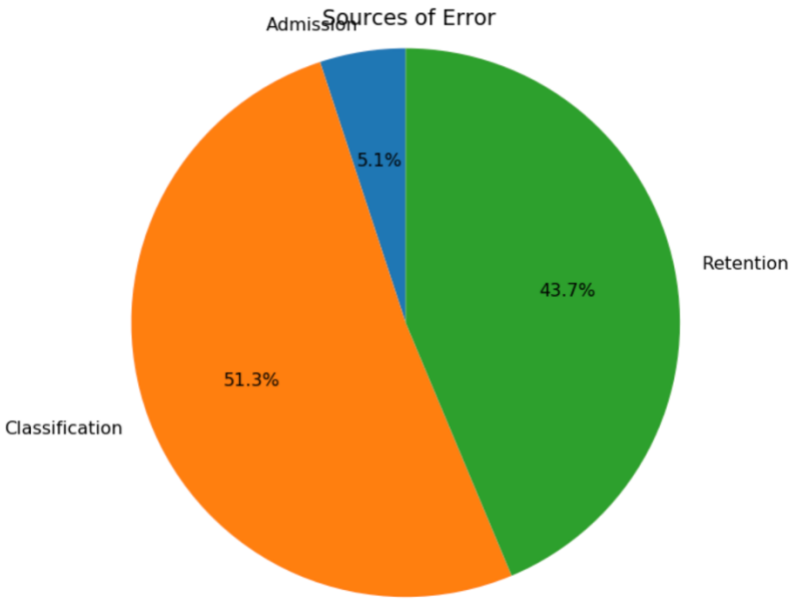

Figure 13

RECOMMENDATIONS & LIMITATIONS

The best recommendation for future analysis is to have a feature that better describes student behavior and qualities, rather than major, courses enrolled, and instructor. Likewise, the

lack of quality data greatly impacted the performance of classifiers, as there were not enough features to adequately differentiate between the two target classes. Furthermore, access to more quantitative and continuous data type features will help not only with better visualization in EDA, but also help with dimensionality reduction by avoiding the need to one-hot encode nominal and categorical data. Additionally, this study only looked at a fraction of the total available data. A more thorough analysis and data mining of the entire dataset could allow classifiers and predictive models to have better fitting results with more samples. Other alternative models and methods need to be explored, such as the use of deep learning neural networks (DNN) that can handle larger datasets and more complex relationships for classification purposes or recursive neural networks (RNN) in regard to time series forecasting. An additional change to research methodology is approaching this problem from a strictly forecasting perspective may offer better results by avoiding the barriers and errors observed with classification. Lastly, due to the complications of the world-wide pandemic at the time of this research, access to financial data regarding instructor contract and data of initial number of groups and student enrollment were not available.

CONCLUSION

Overall, this study fulfilled its goal to create a predictive model for ITCH. This was done through a three-pronged approach. The first component was an SVC binary classifier developed to predict which students would pass or fail each course based upon instructor, major, number of times the student has repeated the course, assigned group, and semester which the course was taken in. The most important based off of a Random Forest Regressor was the number of times the student has repeated a course (0.07), and the semester which the student was enrolled (0.06). The SVC was evaluated on several metrics which scored 69.75% on accuracy, 70.45% on precision, 81.20 % on recall, 54.4% specificity, 75.43% F-1 score, 36.52% Kappa, and 67.84 AUC. The next component was to predict the retention rate of students who passed and failed for the next semester as some student's dropout between semesters. This was accomplished using univariate OLS linear regression model. The final component was to predict the number of new students and returning students from prior semesters using Holt-Winters Exponential Smoothing. The resulting model was fairly accurate at predicting class enrollment for DifCal, IntCal, and DifEqs with a few tens of students of error. However, this was due to a cancelation between overestimates and underestimates of the models. The majority of the error was observed in VecCal, but the source of

the error is attributed to the high false positive rate of the classifier for IntCal and the retention prediction model. Even though the results are not ready for implementation, these results set a baseline for ITCH and the ability to explore other methods.

BIOGRAPHY

Fernando Zambrano is a graduate student in the Data Science Program at the George Washington University. He received his B.A degree from the Elliott School of International Affairs in International Affairs and Security Policy with a second major in Geography and Geographic Information Systems. His professional experiences range from internship on Capitol Hill, Embassy of Mexico, and trade associations to the Smithsonian National Ari & Space museum and bartending in Washington D.C. He is currently a weightlifting coach and employed by Apple Inc. Zambrano is travel enthusiast and foodie.

Nima Zahadat is a professor of information systems and computer science. Dr. Zahadat has taught at University Systems of Maryland and Virginia as well as the George Washington University in the fields of forensics, data science, information systems, web development, systems engineering, and security. He has an undergraduate degree in Mathematics from George Mason, a graduate degree in Information Systems from George Washington, and a Ph.D. in Systems Engineering and Engineering Management from George Washington. Zahadat's research interests are mobile security, information security, digital forensic, risk management, data mining, and information visualization. Dr. Zahadat enjoys biking, photography, travel, skiing, and writing.

REFRENCES

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), 3-17. doi: doi.org/10.5281/zenodo.3554657

Bhardwaj, B. K., & Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security, 9*(4), 136-140. arXiv: 1201.3418

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1-2), 245-271. doi: 10.1016/S0004-3702(97)00063-5

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. doi: 10.1023/A:1010933404324

Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *Proceedings from the 5th Annual Future Business Technology Conference.* Porto, Portugal, 5-12. ISBN: 978-9077381-39-7

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895-1924. doi: 10.1162/089976698300017197

Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*(1), 140-160. doi: 10.1016/j.compedu.2004.05.003

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182. doi: 10.1162/153244303322753616

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies, 13*, 61–72. doi:10.2478/cait-2013-0006

Kaur, P., Singh, M., & Singh Josan, G. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science, 57*, 500-5-8. doi: 10.1016/j.procs.2015.07.372

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence, 18*, 411-426. doi: 10.1080/08839510490442058

Lavilles, R. Q., & Arcilla, M. J. B. (2012). Enrollment Forecasting for School Management System. *International Journal of Modeling and Optimization, 2*(5), 563-566. doi: 10.7763/IJMO.2012.V2.183

López, M.I., Luna, J.., Romero, C., & Ventura, S., (2012). Classification via clustering for predicting final marks based on student participation in forums. *Proceedings from the 5th International Conference on Educational Data Mining*, Chania, Greece, 148–151. doi: **ERIC Number:** ED537221

Luan, J. (2002). Data Mining and Its Applications in Higher Education. *New Directions for Institutional Research, 113*, 17-36. doi: 10.1002/ir.35

Merceron, A., & Yacef, K. (2008). Interestingness Measures for Association Rules in Educational Data. *Proceedings from the 1st Interntional Conference on Educational Data Mining*, 1-10. Corpus ID: 6145782.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. *International Journal of Computer Science Issues, 7*(1), 10-18. ISSN: 1694-0784.

Salley, C. D. (1979). Short-Term Enrollment Forecasting for Accurate Budget Planning. *Journal of Higher Education, 50*(3), 323333. doi: 10.1080/00221546.1979.11779969

Yadav, S. K., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. *World of Computer Science and Information Technology Journal*, *2*(2), 51-56. ISSN: 2221-0741