International Journal of Institutional Research and Management
International Institute of Applied Informatics
2018, Vol.2, No.1, P.1 - 17

Modeling in R and Weka for Course Enrollment Prediction

Amanda Watkins[*], Adam Kaplan[†]

Abstract

Predicting course enrollment is a common university resource planning problem. California State University Northridge (CSUN) faces many unique challenges when predicting student enrollment in its undergraduate Computer Science (CS) and Computer Information Technology (CIT) courses. In this paper, we discuss the design of an enrollment prediction tool which applies three Time Series models using R and four Time Series models using Weka utilizing a database of 19 semesters of enrollment data. The seven different models are tested against varying amounts of holdout data to see which can best predict enrollment for undergraduate CS and CIT courses to within one standard class size of 25 students. Predictions on holdout data are compared both in modified form, with numbers rounded up and negative values zeroed out, and unmodified form. All models were most accurate when predicting three semesters of holdout data using the maximum available enrollment data from Spring term of 2010 to Spring term of 2015 for training. The best resulting predictions were accurate within one standard class size of 25 students for 93.5% of Computer Science Department (CSD) courses, and the worst predictions were accurate within one standard class size for 77.4% of CSD courses.

Keywords: R, Time Series Forecasting, Weka, Machine Learning.

1 Background and Motivation

1.1 Introduction

Predicting course enrollment is a common university resource planning problem. In order to know how many sections of each course to offer, the student demand must be predicted. Failure to accurately predict the number of students wanting to enroll in particular courses has consequences. If the prediction of student enrollment is too high, sections may end up being cancelled due to having insufficient numbers to warrant offering the course. Consequently, students in cancelled sections would need to revise their schedules at the last minute, when other course options may no longer be available. Professors for the cancelled sections would be an under utilized resource and may not meet their required course load. Conversely, if the prediction of student enrollment is too low, then additional faculty need to be found on short notice. This could result in inexperienced or under-prepared course instruction for above average class sizes.

[*] California State University, Northridge
[†] California State University, Northridge

The overall schedule and physical classroom assignments might also need to change in order to accommodate the additional sections, potentially causing conflicts with student and professor schedules. Additionally, planning for course tutors and other learning support, including career counseling, can be negatively affected by inaccurate course enrollment predictions. Ultimately, if a department decides not to accommodate unforeseen course demand, then students could be set back in their academic schedule which could cause them to drop out of the program.

In particular, the Computer Science Department (CSD) at California State University Northridge (CSUN) faces a unique set of needs when creating the schedule of undergraduate courses each semester, owing to a number of factors. The California State University Statistical Reports show that out of the 23 California State University campuses, CSUN consistently has one of the largest overall student bodies as well as the largest numbers of first-time freshmen and new undergraduate transfers every Fall [3]. When it comes to enrollment numbers, CSUN is among the top five universities in all of California, with its enrollment numbers consistently increasing [4]. CSUN's overall enrollment from Fall semester of 1993 to Fall semester of 2014 increased roughly 49%, with the College of Engineering and Computer Science (CECS) experiencing an 82% increase in full-time enrollment [20]. As of 2015, the CECS has the fifth-largest headcount among CSUN's eight colleges [20] The composition of CSUN's new student enrollment for the Fall semester of 2015 was over 51% transfer students, with CSUN's overall undergraduate enrollment consisting of over 18% part-time students. For all 23 California State Universities for the Fall semester of 2015, the student body comprised of a little over 44% transfer students and just over 14% part-time students [3]. Both at CSUN and for all 23 CSUs approximately 4% of all Fall 2015 undergraduate students enrolled as Computer Science, Computer Engineering, or Computer Information Technology majors [3]. While CSUN has the same ratio of undergraduate Computer Science and technology majors that CSUs overall have, it has an overall larger student body with more transfer and part-time students enrolled. When deciding how best to forecast enrollment in undergraduate Computer Science courses, factors such as this can help determine the suitability of different methods and configurations. In the best practices for enrollment modeling, there is the notion that there is no one-size-fits-all approach to projecting postsecondary enrollment due to diverse cultural, financial, and political contexts as well as different challenges with increasing or decreasing enrollment trends, which supports the need for CSUN to have a custom course enrollment prediction system that is the most appropriate based on CSUN's particular contexts and trends [19].

The enrollment planning circumstances at CSUN are unique, since at the time courses are scheduled it is unknown how many students will be classified as majors or minors in Computer Science, Computer Engineering, Software Engineering, or Computer Information Technology. Also, it is unknown how many students are interested in which courses, how many students will be repeating courses due to failing grades, how many new transfer students will be enrolling, and how many continuing students will be returning. Currently, planning is done for two semesters into the future by the Computer Science Department Chair, with total student enrollment being estimated from past enrollment, and then estimating the number of course sections to offer by dividing the total course enrollment estimate by 25. While pre-semester surveys or filing enrollment plans could be instituted to help gauge course interest, student course interest could change for many reasons, rendering such endeavors likely high effort with low accuracy. Possible reasons that student reported future course enrollment plans could prove inaccurate are: students not considering course prerequisites, students not understanding and/or changing

their priorities and goals, students relying on incomplete or invalid sources of information resulting in less than optimal course selection, students selecting courses due to one characteristic of the course causing them to ignore other course or program requirements, students not being able to take desired courses due to time conflicts or overlap after the actual schedule is made, students not participating in the future course selection reporting process, or students not being able to take desired courses due to the final exam time after the actual schedule is made [11]. Unfortunately, there is no solution to knowing concretely before course scheduling how many students will fail courses the previous semester, how many new transfer students will enroll in the upcoming semester, or how many students will choose not to continue their education at CSUN. This means there is no way to gather concrete numbers for enrollment, so enrollment numbers must be predicted. To be useful the predicted numbers must be more accurate than the current educated guess method, with an acceptable bound being an error equal to or less than 25 students. The bound of 25 students is chosen since in 2015, 73.6% of classrooms at CSUN had enrollment between 20 and 49 students, making it an acceptable value for being off by one average class enrollment size [22]. Jacaranda Hall, where all Computer Science and Computer Information Technology classes are held, has 29 classrooms, where 22 have the capacity to accommodate 25 or more students [20]. This means that in the case that the predicted enrollment is underestimated by 25 students, the majority of classrooms available could handle the student overflow as a new course section. Contrarily, if the predicted enrollment is overestimated by 25 students, it would mean culling one course section.

Course planning is also complicated due to the large percentage of transfer and part-time students. CSUN is not a typical cohort style university where most students start as full-time, first-time freshmen then continue in lockstep with their peers until graduation. In the Fall of 2013, 81% of undergraduate students had an average age of 24 or younger and 19% had an average age of 25 or older [14]. The percentage of first-time students who began a bachelor's degree program at CSUN in Fall of 2013 and returned in Fall of 2014 was 77% for full-time students and 40% for part-time students [14]. This retention rate is much lower than that of neighboring University of California Los Angeles (UCLA), which was 97% of full-time bachelor degree students returning and 62% of part-time students returning [14]. Graduation rates for undergraduate students at CSUN who began in the Fall of 2006 are 14% for students graduating after four years, 48% for students graduating after six years, and 55% for students graduating after eight years [24]. The data supports the supposition that CSUN undergraduate students, when compared to other California State University (CSU) or University of California (UC) students, are less likely to continue their studies after their first year and more likely to be transfer students that are older, attend part-time, and take longer to graduate, making the CSUN student body less likely to be organized into groups of students that progress through educational programs at the same rate. For comparison, UCLA has a more traditional cohort program. In the Fall of 2015, UCLA's new undergraduate student enrollment was only 35% transfer students, with only 2% of the overall undergraduate student population enrolled part-time [24]. In the Fall of 2013, 95% of undergraduate students had an average age of 24 or younger and only 5% had an average age of 25 or older. Graduation rates for undergraduate students at UCLA who began in the Fall of 2006 are 71% for students graduating after four years, 92% for students graduating after 6 years, and 93% for students graduating after eight years [24]. Interpreting the data, UCLA students tend to enroll directly after high school and continue in their educational program full-time as a cohort until graduation. Due to this, UCLA's enrollment predictions likely employ forecasting models that would work well with a cohort oriented student

body and that therefore would not be well suited for predicting enrollment at CSUN.

1.2 Organization of This Paper

In this paper we document the design of a of an enrollment prediction tool which applies three Time Series models using the R forecast package [9] and four Time Series models using Weka [7]. This tool uses a database of 19 semesters of historical enrollment data to both train and validate predictions.

Section 2.1 discusses how we selected our Time Series models for our prediction tool, and Section 2.2 demonstrates how we implemented our tool as a Java application using Weka and R libraries, and how we used it to measure prediction accuracy. The accuracy of our predictions, validated against actual enrollment headcounts, are discussed in Section 3. We compare to related work in Section 4.1, and finally conclude in Section 4.2.

2 Methods and Technical Solutions

2.1 Model Requirements and Selection

Many studies address the prediction of overall enrollment at postsecondary institutions, or in kindergarten through 12th grade, as well as the different variables that affect those enrollments. For overall postsecondary enrollment prediction, Time Series models may possess a higher ability to capture the effects of influential variables. However, they may also obfuscate the influence of different variables whose movements are correlated over time. In particular, we identified four studies that specifically aim to address predicting student course selection at the postsecondary level [2, 11, 12, 16]. The approaches vary, and include Variable-Work Models, the Analytical Hierarchy Process, Adaptive Models, Neural Networks. The first three of these cannot be easily automated nor can they be applied within a Time Series framework. However, Neural Networks such as Multilayer Perceptron [21] can be used in this context. Consequently, we include this model in our study.

The CSD at CSUN needs an accurate, easy to use, easy to easy to interpret, easy to maintain, easy to enhance, free tool to aid in planning undergraduate course resources one year in advance. The tool must create different types of predictive models to forecast total student enrollment for undergraduate Computer Science courses at CSUN. At least one model's predictions for each course should be accurate to within 25 students, or one average class size, for the tool to be useful. Results of the tool should be the predictions generated by each model along with the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) in an easily readable spreadsheet file format, where the errors for each model of each course indicates which model will be likely to best forecast the total enrollment for that particular course. The tool should be easy to maintain and to enhance. This means requiring only an undergraduate Computer Science student level of knowledge as well as only utilizing free software technologies that are popular and well-documented.

The overall approach in predicting future course enrollment for CSUN undergraduate Computer Science courses is to first select an appropriate method. The most suitable forecasting method to model enrollment data was determined to be *Time Series Analysis*. Time Series Analysis assumes that the future depends on the present, and that the present depends on the past. This is

certainly true of student course enrollment, since courses have prerequisites that must be met before a student can enroll. These prerequisites can include classification in the appropriate major or minor program of study, completing required coursework with a particular grade, not completing the course previously more than a certain number of times, and other restrictions of this nature. The data used to build a Time Series model is a collection of observations made sequentially in time, such as course enrollment totals for each academic semester and year. There are many different statistical models that can be used to analyze a Time Series - from the simplest mean model where the predicted next value is equal to the historical sample mean to a more complex ARIMA [9] model where the predicted next value is dependent upon the weighted sums of recent values and error measurements. Which model is the most accurate in predicting CSUN undergraduate Computer Science course enrollment is determined by building models using a subset of data and then comparing their predictions against the actual values which were withheld.

Weka [7] and the R forecast package [9] are used to create the models as a Java program, since they are popular, free, and well-documented. Weka (Waikato Environment for Knowledge Analysis), is a free, non-commercial, open source suite of machine learning algorithms written in Java for data mining tasks. The Weka Time Series modeling environment must be installed separately using the package manager. R is a programming language and software environment for statistical computing and graphs that is free under the GNU general public license. R's forecast package contains methods and tools for displaying and analyzing univariate Time Series forecasts. The Time Series Models available in Weka are: Gaussian Processes, Linear Regression, Multilayer Perceptron, and SMOreg. The Time Series Models available in the R forecast package are: ARIMA, ETS, RWF, Meanf, Naïve, SNaïve, HoltWinters, DSHW, BATS/TBATS, LM/TSLM, StructTS, and NNetar. In this work we select the Gaussian Processes, Linear Regression, Multilayer Perceptron, SMOreg, ARIMA, ETS, and RWF models. This selection has been made to best match methods employed successfully in the past to predict overall university enrollment, predict online graduate school course enrollments, forecast website visits, predict parking lot occupancy, forecast energy consumption, forecast stock exchange rates, predict on Makridakis forecasting competition data, and predict the market value of residential buildings [1, 5, 7, 8, 9, 10, 11, 13, 15, 18, 19].

2.2 Methodology

The chosen models were created programmatically using a 95% confidence interval, frequency set to the correct seasonality (2 or 3), steps-to-predict set to the correct number per the test, and default settings for all other model parameters. Feature selection was kept to the required minimum of only past courses and their total enrollment. This is because the amount of available data was fixed, so any added dimensionality to the models would decrease predictive power according to the Hughes phenomenon. Each model's total enrollment predictions for the undergraduate Computer Science courses for the upcoming three semesters, along with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) measurements, are output to a spreadsheet file. The measurements of MAE and RMSE were chosen since they are indicators of how close model predictions are to actual data and because they were available in both Weka and R's forecast package. MAE is the average of the absolute errors, where the error is the difference between the predicted value and the actual value. RMSE sums the squares of the difference between the predicted and actual values, averages them, then takes the square root of

the average. Together, MAE and RMSE can find the variation in the errors of a set of forecasts. The RMSE will always be greater than or equal to the MAE, with a large difference between RMSE and MAE corresponding to a greater variance in the individual errors of the predictions. The lower the RMSE and MAE values are, the smaller the error is for the forecasts, and therefore the better the model is at predicting future course enrollments.

To create a more portable and maintainable application, we decided to employ the rJava package and the JRI Java/R interface it includes, allowing the project to be implemented entirely in Java. The project code logic is to first connect to the CSUN prediction database to get the course start date and course enrollment total for all undergraduate COMP and CIT courses offered in all previous academic terms and years. The CSD will maintain this separate prediction database, scrubbed of all sensitive data, that contains historical enrollment information for CSUN. This project is developed using 19 semesters worth of data, from the Spring semester which began on January 19th, 2010 to the Spring semester which began on January 25th, 2016. Note that undergraduate Computer Science courses are only offered during the Fall, Spring, and Summer semesters and not during the Winter semester at CSUN. The data, along with the data column headers, is then output to a temporary CSV file for Weka and R to read from when generating models. Any courses that have total enrollment of zero for all semesters are purged, since creating a model to predict the enrollment for those courses will fail. The program environment is set including paths to the input data, paths to the output data, and what format the input data is in, including the format of the date field. The variables created from the data are utilized to predict the total enrollment of each course for a set number of steps into the future for each selected model, and then output those results along with the MAE and RMSE to a spreadsheet. Since Summer semester data is different from Fall and Spring semester data due to fewer courses being offered and fewer students enrolling, tests were run on all data and also with only Fall and Spring semester data to see if removing Summer data resulted in more accurate predictions. When running tests on all three semesters of data, the lag time and number of semesters in the future to predict is set to three. Without Summer semester data, the lag time and number of semesters in the future to predict is set to two. The lag time sets the periodicity of the data, for example, for monthly data 12 lag steps would make sense and for hourly data 24 time steps would be logical. When creating models, the minimum amount of data required is anywhere from two to three times the lag time. The different tests that were conducted involved holding out different amounts of data from model creation to see which models predicted the held-out data the most accurately, and if that accuracy was increased by removing and not predicting on Summer term data. Historical data that is used for model creation is referred to as *training data*, while historical data that is withheld to compare with the model's predictions is called *holdout data*. The standard amount of holdout data for testing is one-third, or 33%, of the available historical data. Each run of the program was approximately seven minutes long, which is an acceptable runtime for a program that would run once a year to generate predictions for the upcoming academic year.

3 Empirical Evaluations

The predictability of a course appears to be loosely related to the *variance* of the courses. The variance for a particular course can be thought of as the amount that the course enrollment totals for the particular course vary from the average enrollment count for that course. Variance

can be defined as the average of the squared differences from the mean, and is formalized in Equation 1, where μ is the mean of the values x_i in the sample set and N is the number of values in the sample set.

$$\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Equation 1: Variance

Enrollment variance was calculated for each undergraduate COMP and CIT course over 19 semesters. The variance of the least predictable courses is shown in Figure 1 and the variance of the most predictable courses is shown in Figure 2. These results demonstrate that courses having higher variances generally tend to have a higher enrollment and be less predictable. Unpredictability may be further affected by administrative policies, where additional sections of required core classes will be opened up to accommodate demand. Comparatively, elective course (COMP 467, COMP 484, COMP 484L) demand is not similarly accommodated, with elective courses offered depending on instructor availability with students expected to enroll in any elective course with empty seats. New courses, infrequently offered courses, or experimental electives also tend to be less easy to predict simply due to the lack of historical data. Such courses are not included in this study due to their lack of history, and also because they never enroll beyond the capacity of a single class section.

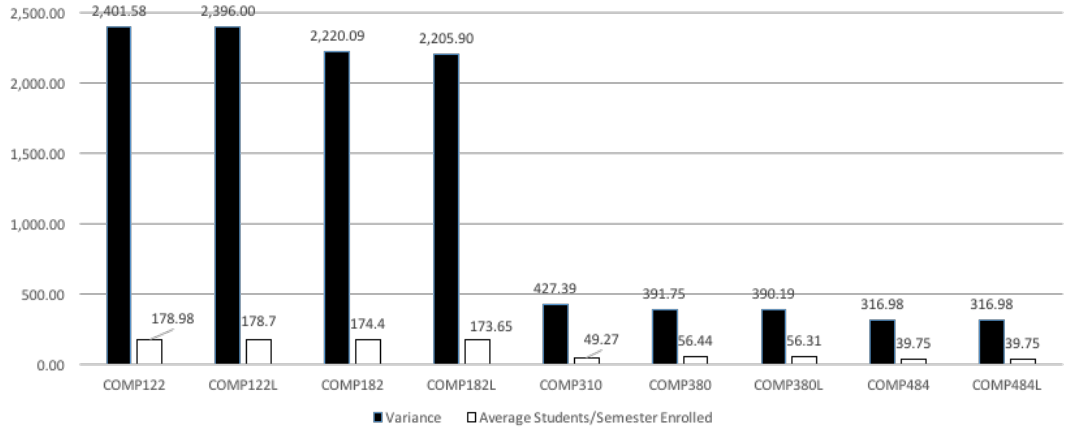

Figure 1: Variance and Average Students/Semester for Least Predictable Courses

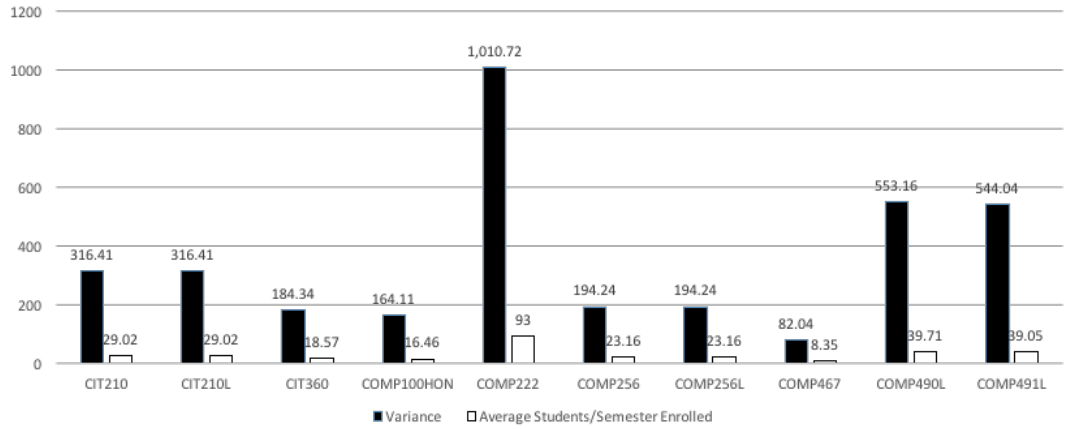

Figure 2: Variance and Average Students/Semester for Most Predictable Courses

Figure 4 shows the MAE and RMSE of each of our chosen Time Series models when predicting 3 semesters ahead using all of the data, whereas Figure 3 shows the errors of our models when predicting 3 semesters ahead with holdout data (Summer semester enrollments) withheld. The models whose predictions were most accurate when comparing modified predictions for Fall, Spring, and Summer semesters without holdout data were: Gaussian Processes, SMOreg, and Linear Regression. This did not align with the models generated using all of the data (shown in Figure 4). This is likely due to overfitting, where the models memorize the in-sample data instead of actually predicting it.

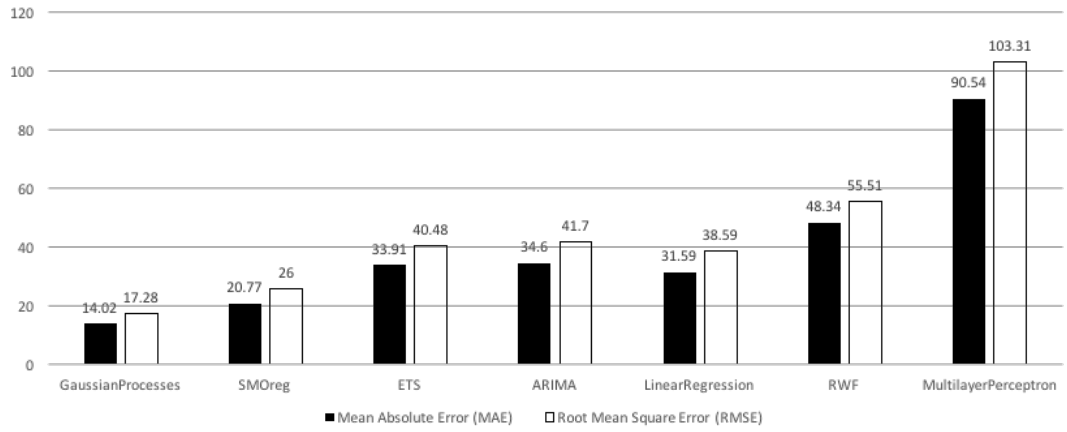

Figure 3: Results of 3 Predictions Ahead on Modified Data

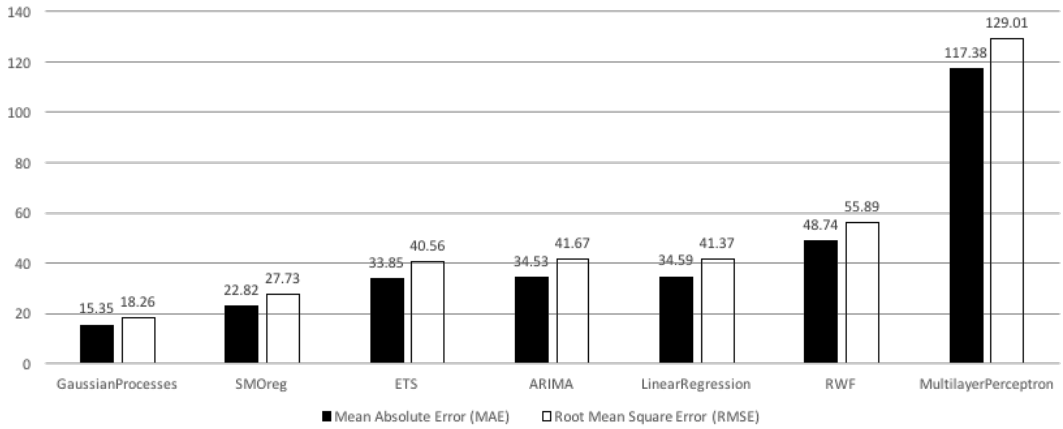

Figure 4: Results of 3 Predictions Ahead on Unmodified Data

Not predicting on Summer semester data (the "modified data" results of Figure 3) caused Fall and Spring semester enrollment predictions for most, but not all, models to become more accurate. ARIMA and ETS models both became less accurate when constructed without Summer semester data. This could be due to them predicting better when there is more data or data that exhibits more seasonality. The models that were the least accurate when comparing modified predictions for all three semesters using holdout data were: Multilayer Perceptron, RWF, and ARIMA. Courses that were the hardest to predict with the largest prediction error did not always align with the courses exhibiting the greatest variance. This could be due to administrative policies on when to open additional course sections, and the *blanket* policy for this research of rounding up the positive predictions to the nearest whole number and setting negative predictions to zero when logically modifying the data. Experienced CSUN planners can eschew the blanket policy, and instead use their experience and intuition to evaluate whether to round up or round down the enrollment predictions for each course. For example, an outsider may think that lower division courses would be more difficult to predict than courses taken further along in the program due to sophomore and senior students being more predictable, when in fact due to administrative policies of accommodating demand for lower division courses and not higher division courses causes the opposite effect. As measurements against holdout data tend to be more accurate than against in-sample data, future work could partition holdout data and use that to calculate the model accuracy, before generating models using all data for calculating predictions.

Examination of the most predictable and least predictable courses as forecast by the best model

(Gaussian Processes) demonstrates that despite applying the blanket policy the worst case scenario is not bad, and applying better intuition can ameliorate the results. Model prediction results will be the most accurate when using as much data as possible to create them, when predicting no further into the future than one season, or three semesters, ahead since the further out a prediction is the less accurate it becomes. A future enhancement would be to evaluate which models best predict using Summer semester data, and use the results to explore a hybrid model method, with different models being used to predict using Summer, Fall, and Spring semester data.

We define the *enrollment error* of our enrollment prediction as predicted enrollment minus actual enrollment. A negative enrollment error indicates underestimation while a positive enrollment error indicates overestimation. Positive predicted enrollment values that are fractional are rounded up to the next whole number. This is due to the fact that you cannot have a fraction of a student, as well as student demand at CSUN being affected by administrative policies. The policy is usually to not accommodate additional student demand, making the actual demand for a course higher than the number of students who were enrolled in the course. Negative predicted enrollment values are treated as zero, since there cannot be negative enrollment. This zeroing method, while introducing some bias, is an interpretation of the results within the context of enrollment. It has also been used successfully in Kaggle competitions, such as Arthur Suilin's first place solution for forecasting web traffic [24]. While negative enrollment predictions could be avoided through either choosing models which cannot yield negative values, modifying model parameters to disallow negative values, or pre-processing data for the additive models via logarithmic or Box-Cox transformations, the goal was to explore the accuracy of the time series models in their basic form. Modifying parameters and adding data pre-processing may be explored in future work. The modified predictions arrived at through rounding up positive fractional predictions and zeroing out negative predictions are considered acceptable as long as the absolute value of the difference between number of actual versus predicted students is less than or equal to 25, which is the size of a typical CSUN class. This metric was chosen since adding or cancelling one class to accommodate enrollment demand is feasible.

We found the Gaussian Processes model to be the best overall predictor, having consistently the lowest MAE and RMSE measures. This may be due to the fact that Gaussian Processes are traditionally used for regression on fixed data sets, and they use Bayesian inference which is good

at dynamic analysis of a sequence of data. Additionally, since Gaussian Processes use probability distributions, they may accommodate the uncertainty of transfer and part-time student schedules better than the other models can. This may be due to their ability to model complex, non-linear functions. The accuracy of model predictions also has much to do with the data used to train the models and its characteristics. Only 19 semesters of data were available, leaving a maximum of 16 semesters for training in order to be able to test predictions for three semesters out. A three semester ahead forecast was used as the minimum future prediction size since the university typically plans one year in advance and because three semesters is considered one season. In general, all models predict better when there is more data to learn from, but some models can be better at short-term predictions and learning on sparse training data, depending on the variance and seasonality of the data.

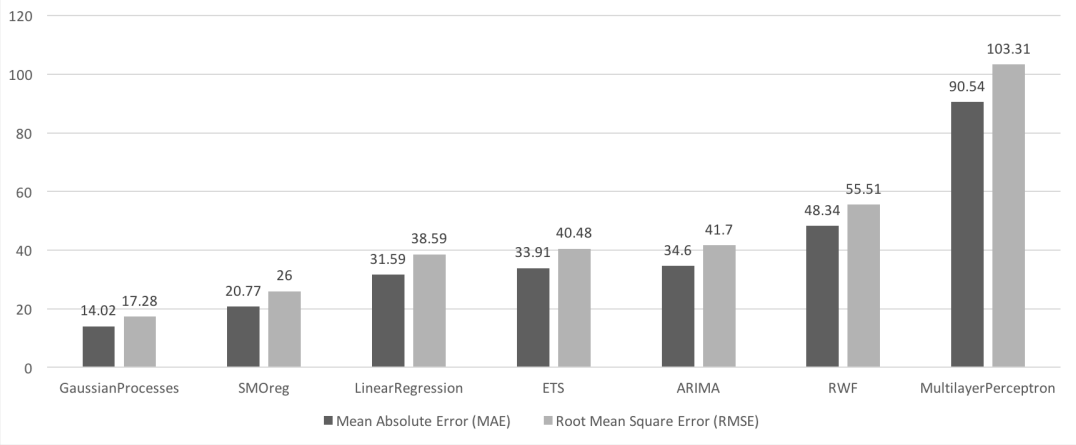

Figure 5: Error for Predicting 3 Semesters Ahead Using Blanket Policy Modified Data

Figure 6 shows a Gaussian Processes model trained on 16 semesters of data beginning with Spring 2010, and predicting three semesters ahead (Summer 2015, Fall 2015, and Spring 2016). Course enrollments are considered to be within 25 students if the absolute value of the difference between actual enrollment and predicted enrollment is less than or equal to 25 students. We compared the results of this model to actual enrollment history from these three semesters. The number of courses out of 62 that were predicted to within 25 students were: 58 for the first step prediction at Summer 2015, 55 for the second step prediction at Fall 2015, and 48 for the third step prediction at Spring 2016. Summer appears to be the easiest term to predict within 25 students, while Spring 2016 is the most difficult term to predict within 25 students. Over all three terms, the most difficult to predict course enrollments were consistently those of the same

two lecture/lab courses (COMP 182/L and COMP 122/L).

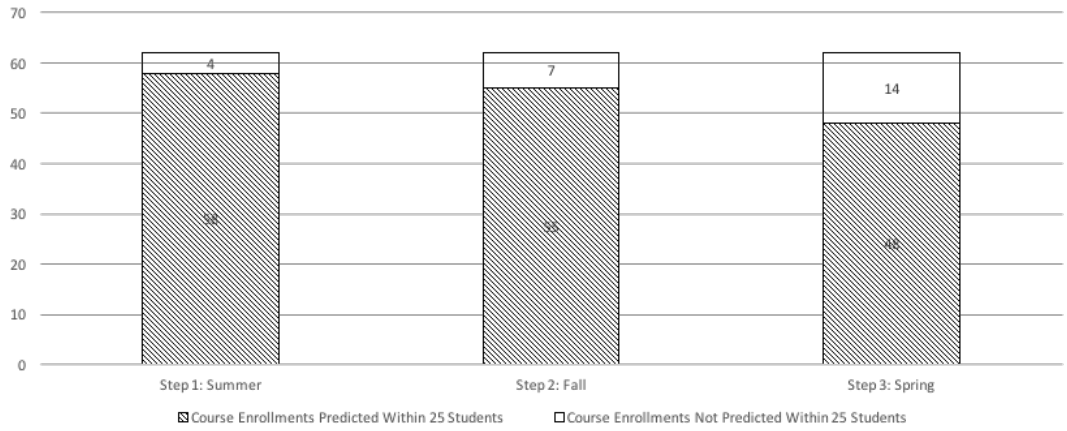

Figure 6: Results of Training with 16 Semesters of Data

Figure 7 and Figure 8 show the results of training the Gaussian Processes model with 10 and 13 semesters of data respectively, predicting nine and six semesters ahead respectively. Once again, we compared the prediction results of this model to those of actual enrollment during the three semesters. These results demonstrate that Spring 2016 is not always the most difficult semester to predict, as Fall 2015 is harder to predict than Spring 2016 in both the 10- and 13-semester trained models.

Figure 7 and Figure 8 also demonstrate the diminishing accuracy of the model as predictions are made beyond the next academic season of 3 semesters. Course enrollments not predicted within 25 students tend to comprise 27% of courses in 5 or 6 semester ahead predictions in the 10-semester trained model, and as high as 47% of courses in an 8 semester ahead prediction. Comparatively, in the 13-semester trained model, course enrollments not predicted within 25 students tend to comprise 39% of courses in the 5 semester ahead predictions. We find it notable that training models on a greater number of semester data does not always yield a better prediction further than a single academic season (3 semesters) into the future. Moreover, training on more semesters of data does not always yield a more accurate prediction of the next academic season, as the 10-semester trained model was able to predict Spring 2014 better than the 16-semester trained model was able to predict Spring 2016. However, this counter-intuitive result may stem from a larger number of unpredictable course sections being offered in Spring 2016 than Spring 2014.

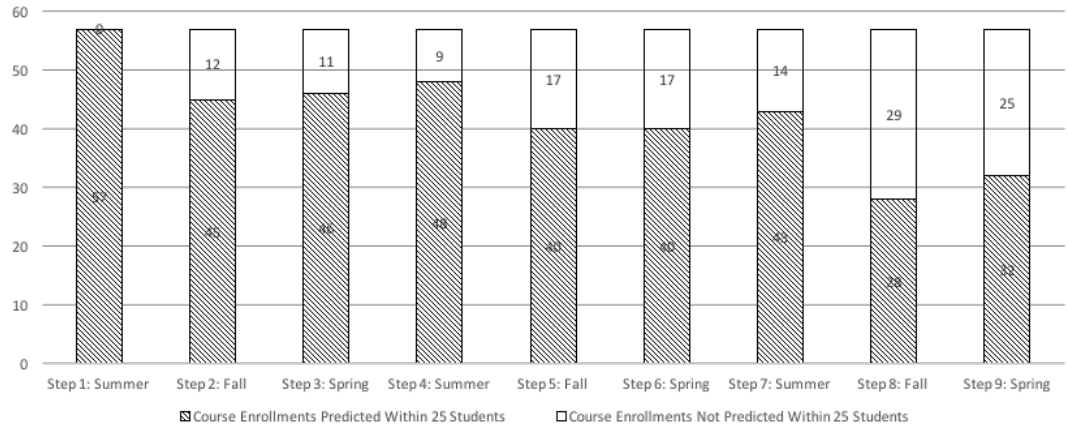

Figure 7: Results of Training with 10 Semesters of Data

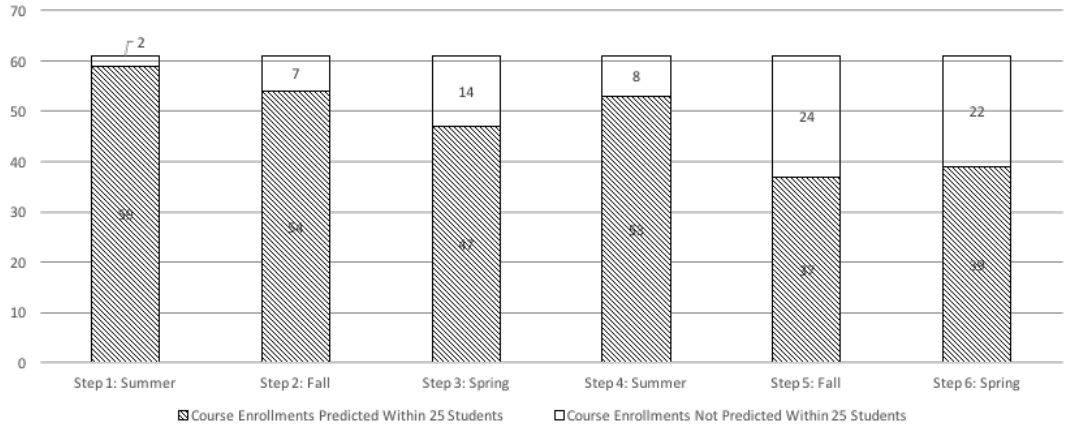

Figure 8: Results of Training with 13 Semesters of Data

For the data set available in the prediction database, it is observed that particular courses tend to only be offered in specific academic terms, making the data more seasonal than linear. Since the historical enrollment data is also accurate, with no missing records, there is also no noise. The trend of the data is that course enrollments tend to go up over time, probably due to increased overall enrollment at the university and in the Computer Science program, and the models need to be able to detect it. Data issues that can affect the predictions could be experimental courses that are infrequently offered, courses that used to be offered but no longer are under the same course name, or new courses that do not have adequate historical enrollment data yet. To know which variables affect the predictions, and to what degree, testing with more data, pre-processing the data to remove outliers, and including supplemental overlay data such as academic (tuition rates, high school graduation rates, community college transfer rates, average grades for a course, CSUN overall enrollment) or economic data (unemployment rates, interest

rates) could provide further insight. We found the majority of COMP and CIT courses can be predicted consistently within an acceptable range of error, and the larger errors when predicting two of the courses can likely be mitigated through training the models with more data and applying intuition when logically modifying the prediction results.

4 Significance and Impact

4.1 Related Work

The problem of predicting course enrollment numbers is not unique to CSUN. The University of Colorado at Boulder uses educated guesses surrounded by considerable error to predict total university enrollment [25]. Their estimated margin of error for large, predictable groups such as undergraduate cohorts is between 2-3%. The inaccuracy of their predictions skyrockets to over 10% when projecting enrollment for smaller groups, such as out of state students or new international graduate studies students, which are more prone to fluctuation. The University of Wyoming conducted an Enrollment Project in order to predict university enrollment one to five years ahead [1]. They found that using four different models to predict the four different populations that comprise their overall enrollment worked well in that it only underestimated the actual enrollment count by 1%, which was likely due to the introduction of a scholarship program [1]. The models used were linear regression for resident undergraduate enrollment, semi-log regression for regional undergraduate enrollment, linear regression for graduate student enrollment, and linear trend regression for all other undergraduate students [1]. The spreadsheet-based matrix-ratio model developed for course enrollment prediction at the University of Missouri-Columbia uses three years of historical student course enrollment data by major and student level to predict how many students in each major at each level (freshman, sophomore, junior, senior) will enroll in a specific course in the upcoming academic year. The model must be updated manually with the latest data to stay current, and its calculations require a ratio of students who will switch majors and the forecasted total university enrollment. The output of the model is 99% correlated with the actual course enrollment results of the university's population of less than 6,000 students [6]. The University of Central Florida updated their enrollment projection model in 2015. The previous spreadsheet model they used was cohort-based, relied on manual model updates for factors based on the judgment of the Institutional Research staff, and was based on student retention rates from the past ten years [1, 17]. The mean absolute percentage error, a measure of how much the actual enrollment numbers varied from the model predictions, of their old method was approximately 0.5% for short term projections and 2% for long term projections [1]. The new University of Central Florida enrollment projection system uses Weka for modeling and R for forecasting and no longer relies on manual updates [17]. The model uses two methods - neural network and regression with different variable selection methods - on the previous Fall headcount, graduate students, first-time-in-college students, new students, and more. The forecast uses the non-seasonal Holt's linear method with damping and exponential smoothing. The error for the new projection framework is a little worse than the old method with a mean absolute percentage error of 2.2% when predicting enrollment for the following year. Oklahoma State University utilizes an Autoregressive Integrated Moving Average (ARIMA) methodology to predict enrollment with a

mean absolute percentage error of 2.11% [5]. The University of Hawaii system also uses ARIMA methodology to predict enrollment with different models for each campus, with a mean absolute percentage error of the models ranging from 2.391% to 6.771% [10].

In comparison to previous enrollment prediction solutions, we present a prediction tool that is free, automated and does not require manual modifications. It is also extremely accurate since its best prediction successfully estimated enrollment to within one standard class size of 25 students for 93.5% of courses, and its worst prediction estimated enrollment to within one standard class size for 77.4% of courses. Since student enrollment patterns may change along with administrative policy, it is advantageous to automatically generate and compare different models each time a new year of enrollment data is available. This is since new information may result in a different model being the best predictor. The generic and automated design also means the prediction tool can be utilized with little modification by other campus departments or even other universities, as long as there is a prediction database with the required information available.

4.2 Conclusion and Future Work

In this paper, we have presented an enrollment prediction tool implemented as a Java application using R and Weka to perform statistical modeling of course enrollment. This tool has been designed to be executed once at the beginning of each academic year by the Chair of the CSUN CSD in order to aid in planning course offerings for the next 3 semesters. A better planning method that does not rely on intuition is in order due to CSUN being home to a number of notoriously unpredictable CS and CIT courses, as CSUN has a larger than average student body, of atypical age and working status compared to other California State and University of California campuses. The code base was designed so that the CSD Chair could easily modify it for greater accuracy, while still allowing undergraduate CSD students to comfortably maintain it, since Java is a core language taught in the undergraduate program.

Using a database of 19 semesters of historical enrollment data, and training our tool on 10, 13, and 16 semesters of this data, we find the Gaussian Processes model to be the best overall predictor. We further find that our tool performs best when trained with the maximum amount (16 semesters) of history, and when predicting one academic season (3 semesters) into the future. Withholding Summer semester data from the training set further increases accuracy. Under these conditions, the best resulting predictions are accurate within one standard class size of 25 students for 93.5% of CSUN CSD courses, and the worst predictions are accurate within one class size for 77.4% of CSD courses.

We propose to further enhance this work by exploring models, model parameters, and data pre-processing that make the best prediction using historical Summer semester data. The results can be used to generate a hybrid model method, where different models can be used to predict Summer, Fall, and Spring semester data. Ultimately these hybrid strategies may be compared to see which combination of models, model parameters, and holdout data yields the highest overall accuracy in course enrollment prediction.

References

[1] K. R. Balachandran and D. Gerwin. Variable-Work Models for Predicting Course Enrollments, *Operations Research, INFORMS*, 3, 1971.

[2] California State University. The California State University Analytic Studies Statistical Reports, *http://www.calstate.edu/as/stats.shtml*, 2016.

[3] California State University Northridge. *CSUN Outreach Publications View Book*, *http://www.csun.edu/sites/default/files/viewbook.pdf*, 2013.

[4] C. Chen. An Integrated Enrollment Forecast Model, *IR Applications*, Association for Institutional Research, 15, 2008.

[5] K. S. Felts and M. Ehlert. *Prediction Model for Course Demand at MU*, Enrollment Management and Institutional Research, University of Missouri-Columbia, 2009.

[6] M. Graczyk, T. Lasota, and B. Trawiski. Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA, *First International Conference Computational Collective Intelligence (ICCCI)*, 2009.

[7] S. Gvaladze. *Evaluating Methods for Time-Series Forecasting Applied to Energy Consumption Predictions for Hvaler (kommune)*, Master's Thesis, Computer Science, Østfold University College, Halden, 2015.

[8] Hyndman, R.J. and Khandakar, Y. Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software*, 26 (3), 2008. Institutional Research and Analysis Office for the University of Hawai'i System. *Enrollment Projections for the University of Hawai'i System Fall 2013 to Fall 2018*, University of Hawai'i Department Report, Honolulu, HI, 2013.

[9] A. Kardan, H. Sadeghi, S. S. Ghidary, and M. R. F. Sani. Prediction of Student Course Selection in Online Higher Education Institutes Using Neural Network, *Computers and Education*, Elsevier, 65, 2013.

[10] C. Kraft. *Planning, Scheduling, and Timetabling in a University Setting*, PhD Dissertation, Mathematical Sciences, Clemson University, 2007.

[11] C. Napagoda. Web Site Visit Forecasting Using Data Mining Techniques, *International Journal of Scientific and Technology Research*, 12, 2013.

[12] National Center for Education Statistics. College Navigator, *http://nces.ed.gov/collegenavigator/*, 2016.

[13] M. D. Nemes and A. Butoi. Data Mining on Romanian Stock Market Using Neural Networks for Price Prediction, *Informatica Economic*, 3, 2013.

[14] I. Ognjanovic, D. Gasevic, and S. Dawson. Using Institutional Data to Predict Student Course Selections in Higher Education, *Internet and Higher Education*, Elsevier, 29, 2016.

[15] P. Ramsey, A. Watts, and L. Sklar. Institutional Knowledge Management Enrollment Projection Model, *Southern Association for Institutional Research*, Savannah, GA, 2015.

[16] M. Reinstadler, M. Braunhofer, M. Elahi, and F. Ricci. *Predicting Parking Lots Occupancy*

in Bolzano, Academic Project, Computer Science, Free University of Bolzano Italy, Bolzano, 2013.

[17] E. Reiss. Best Practices in Enrollment Modeling: Navigating Methodology and Processes, *Southern Association for Institutional Research*, Lake Buena Vista FL, 2012.

[18] Rickes Associates Inc. *California State University Northridge: Teaching, Learning, Office, and Research Space Needs Assessment*, 2015.

[19] J. F. Shepanski. Fast learning in Artificial Neural Systems: Multilayer Perseptron Training using Optimal Estimation, *Proc. IEEE 2nd Intern. Conf. Neural Nets*, 1988.

[20] U.S. News. U.S. News Colleges California State University Northridge 2016 Overview, *http://colleges.usnews.rankingsandreviews.com/best-colleges/csun-1153*, 2016.

[21] University of California. The University of California at a Glance, *http://universityofcalifornia.edu/sites/default/files/uc-at-a-glance-mar-2016.pdf*, 2016.

[22] University of California Los Angeles. UCLA Academic Planning and Budget: Campus Statistics for Enrollment, *http://www.aim.ucla.edu/enrollment2.aspx*, 2016.

[23] University of Colorado Boulder. CU Boulder: Planning, Budget and Analysis - Enrollment Projections, *http://www.colorado.edu/pba/enrlproj/*, 2015.

[24] Web Traffic Time Series Forecasting: Forecast Future Traffic to Wikipedia Pages, *https://www.kaggle.com/c/web-traffic-time-series-forecasting/discussion/43795*, 2017.