Optimizing Academic Scheduling with Machine Learning at Lake Tahoe Community College

Michael Atkinson

Abstract

In order to optimize the decision making abilities of the scheduling office at Lake Tahoe Community College various machine learning modeling frameworks were explored in an attempt to accurately forecast an upcoming Academic Year's section level enrollment counts. Predictors such as course name, day offering patterns, offering times, previous year's enrollment, moving average enrollments, modality, and course type were used to make predictions on section level enrollment. The models used were; Lasso Regression, Random Forests, Gradient Boost, and Neural Networks. Each model was trained on Online and Face to Face sections offered from 2012 to 2017 and then tested on unseen data from the 2018 Academic Year. By using the Mean Squared Error as a measure of fit, each models' prediction accuracy was recorded across the four academic terms comprising the year. Ultimately MSE was minimized using Lasso Regression and Random Forests, however the best score was achieved when an average was taking of the Gradient Boost and two aforementioned models. The results achieved using these models provided a great deal of accuracy and warrant future research into implementing them at LTCC.

Background Literature Review

When attempting to develop forecasting models for both long and short term enrollment patterns there are many possible options for an analyst to explore. In *An Integrated Enrollment Forecast Model* 1. Nine commonly used models are illustrated as potential candidates for an institution projecting future enrollments. These models are as follows;

- ***Subjective Judgement -***This method is 'Non Analytical' and can be implemented when other more rigorous methods are not available

- ***Ratio Method-*** this method takes into account the ratio of previous high school graduates to first year college enrollments and uses that to project future enrollments

- ***Cohort Survival Study-*** This method estimates future enrollments by multiplying the survival rate of the previous cohort by the cohort size of the previous year

- ***Markov Transition Models-*** This method predicts the probabilities of future occurrences of an event based on currently known probabilities.

- ***Neural Network Model-*** This method is a machine learning technique in which previous data is given to the model as a training set and then future predictions are made given new values on the predictor space

- ***Simulation Method-*** This is a complex mathematical model in which relationships between various component inputs are defined and then tweaks to input variables are simulated to test 'what-if' Scenarios

- ***Time Series Analysis-*** Time series analysis is a collection of measurements gathered at equal spaced time periods. Data points close in time are accepted to be highly correlated with one another. From these correlated relationships forecast are made. In the case of forecasting enrollments in higher education, one might collect term on term or academic year on academic year measurements and use a Time Series model to forecast future enrollments.

- ***Fuzzy Time Series Analysis-*** This method use a mix of Fuzzy Set Theory and Time Series Analysis to handle more complex and nonlinear time trends

- ***Regression Analysis –*** Regression Analysis used by fitting a set of key indicators to predict enrollment through the idea of a 'line of best fit'. Common methods implored in this model are Linear and Piecewise Regression. Linear Regression for the continuous impact of the predictor space whereas piecewise regression uses multiple break points in which the relationship between predictors and response change.

Forecasting Community College Enrollments [2], a publication by Hanover Research suggests that a "scan of the literature" shows that the most popular models used for to forecast college enrollment are Time Series and Regression Analysis. In their paper they go on to discuss some of the successes of the various models listed in use cases at a range of colleges. As an example they explain the use of an Autoregressive Integrated Moving Average Model that was used at Monroe County Community College in Michigan that used the unemployment rate and tuition cost to predict credit hour enrollments over a 32-year span. This use case reported the model to be "an excellent fit" and recommended that Monroe County Community College continue to use it in the future.

The publication by the Hanover group goes on to detail the troubles that Metropolitan Community College in Missouri had implementing a model used by St. Charles Community College saying that the success St. Charles had did not translate to Metropolitan because of its size and number of Counties.[2] This result illustrates that for good predictive results there is no one size fits all approach and that each institution is likely to be best served by finding which models and predictors work best for their use case.

Enrollment Forecasting for School Management System[3] details the process of integrating a model into a school's management information system. This paper provides an overview of defining a measure of fit for the model, in the case of the paper they used the Mean Absolute Percent Error (MAPE) and then comparing model performance by selecting the model that minimizes the MAPE. The models they use for comparison are a three year moving average, an exponential smoothing model, and a double exponential smoothing model. In both forms of exponential smoothing models, the authors discuss the process of adjusting tuning parameters to optimize model performance.

In the research conducted at Lake Tahoe Community College many elements from the aforementioned research were applied. My research looked at popular models such as regression and neural networks, selected predictor variables that are unique to the specific use case of the college, and assessed models using an objective measure of fit as scored across all candidate models.

Introduction

During the 2019 academic year the department of Institutional Research and Effectiveness at Lake Tahoe Community College changed the way academic scheduling was performed with the development of the Student Enrollment Management (SEM) dashboard. Previously scheduling had been performed in a rather archaic fashion and was anything but data informed decision making. The SEM dashboard, or as it became known around campus, 'The Tool' allowed the scheduling office to easily look at department and course level enrollment trends over previous years, assess high demand times and dates, and look at common concurrent enrollments for a given course.

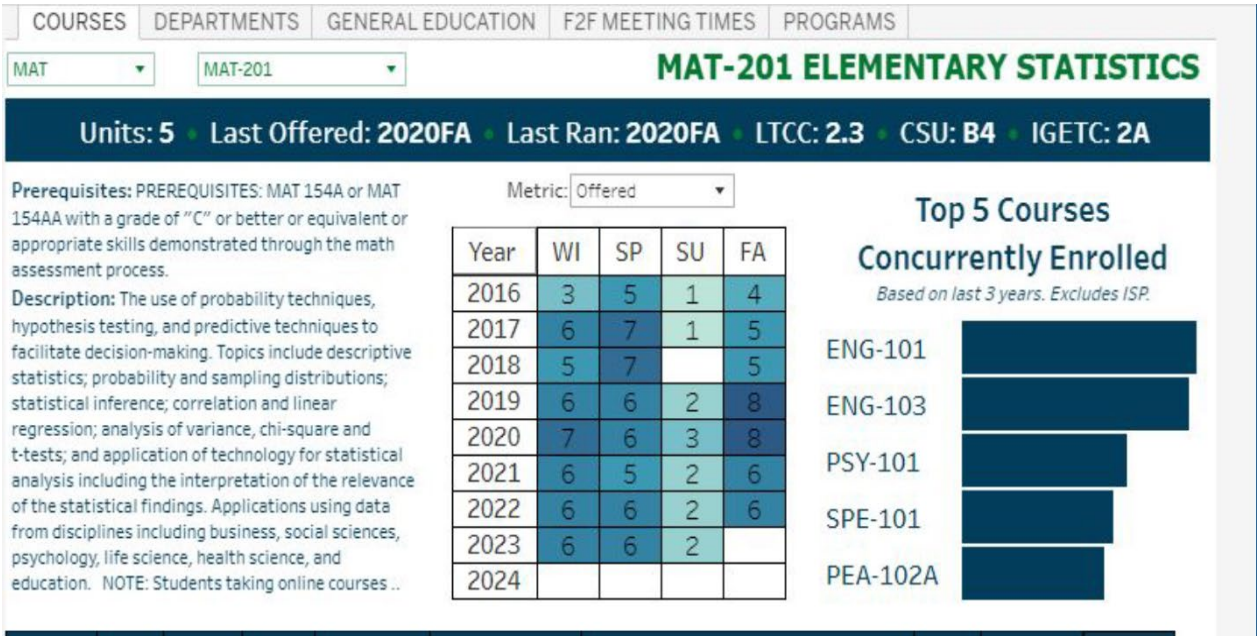

Figure 1. SEM Dashboard Courses tab illustration information on Mat-201

Although the SEM dashboard was major step forward to a more data informed approach to scheduling its current state is limited to providing a descriptive analysis of past enrollments. To further the ability of Lake Tahoe Community College administrators to optimize the scheduling process this paper aims to explore the use of machine learning techniques and methodologies to successfully forecast section level enrollments for upcoming academic years. Being able to accurately forecast future enrollments at the section level can

improve budgeting decisions based on future revenue streams, maximize FTES and increase the efficiency of campus resources.

In order to produce models that make accurate predictions much consideration was put into determining the predictor variables to be used in the selected machine learning algorithms. While much of the previous research in enrollment forecasting focuses on predicting total enrollments for a term or academic year using variables not localized to the institution, (unemployment rates, highs school graduation rates, local economic data) the predictors used for the modelling conducting in my research at LTCC focuses on an entirely localized predictor space that can be gained by tapping into the transactional database at the college. The predictor variables used in the analysis were some of the traditional variables used in previous research such as previous years' enrollment and moving average enrollment. However, other localized variables such as day offering pattern and time offerings were also taken into consideration. Also, course modality, course type, term, previous section cancellation and individual course titles were used in the predictor space. The rational for this deviation from the published literature that takes a more local view was in part due to availability of data and part due to objective of the research. Since the stated goal of the research is to provide accurate forecasts for upcoming section level enrollments and provide tools for schedulers to make more informed decisions, it was a natural fit to use variables that are readily available to the institution and can also answer specific questions on the relationships of certain variables and their impact on enrollment. Questions such as, do Fall sections have higher enrollment than Winter? Do day pattern offerings make a noticeable different in section demand? Does math have more enrollments per section than English? The answers to these questions might vary widely from institution to institution, however if analysis shows that for Lake Tahoe Community College consistent patterns emerge in these predictors then these relationships can be leveraged to produce accurate models.

In order to develop useful models in enrollment forecasting data was collected from the 2012-2018 academic years. After wrangling the data, an exploratory data analysis was conducted to determine the potential predictive power of certain variables in an aim to guide the variables model selection process. For the numeric predictor variables, this was done by visualizing the data, looking at Principle Component Analysis bi-plots to determine which predictors aligned with one another, analyzing the correlations between the numeric predictors and the response variable and then by analyzing variable importance in the tree based

models used in the analysis as well as the coefficients of the Lasso Regression model used on training data. The categorical predictors relationship with the response were explored through various visualization techniques.

After the exploratory data analysis, four models were trained on data from the 2012-2017 academic year and then tested on the 2018 academic year. This was done in order to assess how each model would perform given an unseen data set which, if implored in a production scenario would be the task of the model. After the model training, each model made predictions on the four terms comprising the 2018 academic year and the Mean Squared Error on the test year was used as the objective measure of fit for each model. The four models used for comparison were Lasso Regression, Gradient Boosting, Random Forests, and a Feed Forward Neural Network. After these models made their predictions an ensemble model that averaged the predictions from the three best individual models was used for further comparison.

Data Collection, Feature Engineering and Final Dataset

The data was collected by accessing Lake Tahoe Community College's transactional database via SQL queries. LTCC has a database that records course, section, enrollment, session date and time, session modality (online vs face to face), and a host of transactional data since the academic year 2012. This provided a large amount of potential variables to pick from model selection.

Ultimately, the data collected was limited to the online and face to face section offerings from 2012 through 2018. Although LTCC does offer off campus sections for Firefighter training as well as Incarcerated Student sections throughout the Northern California prison system, these sections were decided to be left out of the first preliminary modelling attempts. This decision was made based on multiple factors. Firstly, the on campus sections will more likely capture the enrollment patterns of more traditional students. Secondly, since much of the predictor space relies on schedule date and time variables, it was decided that looking at online and face to face sections would capture the trends given the local community. Lake Tahoe is a tourist based economy in which many residents work atypical hours. Additionally, one of the immediate goals of enrollment forecasting is to optimize on campus resources and this requires the analysis to be focused on the on campus and online enrollment patterns of the more traditional community college student. Lastly, off campus section offerings often lack normal date and time schedule patterns and are provided as independent study.

An issue that arose in the original data set was the problem with section day scheduling patterns. In LTCC's database section session offerings are coded as binary variables across the days of the week. This means that each day Monday-Sunday will have a yes or no for a given section. One of the predictors of interest was day offering pattern which is a unique variable to each section. This required some feature engineering to turn the one hot encodings across each day to the unique day pattern for the section. For example, if a section had an offering on Monday, Wednesday, and Friday, these three columns of variables were turned into one unique pattern such as MonWedFri for each section. Online sections in the data that did not have any day pattern were coded as OL for online.

The final dataset prior to exploratory data analysis was comprised of n=3785 observations of section offerings from 2012-2018. The variables gathered across each section were as follows.

Numeric Variables

- ***Active Enrollment Count -*** The response variable of interest which was that sections count of active enrollments as of the session date for the given term.

- ***Section Capacity-*** This variable is the mandated maximum capacity for a given section.

- ***Active Enrollment Previous Year-*** This was the section count for the same section offered during the same term in the previous academic year.

- ***3-year Active Enrollment average-*** This variable is the rolling 3-year average for a section and term. For example, Eng-101 in 2018 Fall would have the average section enrollment for the Eng-101 offered in Fall 2015-2017.

- ***Total Average Active Section Enrollment-*** This was the total average section enrollment for previous year's section. For example, Eng-101 in 2017 Spring would be the average of enrollments for all Spring Eng-101 sections from 2012-2016.

- ***Drops Previous Year-*** This variable is the count of student drops prior to census in the same section in the previous year.

- ***Withdraws Previous Year*** – This is a count of post census withdraws the resulted in an enrollment grade of W for the previous year.

- ***Cancelled Sections Previous Year-*** This is a count of the number section offerings for the course that were cancelled the previous year.

- ***Current term Section Offerings-*** This is a count of the number of sections offered for the course in that current term. For example, if the Course Eng-101 had 3 Fall section offerings the value would be 3.

- ***Number of programs-*** This is a count of the number of active programs that are offered in the college catalog year that a course is tied to. For example, if Eng-101 is a requirement for a 2 business programs, 5 English programs and 3 humanities programs the value for Eng-101 sections would be the sum of these offerings, 10.

Categorical Variables

- ***Day Pattern-*** A unique day offering pattern to each section. If there was no day pattern, then the class was coded as OL for online

- ***Face to Face/Online-*** A binary that indicated if a section modality was Face to Face or online

- ***Academic Term-*** Factor that indicates if the section offering was in the Fall, Winter, Spring, or Summer term

- ***CTE -*** A binary variable that indicates if the section is classified as Career and Technical Education or not

- ***Course-*** Every section offering has an associated course. For example, section Eng-101-01 has a course of Eng-101

- ***Dept-*** What department the section/course belong to. For example, Mat -202 is in the Math department

- ***Offering Times-*** This variable is many one hot encodings that indicate if a section had a start time in set hour intervals. The intervals were from 8 am to 8 pm and then anything before or after got its own coding. Sections that have multiple offerings will have multiple codes over their individual session start time

Exploratory Data Analysis

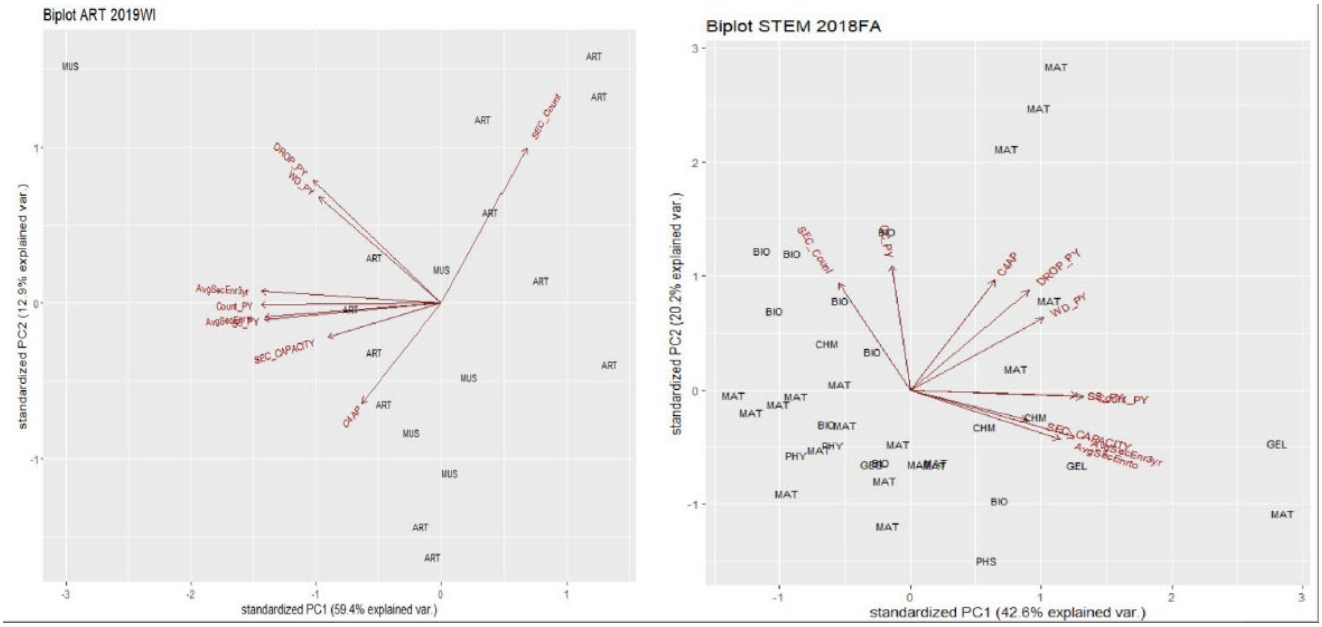

Figure 2. Principle Component Analysis bi-plots on Numeric predictors show which numeric predictors are aligned provide guidance on model selection. This PCA showed that the multiple time series variables; previous year enrollment, 3-year enrollment average and total enrollment average all had arrows that faced in near similar direction on PCA bi-plots suggesting that they provide equal signal to the model.

Figure 2 shows Principle Component bi-plots for the numeric predictors in the dataset. A visual PCA analysis was done to make model selection choices. Each arrow in the plot represents the direction the given predictor lies in the space of the two principle components that account for the largest portion of the variance in the data. If a subset of numeric predictors aligns on the two PC space, this provides evidence that each predictor in the subset could be correlated with one another and thus provide and equal amount of signal to the given models. From the bi-plots it can be seen that previous year enrollment, three-year average, and total average have similar directions. In order to simplify the model, it may be beneficial to leave one or two of the three mentioned predictors out.

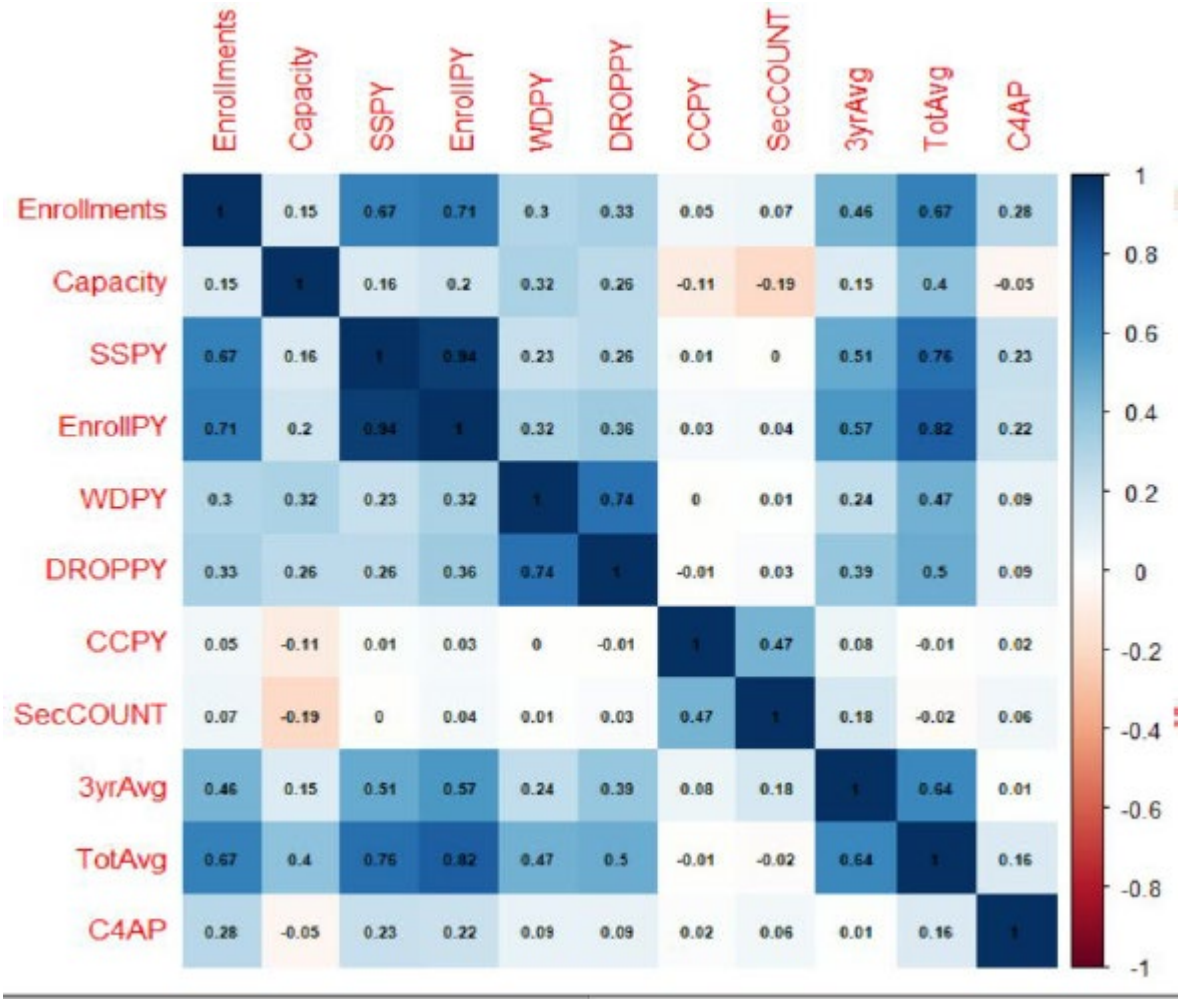

Figure 3. Numeric Predictor Correlations show the correlation coefficient between all numeric predictors. This plot provides a preliminary look into the relationship of the response variable to the numeric predictor space and also shows comparisons of numeric predictors with one another.

Figure 3 shows the correlation matrix of all the numeric variables, including the response which is in the first position of the matrix. The correlation matrix shows that the enrollments are correlated strongly with previous year enrollment and the total average enrollment. Since there is less correlation between enrollment and the three-year average, the three-year average predictor was dropped during model comparison.

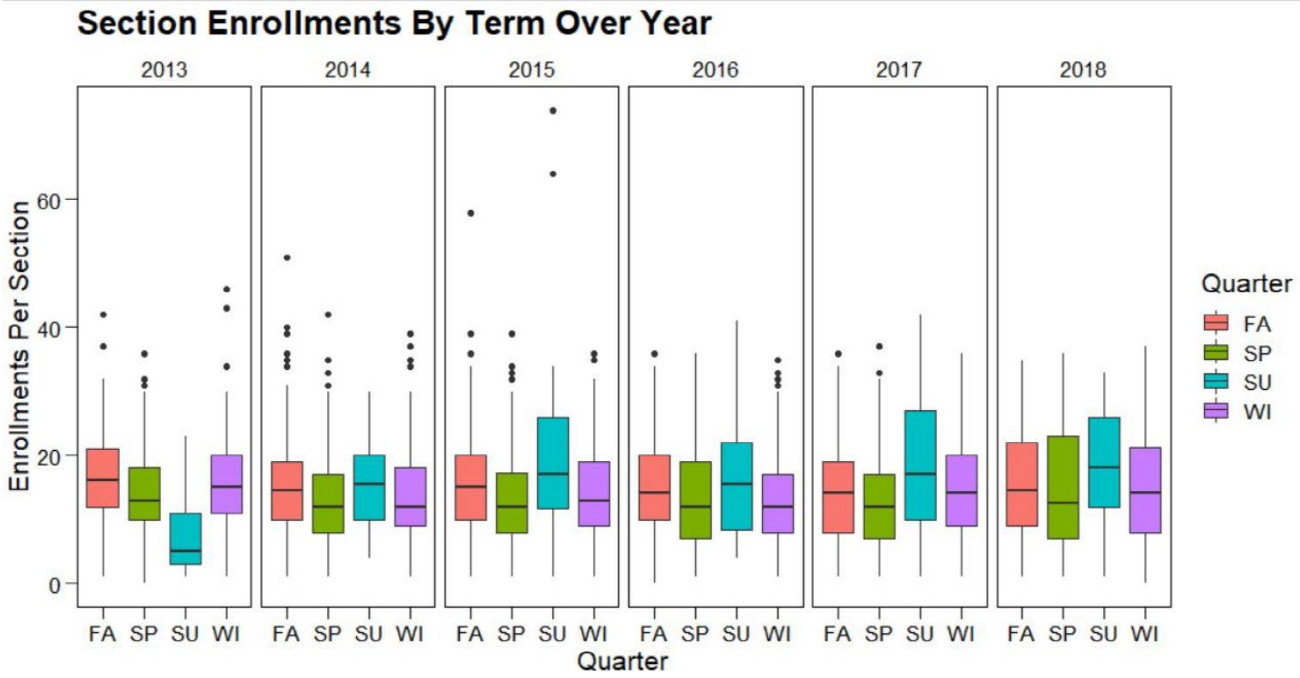

Figure 4. Boxplots showing the average enrollment count per section in the 4 terms over each academic year from 2013-2018. This plot shows that in each term the average section enrollment remains fairly consistent year on year with the exception of some summer section offerings.

Figure 4 shows the distribution of the response variable, active student enrollment broken out by term and then year. Each quarter appears to be mostly consistent year on year with the exception of summer. The distribution across each term does look pretty equal in the Winter Spring and Fall.

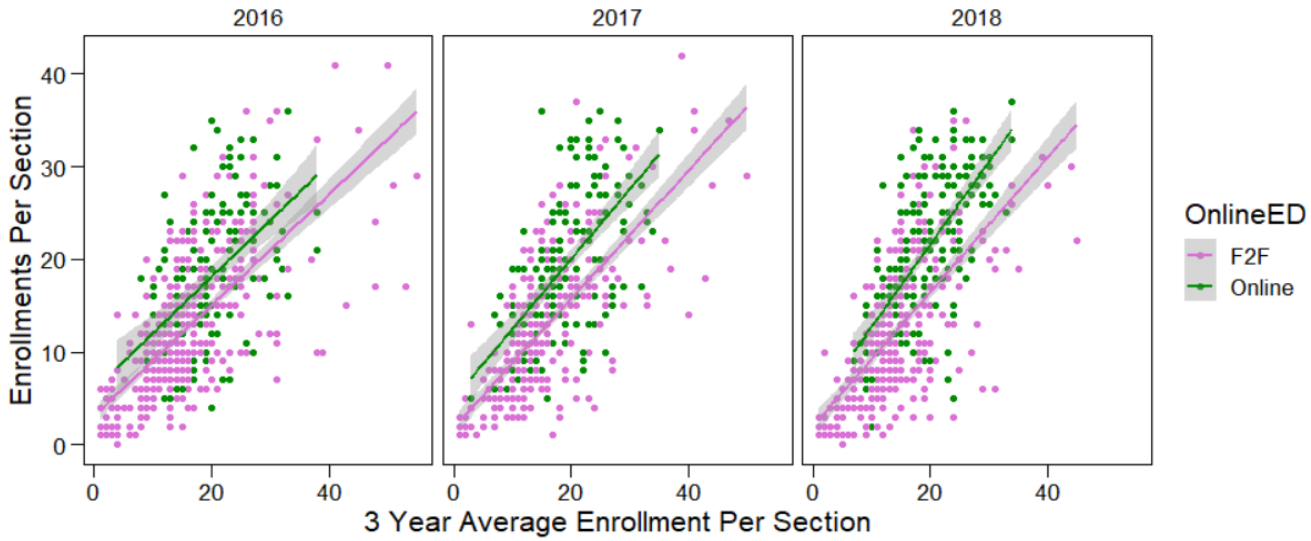

Figure 5. Simple linear regression models of Active Enrollment on 3-year average broken out by the online vs face to face sections show a clear difference between the modalities

Figure 5 is a visualization of a simple linear regression fit of the response variable, section enrollment on the previous three-year average for both Online and Face to Face modalities. There appears to be a noticeable difference in the equation describing the line of best fit for each modality. The slopes for each modality appear to be similar, however the online modality seems to have a noticeable shift suggesting that there is a modality effect that the forecasting models can use to make more accurate predictions.

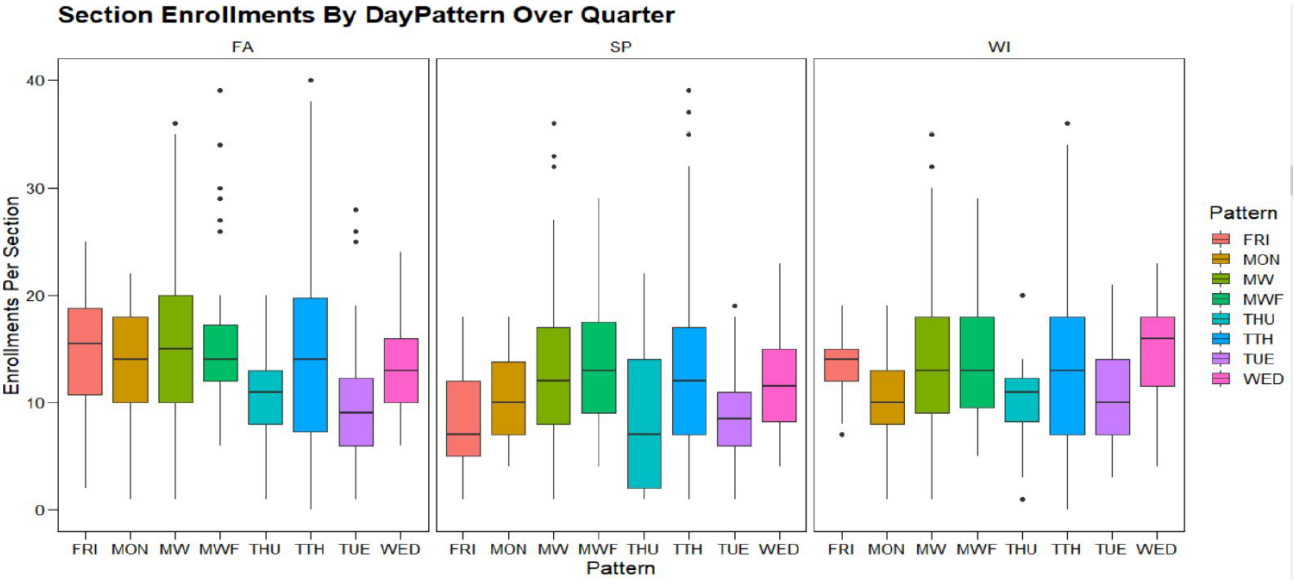

Figure 6. Boxplots of section enrollments of some of the more common day offering patterns in the dataset over the 3 primary terms (Fall, Winter, Spring)

Figure 6 shows the distribution of enrollments across primary terms and 8 of the 69 day patterns. The most common patterns at LTCC are the MWF, MW, and TTH pattern. Over each term these day patterns show a consistent distribution of active enrollment count. In the Fall term the TTH pattern looks to have a much wider distribution than the MWF pattern. This suggests that there may be an interaction effect between term and day pattern that could be engineered into the predictive models.

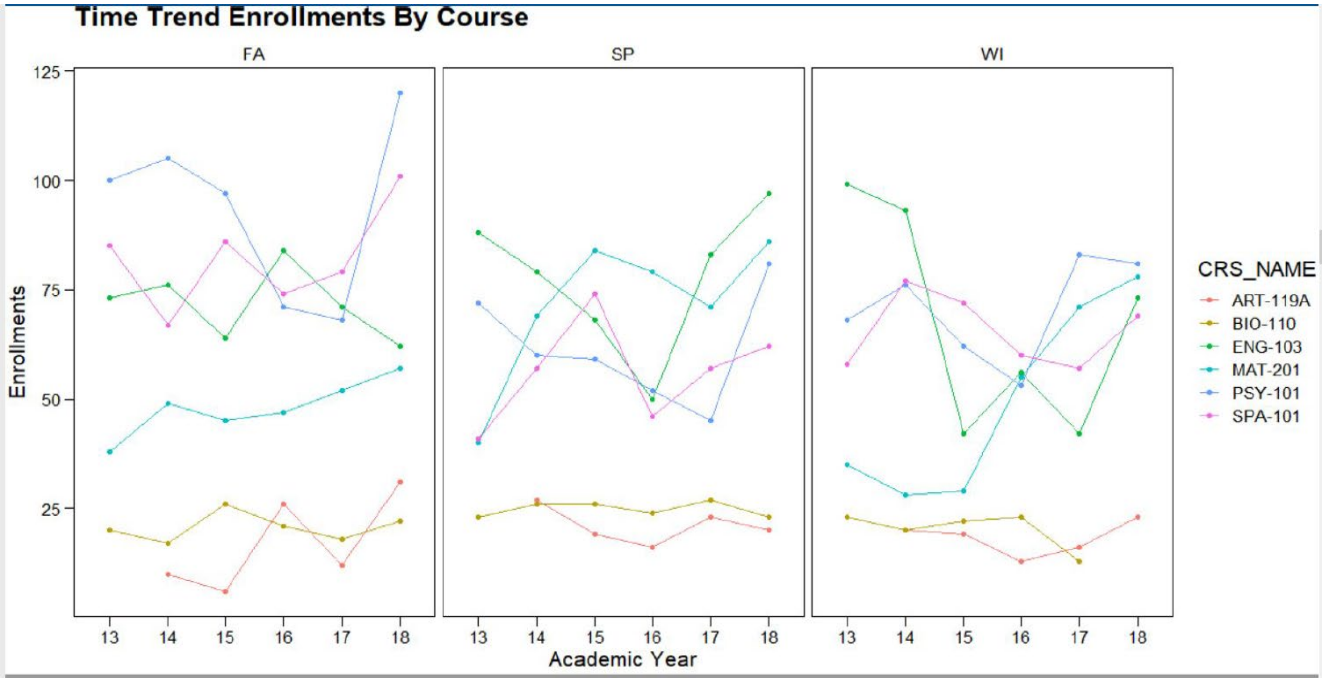

Figure 7. Time series trends of the enrollments aggregated at the course level for 7 sample course offerings

Figure 7 shows the time series trend of the course enrollments from term to term over the set of academic years. It is important to note that in this plot the Y axis is not the section enrollment, but the aggregated enrollments across course level offerings. In the predictive models the course will be used as a categorical predictor with over 150 factors. This plot shows some of the more popular course offerings. The plot shows that BIO-110 and ART-119A have consistent enrollments over primary terms and across the academic years. Eng-103 appears to

have a high variability in the Spring and Winter terms, but is more consistent in the Fall. We also see that MAT-201, PSY-101, ENG-101, and SPA-101 have higher enrollment counts than ART-119A and BIO-110. This is likely due to the fact that the latter courses are more major specific while the former are General –ED courses that are required in most academic programs offered at LTCC.

Preliminary Modelling and Variable Importance

After exploring the variables using data visualization techniques preliminary models were run on the entire dataset to assess the importance of the predictors. This was performed by analyzing the coefficients of the LASSO regression model and the variable importance of the Random Forest Regression.

Lasso Regression

The LASSO regression model creation process is similar to that of traditional linear regression model with a slight adjustment. In the traditional linear regression process the coefficients on each predictor are determined by finding the values that minimize the RSS of the model. The LASSO approach introduces a lambda tuning parameter that attempts to limit the inflation of model coefficients. This is shown mathematically in the equation below. Using the L1 penalty of the LASSO allows for the model to zero out variables that may not provide value to the model. This serves as a built in version of model selection. Looking at which variables had their coefficients eliminated by the lasso can help determine which predictors are providing signal to the model.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|. \quad (4)$$

As a general note the LASSO was used as a candidate model over traditional linear regression because of the LASSO's ability to handle a greater number of predictors. In the LTCC dataset many of the predictor variables have many levels and each level needs to have a one hot encoding that greatly increases the size of the predictor space.

A LASSO regression model was created using the entire data set. First an optimal lambda cost parameter was determined via cross validation. This value was determined to be 0.00387. In figure 8 the MSE of the cross validation process is plotted as a function of the log of the lambda parameter.

Looking at the coefficients of the LASSO model 245 predictors were completely eliminated. The only numeric variable that was eliminated was the number of cancelled sections the previous year. The other 244 were various levels of categorical predictors. Many of the individual courses were eliminated as well as less common day and time patterns.

The LASSO model did not zero out 473 variable coefficients. The values of the coefficients ranged from -2.13 to +1.46. Many of these variables were individual courses and day offering patterns. Interestingly, the minimum valued

coefficient was on DAN-131AE a dance course and the maximum was for the PEA-100A course. This means the LASSO model will add 1.46 to the prediction of the enrollment when it sees PEA-100A and subtract 2.13 when it sees a section of DAN-131AE.

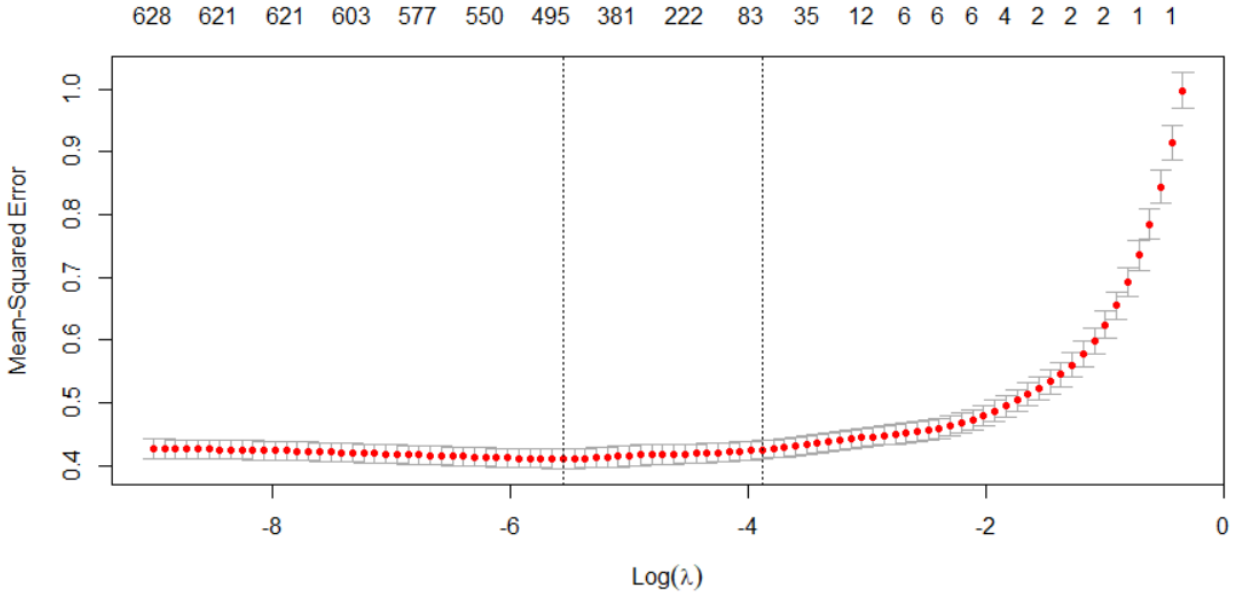

Figure 8 CV MSE score versus log of the lambda parameter in the Lasso Regression

Random Forest

The Random Forest Regression model is made by aggregating many decision trees made from a subset of the total dataset. In each tree of the Random Forest a randomly selected subset of the total predictor space is used to create a single decision tree. A commonly used practice is to set the size of this subset by taking the square root of the total number of predictors. In the Random Forest Regression setting this allows for the calculation of a percent increase in Mean Squared Error (MSE) for the trees in the forest that do not contain a particular predictor. This percent increase in MSE represents the average increase to the MSE on the out of bag samples of the data for the trees in which the given variable was left out. In other words, the percent increase in MSE aims to show how much worse the model would do without the given predictor. A predictor with a larger percent increase in MSE means that it has more importance in creating an accurate model than a predictor with a smaller percent increase in MSE. This allows the

analyst to have insight of the value of each predictor and guide the process of model selection.

A Random Forest was created on the entire data set consisting of 1300 trees. It is important to note a computational limitation to this analysis. The course variable was passed into the Random Forest with a pre one hot encoding done for each level of the variable. This is due to the limitation of the Random Forest function in R that has a limit on the number of levels of a factor. Figure 9 shows the results of the top 10 variables from the variable importance analysis.

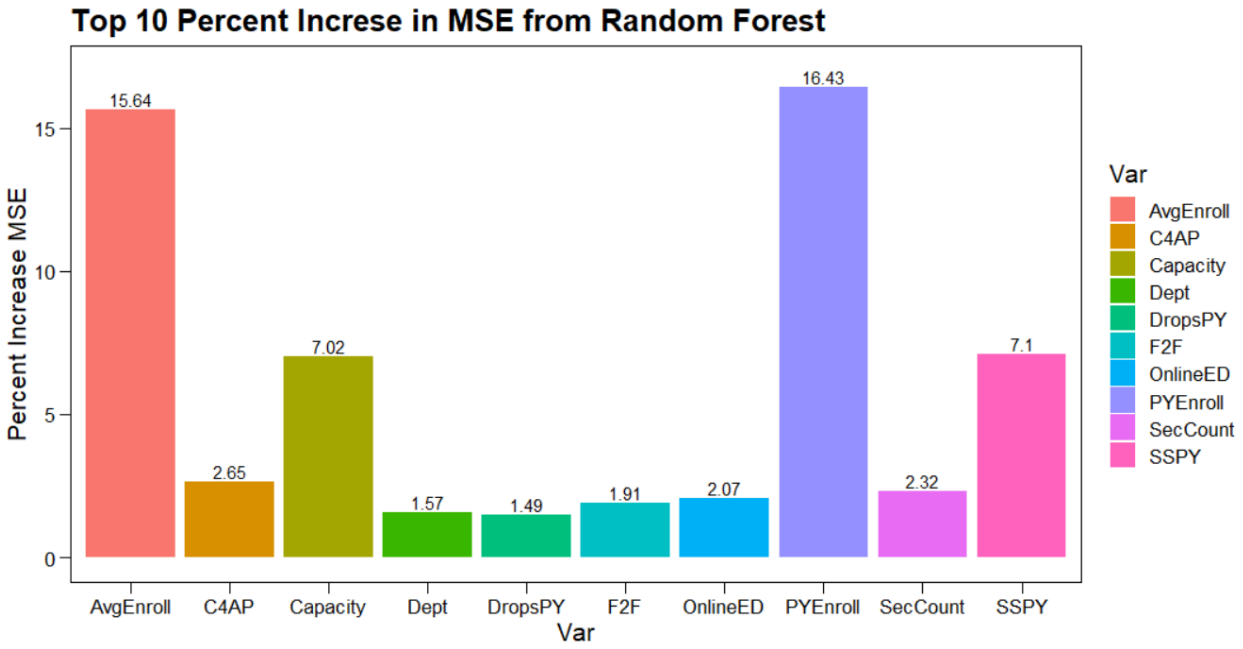

Figure 9. Bar plot of Variable Importance Increase in MSE for Random Forest model of top 10 predictors by percent increase in MSE

From Figure 9 we can see the TOP ten predictors that had the biggest Percent Increase in MSE. The total average enrollment and previous year enrollment turned out to have the highest values of percent MSE increase. Figure 9 also shows many of the other numeric predictors landed in the top 10 of variable importance. We can see that the categorical predictors 'Department' and 'OnlineED' were in the top 10 in variable importance as well. This finding provides further evidence that modality can add some predictive power to modelling efforts as was also seen in exploratory analysis.

After performing exploratory analysis and analyzing variable importance in preliminary it was decided that the 3-year rolling average predictor would be

removed from the model comparison stage. This was because of it showing high correlation with both the previous year's enrollment as well as the total average enrollment. It was also determined to remove the department level categorical variable as because it is a superset of the course variable. These were the only variables that were removed and after the one hot coding of all the categorical variables there were a total of 746 predictors.

Model Comparison

Four different models were used for model comparison. The models used were LASSO Regression, Random Forest, Tree Based Gradient Boost and a Feed Forward Neural Network. Since the goal is to be able to implement this on future versions of academic schedules prior to an enrollment the data was split and trained on the academic years 2012-2017. Then every term in the 2018 academic year was used as the test set of the data. The Mean Squared Error of each model on the unseen test data was recorded and compared on both a term by term and aggregated basis. Then an ensemble model was created using the average prediction of the Random Forest, Gradient Boost, and Lasso Regression. The MSE was then scored for the Ensemble model on all 4 terms as well.

Preliminary cross validation was run to optimize model tuning parameters for the tree based models and the Neural Network. The Neural Network performed best with 2 hidden layers and 2 output layers. The Random Forest did noticeably better after 1300 trees were created. The gradient boost did best by setting the eta learning rate parameter to 0.4 and running 160 iterations of tree boosting. The LASSO model did not need preliminary parameter tuning as the 'glmnet' package in R allows for a grid search of lambda parameters to be done when making predictions. By default, the prediction of the LASSO gives 100 different versions with varying lambda values. For each test the prediction of the LASSO was the version that minimized the MSE when compared to the other 99.

During testing, the Neural Net was wildly inconsistent on some test observations. This is due to the random nature of the starting guess for parameters used in the optimization steps of the Neural Net algorithm. This resulted in differing results when the Neural Net would be run on the same data. Also, while the Neural Net would do well on many of the observations, sometimes it would predict negative enrollments for a section and occasionally it would predict values past the listed capacity. This was dealt with by converting the predictions from the Neural Net that were negative to 0 and capping the maximum of the prediction with the section capacity.

Figures 10-13 below show a term by term analysis of how each of the models performed in each term of the 2018 academic year. Each plot represents the term the models made predictions on and each quadrant plots the actual observed enrollment counts on the vertical axis vs the model prediction on the horizontal axis. The diagonal line represents where the model would perfectly match the true

observation. In the subtitle in each quadrant, the Mean Squared Error is displayed for each model on the given term. The scatterplot points are color coded by absolute residual error on a gradient color scale with the darker blue points having a smaller error while the lighter points have a higher error. The course names are displayed for data points that are in the top 5% of errors. This was done in order to see if the same courses posed problems across the models.

Figure 14 displays the distribution of the errors for each model over the entire 2018 academic year in a histogram.

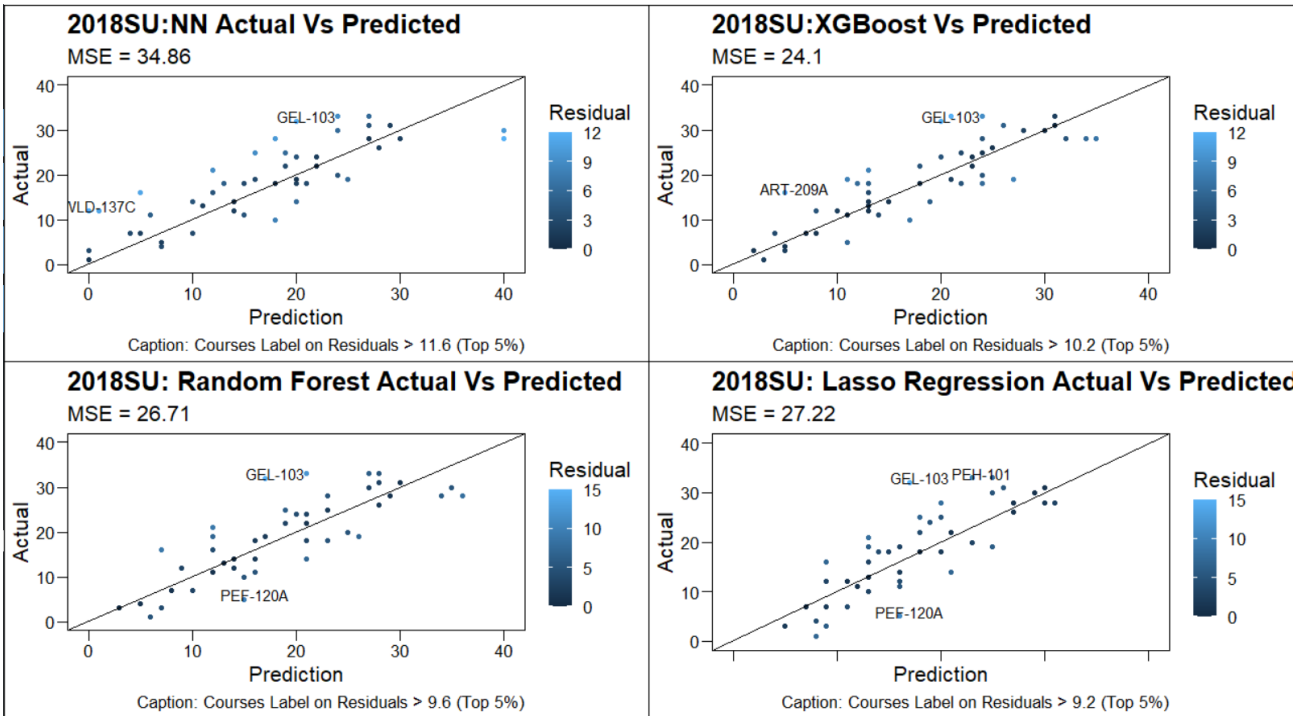

Figure 10. Comparison of four model Actual Vs Prediction on 2018 Summer data

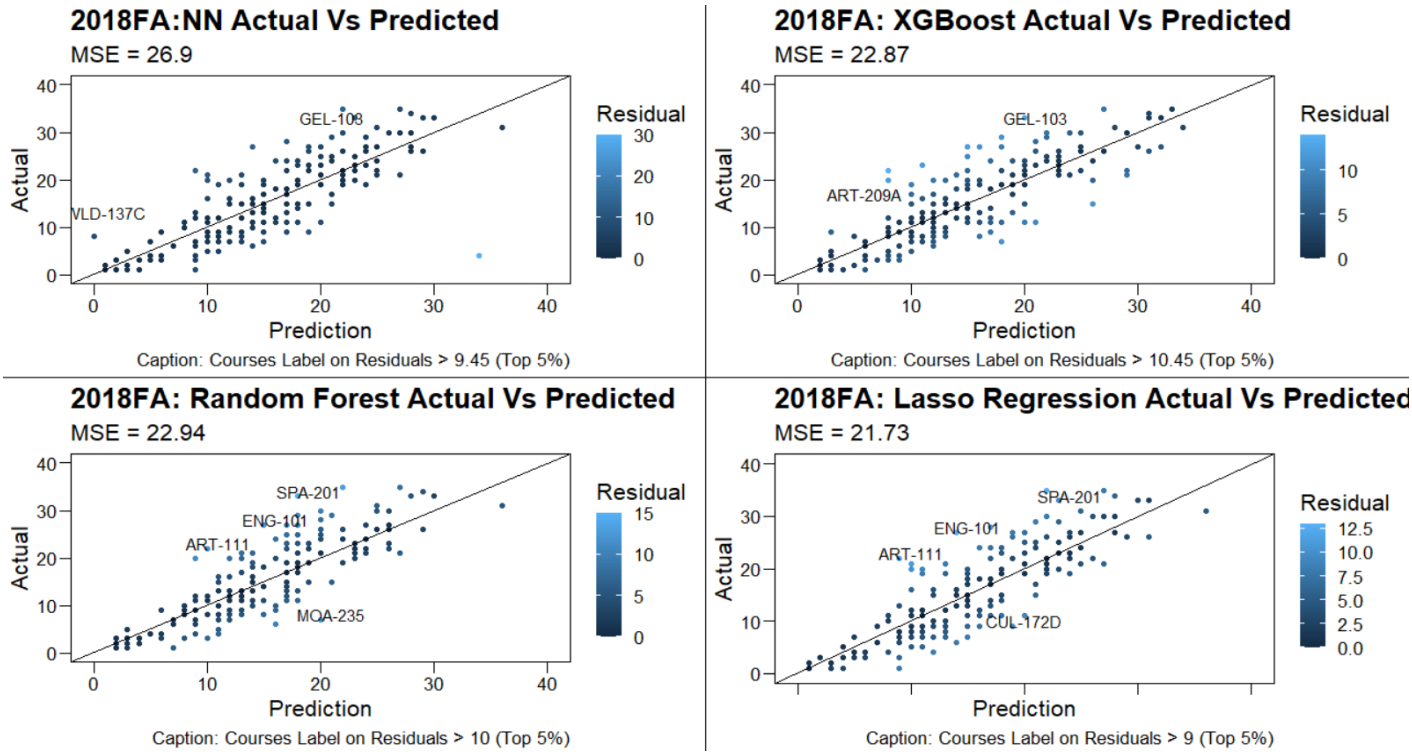

Figure 11. Comparison of four model Actual Vs Prediction on 2018 Fall term

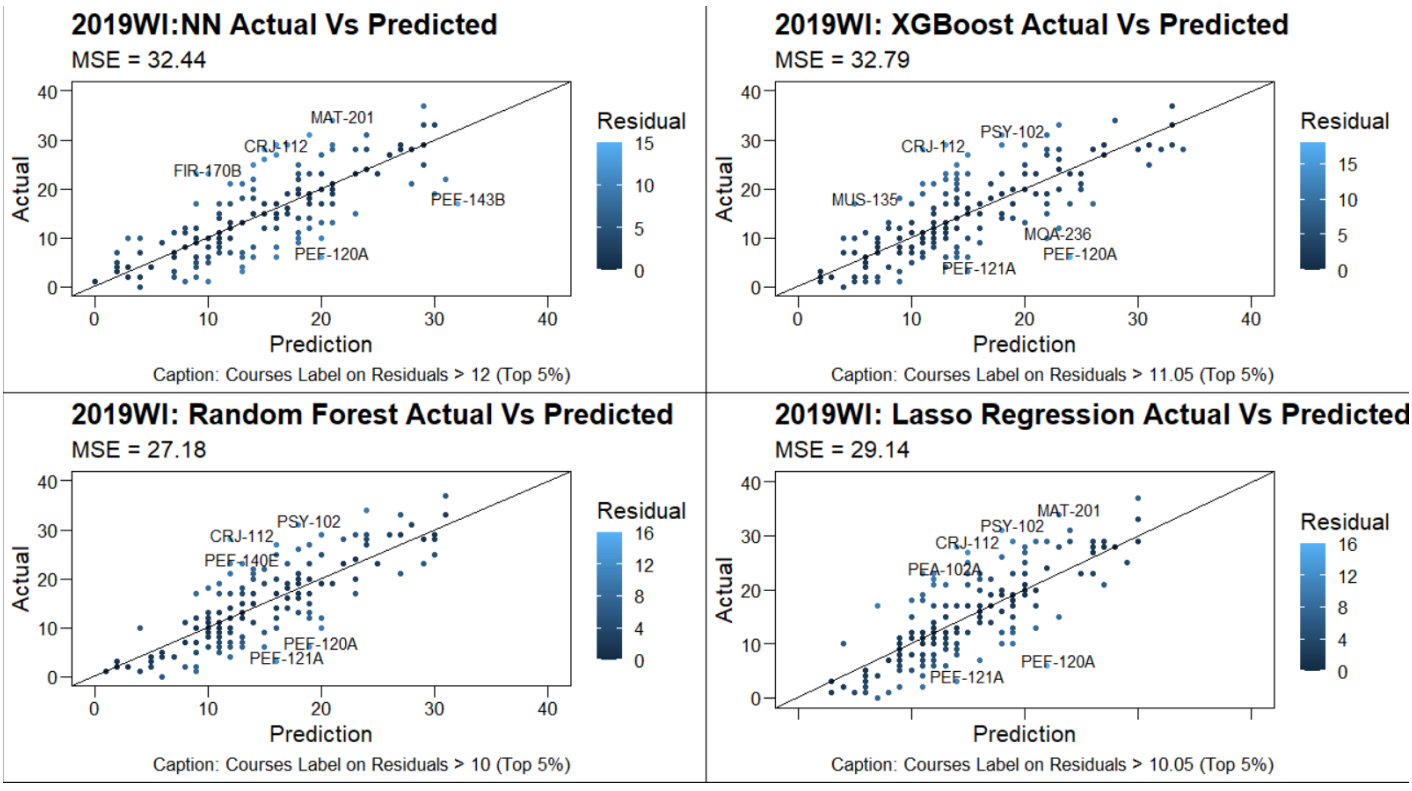

Figure 12. Comparison of four model Actual Vs Prediction on 2019 Winter term

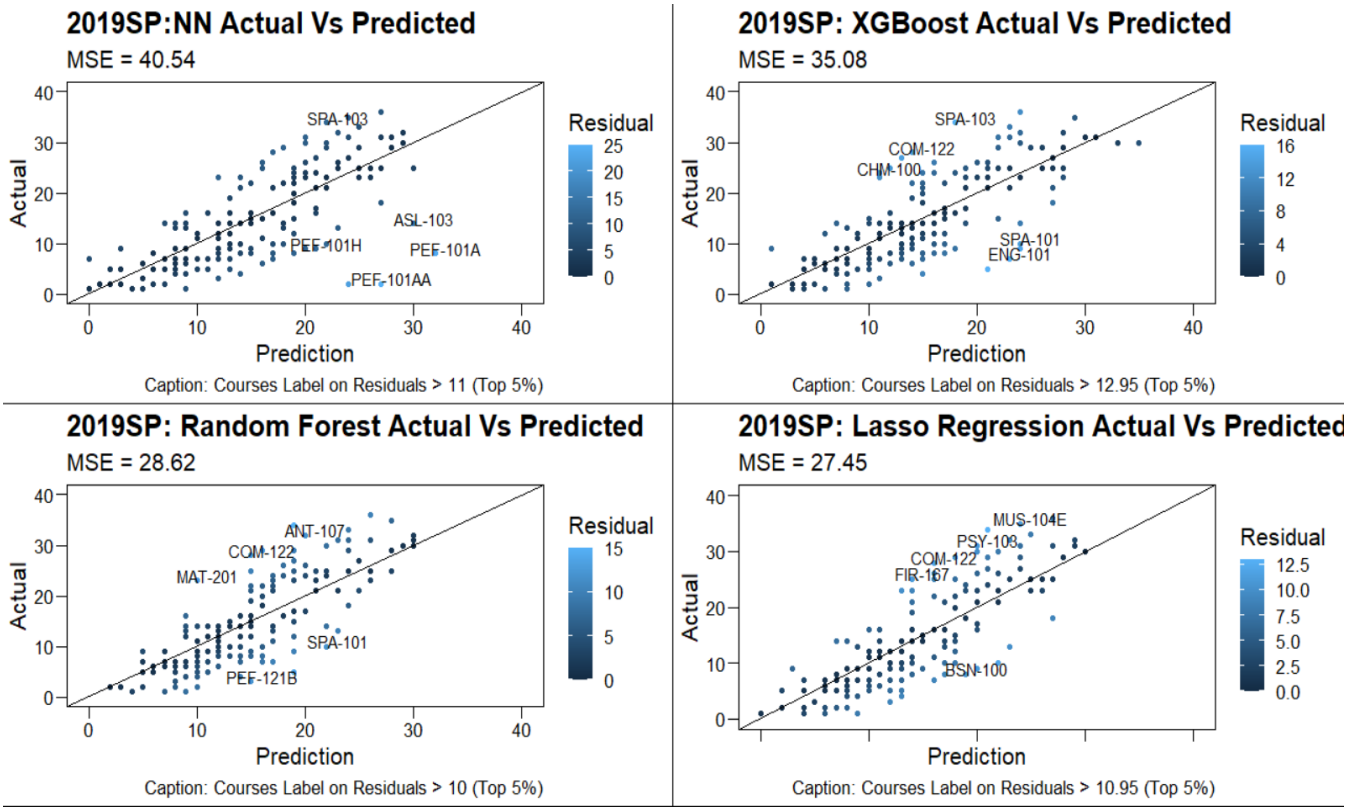

Figure 13. Comparison of four model Actual Vs Prediction on 2019 Spring term

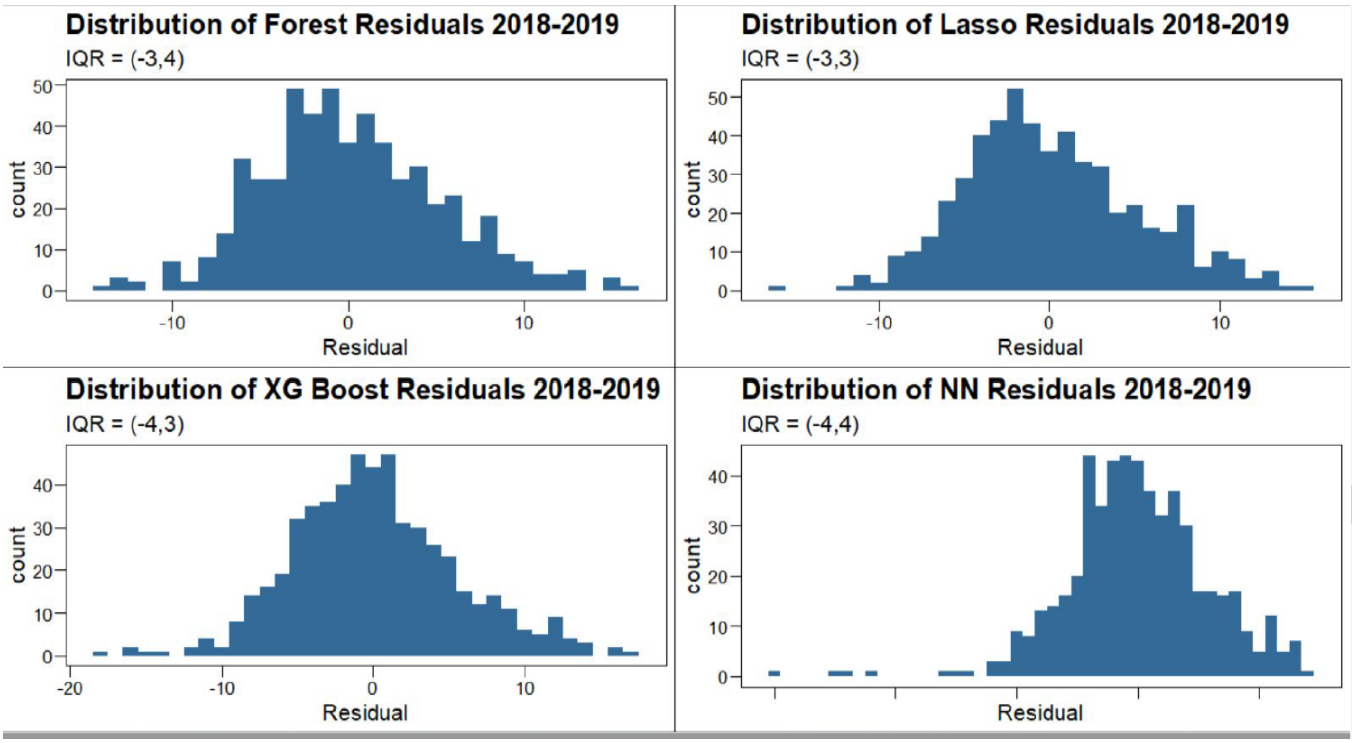

Figure 14. Comparison of four model residuals over the entire academic year

In the 2018 Summer the gradient boost model performed the best with an MSE of 24.1. Although, LASSO and Random Forest were very close with an MSE of 27.22 and 26.71 respectively. Each of the 4 models seemed to have problems predicting the GEL-103 course as its residual was in the top 95% of each model. Also, it is above the diagonal perfect prediction line in each plot suggesting that each of the four models under predicted the course.

In the 2018 Fall term each model posted their best scores. The LASSO had the best MSE with 21.73 with Gradient Boost and Random Forest not far behind at 22.87 and 22.94 respectively. The Neural Network also did its best with a MSE of 26.9 The LASSO and the Random Forest both under predicted the SPA-201, ENG-101, and ART-111 courses, while the Gradient Boost and Neural Network under predicted the GEL-103 course.

In the 2019 Winter Term, Random Forest performed best with an MSE of 27.18. The LASSO Regression was very close with and MSE of 29.14. Each model over predicted the PEF-120A course and under predicted CRJ-112. We also see further similarities between the LASSO and the Random Forest as the had PSY-102, CRJ-112, PEF-120A, and PEF-121A in their top 5 % of residuals.

In the 2019 Spring Term LASSO Regression did the best with an MSE of 27.45 and Random Forest was a close second with and MSE of 28.62. Gradient Boost, Random Forest, and LASSO Regression all under predicted COM-122 to the point in wind up in the top 5% of absolute error in each model.

In Figure 14 the distribution over the entire academic year is shown for each individual model. LASSO, Random Forest, and Gradient Boost all show ideal residual behavior as they have unimodal error distributions that are centered around 0. The Neural Network shows similar behavior however there are a handful of residuals that were extremely negative. Figure 14 also displays the Interquartile Range for the errors of each model. This interval represents the range in which 50% of the residuals lie. LASSO Regression posts the best range as it is the smallest at (-3,3).

The residual analysis plots in figures 10-13 show that Random Forest and LASSO regression perform the best over the four terms. Their MSES were very close to one another and it can also be seen their bigger residual misses were on similar courses. Since figure 14 shows the LASSO regression model to a slightly smaller variance than the Random Forest, I conclude that the LASSO was the best single model for predicting section level enrollments.

After analyzing each single model an ensemble prediction was made by averaging the predictions of the LASSO, Random Forest and Gradient Boost. Figure 15 shows how the ensemble performed on each of the four terms in the academic year and Figure 16 illustrates the ensemble's residuals over the whole year.

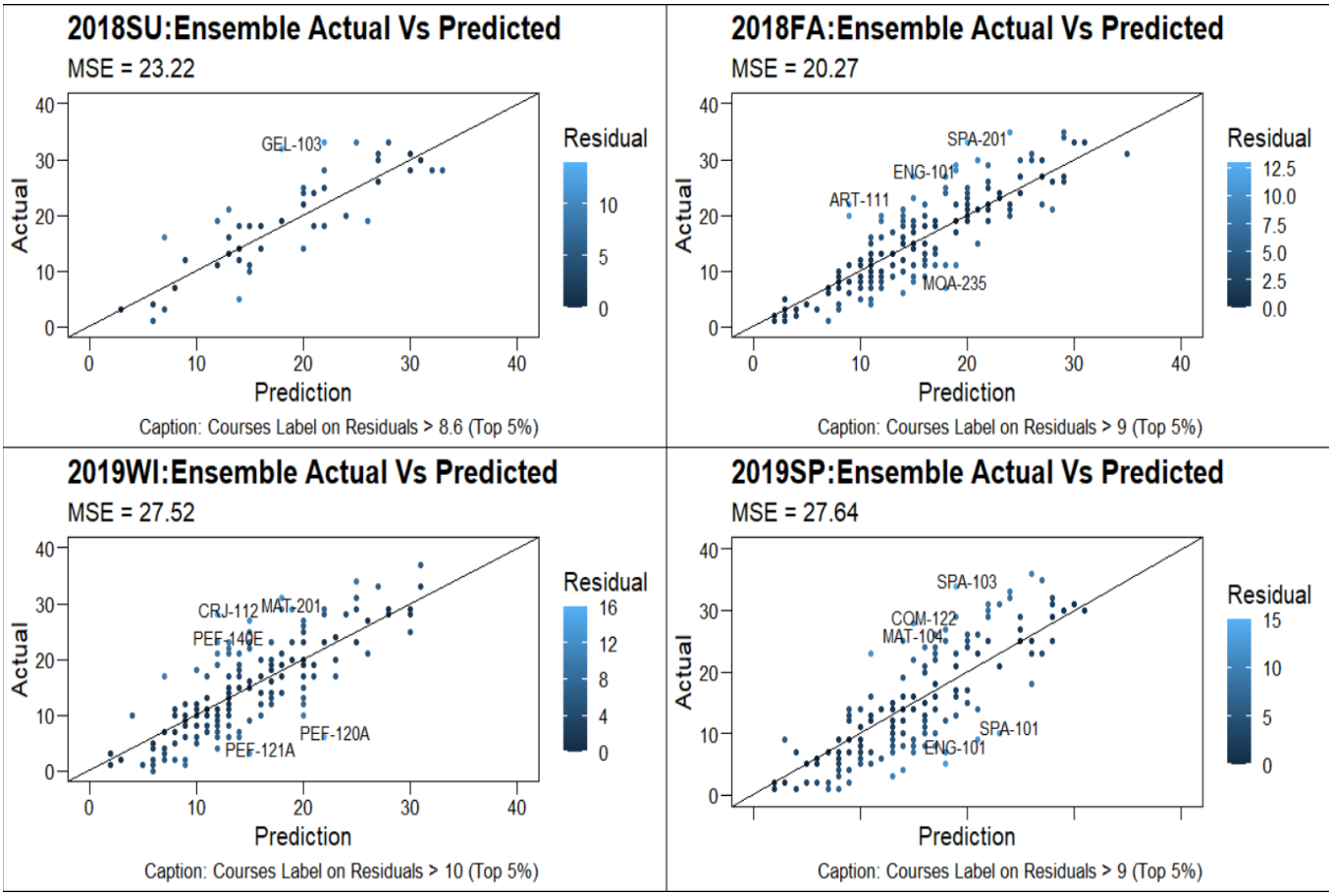

Figure 15. Comparison of Ensemble model across all academic terms

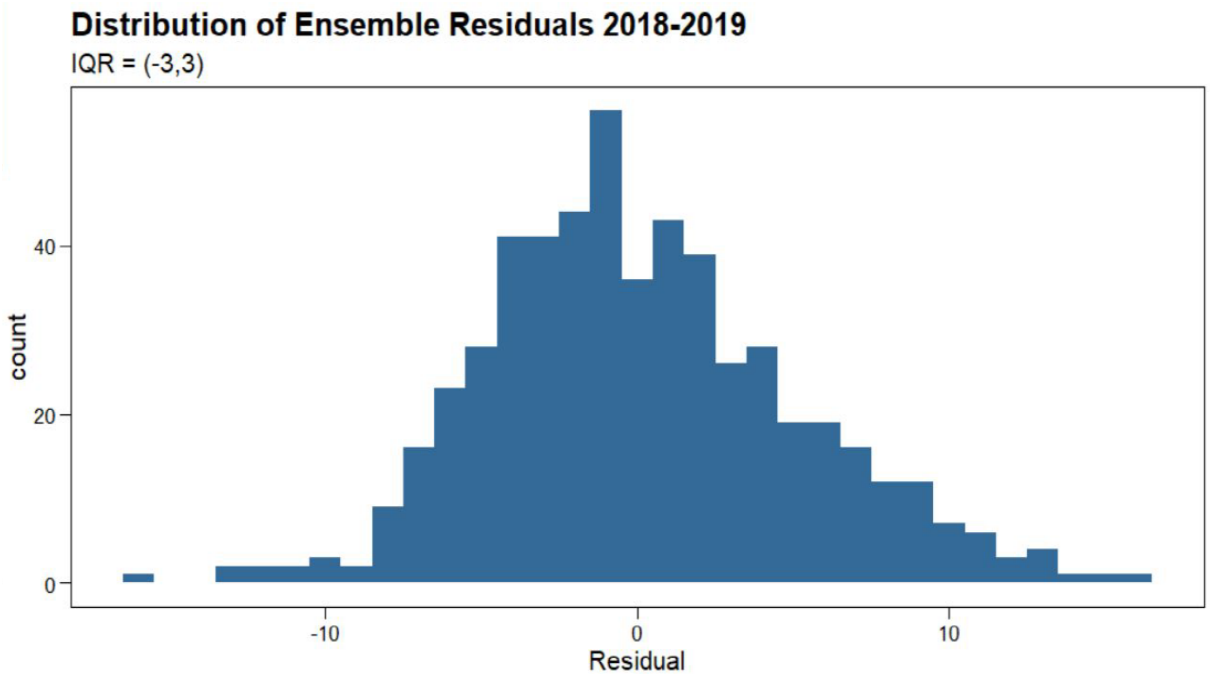

Figure 16. Residual of Ensemble predictions over 2018 academic year

The ensemble model appeared to outperform each individual model. In the 2018 Fall, the ensemble predictions posted the lowest MSE score of all models over all terms with a value of 20.27. The ensemble also did the best MSE score for spring and summer terms and only was a few tenths over the score Random Forest posted on the winter term. From Figure 16, it is also shown that the residuals for the ensemble model have ideal behavior as they are unimodal and centered around 0. The ensemble posted and error IQR of (-3,3) and the variance looks slightly tighter than that of the LASSO regression.

Conclusions and Future Exploration

From the exploratory analysis, variable importance analysis, and model comparison it appears that the strongest predictors of section enrollment are previous year enrollments, total average enrollments, course modality, and course name. None of the date and time variables seemed to show up in any of the aforementioned analysis as standout predictors. However, in future analysis it may be worthwhile to explore the effect of interactions between course, date, and time patterns as a way to add some more signal to the model.

Using the 2018 Academic Year as a test data set, the LASSO, Random Forest and Gradient Boost model all performed relatively well in predicting section level enrollments for Lake Tahoe Community College sections. However, given its performance on the 2018 Academic Year test set, the ensemble model will be the best choice moving forward with implementation of a forecasting system to aide in the academic scheduling process.

References

[1] Dr. Chau-Kuang Chen, "An Integrated Enrollment Forecast" *IR Applications Using Advanced Tools Techniques and Methodologies* Volume 15 January 2008

https://www.researchgate.net/profile/Chau_Kuang_Chen2/publication/280051773_An_Integrated_Enrollment_Forecast_Model/links/55a68ed008ae51639c572ec1/An-Integrated-Enrollment-Forecast-Model.pdf

[2] Hanover Research Group "Forecasting Community College Enrollment" *Hanover Research* October 2016

https://www.imperial.edu/docs/research-planning/7931-forecasting-community-college-enrollment/file

[3] Rabby Q. Lavilles and Mary Jane B. Arcilla, "Enrollment Forecasting for School Management System" *International Journal of Modeling and Optimization,* Vol 2, No.5, October 2012

http://www.ijmo.org/papers/183-E018.pdf

[4] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.