INTERNSHIP REPORT

Student Enrollment Prediction using Machine learning

MIS 64092 - Internship in Business Analytics

Abhishek Sau

Table of contents

1. Executive Summary

Higher education institutions are moving towards machine learning algorithms that can help them identify and predict student enrollments. The results from the enrollment prediction algorithms can be used in multiple ways like based on the likelihood of a student to enroll, university can decide if scholarship should be awarded to the student or not. However, in this paper, we will only discuss the application of machine learning algorithms to predict how likely prospective students are to enroll and try to identify what attributes to enrollment of a student. This paper is based on algorithms used on the data procured from Kent State University (KSU)'s Enrollment Services, post-pandemic.

The goal of this internship was to develop a classification model to predict the likelihood of a student enrolling in the undergraduate program of KSU. More importantly, generation of actionable insights and identification of key attributes was an important part of this paper.

For this study, I merged all the data from the four different files based on the 'ID' attribute, and used an inner join to retain only the common rows. Primarily, I identified the target variable, TMST_15, i.e. 0 if the student enrolled in an undergraduate program else, 1. I noticed that the classes are imbalanced, i.e. there are more students who did not enroll than the ones who enrolled. Secondly, I carried out data preprocessing to drop duplicate student IDs, highly correlated columns and imputed NULL values. After preprocessing, I used the SMOTE method to address the class imbalance after splitting the data into test and train sets. Once I dealt with the imbalance, I tested 3 different models on the data in order to find the optimal model. I found out that Random forest was the best performing model. The Random forest model trained on this data gave an ROC AUC of 0.925 and precision score of 0.9547.

1. Business Problem

The need to use predictive analytics in college enrollment is much higher now than it used to be. The yield rate of an average college has dropped over 17 percent points over the past decade. One of the reasons is that high school students are applying to more colleges than before. This has considerably increased the competition among educational institutions to convert an admitted student to an enrollee. This competition in turn has led colleges to drive more efforts in terms of recruitment and provide financial aid to more and more students.

This predictive analytics task was created to predict enrollment of students in the upcoming years, based on the data collected related to admissions, Financial Aid provided by the institution, federal financial aid and recruitment drives used by the institution. All the data used has been collected post pandemic in order to cope with the change in enrollment patterns due to COVID-19.

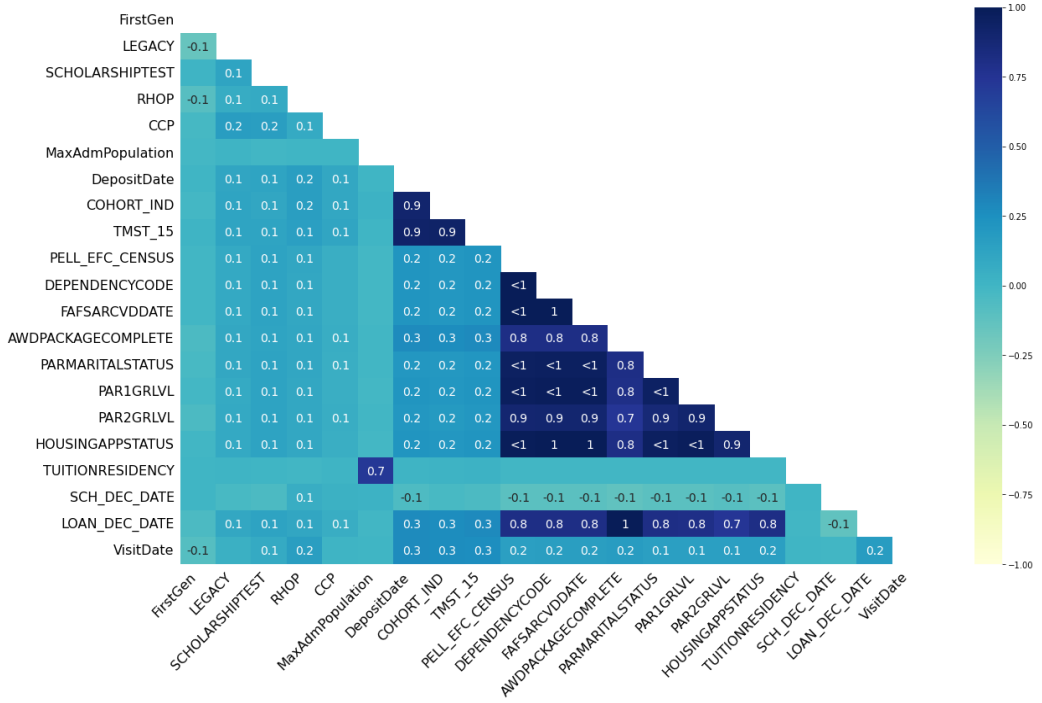
2. Data Preparation

2.1 Data Understanding

For this modeling task, I was given four separate datasets that fetched data from different sources, namely, Admissions data, Financial Data, Funds related Data and Salesforce data.
I merged all these data sets using an inner join on the 'ID' attribute. This merge gave me a training dataset that comprised 15979 rows and 41 columns. This included demographic information about the students, Scholarship, Parents background, High School background, FAFSA records etc.

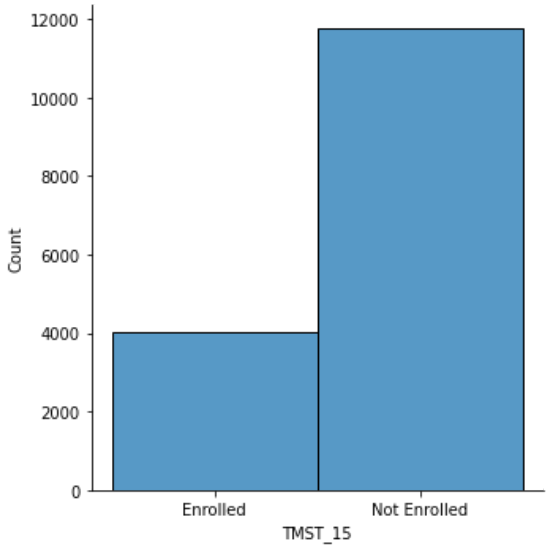
The following gradients were identified in the data upon observation:

1. The data had the majority of missing values of Missing Not At Random(MNAR) type. One such attribute was 'RHOP', which had missing values in it if a student was not admitted to Honors college. If the student was admitted the attribute mentioned 'RHOP'. Processed this attribute to show missing values as 'N' and 'RHOP' was replaced by 'Y'.



The values in the heatmap range between -1 and 1. A value of -1 indicates a negative correspondence: A missing value in feature A implies that there is not a missing value in feature B. Finally, a value of 0 indicates that there is no obvious correspondence between missing values in feature A and missing values in feature B. This is (more or less) the case for all the remaining features. If the value is 1 it means that both feature A and feature B are positive correspondence. The above plot shows that missing values in 'FAFSARCVDATE' have high correspondence with missing values in 'HOUSINGAPPSTATUS'.

2. There was an imbalance in the ratio of the target variable, where the number of students who did not enroll was 11761 and those who enrolled was 4039, after removing duplicates in the original data.



2.2 Target Modeling

The task is to identify the students who are most likely to enroll in an undergraduate program at Kent State university who have been given an admit to one of the courses.

There could be various other possible outcomes in addition to the identification of yield. For example, what are the factors that usually lead a student to not enroll, targeted scholarship given to students to make them more likely to enroll. Predicting enrollment of a student to pursue admission to a college or course, could help the University to properly manage resources and fairly distribute scholarships to underprivileged but deserving students. In a gist, it can be used by the enrollment department to help the university with the budgeting and financial aid planning related to admissions.

2.3 Feature Engineering

After removing the duplicated rows and imputing missing values for most features, I observed that some of the attributes with date format had more than 20% missing values and were also highly correlated to some other attributes.

Then I identified the categorical and numerical attributes and separated them, since they were to be handled differently as mentioned below:

- Numerical features were imputed with the mean of the rest of the values present in the attribute.
- Categorical features had to be one hot encoded and label encoded as needed. One hot Encoding was used on categorical variables which were nominal i.e. did not have an order or preference

associated to them. The categorical variables that had ordinality associated to them were label encoded to maintain the order. The categorical variables which had missing data were imputed using the mode of that column.

3. **Modeling**

I applied the following 3 models:

- Weighted Random Forest
- Weighted Support Vector Machines
- GradientBoostingClassifier

I trained these models on 70% of the data and allotted 30% data for out of sample testing, with 5 fold cross validation, when needed, to optimize the model.

Since the target class had a high imbalance, I used SMOTE (Synthetic Minority Oversampling Technique) methodology to oversample the minority class. The way SMOTE works is that it selects a random instance from the minority class and then uses the k nearest neighbor algorithm to identify its nearest neighbor in the minority class. A synthetic instance is then chosen from this pool of nearest neighbors created and are then connected in the feature space using a line. The synthetic instances are created as a weighted sum of the two chosen instances.

By applying SMOTE to the training set, I created a new train set with an oversampled minority class and used this to train the models. However, I used the original test set to validate the accuracy of the models. Since this is a case of binary classification, I used ROC/AUC score and precision scores as a validation criterion. Accuracy cannot be used as a dependent measure because of the problem of class imbalance.

Below table shows the details of the models used:

Model	Optimized HyperParameters	Performance Evaluation
Weighted Random Forest	bootstrap= True, random_state= 0, class_weight='balanced', criterion= 'gini', max_depth= 13, max_features= 'auto', n_estimators= 200	**ROC/AUC score: 0.9292** Accuracy: 0.9568 F1 score: 0.9714 Recall: 0.9855 **Precision: 0.9576**
Weighted SVM	C=10, class_weight='balanced', gamma=0.01, kernel='poly'	ROC/AUC score: 0.9032 Accuracy: 0.9015 F1 score: 0.9315 Recall: 0.8997 Precision: 0.9656

XGBoost Classifier	criterion='mae', learning_rate=0.01, loss='deviance', max_depth= 8, max_features='sqrt', n_estimators=10, subsample=0.5	ROC/AUC score: 0.9218 Accuracy: 0.9468 F1 score: 0.9646 Recall: 0.9731 Precision: 0.9563

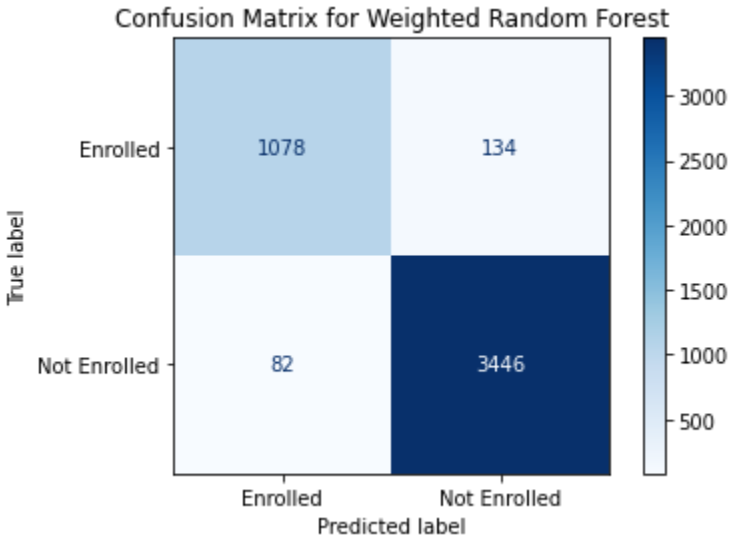
Based on the above table and the Performance validations, I found that the Weighted Random forest model is the optimal model for this task.

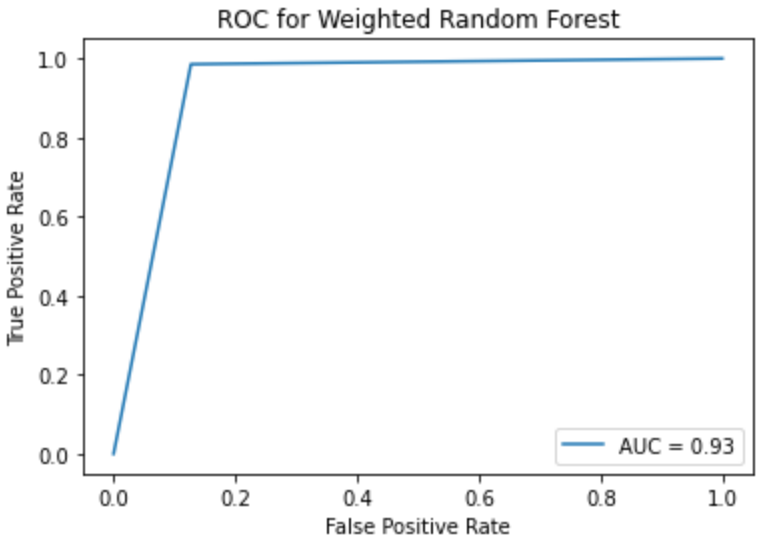
3.1 Optimal Model - Weighted Random Forest

It is an ensemble technique that uses decision trees as base learners. It trains multiple base learners on subsamples created using bagging methodology and averages out the result to improve the prediction accuracy. This bagging helps in creating diversity among the multiple decision trees that are being trained. This algorithm was selected because it performs better in a dataset with severe class imbalance and since it has averaged out the prediction, it also helps control overfitting.

In order to find the optimal value of the hyperparameters for the Random forest model, I used Grid search methodology for cross validation. Since the dataset is not very big, cross validation helped in using the training dataset optimally.

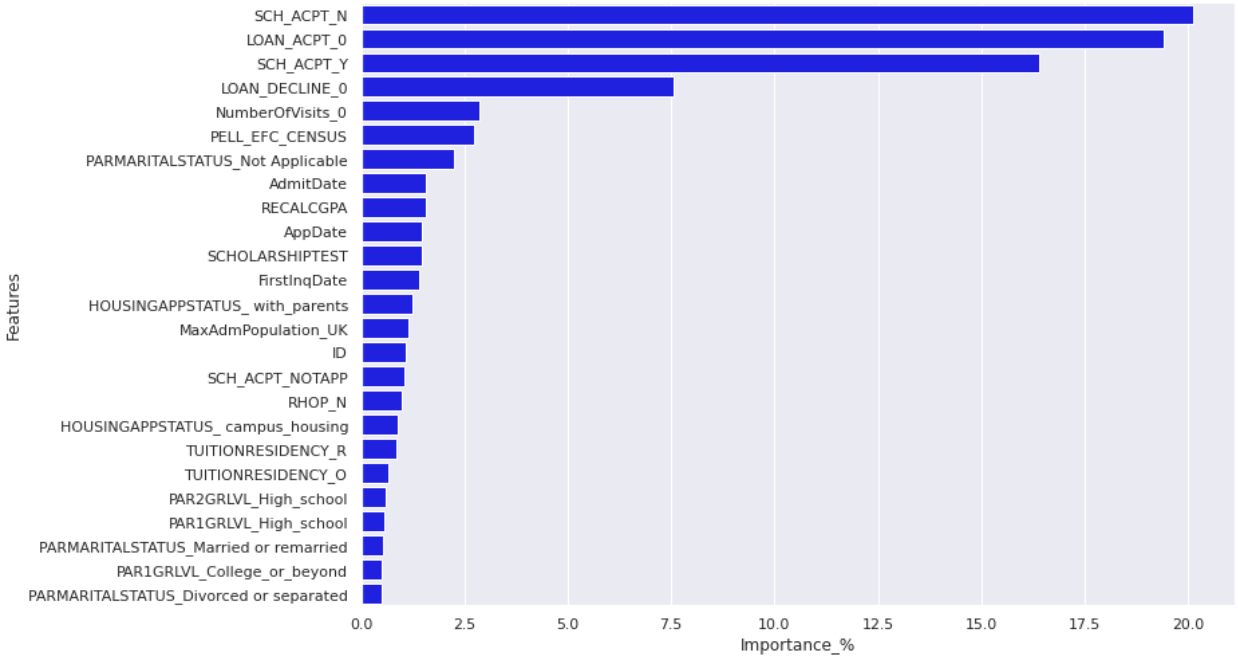
Below mentioned is the confusion matrix and the ROC/AUC curve for the Weighted Random Forest model:





4. Key Performance Indicator Analysis

Random forest has an important attribute associated with it that tells us the importance of features based on the fit. Below mentioned plot shows the importance of features using the feature_importance_ attribute of randomforestclassifier.



In the above plot, the features that have more than 0.5% importance are plotted. By observing the plot, the most important features are Scholarship Acceptance and Loan Acceptance related features. They explain ~64% of the variation in Enrollment.

Another important feature based on the random forest was 'SCH_ACPT_Y' which showed students who accepted the scholarship had a significantly higher impact on enrollment than the students who declined the scholarship.

One of the top 15 attributes is 'TUTIONRESIDENCY_R' which means that the student is a resident of Kent(where the university is located or lives within <70mile radius). This means that the pool for admission will diminish eventually in the local region and will need to increase outreach to different states or locations.

The top 15 attributes in the Feature Importance plot explains ~83% of the variation in the target variable.

	Features	Importance score	Importance_%	Importance_cum
166	SCH_ACPT_N	0.201328	20.132788	20.132788
170	LOAN_ACPT_0	0.193963	19.396345	39.529133
168	SCH_ACPT_Y	0.163906	16.390565	55.919698
171	LOAN_DECLINE_0	0.075651	7.565059	63.484757
205	NumberOfVisits_0	0.028764	2.876406	66.361163
5	PELL_EFC_CENSUS	0.027299	2.729855	69.091018
140	PARMARITALSTATUS_Not Applicable	0.022389	2.238889	71.329907
3	AdmitDate	0.015684	1.568400	72.898307
1	RECALCGPA	0.015637	1.563685	74.461992
2	AppDate	0.014703	1.470325	75.932317
4	SCHOLARSHIPTEST	0.014577	1.457725	77.390041
6	FirstInqDate	0.014210	1.421046	78.811087
161	HOUSINGAPPSTATUS_ with_parents	0.012465	1.246503	80.057590
159	MaxAdmPopulation_UK	0.011387	1.138674	81.196264
0	ID	0.010854	1.085351	82.281615

5. Conclusion

In this paper we shed light on factors that influence the enrollment decision of applicants. We use machine learning methods to measure the level of correlation between enrollment and such factors. In particular we approach the enrollment prediction question from a classification perspective where we need to identify the likelihood of enrollment for a pool of applicants

From the top 5 important features, I observed that 4 of them are related to finances. Hence, financial aid and loans were the most important factors that attracted students to enroll. Students enrolled in a course if they were given financial aid irrespective of their high School GPA. Therefore, financial aid can be used as a controlling factor for increasing the number of incoming students.

The results show that our proposed models can predict enrollment with reliable accuracy using only a small set of features related to student and college characteristics. Another application of this predictive model can assign a percentage to the likelihood of enrollment.

6. References

1. https://machinelearningmastery.com/framework-for-imbalanced-classification-projects/
2. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
3. http://www.chioka.in/why-is-training-set-error-zero-in-svm-and-random-forest/
4. https://mljar.com/blog/how-many-trees-in-random-forest/
5. https://stats.stackexchange.com/questions/189344/choosing-random-forests-parameters
6. https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/
7. https://towardsdatascience.com/heres-what-i-ve-learnt-about-sklearn-resample-ab735ae1abc4
8. https://arxiv.org/ftp/arxiv/papers/2009/2009.00534.pdf
9. https://www.analyticsvidhya.com/blog/2020/11/handling-imbalanced-data-machine-learning-computer-vision-and-nlp/
10. https://machinelearningmastery.com/framework-for-imbalanced-classification-projects/
11. https://towardsdatascience.com/random-forest-for-feature-importance-ea90852b8fc5