Predicting College Dropouts with Survey Responses

Introduction

About a third of students who enroll in a college in the United States end up dropping out (New York Times, 2019). This is a problem for students who dropout - insofar as education increases students' lifetime earnings - as well as society as a whole, because earning a college degree promotes intergenerational mobility (Chetty et al., 2017). This project aims to predict which college students are likely to drop out of college. We have data from EdSights, an EdTech startup, on college students' responses to a survey, as well as whether each student drops out.

The first part of this project will focus on the cleaning of the data and exploratory analysis. The second part describes modeling and evaluation, which is then interpreted in the following section: the descriptive analysis. Finally, we consider ways to improve our survey based on factor analysis.

Business Understanding

Our project adds value to colleges, as student retention increases diversity, improves college rankings, increases funding opportunities for colleges, and reduces loss in revenue associated with students dropping out. Predictive modeling will allow schools to identify which students are going to drop out, enabling them to intervene and increase student retention. Additionally, machine learning will allow us to predict the risk of dropping out before the semester-of, which will give schools a competitive advantage. With descriptive analysis, we can determine the features that play an important role in predicting student dropout. This will make our results easy to communicate to potential clients and facilitate implementation of preventative measures. Lastly, the factor analysis will ensure that the survey improves over time and provides better results.

Data Understanding

Our data comes in the form of survey responses, which are categorical. The survey was initially distributed in the first-year seminar to students. Students who were not in the first-year seminar, received the survey via e-mail. Therefore, we only have data on students who answered the survey (2441 students). We have panel data for students in three universities over the course of 3 semesters (fall 2018 to spring 2019). An instance is a student

from each school described by a feature vector of their responses to survey questions (e.g. I worry a lot about paying for school) for a given semester-year and target variable of whether they dropped out or not. Our target variable is binary, and it equals to 1 if students dropped out and 0 otherwise. Therefore, we have a supervised classification problem.

Data Preparation

Our data contains 2,838 observations and 34 predictor variables. Our data has an imbalanced target variable, since 180 out of 2838 (6.3%) instances have dropout equal to 1. The survey asked two types of questions: 1) Likert scale questions (responses include: Strongly disagree, Disagree, Somewhat agree, Agree, Strongly agree) and 2) categorical questions where students can choose from a small set of responses. Likert scale responses have some notion of ordering amongst the values of the responses. Strongly disagree is closer to somewhat disagree than to strongly agree. Whereas, for nominal categorical responses there is no notion of ordering amongst the responses. For example, for the question 'why did you choose your school?', students could choose from the following six options: 'It was the most affordable option', 'It has a good academic reputation', 'It is close to work/home', 'I like the social scene', 'It offers classes that fit my work schedule', and 'Other'. These responses have no concept of order among them, one of them isn't bigger or better than the other. Therefore, to encode our categorical predictor variables, we used ordinal encoding for Likert scale responses and one-hot encoding for nominal categorical responses.

Not all schools have data for all three semesters; School #1 has data for 3 semesters (Fall 2018, Spring 2019, Fall 2019), school #2 has data for 1 semester (Fall 2019), and school #3 has data for 2 semesters (Spring 2019, Fall 2019). 15 survey questions were asked over all three semesters, while others were only asked for some. Due to this our data contains missing values. To treat missing values for nominal categorical variables, we created another category 'missing' and one-hot encoded this additional category. For missing values for Likert scale attributes, we tested two approaches. We created a dummy variable flagging missing values and 1) replaced the missing values with a neutral response or 2) replaced missing values with the mode of responses. The second approach gave us better results, therefore our final clean data set adhered to this approach. Additional cleaning

was required, such as standardizing the responses (treat 'Strongly Agree' and 'Strongly agree' as the same), combining different spellings, etc. to get our final clean data in the feature vector form with numerical features and no missing values, which most data mining algorithms take as input.

Since our dataset has categorical responses, we cannot use covariance or correlation to visualize association between different variables. Therefore, we used Cramer's V to measure relationships between categorical variables in our data. Cramer's V equals 0 when there is no relationship between the two variables and equals to 1 if there is a strong relationship. Thus, it compares the strength of association between contingency tables. Table 1 below shows the top five variables that have the strongest relationship with the target variable, ranked according to this metric. As observed, none of them have a very strong relationship with the target variable. And thus, the problem is not trivial.

Table 1

feature_1	feature_2	cramers
dropout	didyoufailoneormoreofyourcourses	0.361742
dropout	iamconsideringtakingabreakfromsc	0.185032
dropout	iampayingforcollegeoutofpocketbu	0.173752
dropout	whydidyouchoosethisschool	0.140478
dropout	istruggletopayfortextbooksrentut	0.118737

The graph in Appendix A below displays Cramer's V results for all variables, none of which have a strong association with each other or the target variable 'dropout'.

Modeling & Evaluation

As our baseline model, we ran a decision tree on the survey questions that all students have in common using a simple label encoder. The results can be found in Appendix B.1 (56% mean AUC, 0.18% recall). This performs slightly better than randomly guessing the class. Our goal was then to improve upon this model, first by cleaning the whole dataset as described above and then doing the work described below.

The first step of our prediction work was to choose a set of models to run on our cleaned dataset. We were looking for simple models, as interpretability is important to run descriptive analysis and provide relevant feedback to schools. We had a small data set with many missing values, so we wanted a model which guarantees good generalization by reducing variance. Therefore, we chose to run Random Forest, Naïve Bayes, Logistic Regression, and Support Vector Classification. Although clustering could have been efficient in identifying groups of students, which tend to drop out, we decided against it due to lack of interpretability.

Random forest was run with its default parameters because taking the square root of our number of features when sampling the predictors seemed appropriate, given that we have a feature to instance ratio of 34:2838 and want to reduce variance. Naïve Bayes was run with the multinomial distribution as our data set has discrete features. Logistic Regression was run with a linear kernel as we were looking for a simple model. And lastly, Support Vector Classification was run with a linear kernel for the same simplicity reasons.

In order to accurately evaluate the performance of these four models and reduce variance, we randomly permuted the whole dataset and ran five-fold cross-validation. The random permutation was done in order to avoid dropouts being grouped together and, thus, high variance in the performance of our models. The choice of five folds for the cross validation was made because we wanted big enough testing sets to have a sufficient number of dropouts tested.

Within each fold, in order to tackle the strong class imbalance in our target variable (6.3% dropout), we chose to under sample each training set. We kept all of our dropout instances and randomly sampled the same amount of non-dropouts to have a balanced target variable.

Lastly, for each fold and each model we produced the confusion matrix, the ROC curve, the AUC, and the recall (see Appendix B.1 for details). We sought to maximize recall, as we are interested in minimizing the number of actual dropouts being wrongly classified as non-dropouts (False Negatives). Table 2 below shows the mean AUC and associated standard deviation and recall for each model.

Table 2

	Random Forest	Support Vector Classification	Naïve Bayes	Logistic Regression
AUC Mean	0.8365	0.8368	0.8369	0.8468
Recall Mean	0.8037	0.9339	0.9742	0.7590
AUC Std Dev	0.0174	0.0224	0.0277	0.0339
Recall Std Dev	0.0517	0.0209	0.02424	0.0339

First, note that across all four models we get an average AUC of 0.837, which is much better than the AUC of our baseline model (0.56). Now, our goal is to choose the best data mining algorithm between the four listed above. In terms of AUC, performance is comparable across all models, but Random Forest has the smallest standard deviation. While Naïve Bayes and SVC both performed very well in terms of recall, the confusion matrix shows that around 40% of the actual non-dropouts were predicted as dropouts versus 25% for Random Forest and Logistic Regression. Thus, high recall of Naïve Bayes and SVC was largely due to overestimating the number of dropouts. This approach would not generalize well with more data. Thus, the choice came down to Random Forest and Logistic Regression.

We chose Random Forest for three reasons: 1) it provides a better average recall, 2) it has lower AUC standard deviation, 3) makes more sense given our dataset. As previously discussed, our data has a lot of missing values (due to the difference between the surveys), is very small (due to the under sampling), and has few dropout instances. Even though Logistic Regression penalizes outliers in order to reduce variance, it remains sensitive to the big potential changes in data that can come from new schools and many new dropouts. On the other hand, Random Forest reduces the variance arising from our small dataset (by performing bagging) and reduces the relative importance of predictors with missing values (by randomly choosing a subset of the predictors for each newly sampled training set).

Lastly, we ran the procedure above on a smaller dataset, which contained survey answers for Fall 2018 and dropout information for Spring 2019. The goal was to see how accurate our model would be at predicting a

student dropping out the semester before it actually happened. The best result we got was 66% mean AUC and 61% mean recall with Naïve Bayes (see Appendix B.2). These results are promising given that this subset of our dataset only had 72 dropout instances.

Descriptive Analysis

Beyond accurately predicting which students are most likely to drop out, we are interested in having interpretable results. Our model's interpretability helps improve our predictions, facilitates deployment, and detects bias in our model. In terms of prediction, understanding the model's decisions can help us reduce the number of false negatives. In terms of business value, understanding the main features that determine why a student may drop out of college translates into actionable feedback for the universities that use our data mining solution. Finally, detecting bias can help us account for ethical concerns involved in our model's decision-making.

Determining feature importance for a non-linear model is not immediately obvious. For this reason, we rely on SHAP (SHapley Additive exPlanations) to explain the result of our predictive model (Lundberg et al., 2017). SHAP explains the model's classification of a given student based on each feature's contribution to the model's prediction. The measure of feature importance used in SHAP is the Shapley value, which extends the concept of feature weights in linear models to non-linear models. Larger Shapley values indicate higher feature importance.

We choose SHAP's tree explainer (TreeSHAP) over the model-agnostic kernel explainer because it is faster, computes exact Shapley values (as opposed to approximations), and because our main predictive model is a Random Forest (Molnar, 2019).

Table 3 below shows the most important features from the SHAP analysis of the Random Forest classifier's predictions.

Table 3

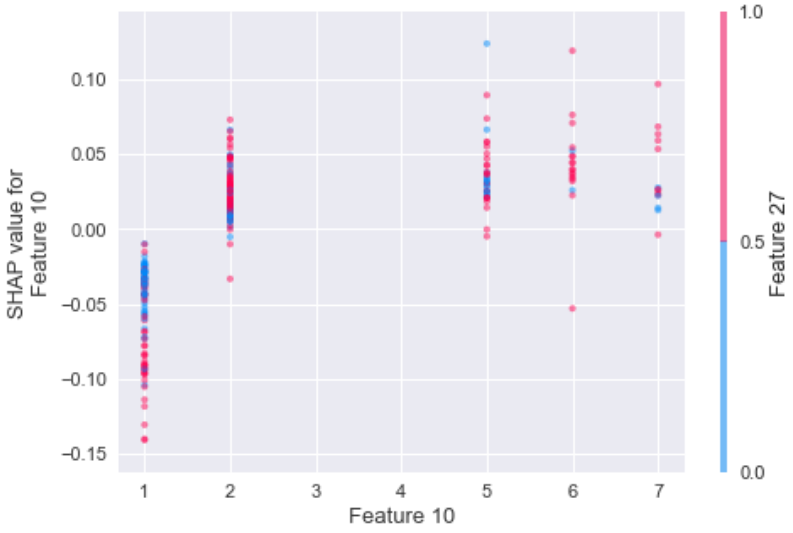
Feature names	Absolute value of average impact on model (SHAP)
"I am considering taking a break from school: Yes."	0.038
"How likely are you to recommend your university?"	0.022
"Do you have children: No"	0.021
"The courses I am taking will help me succeed in my future career"	0.020

To detect possible interaction effects in feature importance, we plot the SHAP values of every feature for every instance in our data. Figure 1 below shows the distribution of each feature's impact on the model prediction. Each point in the scatter plot indicates a single instance of the data. The x-axis indicates the value of the first feature, the y-axis shows its SHAP value, i.e. the feature importance for predicting the positive class, while the coloring indicates the value of a second feature.

Figure below shows the feature dependence plot between for the most important feature ("I am considering taking a break from school: Yes.") and the indicator variable for whether the student has a job.

Figure 1:

Dependence Plot of "I am considering taking a break from school" (feature 10) and "student has a job" (feature 27)



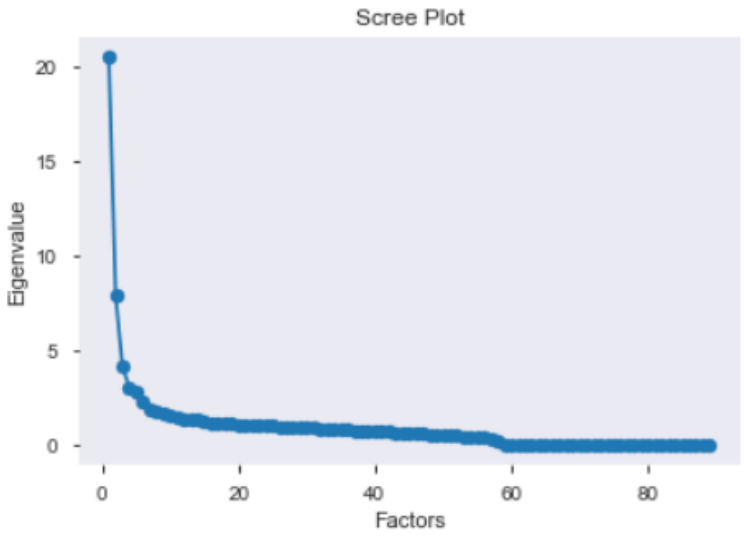
Unsurprisingly, a student's higher inclination towards taking a break from school is important in determining whether they drop out. The coloring indicates the presence of an interaction effect between whether the student works (red) or not (blue). Among the students who are strongly considering taking a break from school, whether they have a job contributes to predicting the positive class (i.e. more likely to drop out of college).

Factor Analysis

Factor analysis groups observed variables (answers in the survey) under different latent variables called factors, based on common patterns of responses. The goal is to interpret the concepts each factor represents. This helps us improve our survey by reducing its size and making the questions more specific. A shorter survey not only encourages students to stay engaged while answering, but also reduces the feature to instance ratio, thus improving our model.

First, we perform an adequacy test—Kaiser-Meyer-Olkin (KMO) test to determine if the dataset is fit for Factor Analysis. The KMO test yields 0.65 on the dataset, which is adequate given that the cutoff is 0.6. To determine how many factors to interpret, we create a scree plot with eigenvalues of the factors (see fig. 2 below).

Figure 2



We observe a significant drop roughly after the first 5 factors. In table 4 below, we can see that the top 5 factors capture 40% of the variance. In fact, the top 2 factors alone account for 31% of the variance.

Table 4: Factor Variance

	0	1	2	3	4
SS Loadings	16.975242	10.977090	3.074909	2.942355	2.228070
Proportion Var	0.190733	0.123338	0.034550	0.033060	0.025034
Cumulative Var	0.190733	0.314071	0.348621	0.381681	0.406715

Then, we perform Factor Analysis and observe the factor loadings (see table 5 below), a matrix that shows the relationship of each variable to the factors. For each of the 5 factors, the variables are ranked in descending order. We take the top few variables and interpret them.

Table 5: Few Rows of Factor Loadings

	0	1	2	3	4
student_id	0.051924	0.020289	0.074902	-0.116808	-0.086761
school	-0.228478	-0.110029	-0.184712	0.182172	0.186851
iaminterestedinthecoursesthatiam	0.040948	0.013458	0.020819	0.494179	0.050839
mycourseloadistoochallenging	0.019326	0.057638	0.081363	-0.191651	0.179209
iamscaredoffailingoneormoreofmyc	0.024737	-0.047560	0.192031	-0.232920	0.253252
thecoursesthatiamtakingwillhelpm	-0.027595	-0.003989	0.039848	0.488826	0.063373
iamconfidentthatiwillgraduateint	0.031366	0.037189	-0.086969	0.400755	-0.073010
myparentsrelativessupportmefinan	-0.055431	-0.046818	-0.358923	0.171569	-0.087116
istruggletopayfortextbooksrentut	0.127331	0.144237	0.342861	-0.229086	0.414344

Top variables for factor 1: balancing work and school is too stressful, I often attend events organized by school, I exercise at least twice a week, I get more than 7 hours of sleep on most nights

Top variables for factor 2: Did you fail one or more of your courses: no, I already have loans, I am paying for college out of pocket: disagree

Top variables for factor 3: Do you live on campus: no, will you be working 20 hours a week or more: yes, do you have any major family commitment: yes, I choose this school because it is close to work/home

Top variables for factor 4: I feel like I fit into my school, How likely are you to recommend your school, I am interested in the courses that I'm taking, the courses that I am taking will help me, If I have an issue I have someone to reach out to

Top variables for factor 5: I live on campus, I struggle to pay for textbooks, rent, utilities, healthcare, transport or food, I worry about paying for school

Factor 1 has to do with lifestyle and time management, which we think affects the stress level. Factor 2 points out a connection between doing well in classes and financial pressures. Factor 3 captures a specific group, which is those who are likely balancing between family, work and school. Factor 4 connects one's positive experiences in school to confidence in their school/education. Factor 5 illustrates the importance of financial stressors.

Taking these findings into consideration, we recommend a few ways of improving the survey. Factor 1 and 2 account for a significant proportion of variance. Therefore, the survey should ask questions regarding lifestyle, time management and stress levels. Factor 3 is connected to this as well. The survey should also include questions directly asking whether financial pressures affect academic performance. In conclusion, the survey should include variations of questions related to important factors identified above and remove questions that do not relate to them. Since the survey has Likert scale questions, a variation of questions can identify when students answer randomly.

Deployment, Risks and Ethical Considerations

Each semester we will apply our model to the new survey responses to predict dropouts. This new data will be fed back into the model once the semester ends, increasing our sample size. Each semester we will send feedback to every school based on the results from the descriptive analysis. Then, we will update the survey for next semester based on factor analysis. We can monitor predictive performance of our model to detect concept drift.

The main risk in our induced model is that of labeling dropouts as non-dropouts (false negatives). In order to mitigate this risk, we chose a model which maximizes recall. Some deployment issues that we might come across are: 1) schools failing to distinguish between students who drop out and those that are taking a leave of absence, 2) introducing missing values to survey responses by adding new questions to our survey that were not asked in previous semesters, and 3) low survey response rates. Since our project identifies which survey questions predict

dropout, we purposefully excluded any demographic information about the students to avoid potential ethical concerns (i.e. using gender and race of students as features).

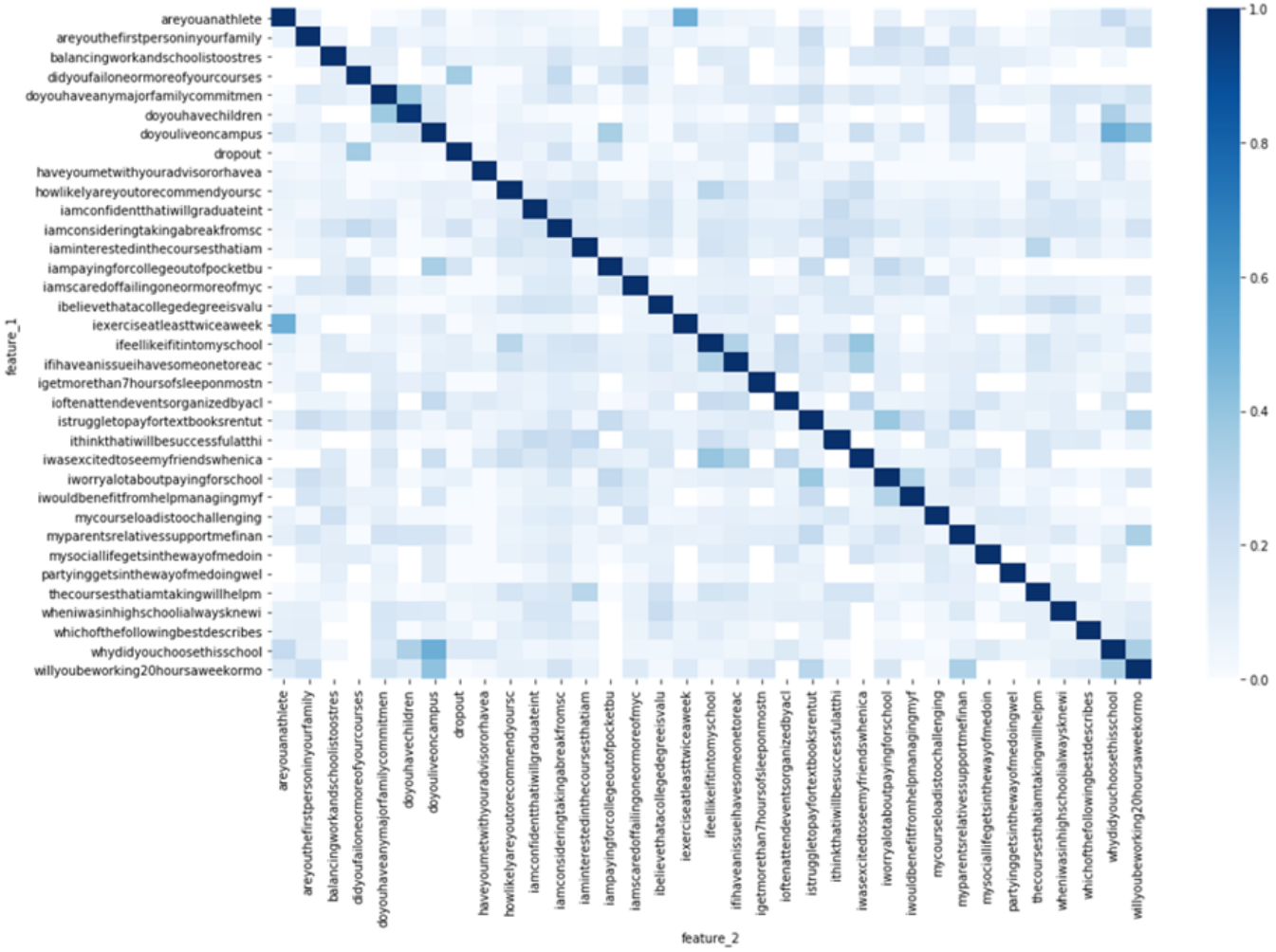
Conclusion

College dropout is a multi-faceted issue with no obvious solution. Our predictive analysis yields good results, but with more data we can improve upon our performance. The results from our descriptive analysis were interpretable and can shed light on the potential determinants of college dropouts. Finally, with factor analysis we were able to identify ways to improve our survey. Although our proposed method can provide a way to mitigate the costs associated with college dropouts in the US, this is an issue with social complexities that, while not the focus of this project, are important to address.

Bibliography

1. Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). *Mobility report cards: The role of colleges in intergenerational mobility* (No. w23618). national bureau of economic research.
2. Leonhardt, David, and Sahil Chinoy. "The College Dropout Crisis." *The New York Times*, 23 May 2019, www.nytimes.com/interactive/2019/05/23/opinion/sunday/college-graduation-rates-ranking.html?smid=pl-share.
3. Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.
4. Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Appendix A


areyouanathlete
areyouthefirstpersoninyourfamily
balancingworkandschoolistoostres
didyoufailoneormoreofyourcourses
doyouhaveanymajorfamilycommitmen
doyouhavechildren
doyouliveoncampus
dropout
haveyoumetwithyouradvisororhavea
howlikelyareyoutorecommendyoursc
iamconfidentthatiwillgraduateint
iamconsideringtakingabreakfromsc
iaminterestedinthecoursesthatiam
iampayingforcollegeoutofpocketbu
iamscaredoffailingoneormoreofmyc
ibelievethatacollegedegreeisvalu
iexerciseatleasttwiceaweek
ifeellikeifitintomyschool
ifihaveanissueihavesomeonetoreac
igetmorethan7hoursofsleeponmostn
ioftenattendeventsorganizedbyacl
istruggletopayfortextbooksrentut
ithinkthatiwillbesuccessfulatthi
iwasexcitedtoseemyfriendswhenica
iworryalotaboutpayingforschool
iwouldbenefitfromhelpmanagingmyf
mycourseloadistoochallenging
myparentsrelativessupportmefinan
mysociallifegetsinthewayofmedoin
partyinggetsinthewayofmedoingwel
thecoursesthatiamtakingwillhelpm
wheniwasinhighschoolialwaysknewi
whichofthefollowingbestdescribes
whydidyouchoosethisschool
willyoubeworking20hoursaweekormo
feature_1
feature_2
1.0
0.8
0.6
0.4
0.2
0.0

Appendix B.1

Baseline Model Results

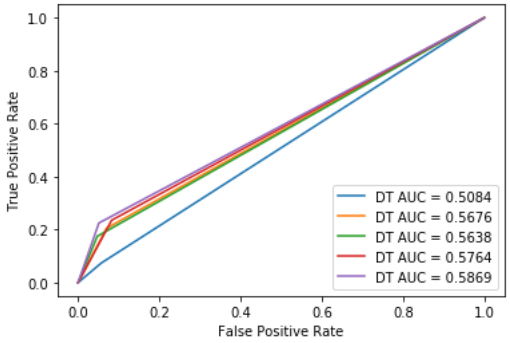
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	510	31
Actual Non Dropout	25	2

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	492	37
Actual Non Dropout	31	8

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	503	25
Actual Non Dropout	33	7

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	489	44
Actual Non Dropout	26	8

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	500	27
Actual Non Dropout	31	9



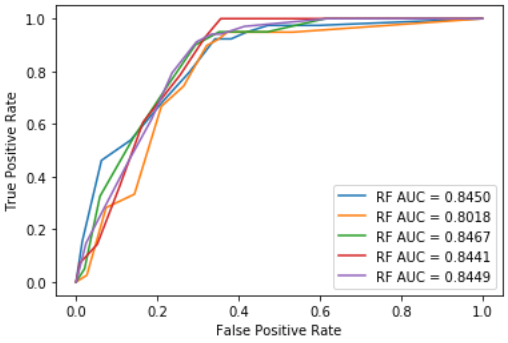
	Decision Tree
AUC Mean	0.5606118337544559
Recall Mean	0.18289927936986758
AUC Standard Deviation	0.02729507272933578
Recall Standard Deviation	0.05816723960834653

Random Forest Results

	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	382	147
Actual Dropout	8	31

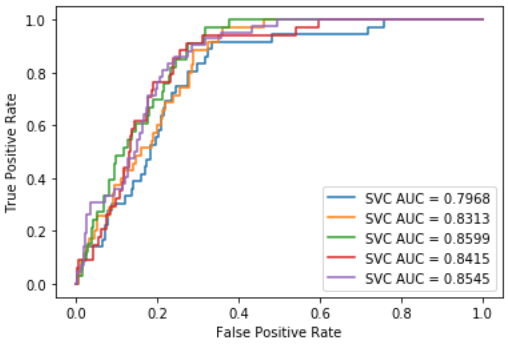
	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	389	140

Actual Dropout	10	29
	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	373	155
Actual Dropout	4	36
	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	401	138
Actual Dropout	6	22
	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	407	126
Actual Dropout	7	27



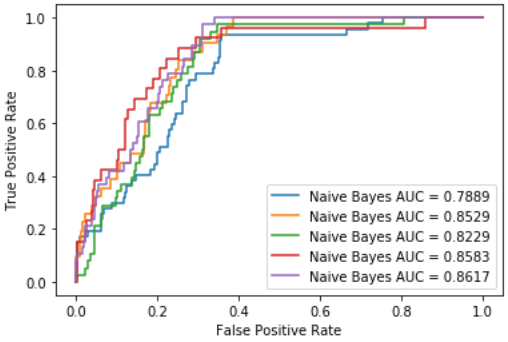
SVC Results

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	301	231
Actual Non Dropout	3	33
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	343	190
Actual Non Dropout	2	33
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	339	196
Actual Non Dropout	1	32
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	374	159
Actual Non Dropout	3	31
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	345	180
Actual Non Dropout	3	39



Naïve Bayes Results

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	327	194
Actual Non Dropout	3	44
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	319	218
Actual Non Dropout	0	31
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	325	205
Actual Non Dropout	1	37
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	339	202
Actual Non Dropout	1	25
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	329	200
Actual Non Dropout	0	38

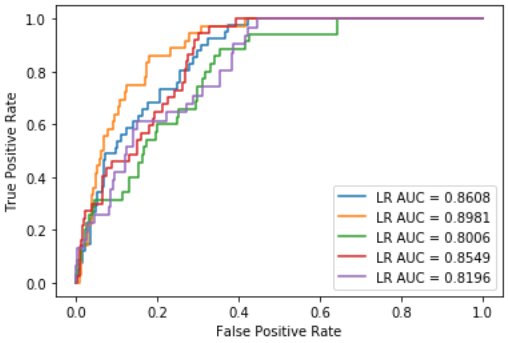


Logistic Regression Results

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	380	147
Actual Non Dropout	8	33
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	417	115
Actual Non Dropout	5	31
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	395	138
Actual Non Dropout	12	23

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	392	138
Actual Non Dropout	10	27

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	366	170
Actual Non Dropout	8	23



Appendix B.2

Entrée [300]:

Random Forest Results for Early Predictions

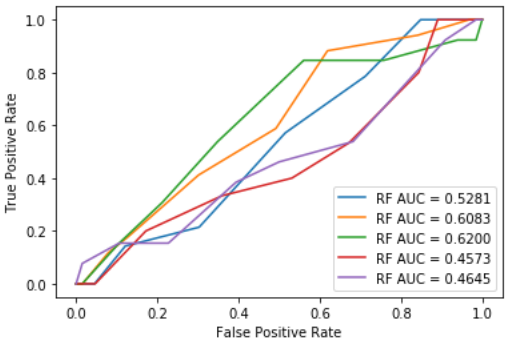
	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	46	20
Actual Dropout	11	3

	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	44	19
Actual Dropout	10	7

	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	43	23
Actual Dropout	6	7

	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	41	23
Actual Dropout	10	5

	Predicted Non Dropout	Predicted Dropout
Actual Non Dropout	40	26
Actual Dropout	8	5



SVC Results for Early Predictions

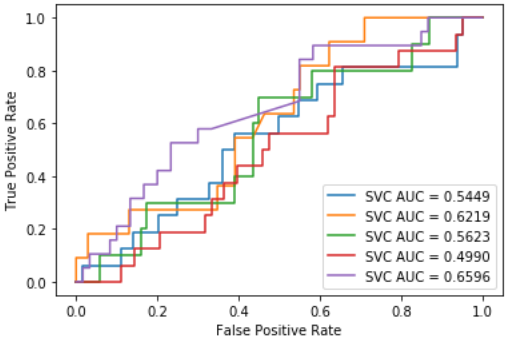
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	38	26
Actual Non Dropout	7	9

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	39	30
Actual Non Dropout	5	6

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	34	35
Actual Non Dropout	3	7

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	38	25
Actual Non Dropout	10	6

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	38	22
Actual Non Dropout	8	11



Naïve Bayes Results for Early Predictions

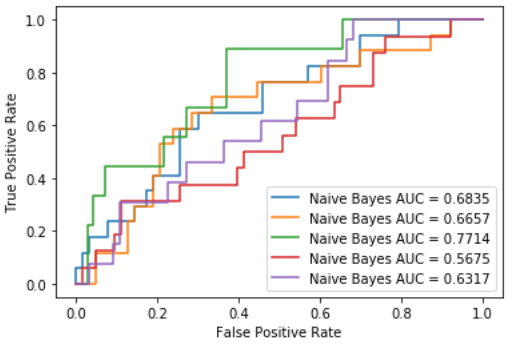
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	39	24
Actual Non Dropout	6	11

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	39	24
Actual Non Dropout	5	12

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	46	24
Actual Non Dropout	3	6

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	34	29
Actual Non Dropout	8	8

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	37	29
Actual Non Dropout	6	7

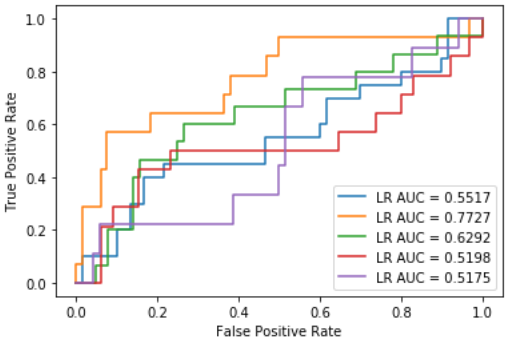


Logistic Regression Results for Early Predictions

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	33	27
Actual Non Dropout	11	9

	Predicted Dropout	Predicted Non Dropout
Actual Dropout	49	17

Actual Non Dropout	5	9
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	46	18
Actual Non Dropout	6	9
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	33	32
Actual Non Dropout	7	7
	Predicted Dropout	Predicted Non Dropout
Actual Dropout	39	31
Actual Non Dropout	6	3



Overall Results for Early Predictions

	Random Forest	Support Vector Classification	Naïve Bayes	Logistic Regression
AUC Mean	0.5356478039933922	0.5775537315270224	0. 663959373959374	0. 5981602286602286
Recall Mean	0. 37649213531566467	0. 5523803827751196	0. 6116138763197586	0. 5052380952380953
AUC Standard Deviation	0. 06880092873432823	0. 05686834803094961	0.0667625653330542 2	0.09617962865254306
Recall Standard Deviation	0.10556309410101014	0.10403031306763284	0.0787180050731014	0.09617962865254306

Appendix C

See below for the list of group members and their respective contributions:

- Tinatin Nikvashvili – Data cleaning and exploration
- Camille Taltas – Modeling and evaluation
- Francesca Guiso – Descriptive analysis
- Christine Shen – Factor Analysis
- Report write-up and general strategy was collaborative