Dropout Analytics: Understanding Key Drivers for Dropout Students

Leena Kang, Ethan Lin, Krystal Qiu, Mackenzie May, Samuel Mai

1

Abstract. *Higher education is an important step for each student's future career [1]. Therefore, understanding the main factors [3] that contribute to whether or not a student drops out of college is essential to improve academic success. Several variables were considered [2] through exploratory analysis and hypothesis testing [3.3] in order to create a model [4] to predict a student's likelihood of graduating from a college/university.*

1. Introduction

In an ideal world, every student has an equal chance to be successful in their educational pursuits. Academic success is strongly correlated with success in adulthood, so when students drop out of school, it reveals patterns of inequality in our educational system that must be addressed. Although dropout rates in the United States have declined in the past decade, we can leverage data to analyze the leading causes of dropouts in order to make better informed solutions to further reduce dropout rates. With this project, we hope to identify which factors are significantly correlated with academic success. These findings, if the factors are controllable, would be able to provide future students and/or parents with guidance on positive actions that can be taken to help level the playing field. In addition, the findings can help to reduce dropout rates, by influencing education policy makers to make adjustments to policy and informing parents how to identify if their children are at the risk of dropping out of school. With our findings, we aim to answer the question: "What causes students to drop out of school?"

The analysis of dropouts is not a new topic. The 2013 research article "Understanding why Students Drop Out of High School, According to Their Own Reports," by Jonathan Jacob Doll, Zohreh Eslami, and Lynne Walters, highlights how "research on school dropout extends from early 20th-century pioneers until now, marking trends of causes and prevention." The article goes on to break down the three main factors that play a role in dropping out. "Push factors include school-consequence on attendance or disciple. Pull factors include out-of-school enticements like jobs and family. Finally, falling out factors refer to disengagement in students not caused by school or outside pulling factors." The article also found that pull factors play the largest role in drop out rates, while push factors have been playing an increasingly larger role. It is theorized that the No Child Left Behind Act of 2001, which increased the standards of education, also led many students to drop out due to feelings of academic inadequacy. Another key insight was the gender disparity, with male students reporting the highest rates of push factors and female students reporting the highest rates of pull factors.

Another analysis of dropouts was found from Dr. Imed Bouchrika's 'High School Dropout Rate Is Decreasing but Race, Income Disability Issues Persist in 2024.' It reports on how dropout rates are highly correlated with the issues of race, sex, and socioeconomic status. In particular, American Indian/Alaska Native students have higher drop out rates than

any other racial group, disabled students are two times more likely to drop out than non-disabled students, and male students are more likely to drop out than female students. There is also the strong negative correlation between family income level and dropout rates, although the rate gap between the richest and poorest has been shrinking in recent years. Similar to the previous article, this article specifies school and family as two of the main reasons for dropping out, but it also cites employment as a key factor.

Our aim is to determine if we can anticipate which students are at risk of poor academic performance early on, based on academic, demographic, and socio-economic factors, and potentially suggest data-driven, optimal solutions to ultimately reduce dropout rates. To do this, we also plan on performing various hypothesis tests to identify the significant differences between graduate and dropout students, while addressing potential biases in the dataset itself. We would expect the data to suggest that children from lower income families have a higher proportion of dropout students. However, further hypothesis testing would be necessary to determine whether these differences are significant.

2. The Dataset

The dataset we are using was adopted from the UCI Machine Learning Repository that includes data containing the academic path, demographics, and socio-economic status of each student along with their academic performance. Each row represents a student, and records their academic status at the end of semester as dropout, enrolled, or graduate. With this, we intend to investigate whether any discrepancies between 'successful' and 'unsuccessful' students are statistically significant.

```
df1_raw.head()
```

	Marital Status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality
0	1	17	5	171	1	1	122.0	1
1	1	15	1	9254	1	1	160.0	1
2	1	1	5	9070	1	1	122.0	1
3	1	17	2	9773	1	1	122.0	1
4	2	39	1	8014	0	1	100.0	1

5 rows × 37 columns

Figure 1. The first few rows and columns of the dataset we will be using

3. Identifying Key Drivers

3.1. Methodology

It was provided that preprocessing methods to handle anomalies, null values, and outliers in the dataset were already performed beforehand. Majority of features consist of

multiclass data, all represented by numerical values (excluding the Target). However, features that are related to 'Previous Qualification' consist of 20+ categories (that could be arguably grouped together), for which they are not uniformly distributed. To avoid the risk of overfitting our models, we have recategorized and grouped these features based on our collective interpretation of the documentation of each feature. For the rest of the multiclass categorical variables, we converted each integer value to its corresponding description to provide more clarity on our data. (Note that this will change when we preprocess this data for our models). Additionally, 6 features consist of binary data, for which we converted each variable into a Boolean type.

It is worth noting that the Target variable started with 3 classes– dropout, enrolled, and graduate. Since our research is primarily concerned with students who dropped out of their respective schools or not, we combined said graduate and enrolled students into one, which will represent the class that did not drop out of their schools.

Due to the uneven distribution of the target class, we visualized the relative frequencies of each categorical feature for dropout and graduate/enrolled students to identify potential discrepancies between the two groups.

	Marital Status	Course	Previous qualification	Previous qualification (grade)	Mother's qualification	Father's qualification
0	Single	Animation and Multimedia Design	Secondary	122.0	Basic	Basic
1	Single	Tourism	Secondary	160.0	Secondary	Higher
2	Single	Communication Design	Secondary	122.0	Basic	Basic
3	Single	Journalism and Communication	Secondary	122.0	Basic	Basic
4	Married	Social Service (evening attendance)	Secondary	100.0	Basic	Basic

Figure 2. First 6 features of finalized preprocessed dataset.

3.2. Results

We observed that a larger proportion of dropout students (1) were married or divorced, (2) did not complete high school, and (3) were debtors. Meanwhile, a larger proportion of graduate/enrolled students (4) were displaced and (5) were scholarship recipients. However, these differences could have simply occurred by chance alone; hypothesis testing for each observation is necessary to draw any stronger conclusion about the key drivers for dropout students.

First, we will assume that dropout and graduate/enrolled students each follow a normal distribution. For each potentially significant difference, we performed a 2 sample proportion z-test by the following 2 methods: (1) Assuming that the distribution of dropout and

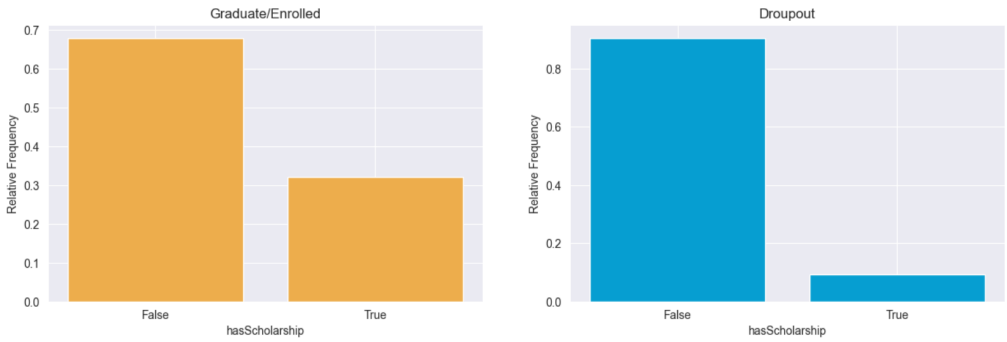

Figure 3. Relative frequencies of marital status separated by target

graduate/enrolled students are from the same distribution, we simulated 500 new samples and generated new difference of proportions to calculate the p-value. To validate our test, we also leveraged the (2) statsmodels module to generate a p-value. We have found that each test resulted in a sufficiently small p-value (all of which are approximately 0.0) where we reject the null hypothesis *(i.e., dropout and graduate/enrolled students are from the same distribution, thus any proportional differences between the two is due to chance)*. Hence, we have sufficient evidence to conclude that the proportional differences found previously regarding dropout students are statistically significant.

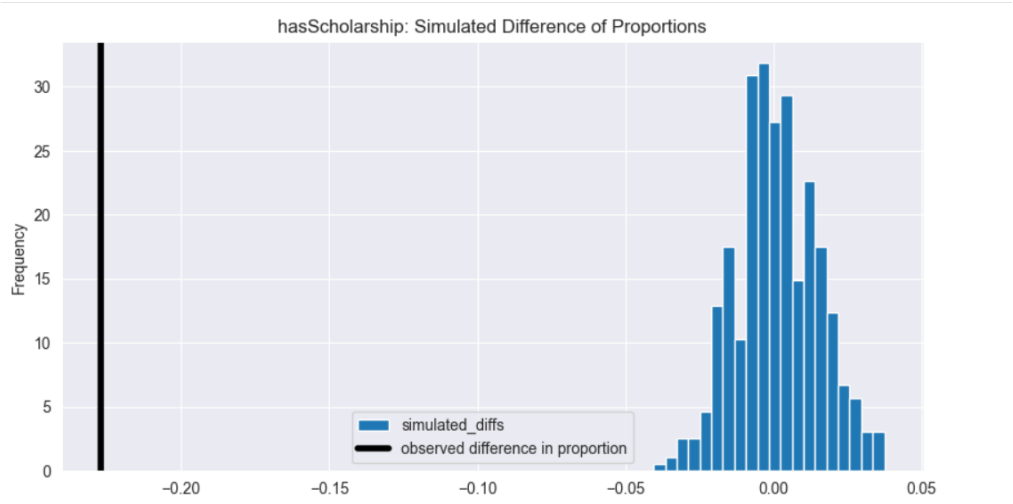

Figure 4. Distribution of simulated difference of proportions of scholarship recipients.

3.3. Course Distribution and Results

Since the type of coursework can be a contributing factor to the academic performance of students, it is worth understanding the distribution of enrolled courses, and investigate whether there are any noticeable differences between the distribution of enrolled courses against the target. To determine if there exists potential skewness of certain enrolled students for each target class, we performed 2 Chi Square Goodness of Fit tests to determine

if the distributions of enrolled courses for both graduate/enrolled and dropout students are uniform. Leveraging the statsmodel package to perform this test, we have found that– assuming that the frequencies for each course being uniform– results in a p-value of approximately 0.0 for both dropout and graduate/enrolled students. Thus, any variance of enrolled courses was not due to chance, and a uniform proportion for each course with each target class is not a good fit.

Course	Advertising and Marketing Management	Agronomy	Animation and Multimedia Design	Basic Education	Biofuel Production Technologies	Communication Design	Equinculture
Target							
Dropout	95	86	82	85	8	51	78
Graduate/Enrolled	173	124	133	107	4	175	63

Figure 5. Frequency table of enrolled courses grouped by the target variable.

This skewness prompted us to compare the proportions for each course against the target class. After performing two proportion z-tests using the statsmodel package, we have sufficient evidence to conclude that a larger proportion of dropout students were enrolled in Equinculture, Informatics Engineering, and Management (evening attendance), whereas a larger proportion of graduate/enrolled students were enrolled in Nursing, Social Service, and Communication Design. Though this does not necessarily entail that Equinculture, Informatics Engineering, and Management (evening attendance) are key drivers for increased dropout rates, this allows us to address potential biases or some relation for the dataset itself.

3.4. Grade Analysis

Another key relationship we want to analyze is the one between grade average in a semester and whether or not the student has dropped out. Under this assumption, it is easy to infer that there would be a significant relationship present, as previous academic success often predicts future academic success.

The grade distribution of dropouts and non-dropouts (removing null values) is markedly different in both semesters. When we look at the average grades of non-dropouts in both semesters, we notice that the distribution is fairly normal. However, the average grade distributions of dropouts are quite different have significant right skews and lower medians.

This indicates that students who would go on to drop out have disproportionately lower grades. This is not surprising, as low academic performance can play a role in all three factors that usually drive drop outs. Push factors occur as students flunk out, pull factors occur as parents seek other options, and falling out factors occur as students get discouraged by their low grades.

Next, we want to look at how factoring in the admission grades can provide a new perspective. While we are uncertain what admission grade represents in the dataset, we believe

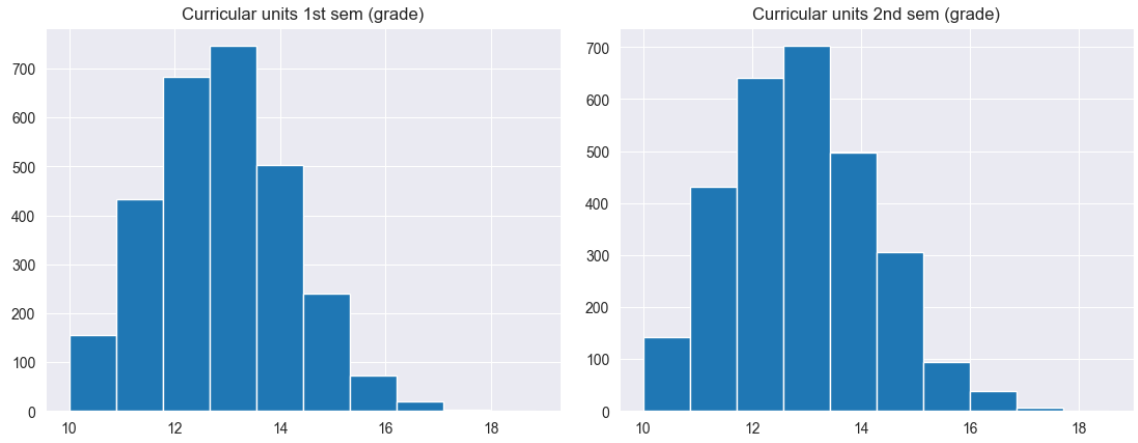

Figure 6. First and second semester grade average distributions of non-dropouts

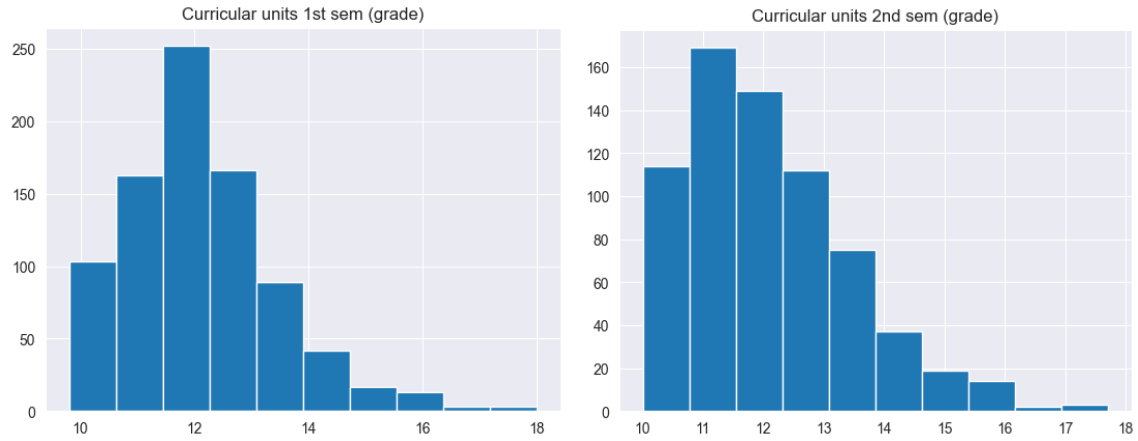

Figure 7. First and second semester grade average distributions of dropouts

it to mean the grade average of the student at a previous academic institution. By considering their past academic status, we can see if significant negative changes in academic status can foreshadow dropping out.

As the data shows, the relationship between admission grades and semester grade averages in dropouts and non-dropouts is quite different. We know that dropouts are more likely to have higher average grades than non-dropouts from the previous graphs, but this graph shows that dropouts occur all across the admission grade distribution, affecting even those who had strong admission grades.

We conducted regression analysis, producing the dark blue and red lines seen in each graph. Non-dropouts have fairly consistent correlation rates between their admission grades and semester grade averages (β_1 is 0.030 for the first semester and 0.025 for the second semester). This indicates that non-dropouts do not experience drastic academic status differences.

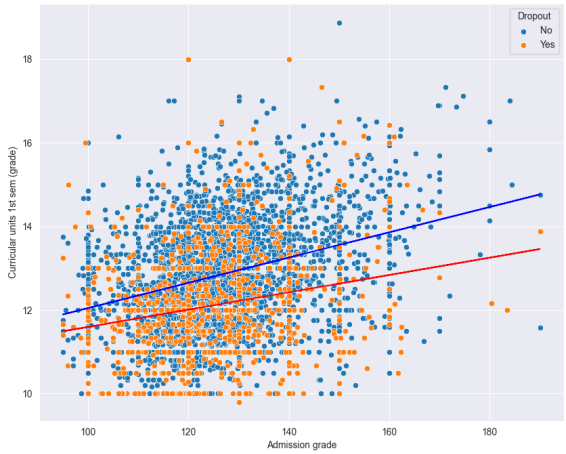

Figure 8. Scatter graph of admission grades and first semester grade averages, separated by dropout and non-dropout

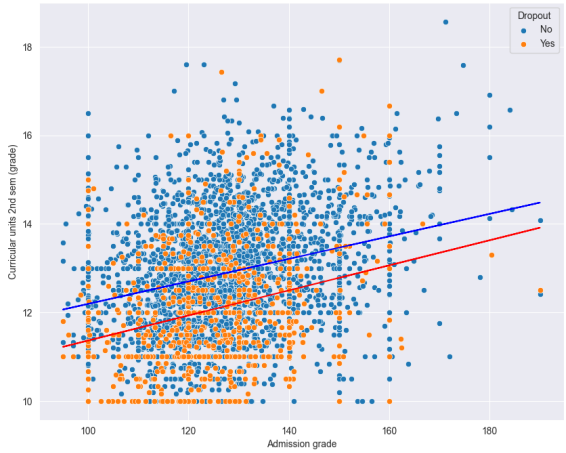

Figure 9. Scatter graph of admission grades and second semester grade averages, separated by dropout and non-dropout

However, the dropout regression line has unusual results. The β_0 value for the dropout regression line in the first semester is actually higher than the β_0 value for the non-dropout regression line (9.523 compared to 9.023), but the β_1 value is considerably lower (0.021 versus 0.030). This pattern is not present in the second semester. Since the relationship is not nearly as positive as the relationship that non-dropouts experience, it can be theorized that dropouts receive significantly lower grades than they are used to during their first semester, which plays a role in their decision to drop out. This phenomenon is worth investigating further.

4. Predicting Student Academic Status

4.1. Logistic Regression Model

We wanted to gain more insight on the predictive power of our features so we ran Logistic Regression on our data. In order to run a Logistic Regression, we needed to preproccess the data in order to run Logistic Regression. We had multiple columns that needed to be transformed. For boolean types, assigned 1 for True and 0 for False. For object types,

we operated on a case to case basis. For ordinal categorical features, we simply assigned numbers in the same ordinal ranking. For binary features, we simply assigned 1 for a category and 0 for the other. Lastly, for non-ordinal categorical features, we used one hot encoding in order to transform them into numerical data as there isn't significant meaning to a numerical translation. One hot encoding consists of creating a bunch of matrices filled with dummy entries except for the position of that specific category. A 1 in that position maps to a certain category.

After we transitioned our columns to numerical values, we set up our model by setting X to our features and y to our target. In our case, X contained all columns except for these columns that we dropped ['Target', 'Course', 'Nationality', 'Marital Status', "Mother's qualification", "Father's qualification"]. We dropped these features since we created one hot encodings for them already, or because there were unknown values that we decided would be better omit in our model entirely. Our target feature was the 'Target' column with a success being 1 = 'Graduate/Enrolled' and a fail as 0 = 'Dropout'. After establishing our X and y, we split 80% of our data for the training set and 20% for the testing set.

4.2. Results and Performance Analysis

After fitting our model with the training set and predicting based on the testing set, we evaluate our model.

```
# Evaluate the model
accuracy = metrics.accuracy_score(y_test, y_pred)
precision = metrics.precision_score(y_test, y_pred, pos_label='1')
recall = metrics.recall_score(y_test, y_pred, pos_label='1')
confusion_matrix = metrics.confusion_matrix(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print(f'Train Accuracy: {train_accuracy}')
print(f'Precision: {precision}')
print(f'Recall: {recall}')
print(f'Confusion Matrix:\n{confusion_matrix}')
```

```
Accuracy: 0.8655367231638418
Train Accuracy: 0.8818875388527833
Precision: 0.8640776699029126
Recall: 0.9384885764499121
Confusion Matrix:
[[232  84]
 [ 35 534]]
```

Figure 10. Evaluation Metrics of Logistic Regression Model

From Figure 10, we can see that our model performed well with an accuracy estimating 0.866, along with a sufficiently high recall and precision score.

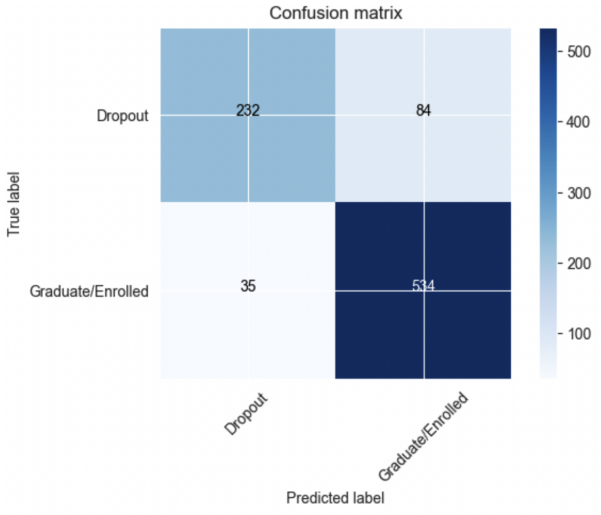

Figure 11. Confusion Matrix visualization

More precisely, we can see that our model correctly predicts the right outcome 232+534/885 times, which coincides with our 0.8665 accuracy above. Our model has a Type I error (*i.e., false positive rate*) of 84/885 = 0.095 and our Type II error (*i.e., false negative rate*) is 35/885 = 0.04.

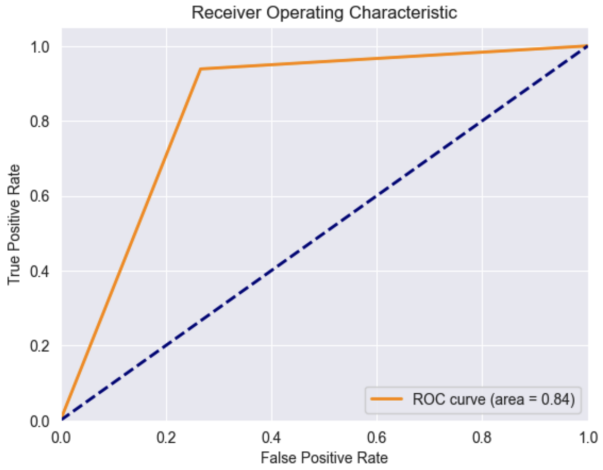

Figure 12. ROC Curve

Our ROC curve further visualizes our model's TPR (true positive rate) compared to it's FPR (false positive rate). The dashed line represents the curve if one was to randomly guess and as we can see our model heavily outperformed that line.

```
train_preds = model.predict(X_train)
train_accuracy = metrics.accuracy_score(y_train, train_preds)
print(f'Train Accuracy: {train_accuracy}')
```

```
Train Accuracy: 0.8818875388527833
```

```
cv_scores = cross_val_score(model, X, y, cv=5)  # 5-fold cross-validation
print("Cross-validation scores:", cv_scores)
print("Mean CV score:", cv_scores.mean())
```

```
Cross-validation scores: [0.88361582 0.88813559 0.86553672 0.88248588 0.85859729]
Mean CV score: 0.8756742592734617
```

Figure 13. Overfit testing

To check for for potential overfitting of our model, we performed tests on the training set to see if there was a noticeable discrepancy between the accuracy of the training vs testing sets, which we found there to be none. Additionally, we cross-validation on the model and all the scores were similar to our model accuracy scores. Therefore, we can have more confidence in our model not being over-fit.

4.3. Random Forest Model - Methodology and Preprocessing

In this section, we are building a Random Forest Classification Model based on our previous findings through observations and hypothesis testing. Throughout this section, we will gradually add features into our model to improve the prediction accuracy, and thus determine if the features have prediction power. In order to perform necessary steps, the dataset will be cleaned differently in this section from the original cleaned dataset. All the necessary binary variables are kept as numerical numbers *(i.e. 0 and 1)*, and all other categorical variables with more than 2 categories are one hot encoded. The target variable preserves similar change as previous; the students who dropped out are in one group, and the rest of students (graduate and enrolled) are kept in another group as a whole. Moving forward, we will call the first group as "drop-out group", and the second group as "regular group". To prevent the model from over fitting, the dataset is split into two parts for further training, as we have 80% training data and 20% test data, similar to the logistic regression.

4.4. Results and Performance Analysis

There are three progressive models in total, (1) baseline model, (2) refined model, and (3) final model. For the baseline model, the features are determined by the key findings in 3.2. Thus the features included are marital status, displaced, previous qualification, debtor, and scholarship. With these five features, our model is able to achieve an accuracy score of 0.6983, meaning it successfully predicts 69.83% of the test data.

However, if we take a look at the second semester grade for both groups, there is a distinguishable difference between their grades. On a 20 points scale, the regular group reaches an average grade around 12.27, but the drop-out group only gets 5.89. This means the second semester grade may have a great effect on the accuracy of the prediction model.

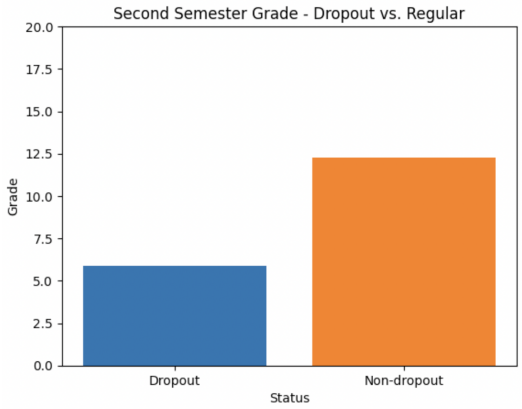

Figure 14. Average Second Semester Grade for Drop-out Students and Regular Students

For the refined model, we are adding this feature in to see if there is any improvement. This addition does improve our accuracy score by around 10%. The refined model accuracy score is now 77.85%. This improvement has shown that the second semester grade has prediction power in terms of deciding if a student dropped out or not.

In the visualization below, it is quite obvious that most students in the regular group have tuition fees up to date. However, by looking at the second graph for the drop out students, 32% of drop-out students do not pay off all the tuition fees.

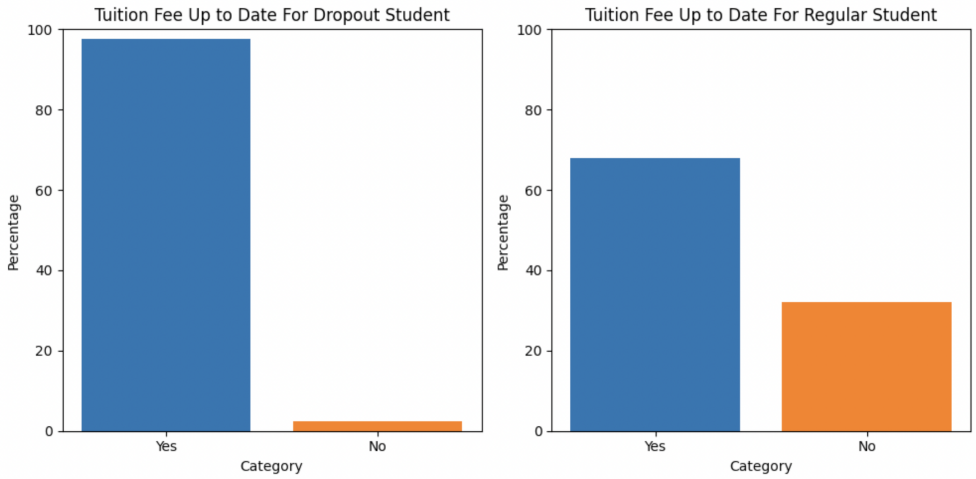

Figure 15. Tuition Fee Up to Date for Drop-out and Regular Student

Moreover, connecting back to the hypothesis testing from earlier, there is sufficient evidence supporting that a larger proportion of dropout students were enrolled in Equinculture, Informatics Engineering, and Management (evening attendance). To have a better understanding of the proportions, the visualization below has shown that 50.7% of the students in the Management (evening attendance) course dropped out at the end.

Thus, in the final model, in addition to the six features, two more features which are "Course" and "Tuition fee up to date" are added. This concludes our model accuracy to 81.24%, which is a relatively high accuracy. In this section, through enhancing the prediction model, we find out that "Course", "Tuition fees up to date", "Curricular units 2nd sem (grade)", "Marital Status", "Displaced", "Previous qualification (grade)", "Debtor", and "Scholarship holder" have prediction powers in deciding if a student is a drop-out.

5. Limitations

5.1. Collection Bias

It is worth addressing that approximately 98% of students in this dataset are Portuguese. Therefore, any conclusion that was drawn and model trained from this data cannot be necessarily generalized towards all students, but instead could be better targeted towards schools that are highly populated with Portuguese students.

5.2. Target

Recall that the Target variable originally was split into 3 classes: Graduate, Enrolled, Dropout. To not lose data while turning the Target variable binary that aligns with the

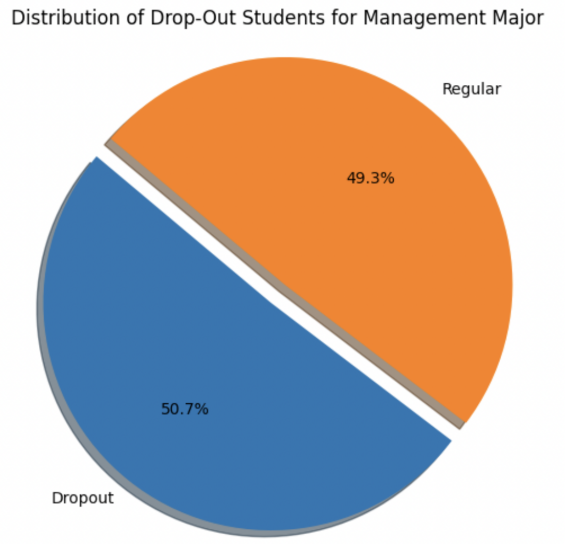

Figure 16. Distribution of Drop-Out Students for Management Major

scope of our research, we have combined Graduate and Enrolled students into one class. However, the Enrolled class of students may not constitute a 'successful' student, that is, it is not guaranteed that these students have graduated from their schools; instead, this was an assumption that we have made about the dataset.

Additionally, the Target was split where 3003 students are classified as Graduate/Enrolled whereas only 1421 students are classified as Dropout. This imbalance of data could potentially affect the accuracy of our models when predicting whether a student has dropped out of their respective schools, and thus can show more bias towards the majority class.

6. Conclusions

Through our analysis we found that the key drivers we explored directly correlated with the success rate of students. The data showed students were more likely to drop out when: 1. married (or once married), 2. debtors, 3. did not receive scholarships, 4. not displaced, 5. exhibited lower grades during the second semester, and 6. dropped out during high school (or equivalent). These variables showed a significant impact on a student's success in graduating from a college or university compared to the many variables we investigated.

The model suggests that these features potentially have predictive power (primarily for Portuguese schools). While there are some limitations within this model, it is still a strong indicator of a student's academic success. This model can educate future students, educators, and parents to improve the graduation rate and educational experience of each student. Furthermore, this model assists in targeting specific students who have the key drivers mentioned above in order to provide the necessary resources and/or assist these students as an effort to decrease the dropout rate.

References

[1] Doll, Jonathan Jacob, et al. "Understanding why Students Drop Out of High School, According to Their Own Reports" SAGE Open, vol. 3, no. 4, 1 Jan. 2013, p. 215824401350383, doi:10.1177/2158244013503834.

[2] Bouchrika, Imed. "High School Dropout Rate Is Decreasing but Race, Income & Disability Issues Persist in 2024" Research.com, research.com/education/high-school-dropout-rate. Accessed 07 Mar. 2024.

[3] M.V.Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V.Realinho. (2021) "Early prediction of student's performance in higher education: a case study" Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer. DOI: 10.1007/978-3-030-72657-7_16

[4] Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.