Agatha Tatang (apt3pa)
Joy Wang (jw7ua)
Yulin Yu (yy3cr)
Maryam Zaher (mz7ey)

Stat 4630 Project: Final Report

1. Executive Summary

This project conducted a case study on the courses on Udemy, an online platform that offers paid and free courses on various subjects. The dataset used in this project covers 3,682 records of Udemy courses from four different subjects: business finance, graphic design, musical instruments, and web design. It also contains important characteristics of courses, including if the course is paid or free, the level of difficulty for each course, the course subject, course price, the number of subscribers the course has, the number of views the course, the number of lectures in a course, the duration of all the course materials, the date and time when a course was published, and the subject of the course.

The first question of interest is “what are the factors that can predict the number of subscribers for a specific course?” Since anyone can create a course, from a creator's perspective, one would want to know what factors could increase the number of subscribers for their own course. Some factors that could influence the number of subscribers for a course are if the course is paid or not, the course price, the number of reviews, the course difficulty, and the duration of the course. To answer this question, we conducted linear regression to find if there exists a strong relationship between the number of subscribers and any characteristics of a course. We also performed tree methods that predict the number of subscribers of a course based on the specific conditions of important course characteristics.

The second question of interest is “what are the factors that influence whether a course is free or paid for those having more than 2300 subscribers?” Courses on Udemy can either be bought or taken for free. This will determine whether the price of certain courses is reflective of course reviews and -- similar to the regression analysis -- allow course creators to increase their value. To answer this question, we used several classification methods, including logistic regression, linear discriminant analysis, and tree methods to predict the probability of a course being paid given its important characteristics.

2. Data Processing & Cleaning

First, our team modified the published_timestamp into a quantitative variable that measures the days since it was published until May 1, 2020. Then the data preprocessing for the regression problem is as followed. First, the entries that had 0 for the variable *num_subscribers* were removed since these observations do not provide useful information on the relationship between the characteristics of courses and the number of subscribers. Then, the categorical

variables were converted into factors and then contrasted so they would be either True or False. The response variable (*is_paid*) was transformed into its power of 0.2. The reason is that the regression model using the original response variable severely violated the assumption of constant variance, and the transformation solved this problem. The data was then split into training data and test data in equal proportion and the seed was set at the value of 199.

For the classification problem, we subset the dataset by only using records with numbers of subscribers greater than 2300 to create a more balanced data set since classification models using an unbalanced dataset often lead to misleading accuracy. Then, the categorical variables were converted into factors and then contrasted. The data was then split into training data and test data in equal proportion and the seed was set at the value of 199.

3. Regression Question

3.1 Exploratory Data analysis

From the scatterplot of the variables and from their calculated correlations, it can be seen that there is a moderate correlation between the number of reviews and the number of subscribers, whereas the other variables are weakly correlated with the number of subscribers. One surprising note is how weak the correlation is between the number of subscribers and price.

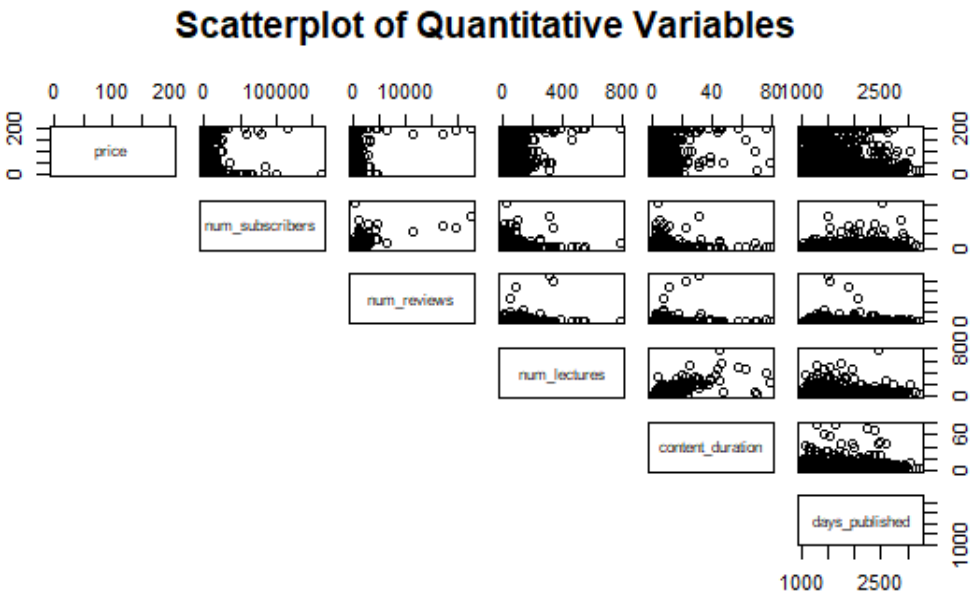

Figure 1: Scatterplot of Quantitative Variables

Figure 2: Boxplot of Subscribers by Is_paid

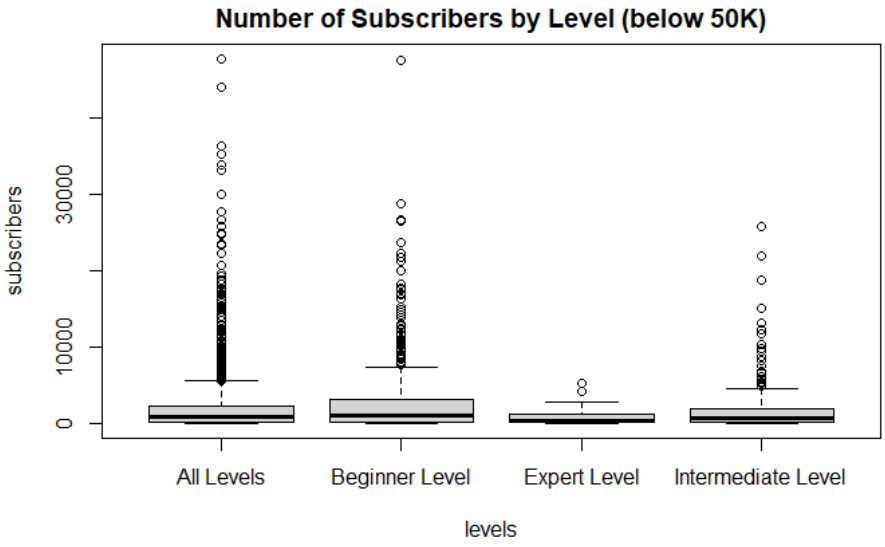

Figure 3: Boxplot of Subscribers by Course Level

From the distribution within the boxplots, it can be seen that the free courses had more subscribers than those that had to be paid for. The boxplots for levels versus number of subscribers and subjects versus number of subscribers were subsetted to show the entries that had less than 50,000 subscribers to better see the distribution. The boxplot for the different subjects indicates that courses about web development have a higher number of subscribers, followed by business finance. As for the different levels, the boxplots indicate that beginner level courses attract more subscribers. These graphical summaries allowed for a better understanding of what variables, or what features of a Udemy course have more subscribers.

3.2 Regression Model

The best linear regression model generated from milestone 3 and its diagnostic is shown below.

Summary Statistics for the Linear Regression Model

```
Call:
lm(formula = num_subscribers2 ~ num_reviews + is_paid + subject +
    days_published + price + level, data = train2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3985 -0.7999 -0.0730  0.7538  4.2216

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.364e+00  1.468e-01  22.913  < 2e-16 ***
num_reviews                8.589e-04  5.778e-05  14.864  < 2e-16 ***
is_paidTRUE               -1.954e+00  1.004e-01 -19.464  < 2e-16 ***
subjectGraphic Design     -4.861e-02  7.979e-02  -0.609  0.54247
subjectMusical Instruments -3.391e-01  7.508e-02  -4.516  6.7e-06 ***
subjectWeb Development     1.271e+00  6.520e-02  19.497  < 2e-16 ***
days_published             8.997e-04  6.434e-05  13.983  < 2e-16 ***
price                      5.372e-03  4.782e-04  11.232  < 2e-16 ***
levelBeginner Level        8.688e-02  5.778e-02   1.504  0.13286
levelExpert Level         -5.479e-01  2.067e-01  -2.651  0.00811 **
levelIntermediate Level   -8.918e-02  8.707e-02  -1.024  0.30585
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.094 on 1790 degrees of freedom
Multiple R-squared:  0.5157,    Adjusted R-squared:  0.513
F-statistic: 190.6 on 10 and 1790 DF,  p-value: < 2.2e-16
```

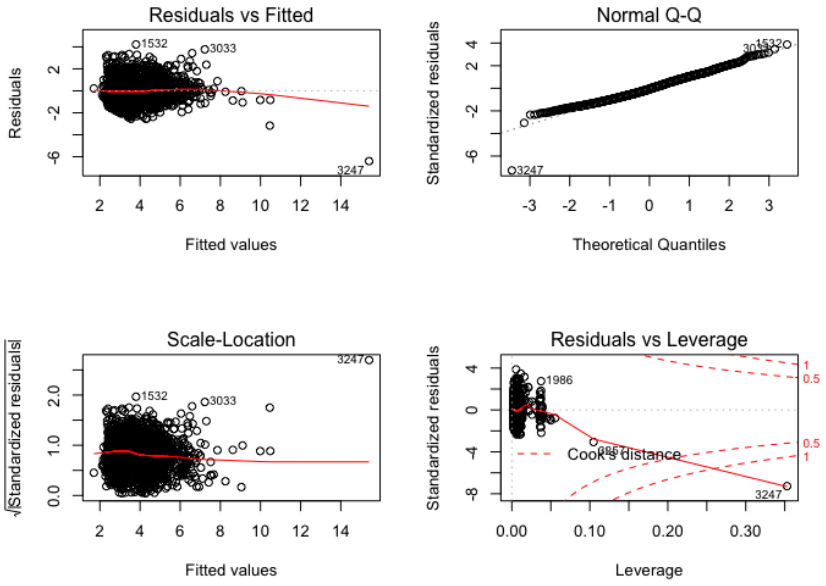

Figure 4: Model Diagnostics for the Linear Regression Model

The residual plot (top left) shows the red line is approximately along the x-axis with a slight curvature. It indicates the form of our model is relatively reasonable. The QQ plot (top right) has the residuals approximately fall along the 45-degree line, indicating that the normality assumption is approximately met. The flat red line along y=1 on the plot of the square root of the standardized residual against the fitted values (bottom left) indicates the satisfaction of the assumption of constant variance, and the last plot shows only one influential outlier. Overall, the regression model meets the basic assumptions.

Because we transformed our response variable y to y^0.2, we transformed y^0.2 back to the original state to better interpret our question:

$$y^{0.2} = \boldsymbol{\beta X}$$
$$\ln(y^{0.2}) = \ln(\boldsymbol{\beta X})$$
$$0.2\ln(y) = \ln(\boldsymbol{\beta X})$$
$$\ln(y) = 5\ln(\boldsymbol{\beta X})$$
$$y = \exp(5\ln(\boldsymbol{\beta X})) = \boldsymbol{\beta X}^5$$

The way that we can interpret our prediction equation is as follows:

Num_reviews: For every Udemy course, the average number of subscribers would increase by $0.0008589318^5 \approx 0$ for each additional number of reviewers the course received, by giving other variables as constant.

is_paidTRUE: For every Udemy course, the average number of subscribers would decrease by $1.9544984872^5 \approx 28$ if the course is paid rather than free, by giving other variables as constant.

SubjectMusical Instruments: For every Udemy course, the average number of subscribers would decrease by $0.3390702072^5 = 0.004483725 \approx 0$ if the course subject is Musical Instruments rather than Business, by giving other variables as constant.

SubjectWeb Development: For every Udemy course, the average number of subscribers would increase by $1.2712623706^5 \approx 3.32$ if the course subject is Web development rather than Business, by giving other variables as constant.

Days_published: For every Udemy course, the average number of subscribers would increase by $0.0008996704^5 \approx 0$ for each additional pasting days of publishing, by giving other variables as constant.

Price: For every Udemy course, the average number of subscribers would increase by 0.0053716155^5≈ 0 for each additional dollar that customers paid, by giving other variables as constant.

LevelExpert Level: For every Udemy course, the average number of subscribers would decrease by 0.5479346104^5 ≈ 0.049 if the course is for expert-level rather than for all-level, by giving other variables as constant.

Based on the final model, the number of subscribers of a particular Udemy course depends on number of reviews, days since the course was published, number of lectures in the course, the difficulty level, and its subject.

3.3 Regression Trees

The summary statistic and graphical representation of the regression tree on a training set taken from the data using recursive binary splitting are shown below. The tree is the same as the pruned tree. It has 8 terminal nodes and the residual mean deviance is 0.9278. The predictors that were used in the tree are num_reviews and subject. The test MSE is 0.9427485.

Summary Statistics of Regression Tree Using Recursive Binary Splitting

```
Regression tree:
tree(formula = num_subscribers2 ~ ., data = train)
Variables actually used in tree construction:
[1] "num_reviews" "subject"
Number of terminal nodes:  8
Residual mean deviance:  0.9278 = 1666 / 1796
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.2250 -0.7374 -0.1443  0.0000  0.6784  5.1780
```

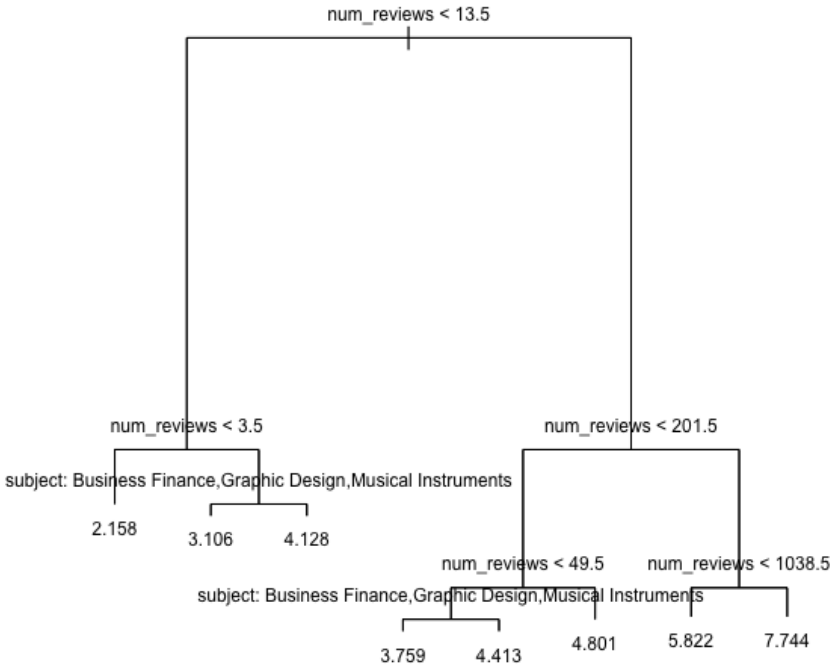

Figure 5: Graphical Output of Regression Tree Using Recursive Binary Splitting

The regression tree using recursive binary splitting answers our question of interest because it shows the predicted number of subscribers for a specific course based on the splits of the predictors num_reviews and subject. Since we transformed the response variable to its power of 0.2, we transformed it back when we interpret it contextually. For example, if num_reviews is less than 3.5 then the predicted number of subscribers is 47 (2.158^5). Num_reviews is the most important factor in predicting the number of subscribers, followed by the subject. Among num_reviews less than 13.5, num_reviews and subject are positively associated with the number of subscribers: the more reviews the higher the number of subscribers. The top number of subscribers are from courses with at least 1038.5 reviews.

Next, we consider using random forests to improve the performance of the tree. There are 8 predictors for random forests, so a new sample of two predictors was chosen as split candidates. Below is a graphical representation of the predictors in order of importance:

Important Predictors of Regression Tree Using Random Forests

	%IncMSE	IncNodePurity
is_paid	18.38733	171.66463
price	26.83590	415.07449
num_reviews	80.65724	1856.73882
num_lectures	22.39350	281.79633
level	15.10376	93.14959
content_duration	22.29737	231.57679
subject	51.72872	636.37959
days_published	32.59195	434.64077

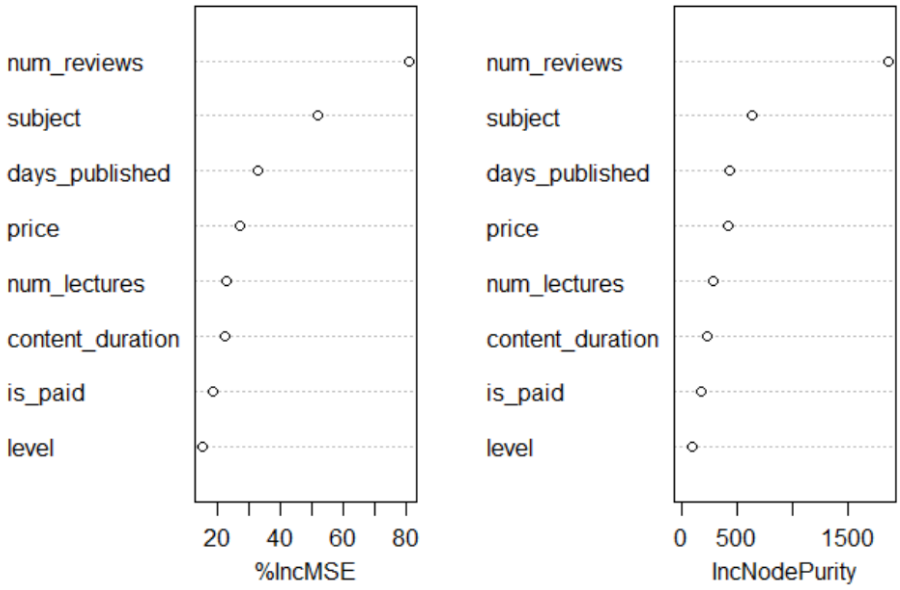

Figure 6: Display of Variable Importance, Using Random Forests

The graphical and numerical output of random forests answers our question of interest as it shows the predictors and their importance in predicting the number of subscribers. From there, it can be seen that the number of reviews, the subject, and how many days it has been published are the three most important factors in the number of subscribers.

3.4 Summary of Findings

According to the table below, the linear regression model has the highest test MSE, followed by the pruned regression tree and then random forests. Random forests produce more accurate predictions as it has the lowest test MSE out of the three.

	Overall Test MSE
Linear Regression	1.491749
Pruned Tree	0.9427485
Random Forest Tree	0.7514318

Table 1: Test MSE with the linear regression model, regression tree, and random forests.

The results from the three methods indicated that they share some common variables that are classified as significant. Two that particularly stood out were subject and the number of reviews as they both appeared in all three. Those two variables also hold the most importance in the regression tree and random forests. The variables if a course is free or not and course level are in the linear regression model, but are not important variables in the regression tree or random forests.

The linear regression model has relatively good interpretability as the model contains significant variables, but it is less accurate than the other two as seen from its test MSE. The regression tree has better interpretability as it gives a clear graphical representation of what variables contribute to the number of subscribers, and the number of subscribers at particular values of the variables. It is also more accurate than linear regression. However, as the tree only uses two variables, the importance of the other variables is unknown. Random forest is the most accurate, and its output gives the weight of the variables, which gives a better understanding of what variables can help predict the number of subscribers, making it the better choice in terms of answering the question of interest.

4. Classification Question

4.1 Exploratory Data analysis

First, we created boxplots by grouping the quantitative predictors by whether the course is paid or free. By looking at the following boxplots, we can see that the free courses have more customers compared to the paid courses as the median of the free class exceeds the median of the paid class. Also, the free courses have relatively more number of reviews, number of lectures and shorter time-length of content, as both medians in the paid class exceed the quartile 3 of free courses. These observations are consistent with our prior knowledge regarding free and paid courses. However, while the paid courses

seem to be published longer according to median, it was surprising to see that the two distributions were similar.

Thus, theoretically, if the course has a smaller number of students and reviews, but offers more lectures and longer content duration, the course is more likely to be categorized as a paid course.

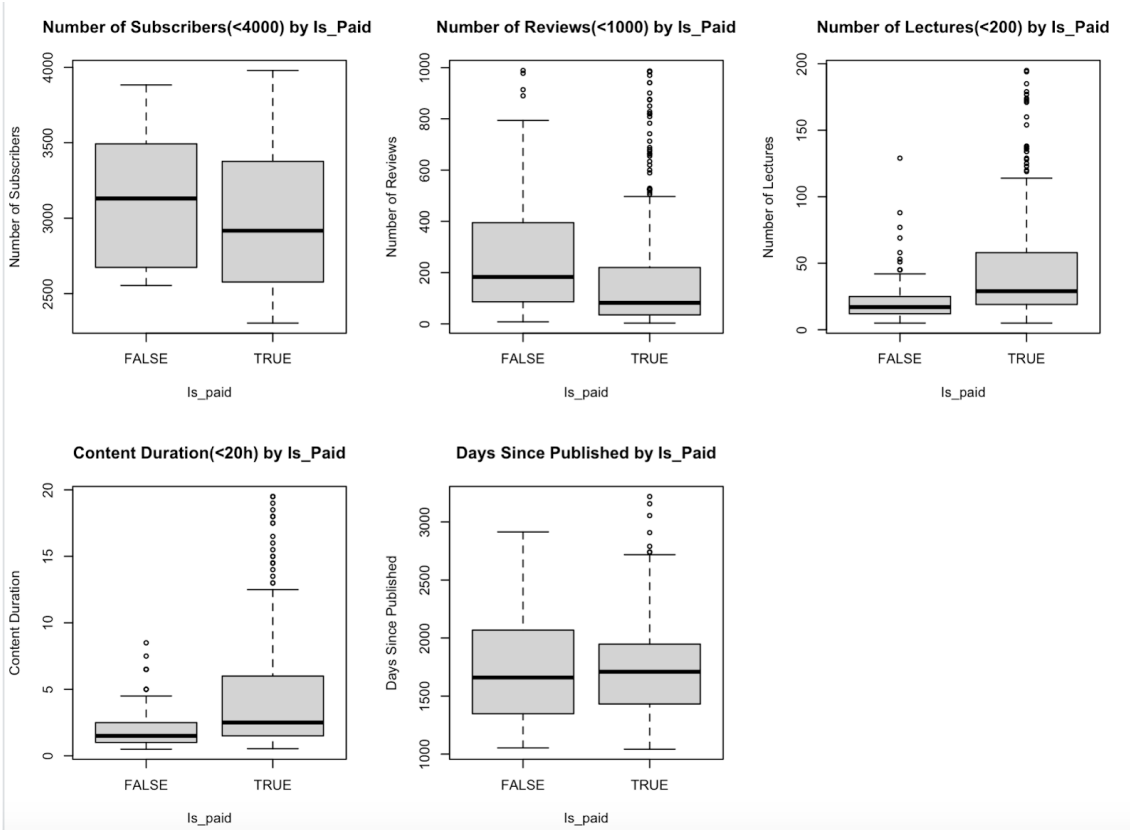

Figure 7: Boxplots of Quantitative Variables by Is_paid

Then, we explored the relationship between the response variable (is_paid) and qualitative variables (course level and course subject). By looking at the distribution of class levels and subjects by class, we can see that most paid and free courses are for all-levels and beginner levels. Expert levels are the lowest proportion of paid and free courses, while intermediate is in the middle, at a proportion of ~6.0%. Also, subscribers are more inclined to take the paid business finance and web development courses, compared to graphic design and musical instruments. Overall, by looking at these plots, we can see that the distribution of course levels and subjects were very different between free and paid courses, indicating that they might be important factors that have influences on whether a course is free or paid.

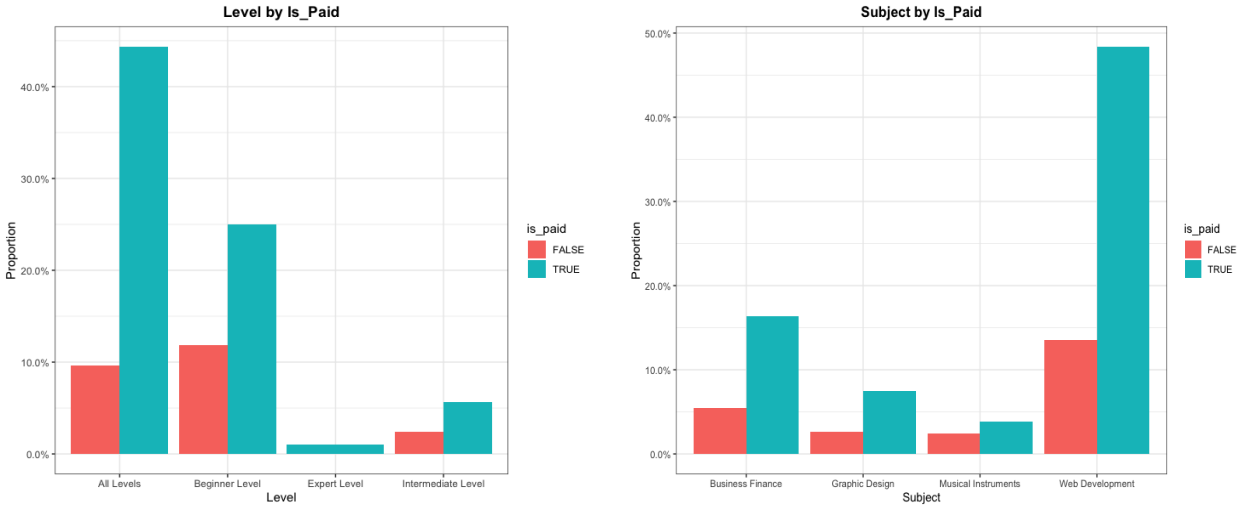

Figure 8 and 9: Distribution of Course Level and Subject by Is_paid

4.2 Logistic Regression Model or Linear Discriminant Analysis

The best logistic regression model generated from milestone 4 is shown below:

Summary Statistics for the Improved Logistic Model

```
Call:
glm(formula = is_paid ~ num_subscribers + num_lectures + days_published,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5924   0.0004   0.4252   0.7613   1.5513

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.298e+00  5.589e-01  -2.321  0.02026 *
num_subscribers -7.929e-05  1.423e-05  -5.571 2.54e-08 ***
num_lectures     4.988e-02  8.326e-03   5.991 2.08e-09 ***
days_published   1.004e-03  3.120e-04   3.219  0.00129 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 546.58  on 495  degrees of freedom
Residual deviance: 435.53  on 492  degrees of freedom
AIC: 443.53

Number of Fisher Scoring iterations: 7
```

Based on this model, Udemy courses that are paid depend on the number of subscribers, number of lectures, days published. The significant variables we used to interpret our equation are as follows:

num_subscribers: The estimated log odds of a Udemy course that is paid increase by exp(-7.929e-05) = 1 for a one-unit increase in the number of subscribers, while holding other variables constant.

num_lectures: The estimated log odds of a Udemy course that is paid increases by exp(4.988e-02) = 1.0511 for the one-unit increase in the number of lectures, while holding other variables constant.

days_published: The estimated log odds of a Udemy course that is paid increases by exp(1.004e-03) = 1.1001 for a one-day increase in the number of days published, while holding other variables constant.

In sum, if the course is longer and has more lectures, then the course is more likely to be paid.

4.3 Classification Trees

The summary statistics and graphical output of the classification tree using recursive binary splitting is shown below. The tree has 19 terminal nodes, with 5 of them being false. The residual mean deviance is 0.5818, and the misclassification error rate is 0.1371. The predictors that were used in the tree are num_lectures, num_reviews, num_subscribers, days_published, level, and content_duration.

Summary Statistics of Classification Tree Using Recursive Binary Splitting

```
Classification tree:
tree(formula = is_paid ~ ., data = train)
Variables actually used in tree construction:
[1] "num_lectures"     "num_reviews"      "num_subscribers"  "days_published"
 "level"            "content_duration"
Number of terminal nodes:  19
Residual mean deviance:  0.5818 = 277.5 / 477
Misclassification error rate: 0.1371 = 68 / 496
```

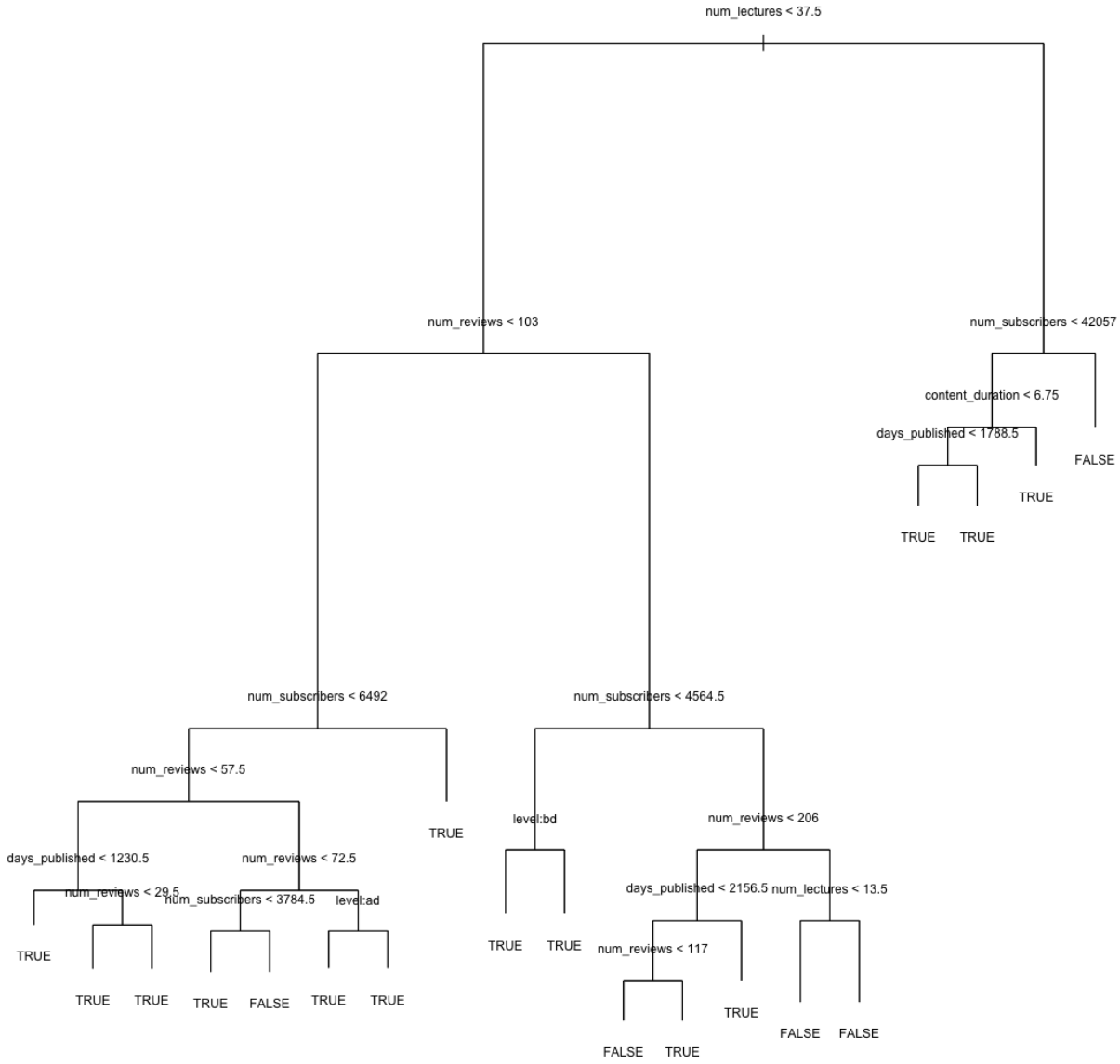

Figure 10: Graphical Output of Classification Tree Using Recursive Binary Splitting

We took the threshold setting into consideration when we chose to present the recursive binary tree over the pruned tree. One practical implication of setting a higher threshold that although we may sacrifice the false-negative rate, which in context meant the course fails to be monetized when it is capable, we believed the cost of false-positive rate -- a free course is set to be a paid course -- is much higher because the course may not be able to get subscribers at the first hand due to an unreasonable price setup. Therefore, we chose the model that has higher precision (or smaller false positive rate) under the threshold of 0.7, which is a recursive binary tree, despite its interpretation being challenging. The recursive binary tree also has a lower mean residual variance (0.5818) than the pruned tree (0.8796).

Given our classification question, the tree not only helps address the important factors but also tells us the important priorities for each predictor. In our case, num_lectures is the most important factor in is_paid, and followed by num_reviews and num_subscibers. Among num_lectures with fewer numbers (fewer than 37.5 lectures), num_reviews and num_subscribers are negatively associated with the class being paid. To be more specific, a course that has less than 42057 for the number of subscribers or less than 103 for the number of reviews is more likely to be categorized as a paid course. And for a course that has fewer num_lectures, more num_reviews, and more num_subscribers, is also more likely to be categorized as a Paid course. However, num_of lectures is still a more important variable.

We used random forests to improve the performance of our trees. Since there are 7 predictors, 2 predictors were chosen as split candidates, as p/3 was used. The figure below shows the most important predictor to predict whether a course is paid or not is the number of reviews, followed by the number of lectures and the number of subscribers based on the mean decrease Gini index.

Important Predictors of Classification Tree Using Random Forests

	FALSE	TRUE	MeanDecreaseAccuracy	MeanDecreaseGini
num_subscribers	23.18934277	0.6507206	16.498063	31.650600
num_reviews	28.89781073	30.3408288	41.099454	42.295956
num_lectures	12.82414440	27.1820519	30.762800	35.804758
level	-0.06522338	11.9697040	9.966358	7.739756
content_duration	3.84753552	25.0610866	25.804087	26.857785
subject	-0.09021071	1.0110426	0.819053	8.143177
days_published	3.58118178	7.0361219	7.865792	26.768314

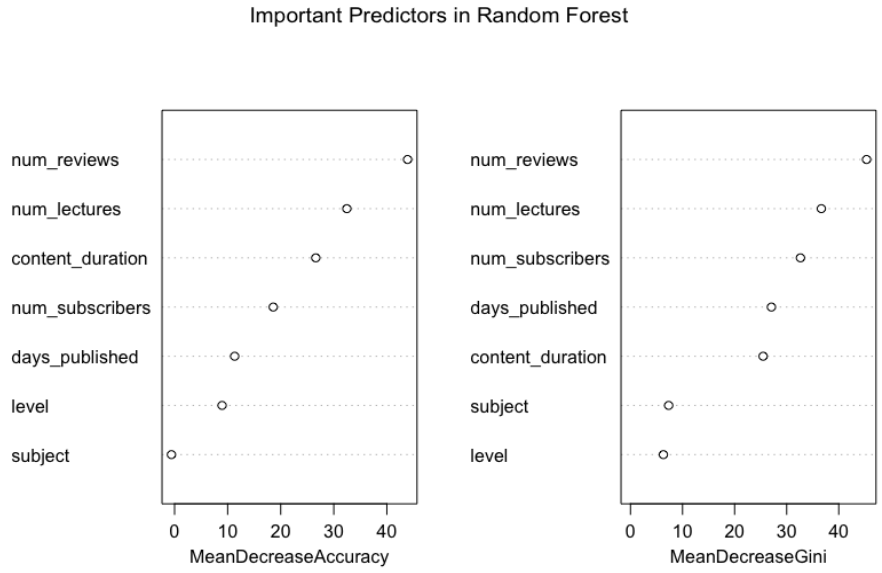

Figure 11: Important Predictors in Random Forest

4.4 Summary of Findings

The confusion matrices from the test data with logistic regression, classification tree, and random forests are shown below:

		Predicted Class	
		FALSE	TRUE
Actual Class	FALSE	24	87
	TRUE	11	374

Table 2: Confusion Matrix of Logistic Regression (Threshold = 0.5)

		Predicted Class	
		FALSE	TRUE
Actual Class	FALSE	41	70
	TRUE	33	352

Table 3: Confusion Matrix of Recursive Binary Classification Tree (Threshold = 0.5)

		Predicted Class	
		FALSE	TRUE
Actual Class	FALSE	56	55
	TRUE	16	319

Table 4: Confusion Matrix of Random Forest (Threshold = 0.5)

Based on the confusion matrices, overall error rates, false positive rates, and false negative rates were calculated and presented in the table below. Under the threshold of 0.5, we can see that Random Forest has the lowest test error rate and false positive rate, while logistic regression has the lowest false negative rate among three models.

	Overall Test Error Rate	False Positive Rate	False Negative Rate

Logistic Regression	0.1976	0.7837	0.0279
Recursive Binary Tree	0.2016	0.6306	0.0857
Random Forest	0.1452	0.4955	0.0416

Table 5: Error Rate of Classification Models (Threshold = 0.5)

However, all three models have very high false positive rates given that our data is still unbalanced. As mentioned in section 4.3, the cost of false positives is higher than false negatives contextually, so we decided to raise the threshold to lower the false positive rate while sacrificing a bit on false negative rate. With experiments, we chose to set the new threshold to 0.7, and the resulting error rates of three models are shown below. We can observe that Random Forest performed the best with the lowest overall test error rate, false positive rate, and false negative rate among the three models.

	Overall Test Error Rate	False Positive Rate	False Negative Rate
Logistic Regression	0.2056	0.2703	0.1827
Recursive Binary Tree	0.2440	0.3153	0.2233
Random Forest	0.1774	0.2342	0.1610

Table 6: Error Rate of Classification Models (Threshold = 0.7)

Comparing the findings from subsections 4.1 to 4.3, we can see that number of subscribers, number of lectures, and days published played important roles in all three models. However, while the number of the reviews is considered as the most important predictor in recursive binary tree and random forest, it wasn't included in the logistic regression model.

In sum, random forest is the best model to answer our question of interest. Given our question that "what are the factors that influence whether a course is free or paid for those having more than 2300 subscribers," the model shows that number of reviews, number of lectures, and number of subscribers are the most important predictors. It also has the best performance with the lowest error rates among the three models.

5. Further Work

If our group had more time to work on this project, we would consider experimenting more statistical models on our data. For the regression problem, we would like to try K-Nearest Neighbors regression since our data does not meet the assumptions for linear regression unless we transform the variable. While this non-parametric method is difficult to interpret, it may perform well since we have much more observations than predictors. For the classification problem, we would also like to apply the KNN algorithm. To potentially improve the performance of tree-based methods, we would perform not only random forest, but also bagging and boosting. Additionally, given our data is very unbalanced, it would be interesting to try resampling methods, such as SMOTE (Synthetic Minority Over-sampling Technique) and Near Miss for undersampling, and explore their advantages and disadvantages.

Besides improving the analyses on current research problems, we would also like to conduct unsupervised learning, such as principal component analysis and clustering to explore if there are subgroups of courses based on the numbers of reviews, numbers of subscribers, price, content duration, and days since published. Such analyses can potentially help content creators to benchmark their course performance, better understand their own market segment, and inform their decision making.