# STAT 340 Group Project Final Report
## Group Name: dolphin
## Group Members:


| Name | NetIDs |
|---|---|
| Tiffany Bahar | tbahar |
| Nabhan Kamarudzman | kamarudzman |
| Eva Lo | mlo23 |
| Jeremy Michael | jmsusanto |
| Jing Kai Ong | jong8 |
| Peter Sykora | psykora |


## Abstract


Excessive alcohol consumption outside of school can impact student academic performance. In this project, we aimed to identify key predictors in student weekend alcohol consumption. Our dataset is *Student Alcohol Consumption: Social, Gender and Study Data from Secondary School Students*, and it contains data from a survey of students in Math and Portuguese language courses from two Portuguese secondary schools. This dataset mostly includes categorical variables and encompasses many aspects of student life, such as academic performance and family information. In the data, weekend alcohol consumption is a categorical variable with five categories ranging from “very low” to “very high”. As these categories are inherently ordered, we employed ordinal logistic regression to fit the data. We also used LASSO and stepwise selection methods to select variables to include in an ordinal logistic regression model. As a comparator, we fitted ordinal random forests to try and improve our final model. To evaluate model accuracy, we calculated the mean absolute deviation and mean squared deviation and compared between models via k-fold cross validation. As a result, we found that our ordinal logistic regression model with stepwise selection performed the best. Based on our model, variables like sex, age, address, family size, father’s job, guardian, travel time to school, study time, activities outside school, nursery school, family relations, going out with friends, and previous school absences were significant predictors of weekend alcohol consumption within these students.

## Section 1: Introduction

Our dataset comes from school reports and questionnaires given to students in Math and Portuguese language courses from two Portuguese public secondary schools from the Alentejo region of Portugal during the 2005-2006 school year. These questionnaires were created by and the data were gathered by Paulo Cortez and Alice Silva. According to Cortez and Silva, the questionnaires were given to supplement school reports since the school reports only contained the grades and number of absences for each student. Furthermore, the data from the questionnaires were combined with each respective student's school report to make two full data sets: one data set for students in Portuguese language courses and one data set for students in math courses. These data were collected because Portugal has overall high student failure rates, especially in Math and Portuguese language courses. Not only does this hinder the success of students in future courses, but it alludes to why Portugal has such a high early school leaving rate compared to the average of the European Union.

Our group used this dataset due to the high number of useful variables contained. With 33 variables, many aspects of these students' lives are covered as these variables range from basic personal information to more complex attributes such as student romantic relationship status. We thought it would be interesting to see what kinds of variables were most important in determining student alcohol consumption.

We believe the reader should care because alcoholism is a serious problem. If there are predictors of high alcohol consumption from an early age (i.e. secondary school) that can be identified, analyses like these could inform intervention programs that combat alcoholism.

## Section 2: Variable Descriptions

***Dependent Variable***

| | |
|---|---|
| **Walc** | **Weekend alcohol consumption (from "1" - very low to "5" - very high)** |

***Independent Variables***

We grouped the variables in the dataset into four groups to allow easier analysis of results.

**Group A: Personal Variables**

| | |
|---|---|
| **sex** | The student's sex ("F": Female or "M": Male) |
| **age** | The student's age, ranging from 15 to 22 |
| **address** | The student's home address type ("U": Urban or "R": Rural) |
| **internet** | Does the student have Internet access at home (Binary: "yes" or "no") |

| | |
|---|---|
| **romantic** | Is the student in a romantic relationship (Binary: “yes” or “no”) |
| **health** | The student’s current health status (Categorical: from 1 - “very bad” to 5 - “very good”) |

**Group B: Academic Variables**

| | |
|---|---|
| **studytime** | The number of hours the student studies in a week. (Categorical: “1”: 1 - 2 hours, “2”: 2 - 5 hours, “3”: 5 - 10 hours, “4”: More than 10 hours) |
| **failures** | Number of past class failures (From 0 to 3, 4 means more than 3) |
| **schoolsup** | Is the student receiving extra educational support (Binary: “yes” or “no”) |
| **paid** | Is the student taking extra paid Math classes (Binary: “yes” or “no”) |
| **nursery** | Did the student attend nursery school (Binary: “yes” or “no”). By definition, a nursery school is, “a school for children usually under five years old.” |
| **higher** | Does the student want to take higher education (Binary: “yes” or “no”) |
| **G1** | The student’s first period grade, ranging from 0 to 20. This is similar to, for example, the first midterm score of a university student. |
| **G2** | The student’s second period grade, ranging from 0 to 20.This is similar to, for example, the second midterm score of a university student. |
| **G3** | The student’s final grade, ranging from 0 to 20. This is similar to, for example, the final exam score of a university student. |

**Group C: Family Variables**

| | |
|---|---|
| **famsize** | The size of the student’s family (“LE3”: Less than or equal to 3 or “GT3”: Greater than or equal to 3) |
| **Pstatus** | Do the student’s parents live together or separately (Binary: “T”’ - living together or “A” - apart).<br>The status of parenthood can emotionally affect students, which might influence how much alcohol they consume as a coping mechanism to their emotional trauma. |
| **Medu** | The student’s mother's education level. (Categorical: “0” - none, “1” - primary education (4th grade), “2” – 5th to 9th grade, “3” – secondary education or “4” – higher education)<br>The mother’s education level might influence the way the student is brought up at home, which in turn might influence alcohol consumption. |

| | |
|---|---|
| **Fedu** | The student's father's education level. (Categorical: "0" - none, "1" - primary education (4th grade), "2" – 5th to 9th grade, "3" – secondary education or "4" – higher education)<br>The father's education level might influence the way the student is brought up at home, which in turn might influence alcohol consumption. |
| **Mjob** | The student's mother's current employment position and status in the workforce. (Categorical: "teacher", "health", "services"- civil services like administrative or police, "at_home", and "other") |
| **Fjob** | The student's father's current employment position and status in the workforce. (Categorical: "teacher", "health", "services"- civil services like administrative or police, "at_home", and "other") |
| **guardian** | The student's guardian. (Categorical: "mother", "father" or "other"). If the student lives with a single mother or father, then "mother" or "father" would be chosen respectively. If the student lives with both parents, one of the parents is chosen to be the guardian. Finally, 'other' is chosen if neither the student's mother or father is their primary guardian. |
| **famsup** | Is the student receiving educational support through their family (Binary: "yes" or "no") |
| **famrel** | The quality of student family relationships (Categorical: from "1" - very bad to "5" - excellent) |

**Group D: Other School Variables**

| | |
|---|---|
| **school** | The student's school ("GP": Gabriel Pereira or "MS": Mousinho da Silveira) |
| **reason** | The student's reason for going to their school. (Categorical: "home" - close to home, "reputation" - school's reputation, "course" - course preference, and "other") |
| **traveltime** | The time required for the student to travel from their living space to school (Categorical: "1" - less than 15 minutes, "2" - 15-30 minutes, "3" - 30-60 minutes, and "4" greater than one hour) |
| **activities** | Is the student involved in extracurricular activities (Binary: "yes" or "no") |
| **freetime** | The student's free time after school (Categorical: "1" - very low to "5" - very high) |
| **goout** | How frequent the student goes out with friends (Categorical: from "1" - very low to "5" - very high) |
| **absences** | The number of school absences the student has (Numeric: from 0 to 75) |

## Section 3: Statistical Question

Question: "*What are the most significant predictors in determining weekend alcohol consumption of the secondary students surveyed who are enrolled in a Math course?"*

Since this dataset consists of a lot of variables, we were curious to see which variables best predict weekend alcohol consumption of secondary students enrolled in a Math course. We chose to examine the students within the Math course specifically because we are studying math-related majors, and the results are more generalizable than if we look at the Portuguese class. More generally, we chose this statistical question because we all recognize the negative effects that alcoholism has as a whole, whether that be the individual effect on the person themself or the effect on others with which this person has relations.

Overall, gaining this insight into what predictors may affect weekend alcohol consumption is not only beneficial to others, such as people who have built careers in better understanding alcoholism and for programs that intervene with those who may be at high risk for future alcoholism, but it is also beneficial to our own insight of our peers. Not too long ago, we also were of the age of these students within this dataset, and so were others within this course. Answering this statistical question will not only allow us better understanding of our peers, but also a better understanding of people in general in determining early age predictors may affect future alcoholism.

## Section 4: Summary of Methods

### 4.1 Variable EDA

With 33 variables in the dataset, we were interested in an initial exploratory analysis of each variable. We were interested in observing the different distributions that each variable carries. For example, the variable "sex" might have a different percentage of M and F in comparison to the variable "Pstatus" of Alone or Together. From this, we decided that each member take up a number of variables of up to 6 variables per person for the exploratory data analysis. Each variable was analyzed through its frequency, distribution, and occurrence of observations. Most variables used boxplots and barplots to visualize the data for each variable. One important analysis was comparing between the Weekday Alcohol Consumption variable (Dalc) and the Weekend Alcohol Consumption variable (Walc), which led to prioritizing Walc as the main response variable.

We looked at how all of the variables are related to the alcohol consumption level during the weekdays and the weekends. After looking at the plots and seeing that there are a lot more variation in the weekend alcohol consumption level, we decided to focus on the weekend alcohol consumption level instead of the weekday alcohol consumption level.

The best visualizing plot which helped us in making this decision was in the exploratory analysis of the variable "sex". In Figure 1 and Figure 2, we are comparing how the alcohol consumption level varies by sex. Since there seems to be a difference between the number of males and females in the sample collected, we decided to look at the proportion, instead of count.

As seen in the both figures, the proportion of females seems to outweigh the proportion of males in lower levels of alcohol consumption.

However, as the level of alcohol consumption increases, the proportion of males seems to outweigh the proportion of females. It would be interesting to look into whether sex actually has an effect on the alcohol consumption level. There seems to be a huge difference between Figure 1 and Figure 2 as well. As we move from weekday to weekend, the proportion of both sex in the lower level of alcohol comsumption decreased significantly, while the proportion of both sex in the upper level of alcohol consumption increased signficiantly. The shift is apparent, which strengthens the idea that Dalc seems to be of better use for a response variable.

Hence, from the plots in each figure, we can reasonably conclude that sex is a significant variable and include it in our model.

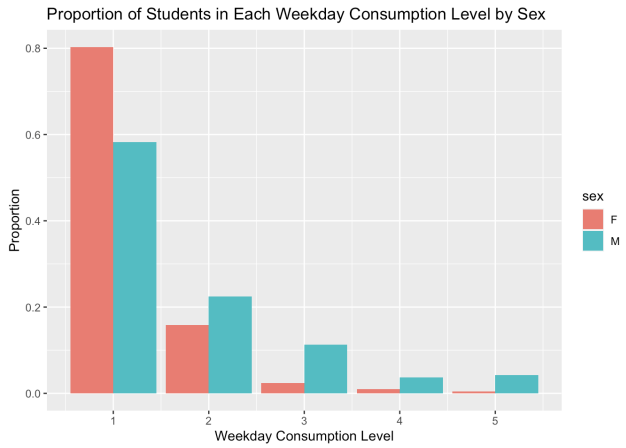


**Figure 1: The Proportion of Students for Weekday Alcohol Consumption Levels by Sex**


**Figure 2: The Proportion of Students for Weekend Alcohol Consumption Levels by Sex**

Naturally, having 33 variables makes it difficult to rely on regular visualization models to decide on which independent variables are most significantly affecting the response variable.

From our initial EDA, our intuition decided that variables like “sex”, “goout” etc. were the variables most likely to have high correlation with Weekend Alcohol Consumption, based on the plots produced comparing each variable. From here, we decided to proceed by running a LASSO regression and a stepwise selection model in deciding which variables are most relevant in order to improve our analysis and performance of our model, which are finally fed into ordinal logistic regression models.

### 4.2 LASSO Regression

As there are 33 variables in the dataset, we utilized LASSO regression to identify which variables seemed to be the most significant to predict weekend alcohol consumption. We used LASSO because it made sure that we only pick the most relevant variables in our model, and reduce the coefficients of insignificant variables to 0.

We began by employing a **multinomial LASSO logistic regression** as the baseline model in initializing LASSO, and we fit the LASSO model five times: once each on all four categories of our variables, and finally on all the variables in our model. For each analysis, we did a 50/50 split of our data and ran cross-validation to find the best lambda value (penalty term), then fit the model with that best lambda value. As our dependent variable (weekend alcohol consumption) has five levels, we had five distinct outputs of significant variables for each output. To determine which variables were most significant, we only chose variables that had non-zero coefficients in at least three of the output levels (for categorical variables, it sufficed if any of its levels appeared in at least three output levels). We then compared these variables with the variables that were deemed to be significant by other models.

After running LASSO on multinomial logistic regression as above and meeting with Keith, we determined that it was better to do variable selection using **ordinal LASSO logistic regression** rather than multinomial, as the outcomes variable of weekend alcohol consumption is ordered, ranging from “very low” to “very high”. Hence, we used the ordinalNet function from the ordinalNet package to run LASSO on ordinal logistic regression models.

When choosing the link function to use, we notice that different link functions work better depending on how the outcome variable was distributed across categories. As seen in the table below, as the outcome was unevenly distributed across categories, we used a complementary log-log link function as it works well with unevenly distributed outcomes, in addition to the default logit link function.

| very low | low | moderate | high | very high |
|---|---|---|---|---|
| 151 | 85 | 80 | 51 | 28 |

**Table 1: Frequency Table of Weekend Alcohol Consumption**

Once we ran the ordinal LASSO logistic regression based on both link functions, we took note of the variables that had a non-zero coefficient and passed that to the ordinal regression model, discussed below in Section 4.5.

### 4.3 Stepwise Model Selection

To examine and identify the most significant variables for a model, we performed stepwise model selection on both linear and ordinal logistic regression. However, we believe that ordinal logistic regression would be more appropriate because our response variable is categorical. Even though fitting a linear regression does work, having a response variable that falls in between the categorical level in decimal forms would be difficult to understand. It fails to provide meaningful interpretations. Hence, we decided to continue our analysis with ordinal logistic regression, and leave linear regression as an exploration process.

The stepwise model selection was conducted in R using the step() function, which chooses models based on AIC. At first, we were uncertain whether AIC would be a proper measure of the models because we wanted to be able to take into account the distance between the predicted and the actual level of alcohol consumption. Due to this reason, we decided that developing a step() function that could take different loss functions would help us analyse the best loss function to use. The idea was to implement a for loop within the function which takes a list of different loss functions and compares each value. This would help improve the performance of the models because we can select the most optimal loss function. Unfortunately, implementing a different step() function proved too difficult and time consuming to implement. We concluded later that AIC is appropriate because it is computed using the maximum likelihood estimate and we couldn't find a better loss function at the time. It is something that should be further explored in the future research questions.

By running the stepwise model selection, we found that both the forward and backward method provided the same final models. These two selections also had the same exact variables selected in the final models with the same exact coefficients. After doing both methods, we tried to make predictions on the same dataset using the models we just found. To see how accurate the models were, we compared the predicted Walc value to the original Walc value. We also calculated the RSS of the models by summing and squaring the difference between the predicted Walc value and the original Walc value. The accuracy of the model selected falls around 49% with an RSS of 512, suggesting that our model is not fitting the data very well. The variables that were considered significant by the stepwise model selection include: sex, age, address, famsize, Fjob, guardian, traveltime, studytime, paid, activities, nursery, famrel, goout, absences.

### 4.4 Ordinal Random Forest

To quote a well-known statistician, "the random forest is magic." Random forests have good predictive power, but the outputs can be harder to interpret and are less intuitive compared to ordinal logistic regression. Nevertheless, we decide to include Ordinal Random Forest as an alternative to the ordinal logistic regression models that we have built. Additionally, and somewhat unorthodoxly, we will take the top few variables that are deemed significant by the random forest models and use them in our ordinal logistic regression model, treating the random forest as another variable selection method.

To predict ordinal variables using random forests, we use the ordfor() function from the ordinalForest package. We will use two different performance functions used to train the random

forest: “equal” and “proportional”. “equal” classifies observations from each class equally accurately (by computing Youden’s J statistic for each class), while “proportional” aims to classify as many observations right as possible (computes Youden’s J statistic for each class, weights by number of observations in class to prioritize classes with more observations).

We then look at the relative importance of each variable in predicting alcohol consumption in the model. We use the Permutation variable importance of the variables to determine how important the variables are to be used in our model. We caveat these findings with the fact that variables with more factors are more likely to be seen as significant in random forest models, due to the fact that they have more categories. We choose variables that have a coefficient with a minimum value of 3 decimal places, which comes out to 12-13 variables for each random forest model, and we will fit ordinal regression models based on these variables too.

We also assess the model accuracy of the random forest models to be compared with the ordinal regression models. As the random forest models take some time to run, we will only run k-fold validation once on both models, and use k = 5 to allow for an 80/20 train-test split and divide our data evenly. The model error rates are shown below, produced by a single 5-fold cross validation. The green box indicates that the error value is lower, and it seems like the random forest model using “equal” performance function is better than the model using “proportional”. We will also compare these error values with the ordinal regression model later.

| | **Model 1 (“equal”)** | **Model 2 (“proportional”)** |
|---|---|---|
| **Zero-One Error (Accuracy)** | 0.3899 | 0.3899 |
| **Mean Absolute Deviation** | 1.1342 | 1.1519 |
| **Mean Squared Deviation** | 2.6278 | 2.7671 |

**Table 2: Comparison between Random Forest Models with Differing Performance Functions**

During the cross-validation process, we also created a confusion matrix for each fold to visualize the patterns in our predictions, and below we provide an example of a confusion matrix from the fold that has the lowest Mean Absolute Deviation. Here, 1 corresponds to “very low” weekend alcohol consumption, while 5 corresponds to “very high.”

We notice that the model tends to predict ‘1’, i.e. “very low” alcohol consumption, and does not really do well identifying high alcohol consumption users. This could pose a problem, as it is important for our model to identify students with high alcohol consumption. To improve this model in the future, we could assign more weights to the higher alcohol consumption categories during the model training process to better predict higher alcohol consumption in students.

| | pred1 | | |
|---|---|---|---|
| ans | 1 | 4 | Row Total |
| 1 | 32 | 3 | 35 |
| 2 | 12 | 1 | 13 |
| 3 | 17 | 2 | 19 |
| 4 | 5 | 4 | 9 |
| 5 | 2 | 1 | 3 |
| Column Total | 68 | 11 | 79 |

| | pred2 | | | | |
|---|---|---|---|---|---|
| ans | 1 | 3 | 4 | 5 | Row Total |
| 1 | 32 | 1 | 2 | 0 | 35 |
| 2 | 12 | 0 | 1 | 0 | 13 |
| 3 | 17 | 2 | 0 | 0 | 19 |
| 4 | 5 | 0 | 4 | 0 | 9 |
| 5 | 2 | 0 | 0 | 1 | 3 |
| Column Total | 68 | 3 | 7 | 1 | 79 |

**Table 3: Confusion Matrices for "equal" (left) and "proportional" (right) Random Forests**

### 4.5 Ordinal Logistic Regression - Final Models

The ordinal logistic regression models were created using the polr() function in R from the MASS package. In total, we had 6 unique ordinal logistic regression models that were created using 6 different predictor combinations, all of which were derived with reason.

Model 1's predictors were chosen based on their p-values evaluated at a significance level of 0.1. Though the polr() function does not output the significance levels of each variable, we were able to derive them by comparing each variable's t-value against a normal distribution. This is the same approach used by other statistical software packages such as Stata. Initially, we created a model using every single available variable. Afterwards, we repeatedly removed statistically insignificant variables until we arrive at a model (model 1) where all the predictors have a p-value of less than 0.1.

Model 2 and Model 3's predictors were derived from the LASSO regression in Section 4.2. Model 2's predictors were chosen based on the default logit link function, and Model 3's predictors were chosen from the complementary log-log link model function. Similarly, Model 4 and Model 5's predictors were derived from the ordinal random forest model in Section 4.4. Models 4 and Model 5 use the variables from the two different performance functions used to train the random forest, "equal" and "proportional," respectively. Finally, Model 6 uses variables that were considered significant from the stepwise model selection, as mentioned in Section 4.3.

Prior to running K-fold to compare the performance of each model, we wanted to evaluate each model through a hypothesis test, where the null hypothesis states that a model without predictors (only an intercept term) performs just as good as a model with predictors. To achieve this, we first created a model using the polr() function with only an intercept term. Afterwards, the anova() function was used as a likelihood ratio test of ordinal logistic regression and to arrive at a p-value. The result was that all 6 models had a significant p-value, suggesting that the predictors certainly had an impact on the model performance.

The use of interaction terms was something that we experimented with and ultimately decided not to include in any of our models. Given the overwhelmingly large number of available predictors, there were simply too many interaction terms to consider.

To ensure that each model is not overfitting the data, we used an 80/20 training/test split and 3 different loss functions which have been previously shown in Table 2. For each observation, the Zero-One loss function outputs a value of 1 upon predicting a correct label, and a value of 0 in all other cases. Using this loss function, we were able to also derive the % accuracy for each model in both the training and test sets. The problem with this loss function is that since we have 5 possible ordered labels, the loss function does not consider the magnitude of the loss. For example, given a true label of 1, we want the loss function to punish a model prediction of 5 more than a model prediction of 2. To overcome this problem, we used mean absolute deviation as our second loss function. This loss function computes the absolute value of the difference between the true label and the predicted label for each observation. The final output is the average of these values across all observations. The last loss function is the mean squared deviation. This is similar to the mean absolute deviation, but with the squared distance rather than its absolute value. In research and mathematical computation, this loss function may be favored as it is differentiable. However, there is no real benefit of using the squared distance over the absolute value in our case. The result can be seen in Figure 3.

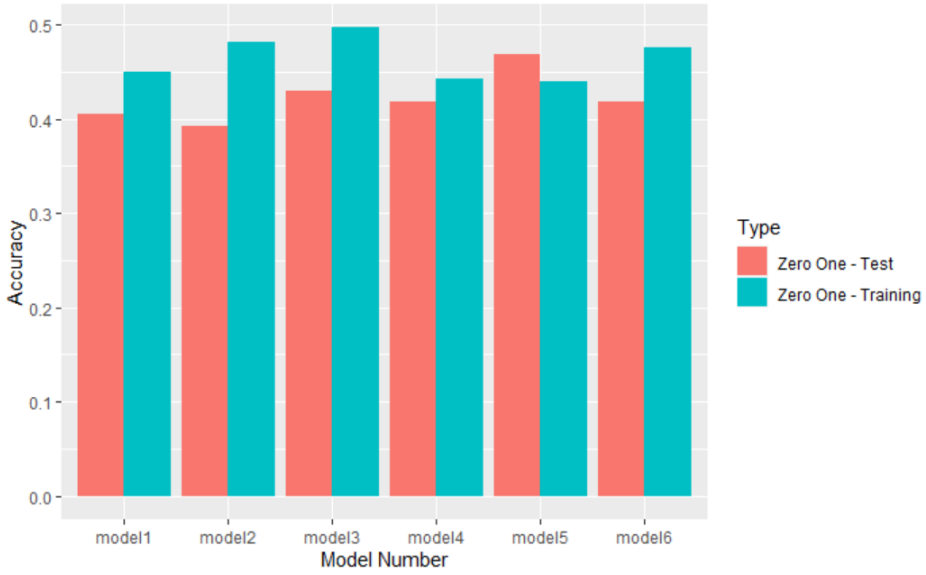


**Figure 3: Accuracy of Each Model Computed using the Zero-One Loss Function**

Figure 3 allows for a comparison of performance of each model. The training and test sets are exactly the same for each model in Figure 3. As can be seen from Figure 3, for almost every model, the accuracy of the training set is greater than the accuracy of the test set, which makes intuitive sense. The lowest model accuracy was model 2 on its test set, and the highest accuracy was model 3 on its training set.

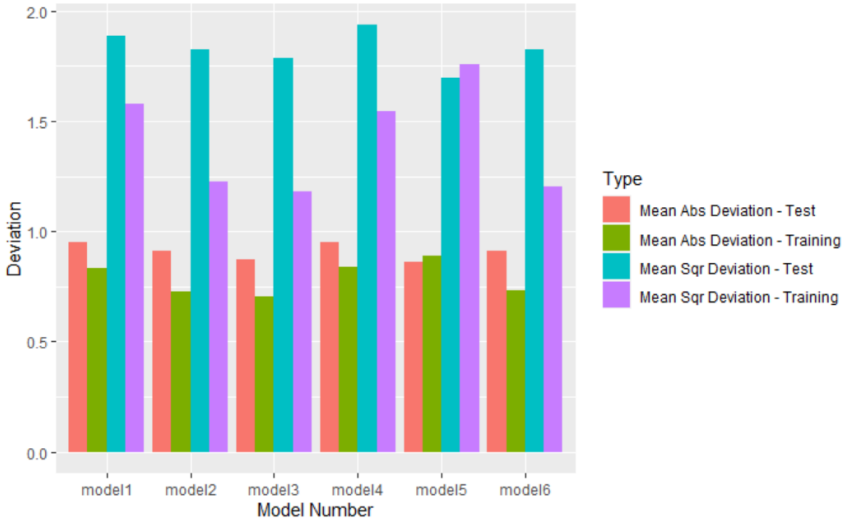


**Figure 4: Mean Absolute and Squared Deviation of Each Model**

From Figure 4, we can see that the mean deviation of each model is relatively similar. The purpose of Figure 3 and Figure 4 is to demonstrate the necessity of the use of k-fold in selecting our final model.

Given the six models from above, we decided to use K-fold cross validation to evaluate these models' prediction error. In using K-fold cross validation, we decided to use 5-fold regularization because we have 395 observations in our dataset so that when we divide our data into 5 folds, we divide it evenly. Furthermore, we also chose to use 5 folds in consideration of our choice of an 80/20 training/test split that we used in our previous reports. Now, in looking at the process of K-fold cross validation, we know that one of the steps is to randomly partition the data into these same-sized subsets. We initially ran K-fold cross validation with K = 5 on a singular random 80/20 training/test split that we had previously been using to evaluate our models; however, we knew that to have stronger evidence of what model had the best prediction error, we should consider averaging due to this randomness in the partitioning of the data. So, we decided to use a Monte Carlo approach, and we created and ran a Monte Carlo experiment that took the average of the averages of each repetition. What we mean is that for each repetition, we aggregated over the K = 5 folds and found the means for each model for that repetition, and then at the end of the experiment, we took the means of each model over all of the repetitions. Let us now show the output of this experiment with 5 repetitions.

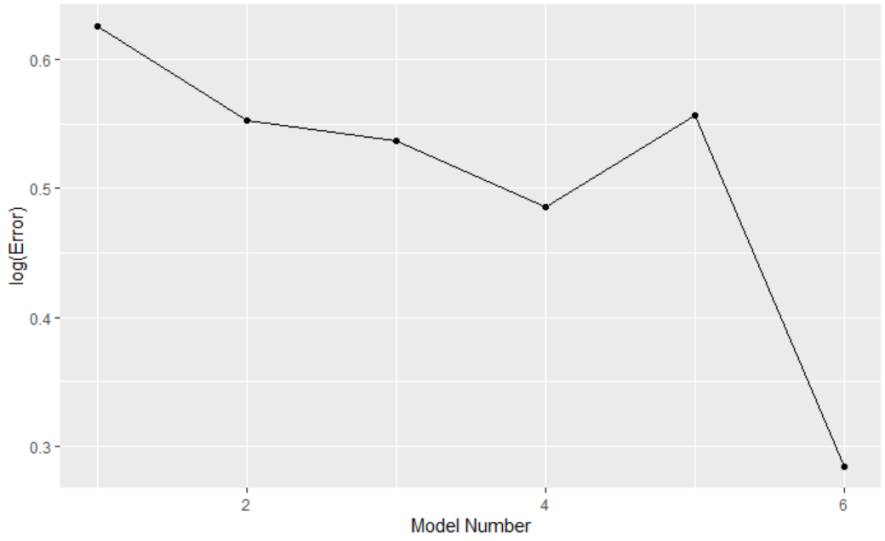


**Figure 5: Result of K-fold Monte Carlo Simulation**

From this, we can conclude that model 6, which uses variables that were considered significant from the stepwise model selection, has the best performance and is thus the final model our group arrived at. The variables used as predictors in model 6 are `sex`, `age`, `address`, `famsize`, `Fjob`, `guardian`, `traveltime`, `studytime`, `paid`, `activities`, `nursery`, `famrel`, `goout`, and `absences`. Moreover, here is a figure that shows the coefficients of the final model.

```
Coefficients:
                  sexM                    age           addressUrban              famsizeLE3             Fjobhealth
            1.20957108             0.07300309            -1.03161652             0.27020068             0.04382604
              Fjobother           Fjobservices             Fjobteacher          guardianmother          guardianother
            1.10287549             1.51601608             0.47258665            -0.54717450            -0.72966465
       traveltime>1 hr traveltime15-30 min traveltime30 min - 1 hr       studytime>10 hrs       studytime2-5 hrs
            4.11475949            -0.35633661             0.20213732            -1.71751275            -0.51393434
      studytime5-10 hrs                paidyes          activitiesyes             nurseryyes       famrelexcellent
           -0.24090446             0.86398485            -0.57453302            -0.77155554            -1.01310161
            famrelgood         famrelmoderate          famrelvery bad                gooutlow          gooutmoderate
           -0.58992569             0.01495893             0.78691079            -2.50240536            -1.10345236
         gooutvery high          gooutvery low                absences
            0.59808427            -2.28881504             0.04085643
```

**Figure 6: Coefficients of Variables in the Final Model**

Note that positive coefficients suggest that an increase in a variable on average would also lead to an increase in weekend alcohol consumption.

### 4.6 Model Performance

The model performance has some variation due to the random nature of the 80/20 train/test split. Here we show one instance of the performance of the final model.

| | **Training Set** | **Test Set** |
|---|---|---|
| **Zero-One Error (Accuracy)** | 0.472 | 0.430 |

| | | |
|---|---|---|
| **Mean Absolute Deviation** | 0.741 | 0.873 |
| **Mean Squared Deviation** | 1.234 | 1.709 |

**Table 4: Performance of the Final Model**

The final model has an accuracy of around 43% with a mean absolute deviation of 0.741.

Total Observations in Table: 316

| actual_value \| predicted_value | 1 | 2 | 3 | 4 | 5 | Row Total |
|---|---|---|---|---|---|---|
| 1 | 105 | 8 | 10 | 4 | 0 | 127 |
| 2 | 46 | 5 | 14 | 1 | 0 | 66 |
| 3 | 22 | 4 | 16 | 13 | 0 | 55 |
| 4 | 6 | 3 | 14 | 19 | 2 | 44 |
| 5 | 1 | 1 | 4 | 8 | 10 | 24 |
| Column Total | 180 | 21 | 58 | 45 | 12 | 316 |

**Table 5: Confusion Matrix of the Final Model on the Training Set**

## Section 5: Findings and Future Steps

In using K-fold cross validation, we found that the Model 6, where the predictors were chosen using stepwise model selection performed the best. The performance of this model can be seen in Section 4.6, and the details of this model can be seen in Section 4.5. We also found that the ordinal logistic regression outperformed the ordinal random forest models based on the loss functions we defined. For the ordinal random forest models that were fitted, we found they tend to over-predict the “very low” alcohol consumption level as compared to higher consumption levels. We hypothesize that this could be caused by a large number of observations having a “very low” alcohol consumption level.

Through these findings, we found potential ways that we could improve our current model. One way we could improve our current model is through introducing weights to the loss function and giving higher weights to higher alcohol consumption levels, as it is more important to identify students with high alcohol consumption. One potential drawback of this, however, is that students who do not have high alcohol consumption could be falsely flagged as such.

Another way that we could potentially improve our current model is through utilizing different values of K when doing K-fold cross validation. Instead of just using K = 5, we could consider K = 6, 7, 8, 9, and/or 10, especially since K = 5 through K = 10, are common choices of K. Moreover, we also think that we could average the errors only once in order to get a more

accurate prediction error for each model. Instead of averaging on each repetition when using the Monte Carlo with K-fold cross validation, we could keep those unaveraged errors and then only average on the complete data across all of the errors for every repetition after all of the repetitions have finished.

A final future step that we could take to improve our model would be to implement a Backwards Stepwise Selection with various loss functions, something mentioned earlier in Section 4.3. Here, the process would be to create an alternative stepwise selection method which takes up different loss functions and compares the values of their predicting errors. Because we want to minimize the loss function, we can see which loss function would fit best with the Backwards Stepwise Selection. Most stepwise selection models are run based on the Akaike Information Criterion (AIC) in determining prediction errors, which might not be suitable for certain predictive models. In this scenario, we can run multiple models with different loss functions which might help us evaluate on the best minimized loss function to use to improve the performance of our model. Implementing such a method in the future can help improve the performance of future models.

Some potential future research questions that we could ask include what are the most significant predictors in determining weekday alcohol consumption. In considering our focus on weekend alcohol consumption, this future question seems feasible as we could simply try running our same analyses using this very similar dependent variable. Moreover, we could also try to ask what the most significant predictors were in determining overall alcohol consumption. In doing this, we would also beg the question as to how to utilize the already existing weekend alcohol consumption and weekday alcohol consumption variables together to answer how the predictors would predict some sort of total alcohol consumption.

Overall, we think that these potential future research questions would shed even more light on what are effective predictors of above average alcohol consumption for students such as those in secondary school. Moreover, we think that answering more general questions such as the latter could aid in more general approaches to combating future alcoholism in students and younger persons alike.

## References

The dataset on *Kaggle*: https://www.kaggle.com/uciml/student-alcohol-consumption
The source article of the dataset: http://www3.dsi.uminho.pt/pcortez/student.pdf
*GitHub* repository for this project: https://github.com/jingkai02/stat340