Predicting Alcohol Consumption

Arda Aysu, Shivakanth Thudi, Francisco Calderon

12/07/2016

Introduction

The goal of this project is to predict the level of alcohol consumption of students based on certain attributes of their lives. We can categorize these attributes as falling into three groups:

- **Demographic factors such as sex, age, family size and occupations**
- **Academic factors such as grades, study time and past failures**
- **Social factors such as the amount of free time after school, internet access, extra-curricular activities and quality of family relationships**

The data we use in this project was collected on students from two Portuguese secondary schools. In our analysis, we treated this problem as both a regression and classification problem (distinctly). We chose to present it as a classification problem with five levels since the alcohol consumption in our data appears as a ranking from 1 (low consumption) to 5 (high consumption). Initially, however, we also performed regression since the class had equidistant ordering.

Data Sources

The data for this project was downloaded from the UC Irvine Machine Learning Repository. The original data was collected on secondary school students during the 2005-2006 school year from two public schools in the Alentejo region of Portugal. Paulo Cortez and Alice Silva compiled this data from two sources: school reports in a paper format, and questionnaires that were used to supplement the previous information. They integrated the data into two datasets, with one based on a Mathematics class and the other on a Portuguese class.

Preprocessing:

We merged the data from the two classes into a single dataset and also added binary variables to indicate whether each observation came from the Math or Portuguese class. While merging the data we found that several students took both classes and belonged to both datasets; consequently, we removed these duplicates. We had a total of 662 unique students and there were 34 attributes associated with each student. The target variable was the "Walc" variable which corresponded to the weekend alcohol consumption. The features were comprised of the 33 other attributes and the categorical features among them were encoded as binary variables.

See the appendix for a detailed description of all the attributes present in our model.

Methods

Before fitting any model for classification, the first consideration should be to pick an appropriate metric and a baseline. We chose 10-fold cross-validated accuracy as the metric to score our models. Since there are five classes, random guessing would net us around 20 % accuracy. However, when we look at the frequency of students according to the level of alcohol consumption, we note that the most frequent level was 1 (low) - around 38 %. So a more appropriate baseline would be 38 % accuracy. Our intention is to build models and evaluate their cross-validated accuracy, and to pick the algorithm that has the highest improvement over the baseline.

Algorithms Considered:

Many algorithms were considered during our analysis since this problem could either be treated as a regression or classification problem.**For the regression approach, we initially chose Linear Regression and K-Nearest Neighbors**. For this step, we did not do any particular feature selection, but instead fed both algorithms all the features, both numeric and categorical (with categorical variables encoded as binary features) to see if there was a linear relationship between the features and the target.

Our intuition was that a higher cross-validated MSE (mean squared error) from linear regression over kNN with the same feature set would imply that a linear relationship existed between the features and the target; the opposite would be true otherwise. For the kNN approach, we considered a wide range of values of k (the number of neighbors) and used the one with the lowest cross-validated MSE. However, the results from regression were less interpretable than desired, so we decided to solely continue with the classification approach.

For classification, we considered K-Nearest Neighbors Classifiers, Linear Discriminant Analysis, Decision Tree Classifiers, Random Forest Classifiers, Ada Boosting and MLP Classifiers. We took advantage of an attribute inside the Decision Tree Classification object called **Feature Importance** that tells us how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means "not used at all" and 1 means "perfectly predicts the target." The feature importances always sum to 1.

We used these feature importances to select and train our models on a subset of the feature space (7 variables). When we used this subset of features (outlined below) rather than all of them, we found that the cross-validated accuracy was higher across all the algorithms. All accuracy scores reported are 10-fold cross validated.

Final Features Used:

- F_sex = 1 if female, 0 if male
- 2_failures = 1 if student has 2 failures, 0 otherwise
- no_schoolsup = 1 if student has school support, 0 otherwise
- 18_G2 = 1 if student has second period grade of 18, 0 otherwise
- 1_Dalc = 1 if student has very low alcohol consumption during workday, 0 otherwise
- 2_Dalc = 1 if student has low alcohol consumption during workday, 0 otherwise
- 3_Dalc = 1 if student has moderate alcohol consumption during workday, 0 otherwise

Results:

K-Nearest Neighbors:

KNN performed really well with an accuracy of 52%. This would be a 39% improvement over the baseline. KNN was initially used with the number of nearest neighbors equal to 5, which gave us an accuracy score of 40%. But by tuning the hyper-parameter through cross validation, we were able to achieve the maximum accuracy of 52% by using 20 as the number of nearest neighbors.

Linear Discriminant Analysis:

Performing LDA on the subset of features we felt were most related to alcohol consumption gave us an accuracy of 50%. Considering that the baseline is 38%, this is an improvement of 31% in accuracy.

Decision Tree Classifier:

The decision tree classifier performed admirably by giving us a 53% accuracy score. That is 39% over the baseline. For this algorithm, we used the 'gini' index as the criterion and set the maximum depth of the tree to 3.

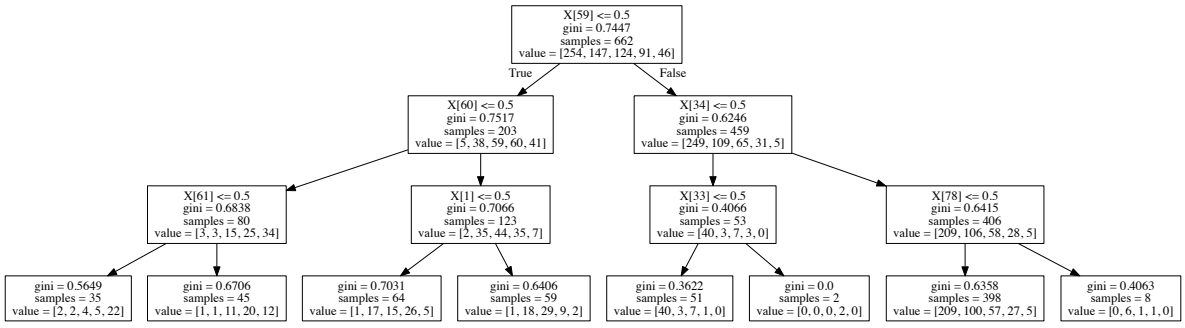

Figure 1: Decision Tree

Random Forest Classifier:

Random forests, which are collections of decision trees, were also strong contenders with an accuracy score of 51% which is 34% over the baseline. We tuned the number of trees to use through cross-validation, and found 25 trees to be the optimal value to use.

Ada Boosting Classifier:

This model improved the baseline by nearly 24%, but it was the worst performing model with an accuracy of 47%. The optimal learning rate was .7, which gave us this accuracy score.

Multi-Layer Perceptron Classifier:

Just to dip our feet into neural network waters, we also considered the MLP Classifier that performed extremely well with 52% accuracy. We used 3 hidden layers with 40 units in each. We had no prior intuition about the hidden layer size to use so we searched through many combinations and found that this hidden layer size gave us the best result.

Conclusion

Model	10-fold Cross-Validated Accuracy
k Nearest Neighbors	0.44
Linear Discriminant Analysis	0.52
Decision Tree	**0.53**
Random Forests	0.51
ADA Boosting	0.47
Multilayer Perceptron (MLP) Classifier	0.52

The model that performed the best was the Decision Tree Classifier with a maximum depth of 3 and the criterion set to the gini index. Other models performed nearly as good, but the decision tree performed the best and is also more interpretable and allows for easier explanation. It improved the baseline performance by 39%, which we believe to be significant.

Ideally, we would want our model to be able to predict with a much higher accuracy what each student's level of alcohol consumption is, since there are many potential benefits and use-cases in doing so. For example, our model could be used by universities and schools to detect which students are most likely to have drinking problems, and whether these are signs of more serious underlying issues. We could also set appropriate thresholds and tweak our model so that it more accurately identifies people with higher levels of alcohol consumption (true positive rate), but this would mean our accuracy on people with lower levels of alcohol consumption would suffer.

It is our belief that an accuracy higher than 53% is needed if were to put such a model into production. This could be achieved by gathering more data about students and engineering more features, which can then be used to augment our present model. However, while it might not be *production grade*, our model could still be used to explore and identify students with high levels of alcohol consumption.

Appendix

Features Considered

- school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- sex - student's sex (binary: "F" - female or "M" - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: "U" - urban or "R" - rural)
- famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian (nominal: "mother", "father" or "other")
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)

- absences - number of school absences (numeric: from 0 to 93)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)