The influence of demographic, social, and school related variableson student's grades - Final Report

Table of Contents:

1. Introduction

This project studied student's grades as potentially influenced to a range of variables. The data set was from Paulo Cortez, University of Minho, GuimarÃ£es, Portugal, http://www3.dsi.uminho.pt/pcortez. Student achievement in secondary education was gathered for two Portuguese schools. The attributes included the student's grades, demographic, social and school related variables. The datasets that were provided included student information for two categories, mathematics and the Portuguese language classes. The data from both datasets were studied for this project. The data was collected by using school reports and surveys. The student's grades are provided for three periods, G1, G2, and G3. These correspond to three school periods with G3 being the final student grade. By analyzing these data sets, correlations between the attributes and the student's grades can be explored with the goal of determining what impacts the student's grades, and therefore what recommendations can be made to improve student's school performance.

Objectives

The work performed for this project seeks to answer the following questions:

1. Which demographic, social, and/or school related variables have a positive effect on a student's grades?
2. Do the demographic, social, and /or school related variables have the same impact on mathematics grades and Portuguese language grades?

Data Set Information:

The data was collected from two Portuguese schools of secondary education. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Data Set Features, Names, and Types

The demographic, social, and school related variables that were investigated include:

Data Set Feature	Name of Feature	Type	Value
School of attendance	school	Binary	**GP** Gabriel Pereira or **MS** Mousinho da Silveira
Sex	sex	Binary	**F** or **M**
Age	age	Numeric	**15 to 22**
Address	address	Binary	**U** urban or **R** rural
Family size	famsize	Binary	**LE3** less or equal to 3 or **GT3** greater than 3
Cohabitation status of parents	Pstatus	Binary	**T** living together or **A** apart
Mother's education level	Medu	Numeric	**0** none, **1** 4th grade, **2** 5th to 9th grades, **3** secondary education, **4** higher education
Father's education level	Fedu	Numeric	**0** none, **1** 4th grade, **2** 5th to 9th grades, **3** secondary education, **4** higher education
Mother's job	Mjob	Categorical	**teacher**, **health** care, civil **services**, **at_home**, or **other**
Father's job	Fjob	Categorical	**teacher**, **health** care, civil **services**, **at_home**, or **other**
Reason for choosing the school	reason	Categorical	Close to **home**, school **reputation**, **course** preference, or **other**
Students guardian (guardian)	guardian	Categorical	**mother**, **father**, **other**
Traveltime to the school	traveltime	Numeric	**1** < 15 min, **2** 15 to 30 min, **3** 30 min to 1 hour, **4** greater than 1 hour
Weekly study time	studytime	Numeric	**1** < 2 hours, **2** 2 to 5 hours, **3** 5 to 10 hours, **4** > 10 hours
Number of past failures	failures	Numeric	n if **1** <= n < **3**, else **4**
Extra educational support	schoolsup	Binary	**yes** or **no**
Family educational support	famsup	Binary	**yes** or **no**
Extra paid classes	paid	Binary	**yes** or **no**
Extra-curricular activities	activities	Binary	**yes** or **no**
Attended nursery school	nursery	Binary	**yes** or **no**
Wants to go on to higher education	higher	Binary	**yes** or **no**
In a romantic relationship	romantic	Binary	**yes** or **no**
Quality of family relationships	famrel	Numeric	From **1** – very bad to **5** - excellent
Free time after school	freetime	Numeric	From **1** – very low to **5** – very high
Going out with friends	goout	Numeric	From **1** – very low to **5** – very high
Workday alcohol consumption	Dalc	Numeric	From **1** – very low to **5** – very high
Weekend alcohol consumption	Walc	Numeric	From **1** – very low to **5** – very high
Current health status	health	Numeric	From **1** – very bad to **5** – very good
Number of school absences	absences	Numeric	**0** to **93**
First period grade (G1)	G1	Numeric	**0** to **20**
Second period grade (G2)	G2	Numeric	**0** to **20**
Final grade (G3)	G3	Numeric	**0** to **20**

Client

There are several potential clients that would benefit from the analyses of the data sets. These include students, parents of the students, the student's teachers, school administrators, and community leaders.

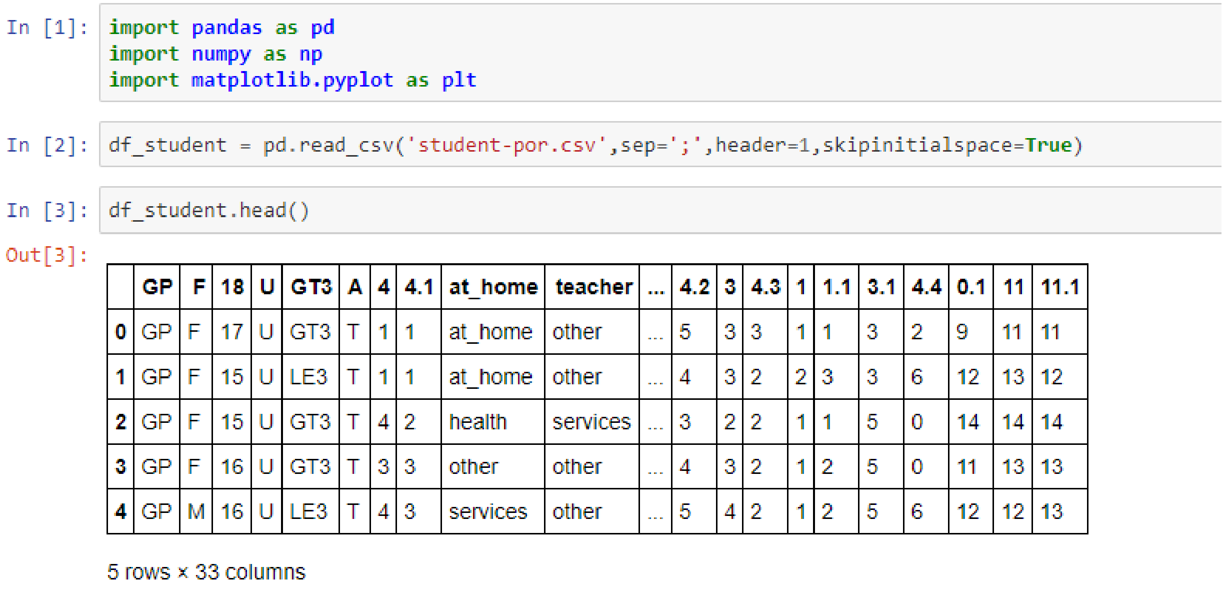
2. Data Acquisition/Cleaning

In this section I will explore the datasets and use preprocessing and data wrangling techniques to prepare the data. This will include the following steps:

1. Loading the data and extracting general info and structure
2. Exploring data types
3. Identifying & dealing with missing values
4. Preprocessing techniques

Data Information and Structure

The data sets were fairly clean when I got them. There was no missing or NAN values. In order to utilize the data sets, I added column headers, and divided the data in to male and female groups. For the Portuguese language data there were 382 male students and 266 female students. For the math data there were 187 male students and 207 female students



```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
```

```
In [2]: df_student = pd.read_csv('student-por.csv',sep=';',header=1,skipinitialspace=True)
```

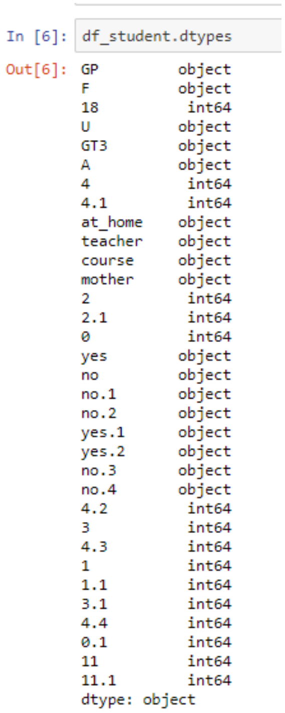
```
In [3]: df_student.head()
```

Out[3]:

	GP	F	18	U	GT3	A	4	4.1	at_home	teacher	...	4.2	3	4.3	1	1.1	3.1	4.4	0.1	11	11.1
0	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
1	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
2	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
3	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
4	GP	M	16	U	LE3	T	4	3	services	other	...	5	4	2	1	2	5	6	12	12	13

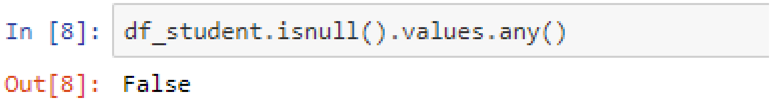
5 rows × 33 columns

General data types:



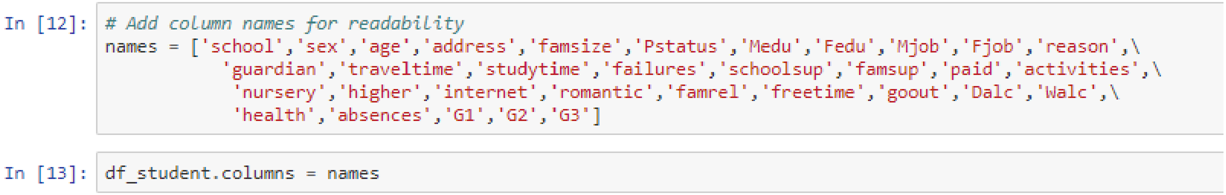
```
In [6]: df_student.dtypes

Out[6]: GP          object
        F           object
        18           int64
        U           object
        GT3         object
        A           object
        4            int64
        4.1          int64
        at_home     object
        teacher     object
        course      object
        mother      object
        2            int64
        2.1          int64
        0            int64
        yes         object
        no          object
        no.1        object
        no.2        object
        yes.1       object
        yes.2       object
        no.3        object
        no.4        object
        4.2          int64
        3            int64
        4.3          int64
        1            int64
        1.1          int64
        3.1          int64
        4.4          int64
        0.1          int64
        11           int64
        11.1         int64
        dtype: object
```

Missing values? This was true for both the language and math data.



```
In [8]: df_student.isnull().values.any()

Out[8]: False
```

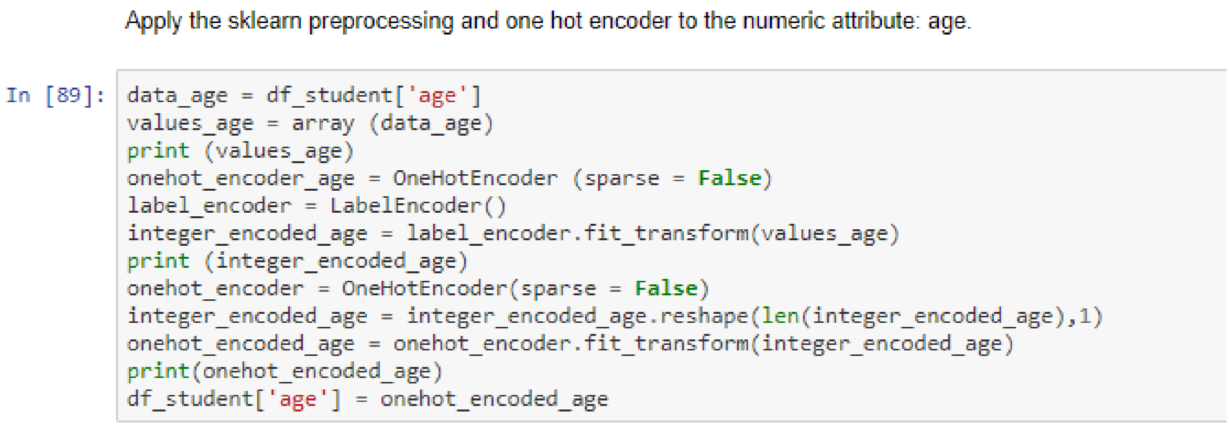
Add column names for readability. Both data files (language and math) have the same attributes.



```
In [12]: # Add column names for readability
         names = ['school','sex','age','address','famsize','Pstatus','Medu','Fedu','Mjob','Fjob','reason',\
                   'guardian','traveltime','studytime','failures','schoolsup','famsup','paid','activities',\
                    'nursery','higher','internet','romantic','famrel','freetime','goout','Dalc','Walc',\
                    'health','absences','G1','G2','G3']

In [13]: df_student.columns = names
```

Preprocessing Techniques

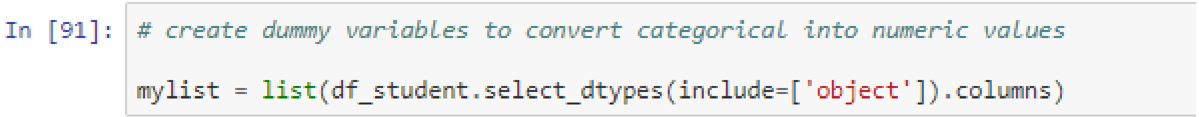
Two preprocessing techniques were used on the data sets. sklearn preprocessing and one hot encoder were used to change the categorical attributes: age, Mjob, Fjob, reason, and guardian.



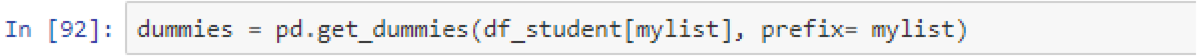
Apply the sklearn preprocessing and one hot encoder to the numeric attribute: age.

```
In [89]: data_age = df_student['age']
         values_age = array (data_age)
         print (values_age)
         onehot_encoder_age = OneHotEncoder (sparse = False)
         label_encoder = LabelEncoder()
         integer_encoded_age = label_encoder.fit_transform(values_age)
         print (integer_encoded_age)
         onehot_encoder = OneHotEncoder(sparse = False)
         integer_encoded_age = integer_encoded_age.reshape(len(integer_encoded_age),1)
         onehot_encoded_age = onehot_encoder.fit_transform(integer_encoded_age)
         print(onehot_encoded_age)
         df_student['age'] = onehot_encoded_age
```

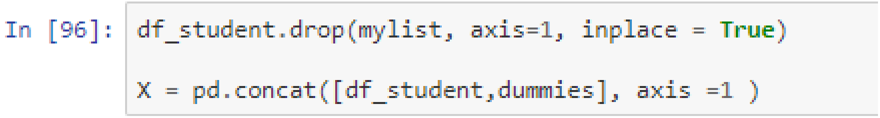
Pandas pd.get_dummies was used to convert the Boolean categorical data into ones and zeros. These attributes included school, sex, address, famsize, Pstatus, schoolsup, famsup, paid, activities, nursery, higher, internet, and romantic. This resulted in the addition of 13 columns to the data set.



```
In [91]: # create dummy variables to convert categorical into numeric values

         mylist = list(df_student.select_dtypes(include=['object']).columns)
```



```
In [92]: dummies = pd.get_dummies(df_student[mylist], prefix= mylist)
```

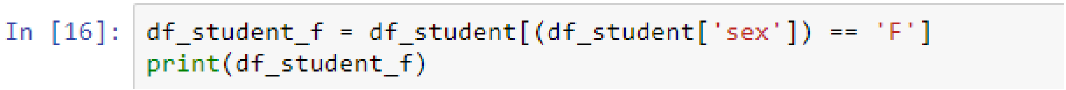


```
In [96]: df_student.drop(mylist, axis=1, inplace = True)

         X = pd.concat([df_student,dummies], axis =1 )
```

3. Data Exploration

Grades for Male vs. Female students

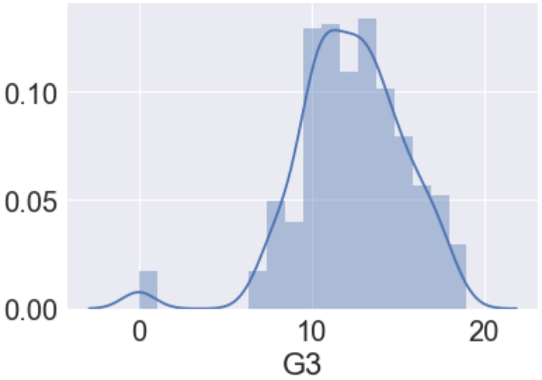
The first comparison that I opted to explore was the difference between the grades of male and female students.



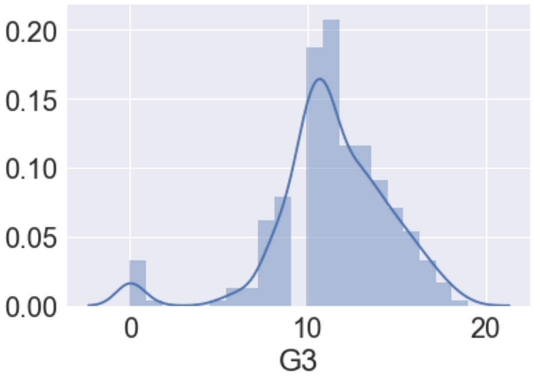
```
In [16]: df_student_f = df_student[(df_student['sex']) == 'F']
         print(df_student_f)
```

The data was separated in to two data frames, df_student_f and df_student_m. Presented below are grade distributions for male and female students for Portuguese language grades and math grades.

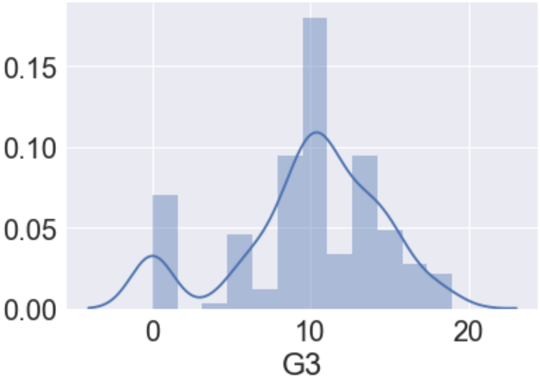
Portuguese Language Grade Distribution for female students



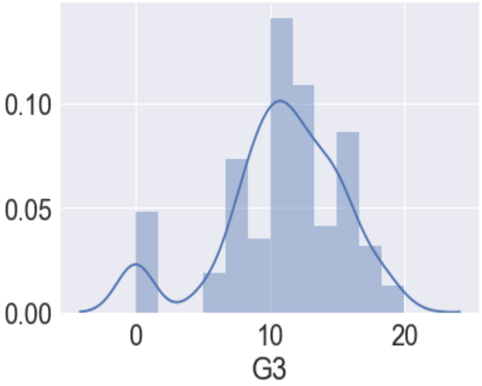
Portuguese Language Grade Distribution for male students



Math Grade Distribution for female students



Math Grade Distribution for male students



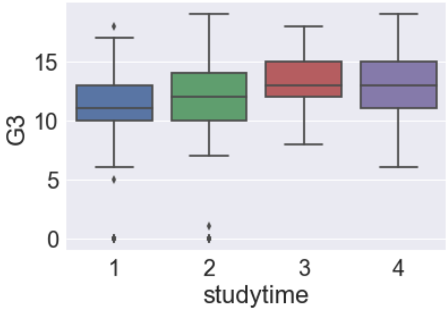
The distributions show a similar trend. The mean for the female grades in Portuguese is 12.3. The mean for the male grades in Portuguese is 11.4. For the math grades the mean for the female and male grades was 9.9 and was 10.9. A t-test was performed to determine if the means are indeed different. For the Portuguese language grades a t-value of -3.31 and a p-value of 0.00095 were calculated. Conducting the same test for the math grades yields a t-value of 2.01 and a p-value of 0.044. Although there is a higher p-value for the math grades, it is still below the 0.05 threshhold. These results indicate that the groups are similar and the data didn't occur by chance. So, there is statistically no difference between male and female grades.

Analysis of Grades vs. Demographic, Social and School Related Attributes

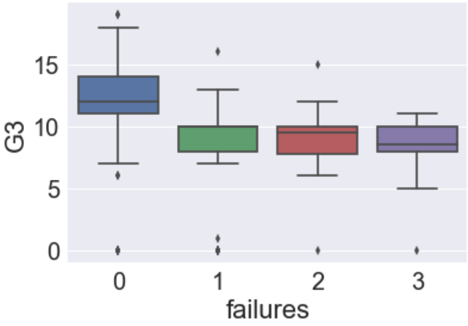
By understanding which demographic, social and school related attributes affect grades, we can possibly improve grade outcomes. Box and whisker plots were created for each demographic, social, and school related attribute. The plots for amount of study time (studytime), number of failures (failures), interest in higher education (higher), quality of family relationships (famrel), and daily consumption of alcohol (dalc) visually showed more variation than the other variables. The box plots for these five attributes are plotted below for the language and math grades.

Portuguese Language Grades

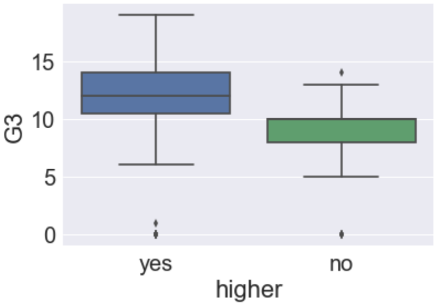
Study Time



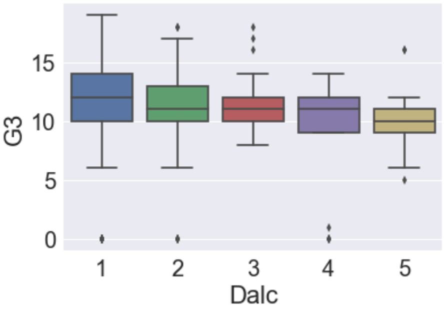
Failures



Interest in Higher Education

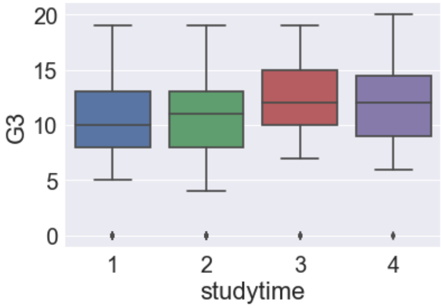


Daily Consumption of Alcohol

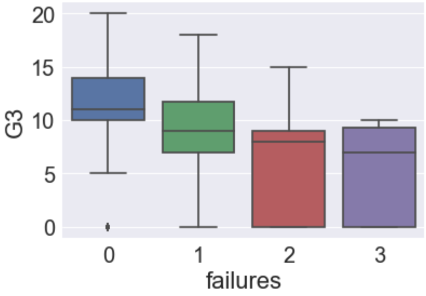


Math Grades

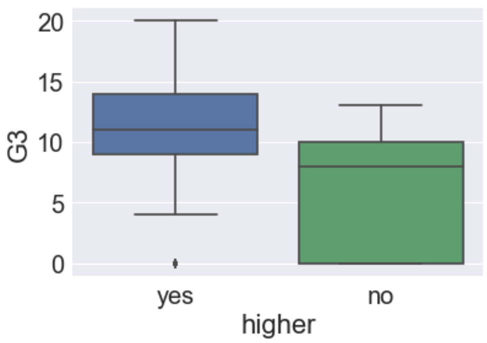
Study Times



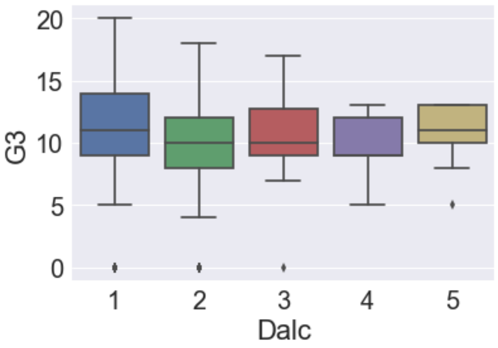
Failures



Interest in Higher Education



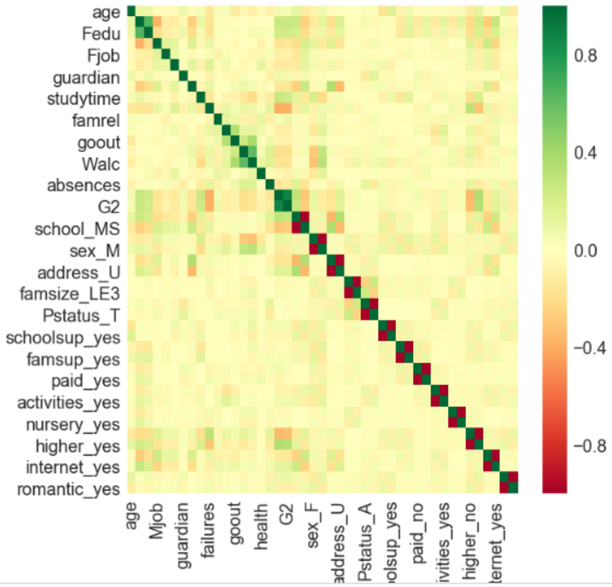
Daily Consumption of Alcohol



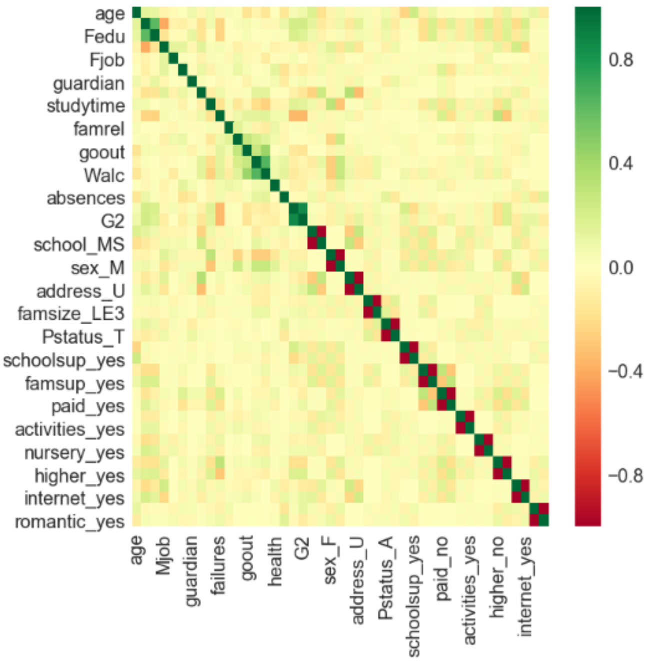
A Pearson correlation was used to test if there is a statistically significant linear relationship between the above attributes and student's grades. The Pearson r-values for Portuguese language study time, failures, interest in higher education, and daily consumption of alcohol were 0.25, -0.39, 0.33, and -0.2, respectively. These correlation values indicate weak positive correlations between study time and interest in higher education with grades, and weak negative correlations between failures and daily consumption of alcohol with grades. The Pearson r-values for math study time, failures, interest in higher education, and daily consumption of alcohol were 0.97, -0.36, -0.05, and -.01, respectively. I think that the weak correlations are a result of possible interdependence between some of the attributes. The Pearson r-value assumes that the variables are independent from one another.

Additional exploration between student's grades and the attributes was conducted by using a correlation heatmap. The heatmaps are displayed below. Not surprisingly, based on the above box plots, the correlation between attributes is fairly weak for both data sets.

Heatmap Correlation for Portuguese Language

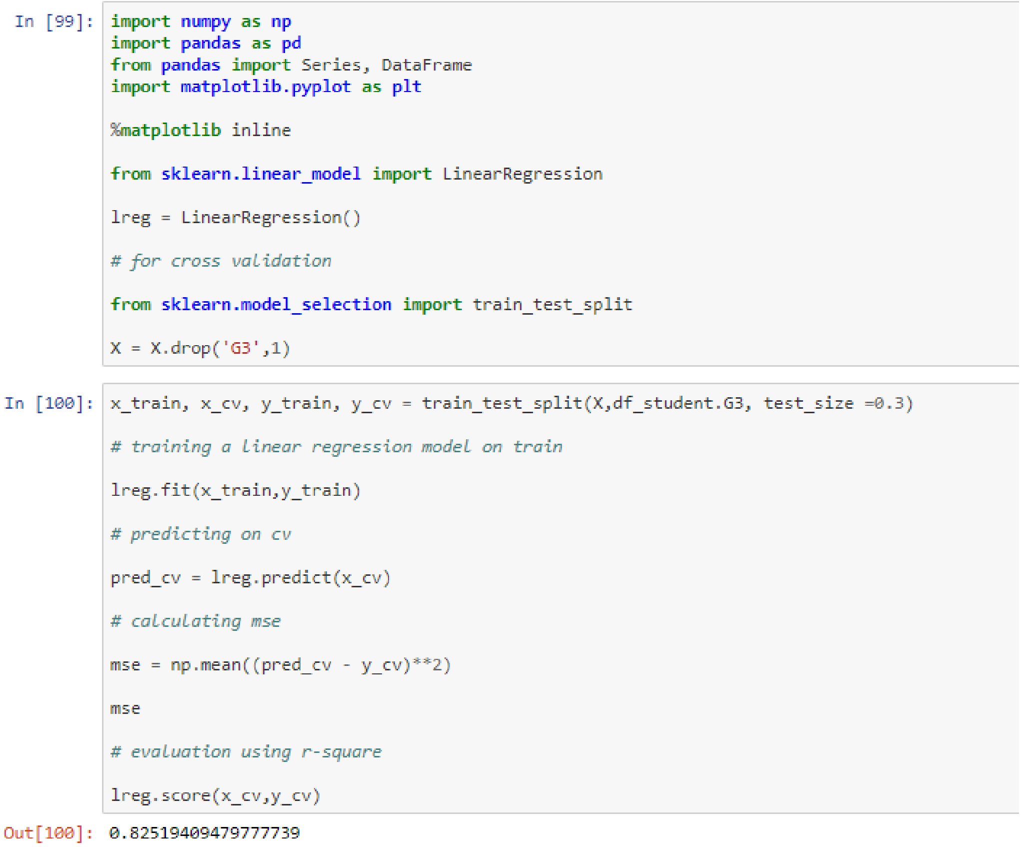


Heatmap Correlation for Math



4. Initial Findings

Following the data exploration, a machine learning algorithm was utilized to create a predictive model. The train_test_split algorithm used a portion of the attribute data, in this case 30%, that was available to perform a linear regression on the selected portion of data, and tested the fit on the remaining data. The following code was used for the language and math data:

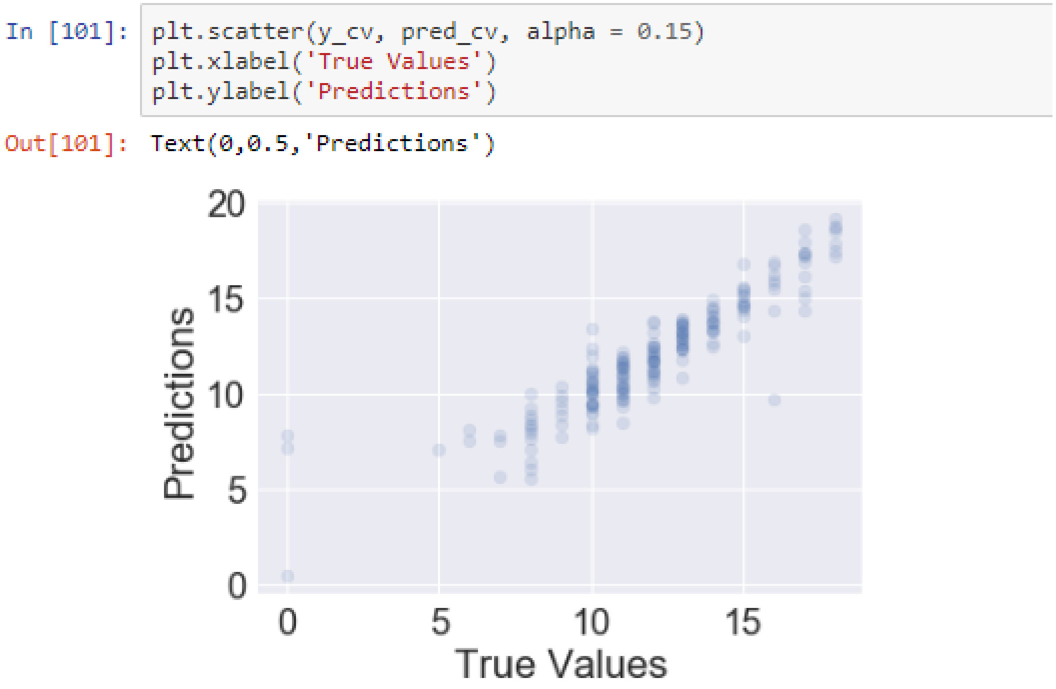


```
In [99]:  import numpy as np
          import pandas as pd
          from pandas import Series, DataFrame
          import matplotlib.pyplot as plt

          %matplotlib inline

          from sklearn.linear_model import LinearRegression

          lreg = LinearRegression()

          # for cross validation

          from sklearn.model_selection import train_test_split

          X = X.drop('G3',1)
```

```
In [100]: x_train, x_cv, y_train, y_cv = train_test_split(X,df_student.G3, test_size =0.3)

          # training a linear regression model on train

          lreg.fit(x_train,y_train)

          # predicting on cv

          pred_cv = lreg.predict(x_cv)

          # calculating mse

          mse = np.mean((pred_cv - y_cv)**2)

          mse

          # evaluation using r-square

          lreg.score(x_cv,y_cv)

Out[100]: 0.82519409479777739
```

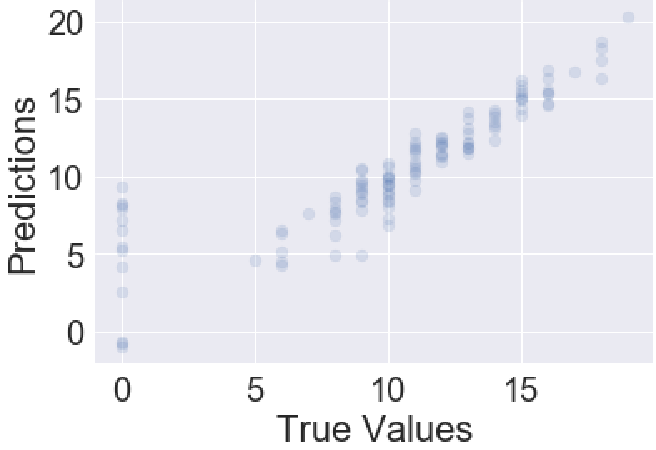
R-Square**:** determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). The value of R-square is always between 0 and 1, where 0 means that the model does not explain the variability in the target variable (Y) and 1 meaning it explains full variability in the target variable. The score of the linear regression model was computed to be 0.83 and 0.77 for the language and math data, respectively. A scatter plot shows the correlation between the true values and the predicted values.

Portuguese Language Data



```
In [101]: plt.scatter(y_cv, pred_cv, alpha = 0.15)
          plt.xlabel('True Values')
          plt.ylabel('Predictions')

Out[101]: Text(0,0.5,'Predictions')
```

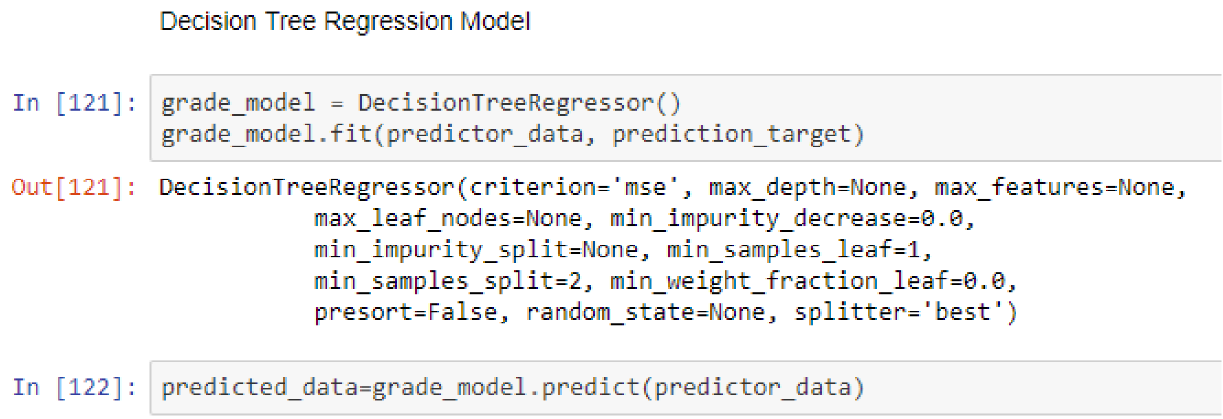
Math Data



5. Additional Model Testing

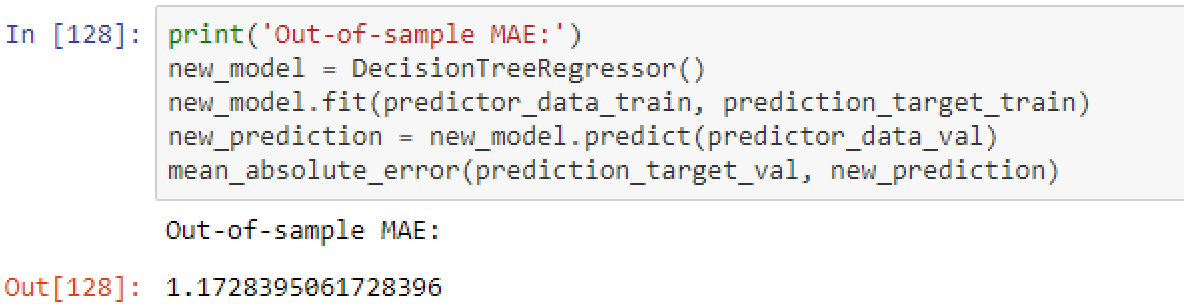
The models were tested using a Decision Tree Regressor.

A decision tree splits the input features into regions and assigns a prediction value to each region. The selection of the regions and the predicted value within a region are chosen in order to produce the prediction which best fits the data. For best fit we mean that it minimizes the distance of the observations from the prediction. The following code was utilized for the decision tree regression model for both language and math data.



Decision Tree Regression Model

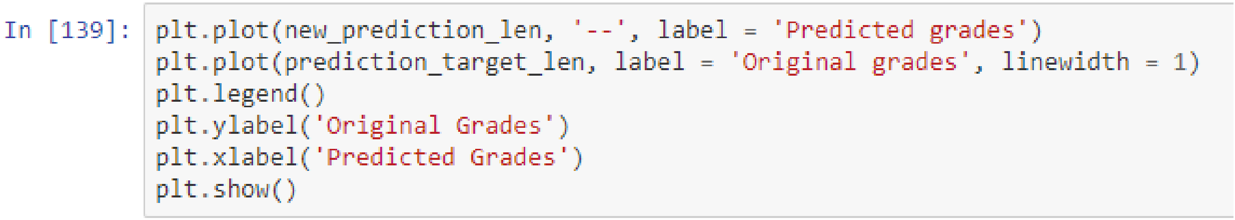
```
In [121]: grade_model = DecisionTreeRegressor()
          grade_model.fit(predictor_data, prediction_target)

Out[121]: DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=None,
                     max_leaf_nodes=None, min_impurity_decrease=0.0,
                     min_impurity_split=None, min_samples_leaf=1,
                     min_samples_split=2, min_weight_fraction_leaf=0.0,
                     presort=False, random_state=None, splitter='best')

In [122]: predicted_data=grade_model.predict(predictor_data)
```



```
In [128]: print('Out-of-sample MAE:')
          new_model = DecisionTreeRegressor()
          new_model.fit(predictor_data_train, prediction_target_train)
          new_prediction = new_model.predict(predictor_data_val)
          mean_absolute_error(prediction_target_val, new_prediction)

          Out-of-sample MAE:

Out[128]: 1.1728395061728396
```

The mean absolute error (MAE) for the language fit was 1.17, and was 1.16 for the math data. MAE of 1 point is a difference of 1 point of between the actual and predicted values. The MAE for both data sets are equivalent and are low.

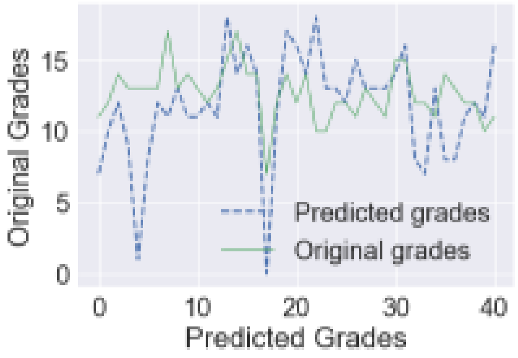
A plot of the predicted grades and the actual grades was created from the following code:



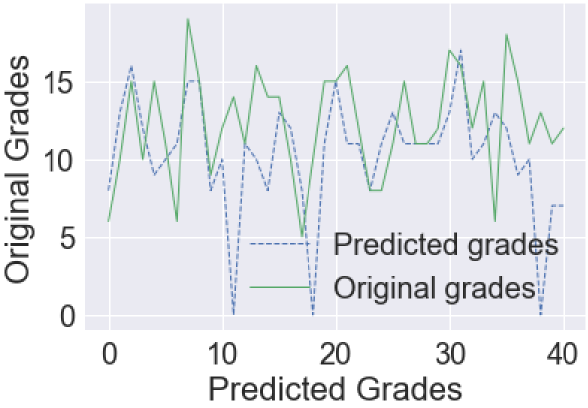
```
In [139]: plt.plot(new_prediction_len, '--', label = 'Predicted grades')
          plt.plot(prediction_target_len, label = 'Original grades', linewidth = 1)
          plt.legend()
          plt.ylabel('Original Grades')
          plt.xlabel('Predicted Grades')
          plt.show()
```

The plots for the language and math data are shown below.

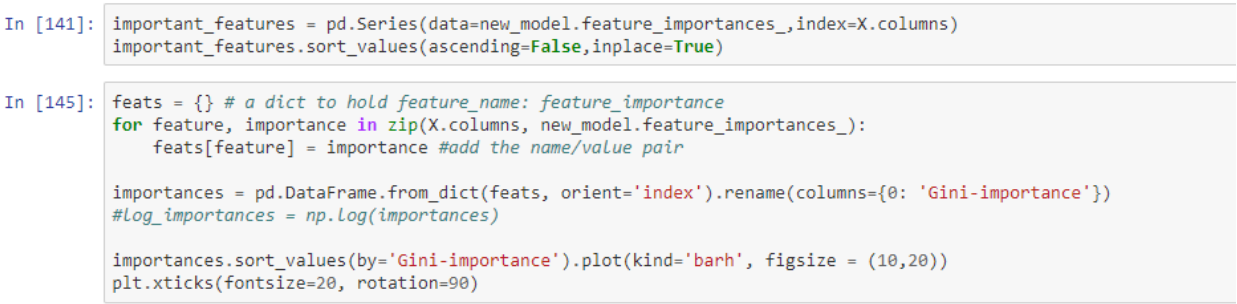
Portuguese Language Data



Math Data



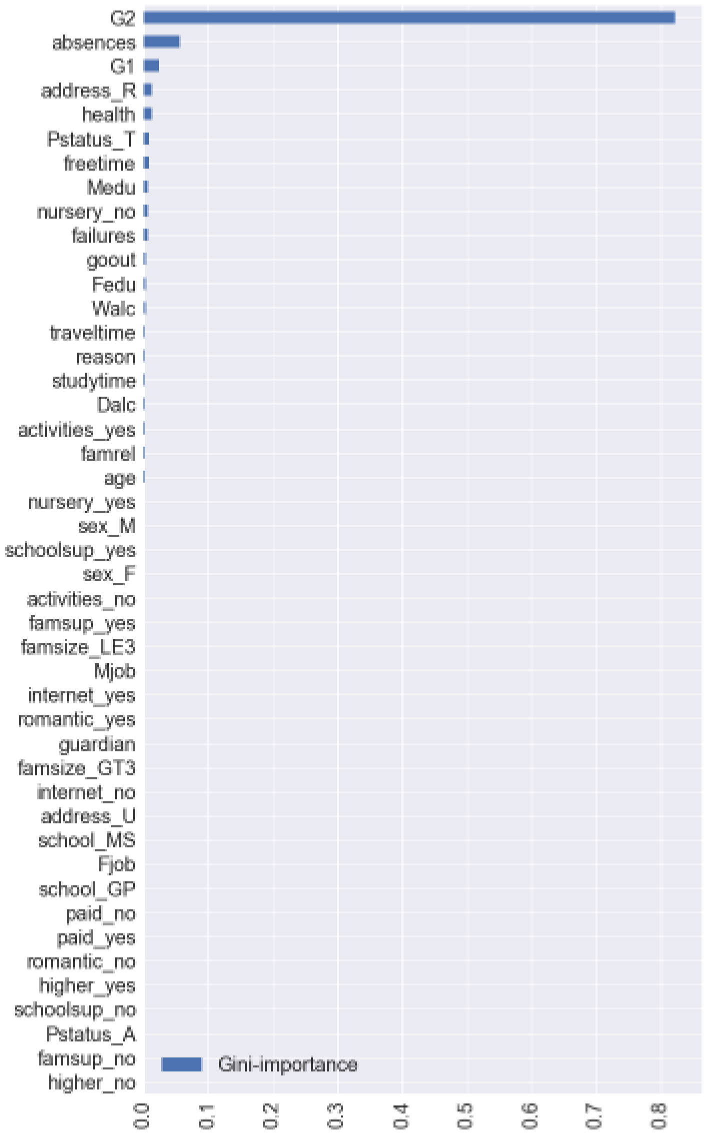
The importance of each feature was produced with the following code:



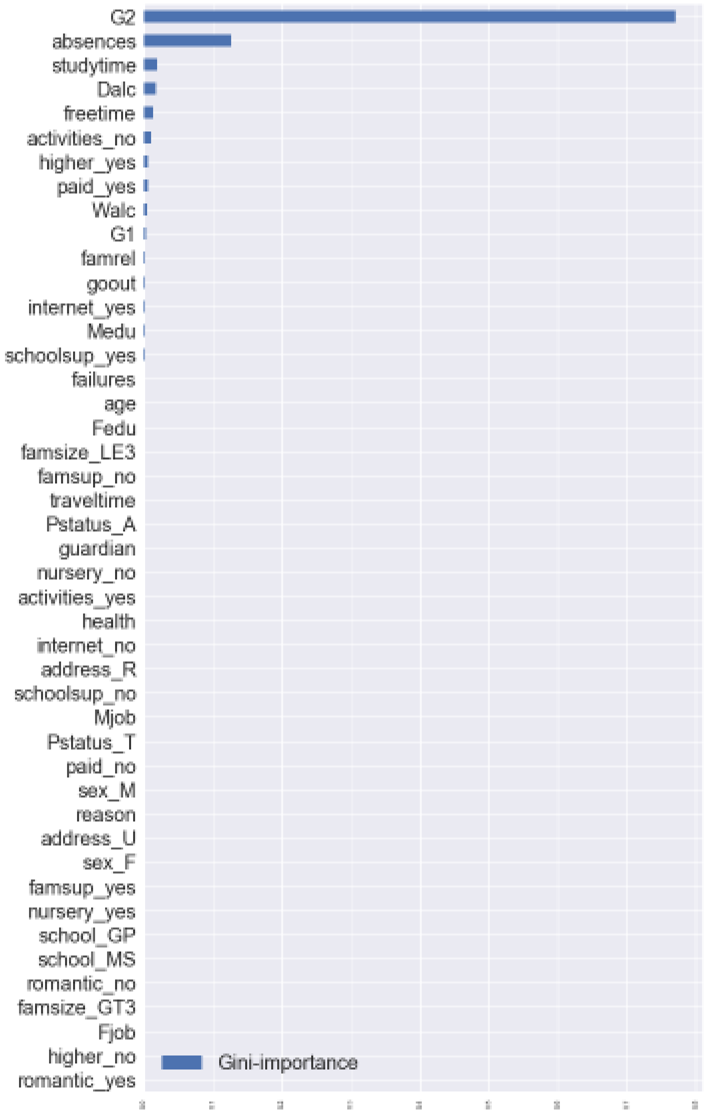
```
In [141]: important_features = pd.Series(data=new_model.feature_importances_,index=X.columns)
          important_features.sort_values(ascending=False,inplace=True)

In [145]: feats = {} # a dict to hold feature_name: feature_importance
          for feature, importance in zip(X.columns, new_model.feature_importances_):
              feats[feature] = importance #add the name/value pair

          importances = pd.DataFrame.from_dict(feats, orient='index').rename(columns={0: 'Gini-importance'})
          #log_importances = np.log(importances)

          importances.sort_values(by='Gini-importance').plot(kind='barh', figsize = (10,20))
          plt.xticks(fontsize=20, rotation=90)
```

The resulting plot of importance and features is shown below for the language data and the math data.

Feature Importance for Portuguese Language Data

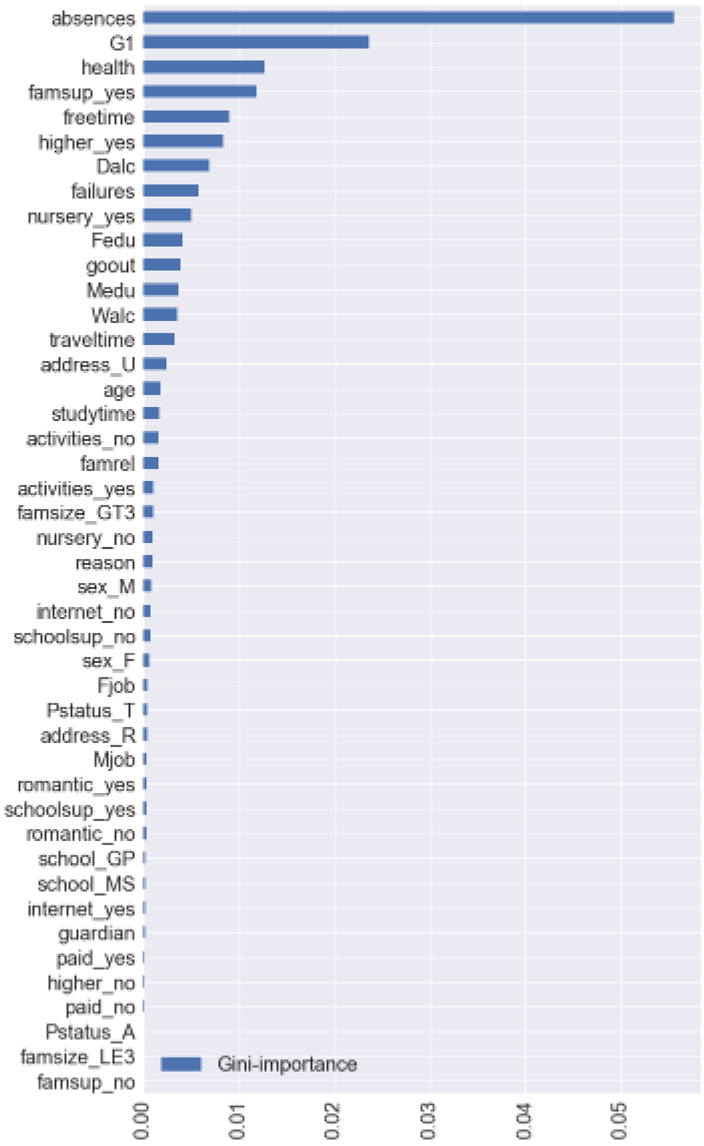


Feature Importance for Math Data

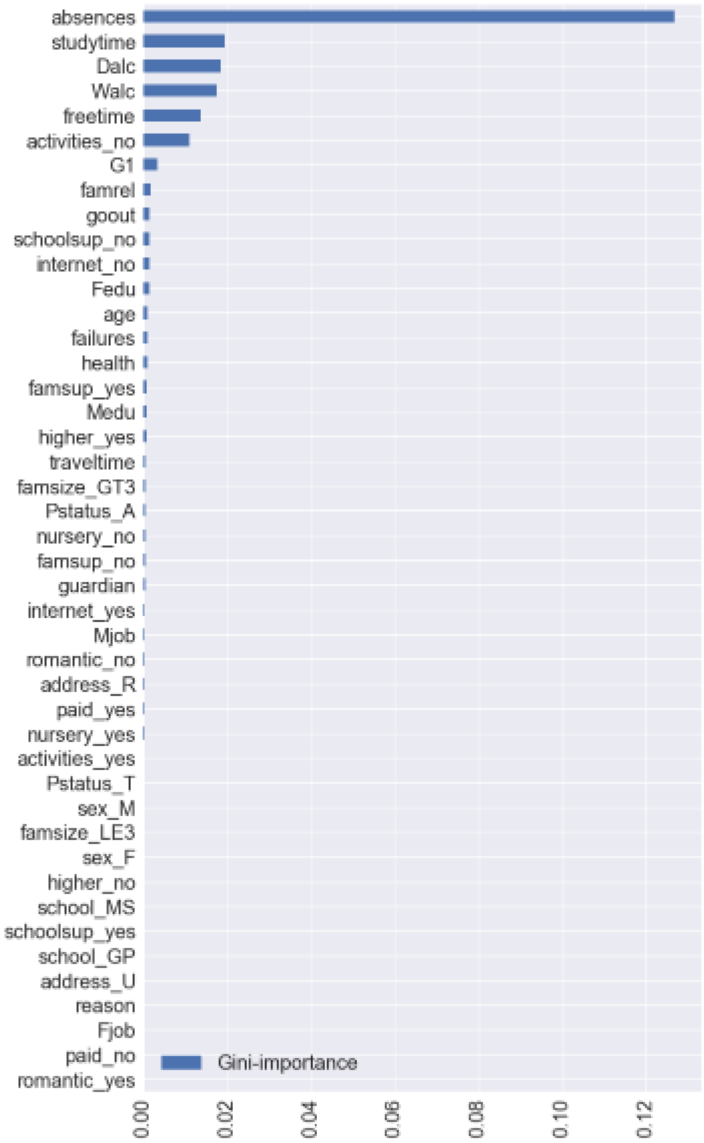


Inspection of the feature importance plots for both the language data and the math data show that the importance of feature G2 far outweighs the importance of the other features. For this reason, the importance data were re-plotted excluding feature G2. These plots are presented below.

Feature Importance for Portuguese Language Data Excluding G2



Feature Importance for Math Data Excluding G2



6. Conclusions and Recommendations

The regression model utilizing the train_test_split algorithm resulted in good correlations between the attributes and the student's final grade with R-squared being 0.84 and 0.77 for the language and math data, respectively. The decision tree regressor also produced good results with the MAE being ~1.17 for both models. The decision tree regressor algorithm allows us to see which features influence the student's final grade. The top importance features are presented in the following table.

Top Importance of Features of Portuguese language and math data

Portuguese Language	Math
G2	G2
absences	absences
G1	study time
health	Dalc
famsup_yes	

The above table lists G2 as the most important feature in predicting student's final grades, G3. After feature G2 the next three top important features, also excluding G1, are also listed in the above table. The fact that G2 is the biggest predictor of final grades, G3 allows us to conclude that past performance is evidence of future performance.

Features Impacting Portuguese Language Grades

For the Portuguese language grades, absences, health, and having family educational support (famsup_yes) were the top features that impacted the student's grades, excluding G1 and G2. Students who don't skip school will have better grades. School absences could impact grades if work performed in school was not completed, or if there was a component to the grade result that required in class speaking time. Absences can also be a result of problems with reliable transportation or lack of familial interest in the student's grades, however these features were considered separately in the data analysis.

Health was the second important feature from the importance analysis. The health feature data was collected as a numeric value from 1 (very bad) to 5 (very good). It is understandable that students who are well slept and well fed will be more alert and successful at learning new things.

Family educational support (famsup_yes) was the next important feature impacting grades. Family educational support could encompass a variety of actions such as help at home with take-home school work, interest in student's assignments, dedication to transporting student to school, and/or supplying the student with tools necessary to complete school work.

Features Impacting Math Grades

The top three features impacting math grades, excluding G2, are absences, study time (studytime), and daily consumption of alcohol (Dalc). With absences having the most impact on school performance for both Portuguese language students and math students demonstrates how valuable class time is for obtaining better grades.

The second feature for the prediction of math grades was study time. This data was collected as a weekly value with 1 corresponding to less than 2 hours, 2 corresponding to 1 to 5 hours, 3 corresponding to 5 to 10 hours, and 4 corresponding to greater than 10 hours. The more time that is spent pursuing a problem solving (math) subject will have a positive effect on student's grades.

Daily consumption of alcohol was the next important predictor. The data was collected as a value between 1 and 5 with the stipulation that only weekday drinking was considered. A value of 1 was 'very low' and a value of 5 was 'very high'. Drinking on the weekdays implies less focus and/or time spent studying. Daily drinking can also lead to disrupted sleeping and eating patterns.

Conclusions

From the analyses of these data sets we can conclude:

1. Class attendance has the most influence on student's grades, whether the student is studying language or math.
2. For an analytical class like math, increasing study time predicts that the student will have a higher math grade.
3. Healthy students are predicted to earn higher grades.
4. Families that provide educational support will yield students with higher grades for language learners.
5. Weekday consumption of alcohol can influence student's grades negatively.

Recommendations

1. Work to prevent obstacles that result in student absence.
    a. Schools can work to provide students with transportation (bussing).
    b. Parents can organize carpools so that a student isn't reliant on only one way to get to school.
    c. Parents can schedule family vacations when the student is on break from school.
    d. Teachers and parents can be more involved with their students to help understand if there are problems at school leading to absences.

2. Increase opportunities for study time.
    a. Parents can organize a place to study where the student is comfortable and has the necessary tools to study.
    b. Parents can offer rewards for increasing study time.
    c. Parents can show interest in what their student is studying and create dialog about the task at hand.
    d. Teachers can communicate with parents if they feel the student isn't studying enough.

3. Make sure that students are healthy and stay healthy.
    a. Schools and school systems can have exercise and wellness programs for students as part of their regular curriculum.
    b. Schools/communities can providing free and reduced breakfast and lunch for students in need.
    c. Parents can provide good nutrition, wellness checks, and appropriate vaccinations.
    d. Schools can teach personal wellness such as nutrition, benefits of sleep, and benefits of exercise.

4. Increase educational support from families.
    a. Teachers and parents can increase communication about what kind of support would benefit the student.
    b. Parents can volunteer time to help in class.
    c. Communities/schools can offer informational conferences to parents to relay the importance of family involvement and educational support.
    d. Parents can identify school as a student's priority.
    e. Parents can stay involved with the student's course work.

5. Do not permit weekday consumption of alcohol.
    a. Parents can stay informed about what their student is doing with their free time.
    b. Parents can be mindful not to provide opportunities for the student to have access alcohol.
    c. Teachers can inform parents if they suspect a student is under the influence, if a student becomes sick, or is getting behind with schoolwork.
    d. Parents can set an example by not consuming alcohol daily.
    e. Schools can let parents know if there are incidents with alcohol at school events or on the school campus.

7. Next Steps

A next step would involve using the trained model for the Portuguese language data with the data from the math file, and evaluate the performance of the model with the other dataset.

8. Resources

- Dataset - http://www3.dsi.uminho.pt/pcortez
- Project Proposal - github link
- Code (IPython Notebook) - https://github.com/hhtdata/SB2018/blob/master/Capstone1_student-Por%20(10).ipynb