Ryan Long
DSC 680-T301
3.1 Project 1: Draft - Milestone 2

Business Problem

The objective of this project is to predict running pace of a youth cross country participant using a multiple regression model given several criteria obtained from seasonal data of a female middle school cross-country season. This information will aid runners, coaches, and parents to focus on healthy and holistic training outcomes.

Background/History

Formalized running teams typically start in the US during middle school or junior high phase of education with participants expanding their understanding of training requirements during this time. Caution during this stage should be exercised to avoid injury, mental fatigue, and potential burnout. Being able to predict pace for maturing runners helps establish realistic expectations and manage the aforementioned risks.

Data Explanation (Data Prep/Data Dictionary/etc.)

Race result data was obtained primarily from an online source, focusing on the 2021 Parochial Athletic League (PAL) cross country results of St. Wenceslaus school in Omaha, NE. The data set was augmented with weather data to approximate the air temperature at the time of the race. Please see the references section for details on data sources. Additionally, a categorical description of the course was subjectively assigned by the author for each meet. The following table outlines the variables of the dataset.

Variable	Description
MEET_LOC	Indicates the location or school host of the meet.
DATE	Date of the race
PLACE	Overall race placement of the participant
GRADE	Grade level of race participant
NAMETOKEN	Tokenized name of the participant
TIME	Final race time for the participant
DIST_M	Distance in meters
DIST_MI	Distance in miles
PACE_CALC	Pace calculated in minutes per mile
SCHOOL	School represented by participant
TEMP_F	Degrees Fahrenheit at the time of the race in degrees
COURSE_TYPE	Description of the course (Flat, Rolling, Hilly)

See Appendix 1 – Image 1 for more details on the dataset.

Four additional features were created from the existing dataset to process the information through the model. TIME and PACE_CALC were transformed into new features from 'timedelta' types to floats (i.e., 07:15 to 7.25). Additionally, DATE and COURSE_TYPE were transformed into integers via label encoding resulting in DATE_LABEL and COURSE_TYPE_LABEL.

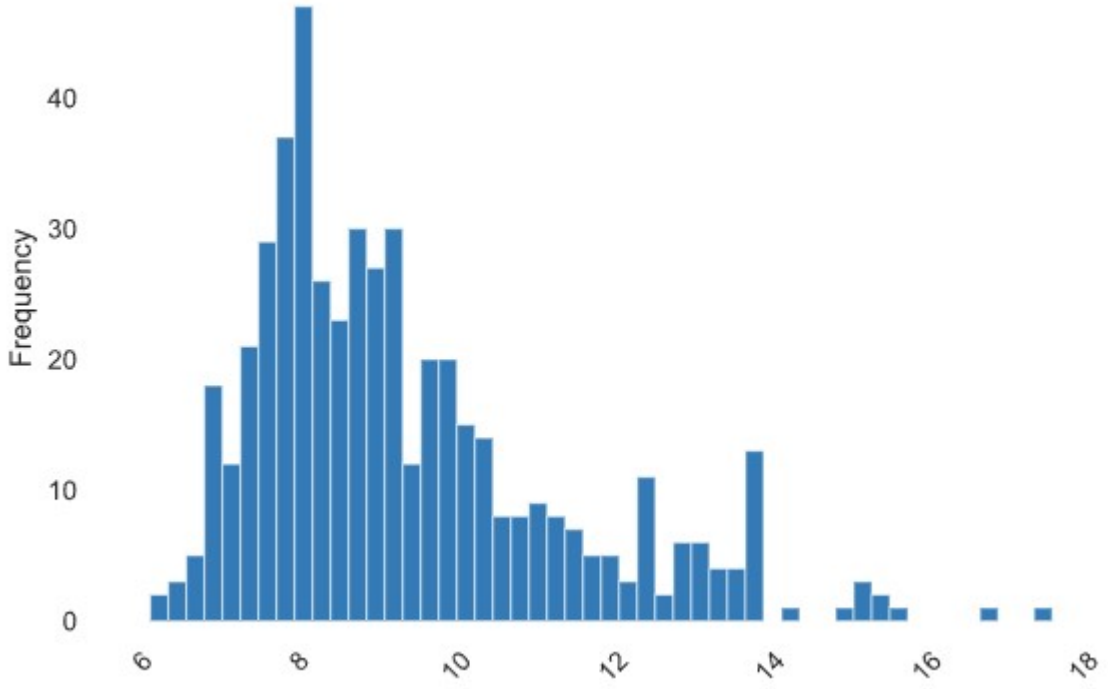
For Exploratory Data Analysis (EDA) the Pandas Profiling package provides an extensive overview of the dataset. The following features played a prominent role in the modeling process.

'GRADE' Frequency

Value	Count	Frequency (%)
7	236	47.2%
8	180	36.0%
6	80	16.0%
5	4	0.8%

As can be seen from the grade frequency illustration, the dataset is heavily weighted towards seventh grade, followed by eighth grade. Sixth and fifth grade only represent 16.8% of the dataset.

'PACE_CALC' Histogram



Pace distribution is heavily right skewed. This is to be expected given the novice level of runners comprising the dataset.

Methods

For language and IDE, Python within Jupyter Notebook was used. Python libraries leveraged included: Pandas, NumPy, Sklearn, Matplotlib, Pandas Profiling, and statsmodel.

The dataset was run through a linear regression evaluation with pace as the dependent while distance, place, date, temperature, grade, time, and course type as the independent variables. The results showed a high R-squared of 0.996. Grade returned the strongest P value of 0.674. See Appendix 1 – Image 2.

The Spearman's correlation heatmap indicates a strong positive correlation for the target PACE_CALC with PLACE, TIME. Lesser but still relevant correlations exist with DATE_LABEL and COURSE_TYPE_LABEL. See Appendix 1 – Image 3.

Next, a features analysis was performed using a random forest classifier. The 'PLACE' feature was used as a proxy for time to perform this analysis. Intuitively, TIME and PACE_CALC indicated the strongest importance, followed by GRADE, TEMP_F, DATE_LABEL, and COURSE_TYPE_LABEL. From this analysis, DISTANCE was dropped from further use in the process. PLACE and TIME were also dropped as the objective is to determine a predicted pace. Both of those features can be inferred from the predicted pace value. See Appendix 1 – Image 4

A multiple linear regression model was created using the DATE_LABEL, TEMP_F, GRADE, and COURSE_TYPE_LABEL. Entering integer inputs for date, temp, grade, and course provided a predicted pace for the runner.

Analysis

The model output follows intuition and coefficient relationships in which inputs for later season, lower temperature, and flat course types predict a faster pace. The grade independent does not follow suit, with a higher-grade (using 8^{th} rather than 6^{th}) integer increasing (slowing) the predicted pace. See Appendix 1 – Code Snippets 2 for additional examples of model outputs. At this developmental stage it should be assumed an older child should have a decrease (faster) in predicted pace and performance due to natural progression and experience. See the 'Model Outputs' section within Appendix 1.

Revisiting the data shows the mean grade for top 10 placing in the dataset was 6.93, given the range of 5^{th} through 8^{th}. Additionally, better performance in terms of lower mean pace is found in the 6^{th} grade

subset (8.26) compared to 7th (9.69) and 8th (9.34). See Appendix 1 – Code Snippets 1 An alternative solution is the removal of the GRADE independent. However, this would decrease the applicability of the model. In its current state, the dataset has disproportionate higher performers in a minority subset.

Conclusion

This model's ability to provide realistic predictions is hampered by the shortcomings of the dataset selected. A diverse distribution amongst the grade level results would likely improve the output to provide a better perspective on the predicted pace improvements over the course of multiple cross-country seasons. With improved underlying data, the predictions will better aid coaches (and parents) focus on healthy and holistic training for multi-season development.

Outside of the issues with the GRADE feature, the model is still suitable for predicting performance improvements over the course of a season. Interpretations could be adjusted as percentage of improvement from one period to the next.

Assumptions / Limitations / Challenges

The largest challenge is the dataset itself. What has been collected is relatively small number of records and caused issues when modeling due to its composition. In its current state, the dataset is rather limiting as it is a very small selection of female middle school runners with stronger performance from a younger and underrepresented subset. Additionally, the model assumes users will input realistic or real-world independent variables. Unlikely scenarios such as negative temperatures, expansive season durations, and/or grade levels incongruent with what was used to create the model will negatively the output with unrealistic predictions.

Future Uses/Additional Applications

Future uses include both further expansion of the dataset to improve the grade-based results. Additionally, this modeling approach given the features could be expanded to other datasets as well.

Recommendations

Recommends include enhancement of the dataset and consideration of other external or environmental factors such as dewpoint and humidity. Interval splits for the race could provide further insight into whether what strategy is optimal given the environmental and course layout.

Implementation Plan

Next steps for implementation would be to expand the dataset to increase the practicality of the output given the limitations specific to grade-level influences. Additionally, controls would need to be

implemented to ensure sensible inputs can be used for the independent variables, such as no negative air temperatures and realistic intervals for dates throughout the season.

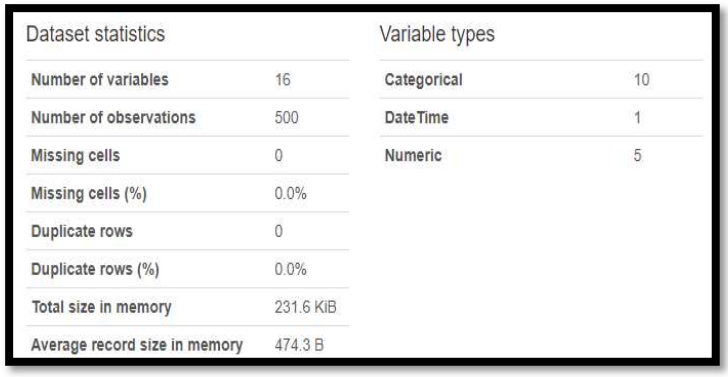
Ethical Considerations

The model is predicting a decline in performance as a runner moves from 6th to 8th grade, there is less opportunities for setting the bar to high and driving runners to burnout or injury. Caution should still be given if the dataset improves the output and what impacts that could have on the mental health of aspiring runners as they set goals for performance.

While the data is publicly available on the internet, names of the individual runners were tokenized/coded as they are minors. The author did not want to be responsible for any unintended distribution of information which wasn't agreed upon to begin with.

Appendix 1 – Images

Image 1



Dataset statistics		Variable types	
Number of variables	16	Categorical	10
Number of observations	500	DateTime	1
Missing cells	0	Numeric	5
Missing cells (%)	0.0%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	231.6 KiB		
Average record size in memory	474.3 B		

Image 2

OLS Regression Results

Dep. Variable:	PACE_CALC2	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.996
Method:	Least Squares	F-statistic:	1.869e+04
Date:	Wed, 22 Jun 2022	Prob (F-statistic):	0.00
Time:	18:40:24	Log-Likelihood:	344.78
No. Observations:	500	AIC:	-673.6
Df Residuals:	492	BIC:	-639.8
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.1101	0.178	67.915	0.000	11.760	12.460
DIST_M	-0.0040	3.46e-05	-114.327	0.000	-0.004	-0.004
PLACE	-0.0010	0.000	-4.299	0.000	-0.001	-0.001
DATE_LABEL	-0.0108	0.007	-1.651	0.099	-0.024	0.002
TEMP_F	-0.0077	0.001	-6.269	0.000	-0.010	-0.005
GRADE	-0.0033	0.008	-0.421	0.674	-0.019	0.012
TIME2	0.5459	0.003	181.963	0.000	0.540	0.552
COURSE_TYPE_LABEL	0.1193	0.010	12.088	0.000	0.100	0.139

Omnibus:	265.446	Durbin-Watson:	0.260
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10032.194
Skew:	1.633	Prob(JB):	0.00
Kurtosis:	24.700	Cond. No.	9.66e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.66e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Image 3 – Spearman's Correlation Heatmap

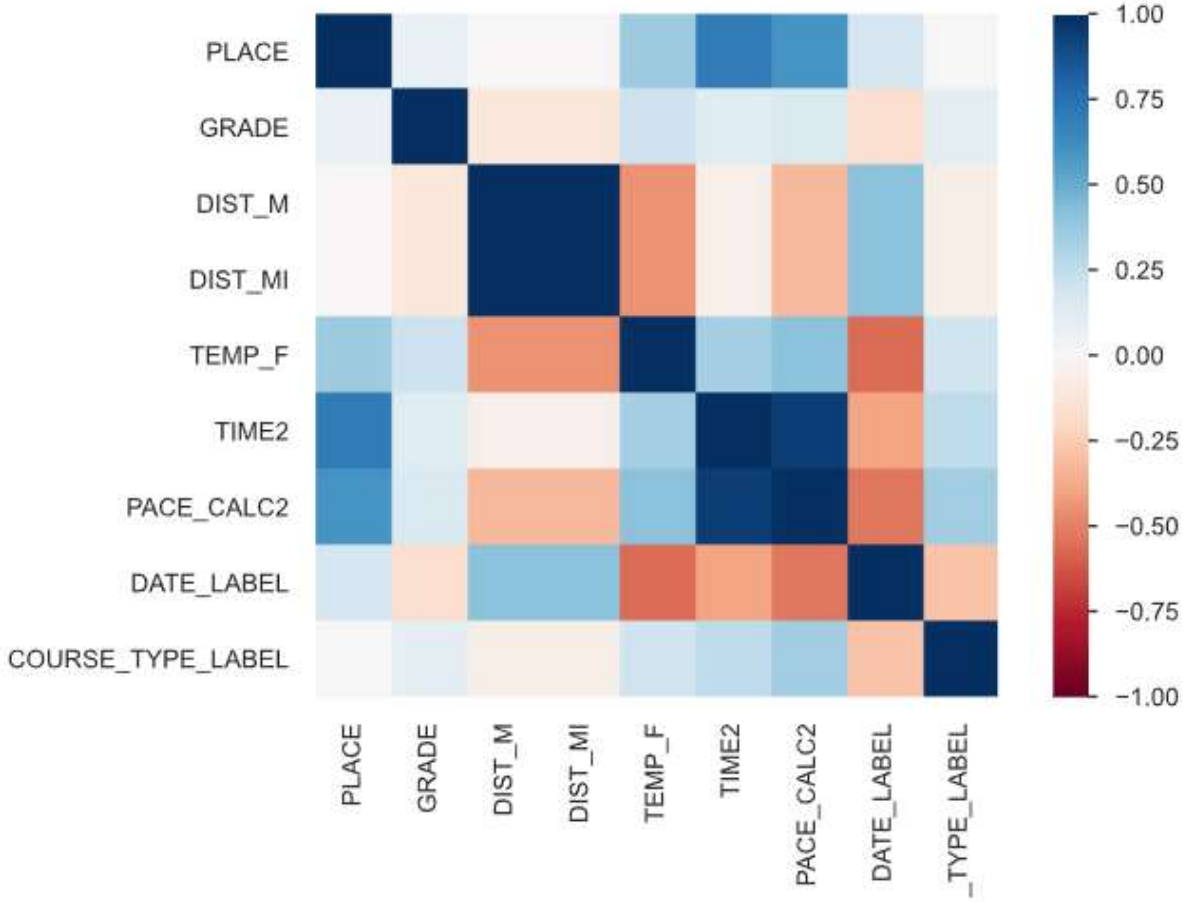
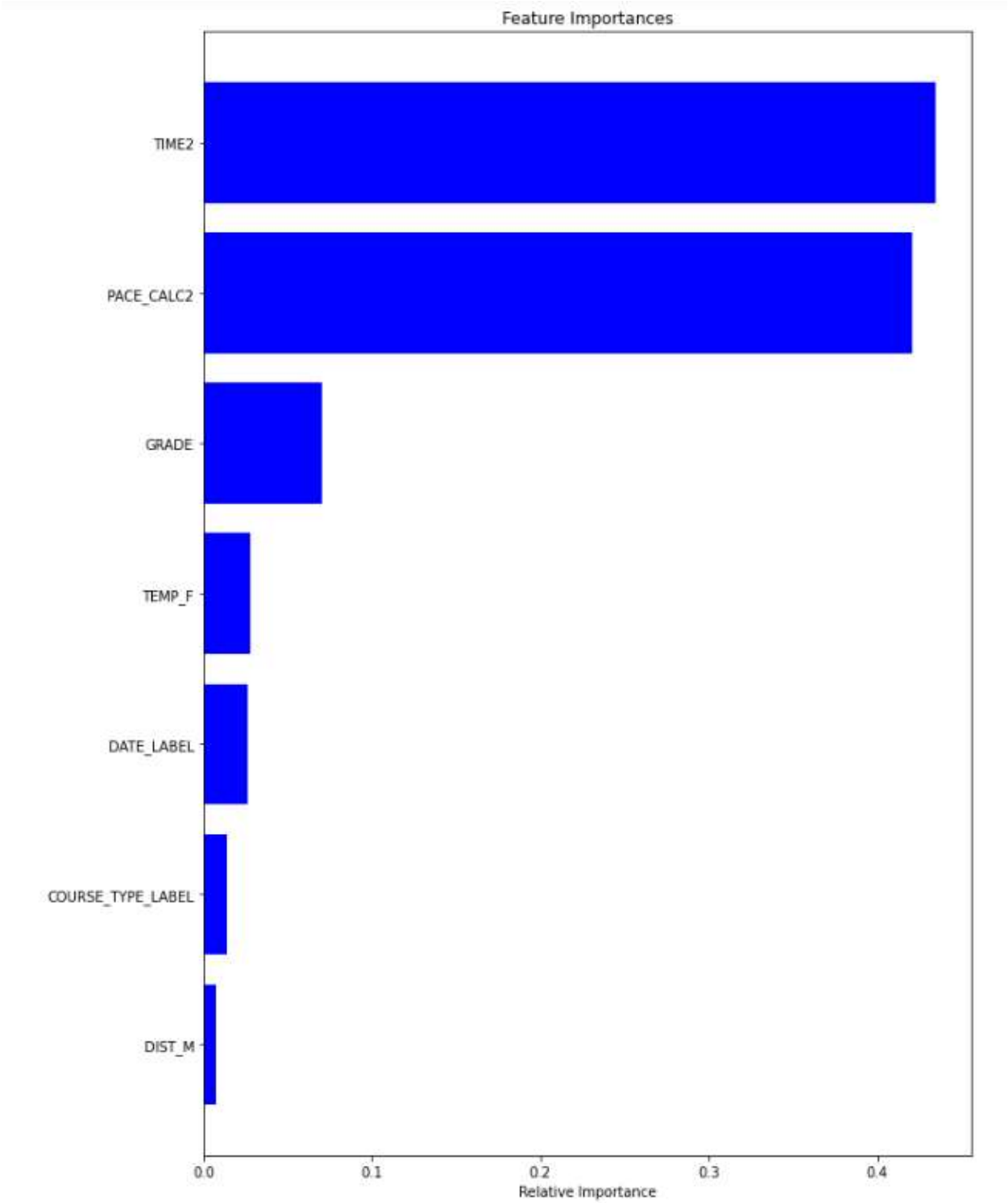


Image 4 – Feature Importance



Code Snippets 1 – Mean Paces

Mean pace of top 10 for each race

```
data.loc[data['PLACE'] <= 10].mean()
```

```
DIST_M               2933.333333
PLACE                   5.500000
DATE_LABEL              2.500000
TEMP_F                 69.000000
GRADE                   6.933333
TIME2                  13.878611
PACE_CALC2              7.588611
COURSE_TYPE_LABEL       1.000000
dtype: float64
```

Mean pace for each grade level

```
eighth_pace = data.loc[data['GRADE'] == 8]
eighth_pace['PACE_CALC2'].describe()
```

```
count    180.000000
mean       9.335833
std        1.925735
min        6.683333
25%        7.979167
50%        8.900000
75%       10.266667
max       17.583333
Name: PACE_CALC2, dtype: float64
```

```
seventh_pace = data.loc[data['GRADE'] == 7]
seventh_pace['PACE_CALC2'].describe()
```

```
count    236.000000
mean       9.693362
std        2.103619
min        6.083333
25%        8.095833
50%        9.141667
75%       10.750000
max       16.750000
Name: PACE_CALC2, dtype: float64
```

```
sixth_pace = data.loc[data['GRADE'] == 6]
sixth_pace['PACE_CALC2'].describe()
```

```
count    80.000000
mean      8.260625
std       1.273563
min       6.533333
25%       7.479167
50%       7.983333
75%       8.683333
max      12.833333
Name: PACE_CALC2, dtype: float64
```

Code Snippets 2 - Model Outputs

The first two cells illustrate predicted paces for the beginning (0) of the season and end (5) of the season for an 8th grader. The third cell illustrates the impact changing course type has on the predicted pace, changing from flat to hilly slows the predicted pace.

```
# Enter independent values to predict pace
DATE = 0 # Date Labels: 0-5 (beginning of season through end of season)
TEMP = 65 # Temperature in F
GRADE = 8 # Grade 5-8
COURSETYPE = 0 # COURSE TYPE: 0=FLAT,1=HILLY,2=ROLLING

predictedpace = regr.predict([[DATE,TEMP,GRADE,COURSETYPE]])
print(predictedpace)
```

```
[9.41096864]
```

```
# Enter independent values to predict pace
DATE = 5 # Date Labels: 0-5 (beginning of season through end of season)
TEMP = 65 # Temperature in F
GRADE = 8 # Grade 5-8
COURSETYPE = 0 # COURSE TYPE: 0=FLAT,1=HILLY,2=ROLLING

predictedpace = regr.predict([[DATE,TEMP,GRADE,COURSETYPE]])
print(predictedpace)
```

```
[7.6146448]
```

```
# Enter independent values to predict pace
DATE = 5 # Date Labels: 0-5 (beginning of season through end of season)
TEMP = 65 # Temperature in F
GRADE = 8 # Grade 5-8
COURSETYPE = 1 # COURSE TYPE: 0=FLAT,1=HILLY,2=ROLLING

predictedpace = regr.predict([[DATE,TEMP,GRADE,COURSETYPE]])
print(predictedpace)
```

```
[8.32764639]
```

These next two cells illustrate predicted paces for the beginning (0) of the season and end (5) of the season for an 6th grader. Note the lower predicted paces using the same inputs as the 8th grade example above. The final cell shows the impact of increasing the temperature on the predicted pace.

```
# Enter independent values to predict pace
DATE = 0 # Date Labels: 0-5 (beginning of season through end of season)
TEMP = 65 # Temperature in F
GRADE = 6 # Grade 5-8
COURSETYPE = 0 # COURSE TYPE: 0=FLAT,1=HILLY,2=ROLLING

predictedpace = regr.predict([[DATE,TEMP,GRADE,COURSETYPE]])
print(predictedpace)
```

```
[9.28166609]
```

```
# Enter independent values to predict pace
DATE = 5 # Date Labels: 0-5 (beginning of season through end of season)
TEMP = 65 # Temperature in F
GRADE = 6 # Grade 5-8
COURSETYPE = 0 # COURSE TYPE: 0=FLAT,1=HILLY,2=ROLLING

predictedpace = regr.predict([[DATE,TEMP,GRADE,COURSETYPE]])
print(predictedpace)
```

```
[7.48534225]
```

```
# Enter independent values to predict pace
DATE = 5 # Date Labels: 0-5 (beginning of season through end of season)
TEMP = 80 # Temperature in F
GRADE = 6 # Grade 5-8
COURSETYPE = 0 # COURSE TYPE: 0=FLAT,1=HILLY,2=ROLLING

predictedpace = regr.predict([[DATE,TEMP,GRADE,COURSETYPE]])
print(predictedpace)
```

```
[8.01754307]
```

Appendix 2 – References

Historical weather. Weather Underground. (n.d.). Retrieved June 24, 2022, from https://www.wunderground.com/history

Track & Field and Cross Country Statistics. Athletic.net. (n.d.). Retrieved June 24, 2022, from https://www.athletic.net/CrossCountry/School.aspx?SchoolID=65181