KodeinKGP: Technology Web 3.0 Society: AI and Metaverse Team: Review Task

Burnt Calories Predictor

Models Used: Linear Regression and RandomForestRegressor

Introduction

The problem statement involves using the given `dailyActivity_merged.csv` dataset to predict the amount of calories burned when the activity of the user is described by the given Features.

EDA Findings:

1. The dataset contains 15 Features and 457 data rows.
2. The Features are as follows:

```
['Id', 'ActivityDate', 'TotalSteps', 'TotalDistance', 'TrackerDistance',
'LoggedActivitiesDistance', 'VeryActiveDistance', 'ModeratelyActiveDistance',
'LightActiveDistance', 'SedentaryActiveDistance', 'VeryActiveMinutes', 'FairlyActiveMinutes',
'LightlyActiveMinutes', 'SedentaryMinutes', 'Calories']
```

3. The dataset was pretty clean with no missing rows, only some abnormalities..
4. One such abnormality was the presence of 61 columns where almost all values were 0 except for `'SedentaryMinutes' and 'Calories'`.
5. This, however, did not need to be handled as the data turned out to be crucial for the prediction.
6. The addition of some correlation columns, like `StepSize`, turned out to be extremely beneficial for the accuracy of the model.

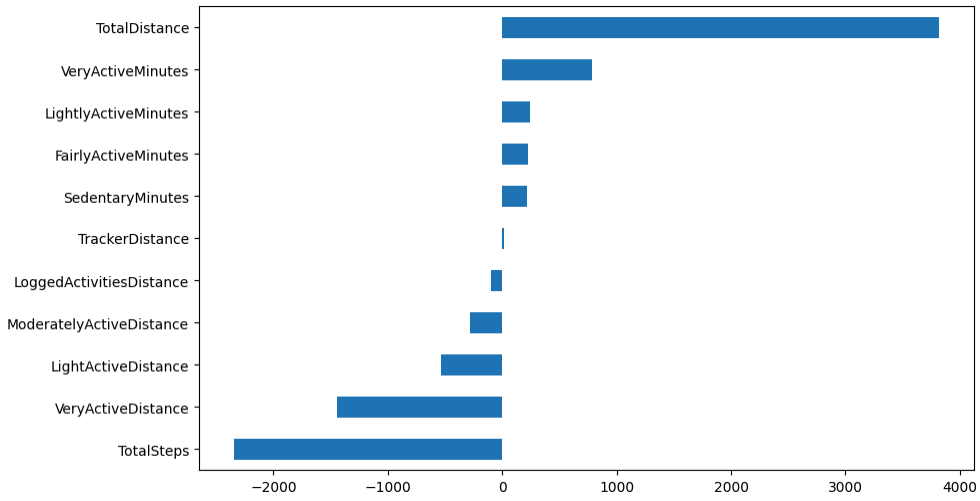
Feature Engineering Choices:

Linear Regression:

1. First, the columns `ID` and `ActivityDate` were dropped as they have no correlation with the `Calorie` value. The `ActivityDate` would have been useful if the weather data were given.
2. The data(except the `y` value) was scaled using the `StandardScaler()` function.
3. The Ridge Regularisation was used to eliminate overfitting.
4. The conclusion of overfitting came due to the R2 and Cross Validation R2 Score of the raw model:

```
Model:            LinearRegression()
R2 Score:         0.7405772017712484
CV R2 Score:      -0.6600870916930928
MAE:              327.2878378828919
RMSE:             425.8159069112786
```

5. The weight of `SedentaryActiveDistance` was very small, hence it was dropped with a negligible decrease in performance, but faster(theoretical) execution time on `Gradient Descent`:
    a. Before Drop: `alpha: 0.027, R2 Score: 0.740, CV R2 Score: 0.503, MAE: 326.904, RMSE: 425.714`
    b. `After Drop:` `alpha: 0.034, R2 Score: 0.7419, CV R2 Score: 0.506, MAE: 326.560, RMSE: 425.406`
6. Weights:



7. We observe that `TrackerDistance` and `TotalDistance` have only 16 rows, where the two have unequal values, hence, we drop one of the columns, here `TrackerDistance`, as it has very little weight.
8. We add a new column, `StepSize`, that is numerically equal to `TotalDistance / TotalSteps,` which improves performance. Since both columns involved have high weights, we cannot drop either of the two columns. This improves the `RMSE` slightly, but we observed that further columns give better results with this column(especially in the `RandomForestRegressor`)
9. We observe that `VeryActiveMinutes` and `VeryActiveDistance`, and `FairlyActiveMinutes` and `ModeratelyActiveDistance` have very similar weight values, so we may be able to drop one of each to reduce computation load, but they prove to be significant.
10. We observe that there are `61 rows` where almost everything except the `SedentaryMinutes` and `Calories` are zero, but this data seems to help the model learn about inactive days, so it needs to be kept in.
11. We add a column, `VeryActiveSpeed`, numerically equal to `VeryActiveDistance / VeryActiveMinutes`, which improves the CV R2 score of the model.
12. Repeated this with Light and Moderate columns, but did not give significant results.

RandomForestRegressor:

1. For this model, the scaling does not matter, and it is demonstrated in the first two cells, as we get identical values.
2. Here, the weights obtained were:

	Feature	Importance
1	TotalDistance	0.299613
11	SedentaryMinutes	0.174248
8	VeryActiveMinutes	0.131019
9	FairlyActiveMinutes	0.083482
2	TrackerDistance	0.076097
10	LightlyActiveMinutes	0.073707
6	LightActiveDistance	0.060951
0	TotalSteps	0.037226
3	LoggedActivitiesDistance	0.020530
4	VeryActiveDistance	0.019460
5	ModeratelyActiveDistance	0.017831
7	SedentaryActiveDistance	0.005835

3. While both the `SedentaryActiveDistance` and the `LoggedActivitiesDistance` show very little weight, the model performed better with `LoggedActivitiesDistance` removed, but going further, we observed that the weight of `SedentaryActiveDistance` becomes zero. Here, we only removed `LoggedActivitiesDistance`.
4. Once again, we attempt to remove one of TotalDistance and TrackerDistance, but here we observe, removing both results in a massive improvement in both scores, and hence we can drop both. With `num_estimators3: 22, max_depth3: 11`, the `R2 score jumps from 0.641 to 0.773`, the `CV R2 Score jumps from 0.402 to 0.502`, the `MAE drops from 378 to 282,` and the `RMSE drops from 496 to 398`.
5. Though this is a very favourable result, the justification for dropping these columns cannot be provided for sure, as both have significant weights and are significantly important, semantically. The only justification possible is that some other feature compensates more than enough for their loss. This column, here, is `StepSize`, which we added in the last section

	Feature	Importance
9	StepSize	0.376857
0	TotalSteps	0.266530
8	SedentaryMinutes	0.126639
5	VeryActiveMinutes	0.056891
6	FairlyActiveMinutes	0.055263
7	LightlyActiveMinutes	0.049686
3	LightActiveDistance	0.047910
2	ModeratelyActiveDistance	0.012011
1	VeryActiveDistance	0.006664
4	SedentaryActiveDistance	0.001550

6. We attempt to recreate the experiment of Step 9 in the last section. In this case, dropping `VeryActiveDistance` and `ModeratelyActiveDistance` gives us a slight improvement in the performance, with less required computation. The best justification is that a combination of `StepSize` and the `corresponding time features` gives the desired result.

7. A similar approach with `LightActiveDistance` gives a similar result. The current weights are:

	Feature	Importance
6	StepSize	0.387700
0	TotalSteps	0.281464
5	SedentaryMinutes	0.128124
4	LightlyActiveMinutes	0.070515
2	VeryActiveMinutes	0.067557
3	FairlyActiveMinutes	0.062137
1	SedentaryActiveDistance	0.002502

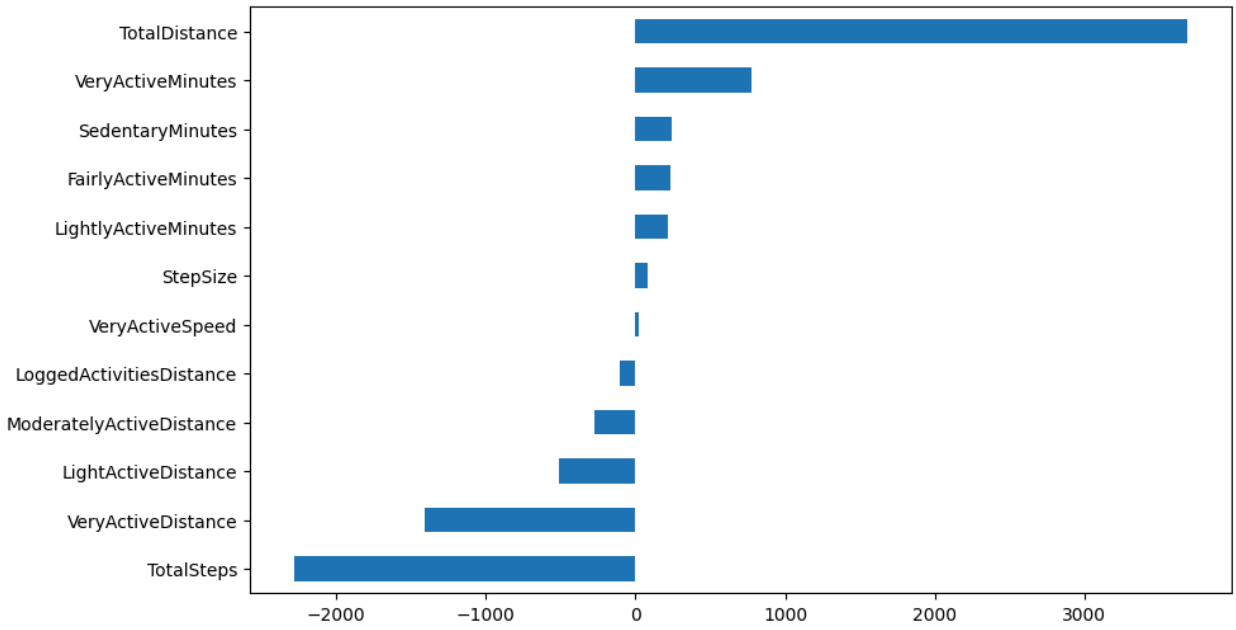
8. Dropping the abnormal rows is destructive for the performance of the model.
9. Dropping `TotalSteps` as well causes massive overfitting of the data, probably because of a smaller number of significant features
10. Adding the `FairlyActiveSpeed` and `SedentarySpeed` columns, similar to the previous speed columns, gives better scores.

Results:

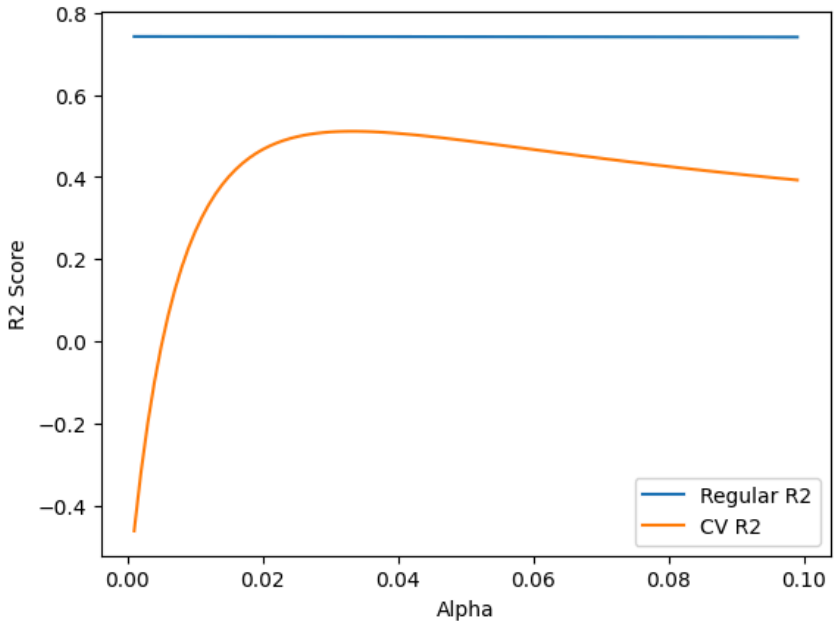
Model	R2 Score	CV R2 Score	MAE	MSE
LinearRegression	0.742	0.511	330.248	424.541
RandomForestRegressor	0.804	0.501	263.362	369.728

Linear Regression:

Plot of Weights:

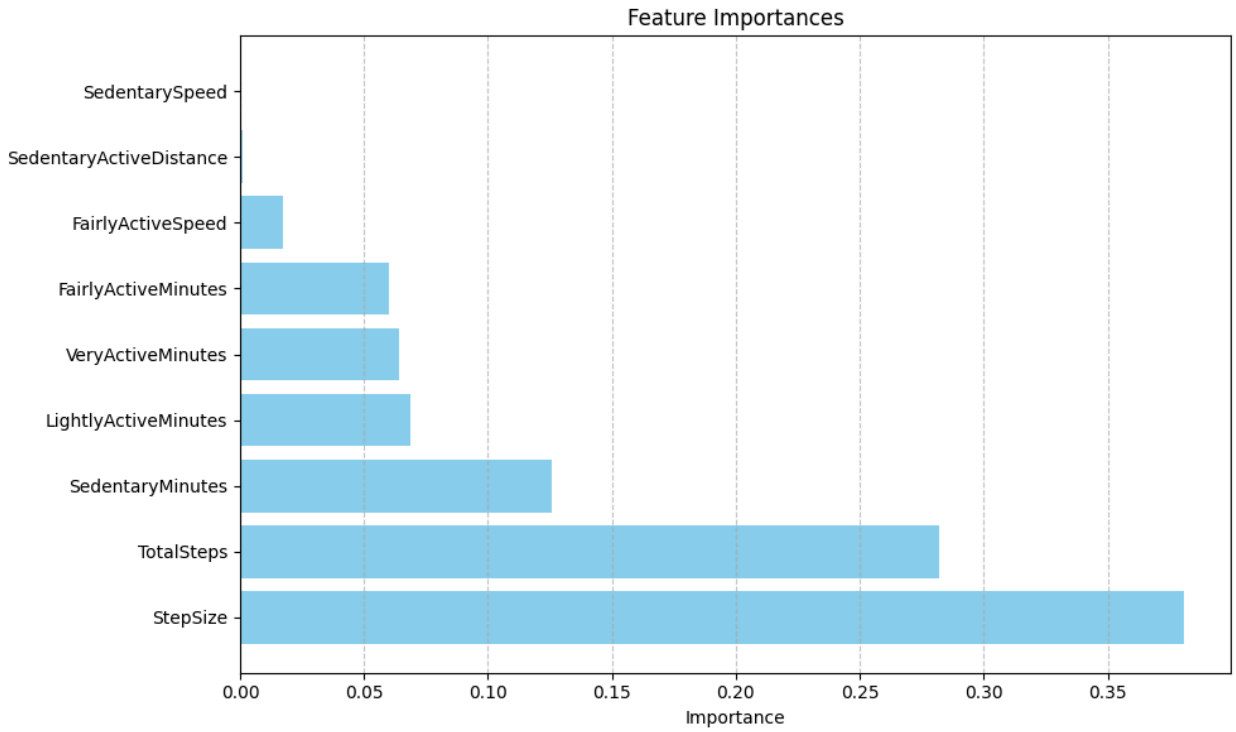


Variation of scores with alpha, in the final step:



RandomForestRegressor:

Plot of importances:



Key Insights:

1. Speed has turned out to be a major indicator of Calories burned, as adding the Speed columns resulted in much better results.
2. `SedentaryActiveDistance`, despite having much less weight, is significant, as the 61 columns with no activity indicate the rough amount of calories burned when activity is minimal.
3. The distance-related fields, like `StepSize` and `TotalDistance`, carried the most weight.
4. RandomForestRegressor provides a better prediction, in general, as compared to LinearRegression.