Multi-Level Statistical Model for Forecasting Solar Radiation

Team Members

- Pratham Nayak - 191IT241
- Aprameya Dash - 191IT209
- Suyash Chintawar - 191IT109

Introduction and Dataset

- Solar radiation relies on weather patterns such as sunshine duration, dispersion radiation, temperature and relative humidity. This fact can be used to create a model which can be used for solar radiation forecasting.
- So, we propose a multileveled statistical model which can forecast solar radiation using other independent factors.
- The dataset has a total of 32,686 samples.
- There are 11 attributes corresponding to each sample
- Radiation is the response variable and the rest of the 10 attributes are the predictor variable.
- We plan to perform an 80-20 split to generate the train and test set.
    - Training data: 26,148 and Test data: 6,538

Fig. Dataset

	UNIXTime	Date	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
0	1475229326	9/29/2016 12:00:00 AM	23:55:26	1.21	48	30.46	59	177.39	5.62	06:13:00	18:13:00
1	1475229023	9/29/2016 12:00:00 AM	23:50:23	1.21	48	30.46	58	176.78	3.37	06:13:00	18:13:00
2	1475228726	9/29/2016 12:00:00 AM	23:45:26	1.23	48	30.46	57	158.75	3.37	06:13:00	18:13:00
3	1475228421	9/29/2016 12:00:00 AM	23:40:21	1.21	48	30.46	60	137.71	3.37	06:13:00	18:13:00
4	1475228124	9/29/2016 12:00:00 AM	23:35:24	1.17	48	30.46	62	104.95	5.62	06:13:00	18:13:00
...	...	...	...	...	...	...	...	...	...	...	...
32681	1480587604	12/1/2016 12:00:00 AM	00:20:04	1.22	44	30.43	102	145.42	6.75	06:41:00	17:42:00
32682	1480587301	12/1/2016 12:00:00 AM	00:15:01	1.17	44	30.42	102	117.78	6.75	06:41:00	17:42:00
32683	1480587001	12/1/2016 12:00:00 AM	00:10:01	1.20	44	30.42	102	145.19	9.00	06:41:00	17:42:00
32684	1480586702	12/1/2016 12:00:00 AM	00:05:02	1.23	44	30.42	101	164.19	7.87	06:41:00	17:42:00
32685	1480586402	12/1/2016 12:00:00 AM	00:00:02	1.20	44	30.43	101	83.59	3.37	06:41:00	17:42:00

32686 rows × 11 columns

Fig. Dataset after cleaning

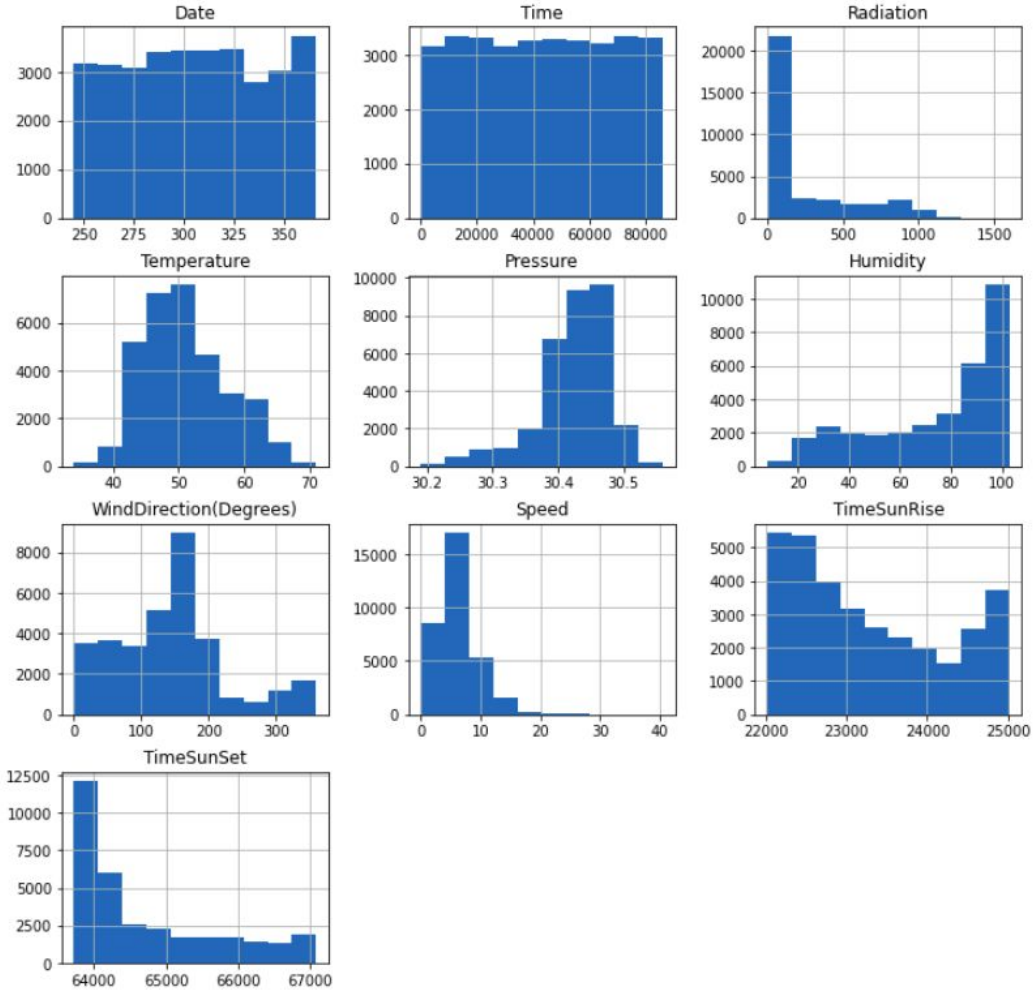
	Date	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
0	273	86126	1.21	48	30.46	59	177.39	5.62	22380	65580
1	273	85823	1.21	48	30.46	58	176.78	3.37	22380	65580
2	273	85526	1.23	48	30.46	57	158.75	3.37	22380	65580
3	273	85221	1.21	48	30.46	60	137.71	3.37	22380	65580
4	273	84924	1.17	48	30.46	62	104.95	5.62	22380	65580
...	...	...	...	...	...	...	...	...	...	...
32681	336	1204	1.22	44	30.43	102	145.42	6.75	24060	63720
32682	336	901	1.17	44	30.42	102	117.78	6.75	24060	63720
32683	336	601	1.20	44	30.42	102	145.19	9.00	24060	63720
32684	336	302	1.23	44	30.42	101	164.19	7.87	24060	63720
32685	336	2	1.20	44	30.43	101	83.59	3.37	24060	63720

32686 rows × 10 columns

Dataset Summary

	Date	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
count	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000
mean	306.110965	43277.574068	207.124697	51.103255	30.422879	75.016307	143.489821	6.243869	23258.431133	64691.463624
std	34.781367	24900.749819	315.916387	6.201157	0.054673	25.990219	83.167500	3.490474	931.122823	995.053346
min	245.000000	1.000000	1.110000	34.000000	30.190000	8.000000	0.090000	0.000000	22020.000000	63720.000000
25%	277.000000	21617.000000	1.230000	46.000000	30.400000	56.000000	82.227500	3.370000	22440.000000	63900.000000
50%	306.000000	43230.000000	2.660000	50.000000	30.430000	85.000000	147.700000	5.620000	23040.000000	64260.000000
75%	334.000000	64849.000000	354.235000	55.000000	30.460000	97.000000	179.310000	7.870000	24000.000000	65340.000000
max	366.000000	86185.000000	1601.260000	71.000000	30.560000	103.000000	359.950000	40.500000	25020.000000	67080.000000

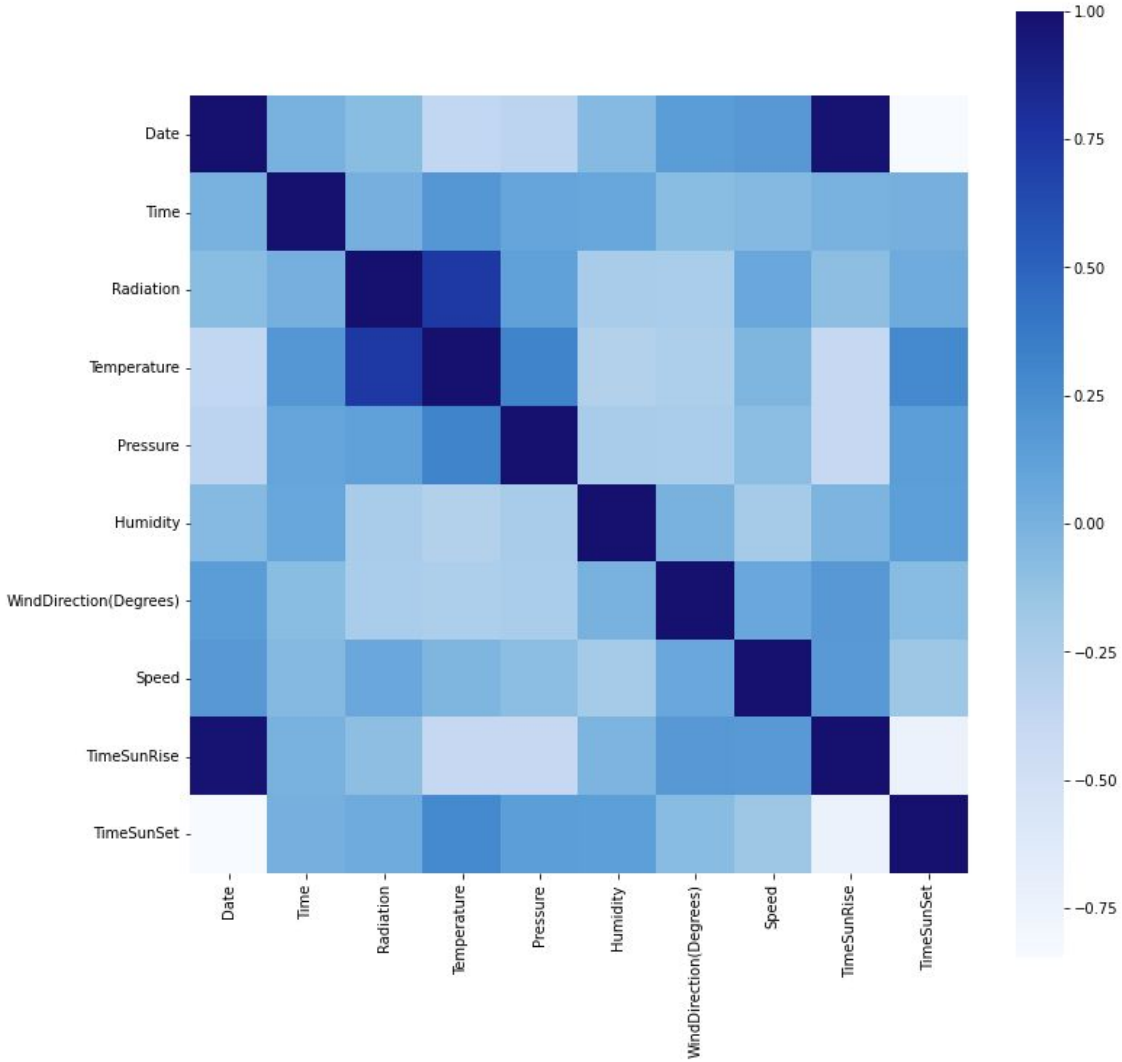
Fig. Histogram of different features of dataset



Correlation Matrix

	Date	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
Date	1.000000	-0.007094	-0.081320	-0.370794	-0.332762	-0.063760	0.153255	0.174336	0.981939	-0.847401
Time	-0.007094	1.000000	0.004348	0.197227	0.091066	0.077851	-0.077956	-0.057908	-0.006639	0.008038
Radiation	-0.081320	0.004348	1.000000	0.734955	0.119016	-0.226171	-0.230324	0.073627	-0.092850	0.045688
Temperature	-0.370794	0.197227	0.734955	1.000000	0.311173	-0.285055	-0.259421	-0.031458	-0.380968	0.285131
Pressure	-0.332762	0.091066	0.119016	0.311173	1.000000	-0.223973	-0.229010	-0.083639	-0.380399	0.146884
Humidity	-0.063760	0.077851	-0.226171	-0.285055	-0.223973	1.000000	-0.001833	-0.211624	-0.023955	0.135243
WindDirection(Degrees)	0.153255	-0.077956	-0.230324	-0.259421	-0.229010	-0.001833	1.000000	0.073092	0.176929	-0.068040
Speed	0.174336	-0.057908	0.073627	-0.031458	-0.083639	-0.211624	0.073092	1.000000	0.167075	-0.159400
TimeSunRise	0.981939	-0.006639	-0.092850	-0.380968	-0.380399	-0.023955	0.176929	0.167075	1.000000	-0.738271
TimeSunSet	-0.847401	0.008038	0.045688	0.285131	0.146884	0.135243	-0.068040	-0.159400	-0.738271	1.000000

Fig. Heatmap of correlation matrix



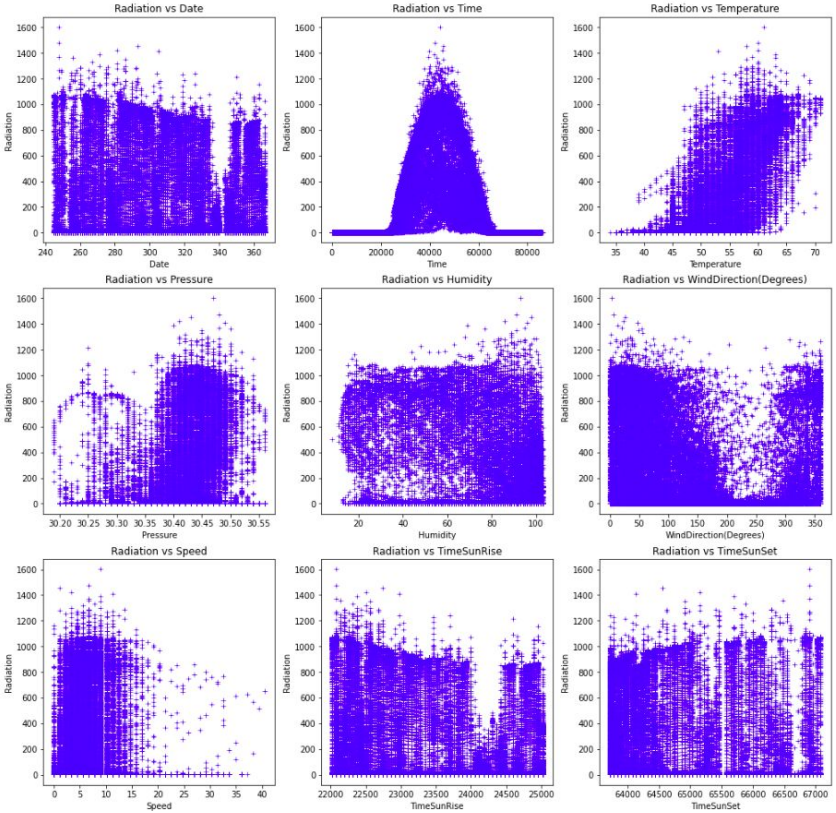


Fig. Scatterplot of different features of dataset against Radiation

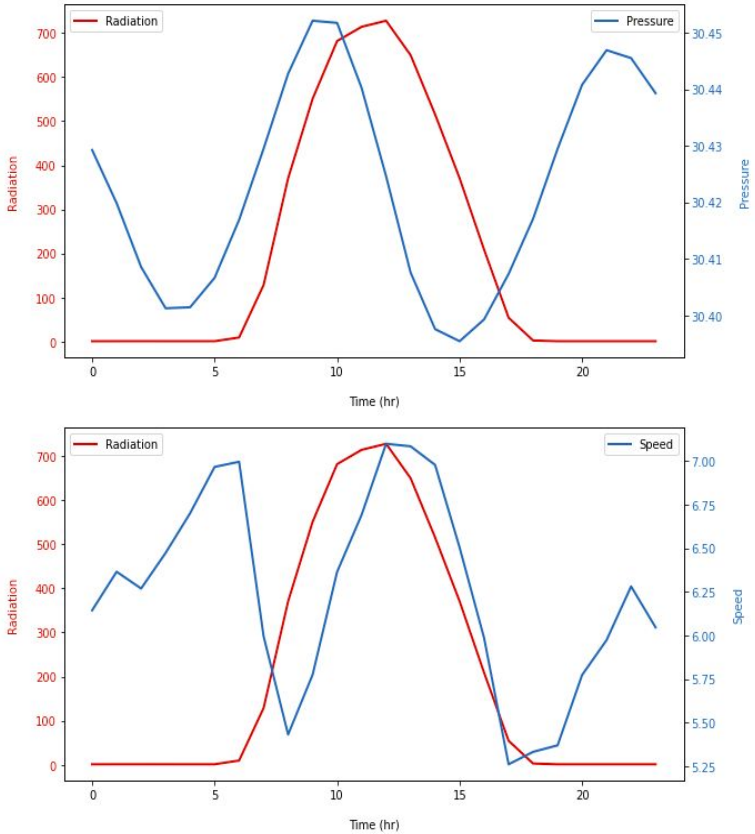


Fig: Hourly Averages

Inferences

- The Unix Time feature is redundant as it can be derived using the Date and Time, hence it is dropped.
- Date helps to determine seasonal factors and Radiation depends on seasonal changes.
- The scatter plot indicates that radiation depends on the time of the day. Radiation is low during the early morning, approaches maximum value around noon, and gradually decreases by the end of the day.
- The correlation matrix indicates that Temperature has the highest correlation with radiation.
- There is a high correlation between temperature and pressure, thus pressure has a transitive dependency with radiation.
- Humidity and Wind Direction have high correlation with Radiation as shown by the correlation matrix.
- The hourly average plot of Wind speed indicates that whenever the Radiation is high, the wind speed is usually high as well.
- Time of sunrise and sunset have very low correlation values with radiation, and neither do they show any relation with Radiation in the scatter plot. Therefore these features are dropped.

Results: Correctness of feature selection

Standard models	Test R^2 score (without FS)	Test R^2 score (with FS)
SVR	0.137	0.366
Linear Regression	0.626	0.625
MLP Regressor	0.817	0.825
Decision Tree Regressor	0.872	0.874
Gradient Boost	0.884	0.883
XGBoost	0.930	0.931
Random Forest	0.938	0.938

After dropping the features that are judged as irrelevant by the statistical techniques, the R^2 scores have either improved or have remained almost the same.

As dimensional reduction is always advantageous we conclude that the feature selection is correct and the inferences derived regarding the dependencies of each selected feature with response variable are true.

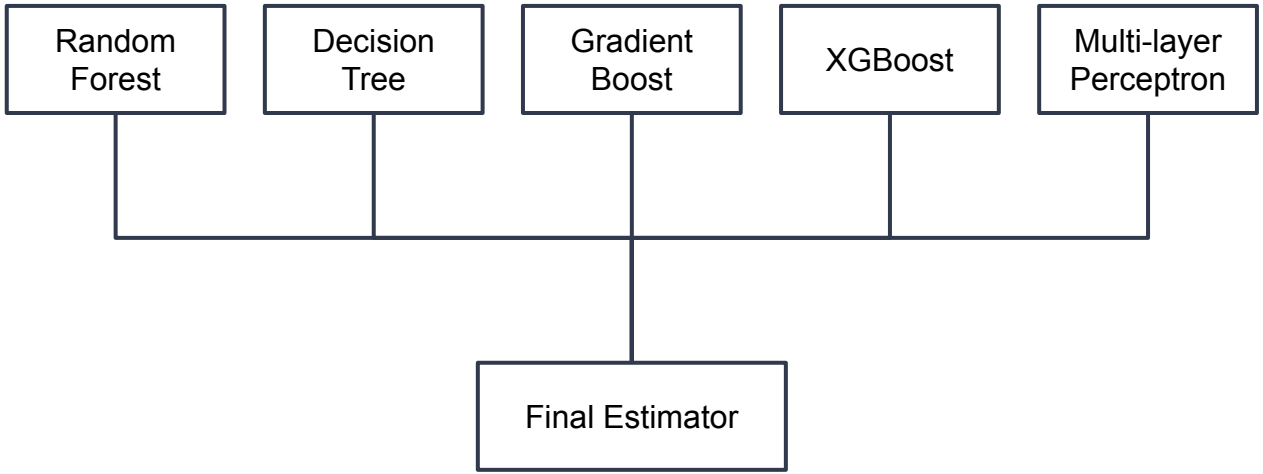
Results: R^2 score of single level models

Standard models	Train R^2 score	Test R^2 score
SVR	0.365	0.366
Linear Regression	0.619	0.625
MLP Regressor	0.824	0.825
Decision Tree Regressor	1.000	0.874
Gradient Boost	0.891	0.883
XGBoost	0.969	0.931
Random Forest	0.990	0.938

It is observed that among individual standard regression models, tree based models such as Decision Tree, Random Forest, Gradient Boost and XGBoost Regressor perform well.

Therefore these are used as the base level regressors for 2-level models.

Results: Architecture of 2-level models



- The best-performing standard models are taken in the base layer of the two-level model.
- Then, all these standard models are used as final estimators one-by-one.

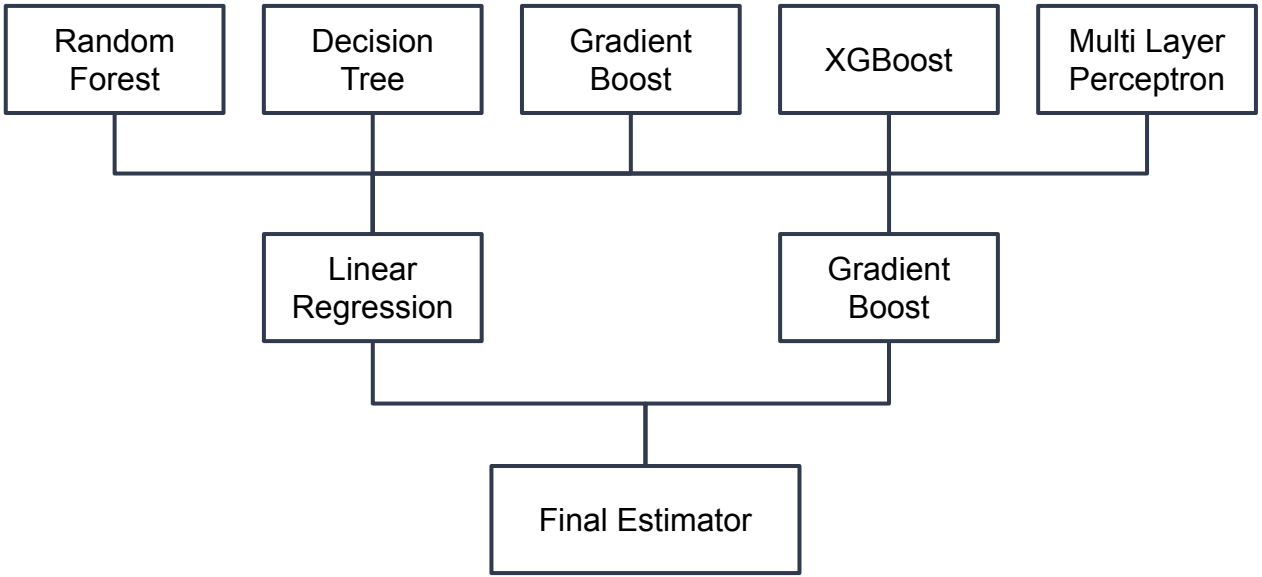
Results: R^2 score of 2-level models

Final Estimator	Train R^2 score	Test R^2 score
SVR	0.980	0.932
Linear Regression	0.987	0.940
MLP Regressor	0.986	0.939
Decision Tree Regressor	0.934	0.876
Gradient Boost	0.982	0.941
XGBoost	0.977	0.935
Random Forest	0.981	0.937

It is observed that two-level models give comparatively better performance when compared to the standard models.

Here the best R^2 score is obtained when Gradient Boost is used as the final estimator.

Results: Architecture of 3-level models



- The two best-performing two-level models are then combined with 3 different standard models as final estimators for the three-level models.

Results: R^2 score of 3-level models

Final Estimator	Train R^2 score	Test R^2 score
Linear Regression	0.986	0.941
MLP Regressor	0.985	0.941
Gradient Boost	0.986	0.942

When we move from 2-level to 3-level models, there is a slight improvement in terms of R^2 score (approx 0.001), however the training time is several folds greater than 2-level models.

This shows that multilevel statistical models can perform better prediction as compared to individual standard models. However, in this particular case, 2-level models are preferred.

THANK YOU