Preprocessing performed

- Calculate distance from longitude and latitude.
- Take difference between longitude and longitude
- Drop pickup/drop longitude and latitude.
- Take dates as different features like(year month and day).
- Take time as in shift(day[0600-2359] and night[0000-0600])

Gradient Boosting classifier

When we try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are **noise, variance, and bias**. Ensemble helps to reduce these factors (except noise, which is irreducible error) Boosting is an ensemble technique in which the predictors are not made independently, but sequentially(that's why training is slow).

- Reduce bias and variance.
- Sequential classifier.
- Sometimes overfits.

It is ensemble of weak prediction models, typically decision trees.
The intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better.

GBTs build trees one at a time, where each new tree helps to correct errors made by previously trained tree. With each tree added, the model becomes even more expressive. There are typically three parameters **- number of trees, depth of trees and learning rate, and the each tree built is generally shallow.**

GBDT are better learners than Random Forests and even linear regression.
Gradient boosting Algorithm involves three elements:

- A loss function to be optimized.
- Weak learner to make predictions.
- An additive model to add weak learners to minimize the loss function.

Linear regression

A basic assumption of linear regression is that sum of its residuals(the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the **residual** (e)) is 0, i.e. the residuals should be spread randomly around zero.

It is simple even when it doesn't fit the data exactly, we can use it to find the nature of the relationship between the two variables.

Prone to **outlairs**(After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an ***outlier***. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an ***influential observation***. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line.

Some major differences.

- Linear Regression Prone to outlairs but GB is not that prone as in some trees it misses the outlairs.
- Only regression line gives relationships between data but in GB it collection of which gives better result.

XGBoost and Gradient boosting are almost same

Top 3 score:

Gradient boosting : 3.72
Xgboost : 3.75
Linear Regression : 6.39353

Results:

submissions_gradient.csv 28 minutes ago by Ritesh Gupta add submission details	3.72070	☐
submissions_gradient.csv a day ago by Ritesh Gupta add submission details	3.75735	☐
submissions_xgb.csv a day ago by Ritesh Gupta add submission details	3.75307	☐
submissions_linear.csv a day ago by Ritesh Gupta add submission details	6.39353	☐
submission.csv 9 days ago by Ritesh Gupta gradient_modify_long_lat_dif	5.80309	☐
submission.csv 9 days ago by Ritesh Gupta gradient_modify	8.01681	☐
submission.csv 9 days ago by Ritesh Gupta gradient_modify	3.80858	☐
submission.csv 9 days ago by Ritesh Gupta gradient_boosting	3.79072	☐
submission.csv 9 days ago by Ritesh Gupta gradient_boosting	3.79853	☐
submission.csv 9 days ago by Ritesh Gupta gradient_boosting_est_1000_lr_0.1	4.15131	☐

submission.csv 12 days ago by Ritesh Gupta SGD	175126578042498000.00000	☐
submission.csv 12 days ago by Ritesh Gupta ADABOOST	7.84438	☐
submission.csv 13 days ago by Ritesh Gupta ADABOOST	7.45995	☐
submission.csv 13 days ago by Ritesh Gupta ADABOOST	7.46120	☐
submission.csv 13 days ago by Ritesh Gupta ADABOOST	7.46120	☐