**IPL SCORE PREDICTION**

The data set I took is “ipl.csv” downloaded from the Kaggle website. IPL(Indian Premiere League) is a famous cricket league which is as popular as International cricket.The dataset contains record per each ball that is bowled in that particular match.

**Objective: -** The main objective of this assignment is that given a game situation(not lessthan 5 overs) our algorithm will predict the final score of the innings that the batting team can score.

**Preprocessing: -**

1. I found that venue column is an important feature from the Bar plot as each venue has different scoring rates.
2. From the correlation matrix plot I filtered the important features.
3. I have considered the teams which are there from long time and ignored teams which played 1-3 years
4. Performed one hot encoding with respect to columns venue,bat\_team and bowl\_team
5. Merged the duplicate columns like mumbai1,mumbai2…etc.
6. Since it is a time series data, I have done train test split based on time

**Modelling:-**

Since our problem statement is a regression problem, I have tried multiple regression algorithms.Find the below statistics for each model. I have chosen KPI(Key Performing Indicator) as Mean Squared Error.

1. Linear Regression

MAE: 11.896356924448696

MSE: 256.3724115968118

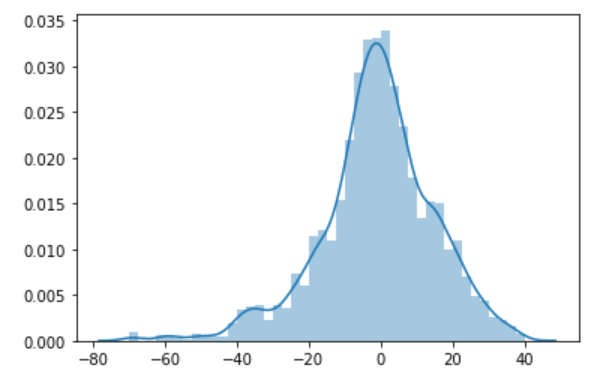
RMSE: 16.01163363298111

1. Ridge Regression

MAE: 11.896356924448696

MSE: 256.3724115968118

RMSE: 16.01163363298111



The above plot is the distribution of difference in actual and predicted scores.

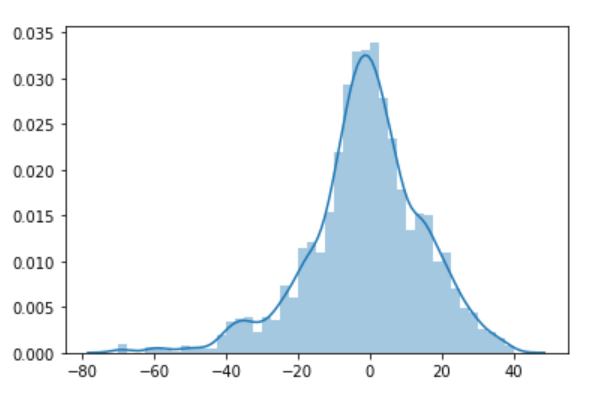
If you observe the above plot, the distribution is almost normally distributed and almost 95% data have the score difference within 20 runs. Which means this model is working fairly well.

1. Ridge Regression :

MAE: 11.876143652607277

MSE: 254.85371136977776

RMSE: 15.964138290862358



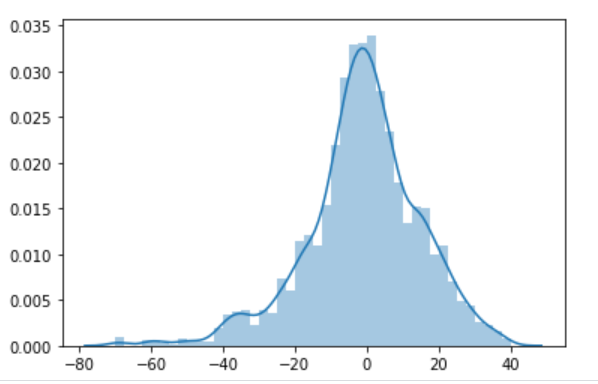
For the Ridge regression, the MSE has reduced compared to previous models.Even the distribution Plot also better distributed compared to previous models.

1. Lasso Regression

MAE: 11.806727584659896

MSE: 250.96193715531106

RMSE: 15.841778219483793



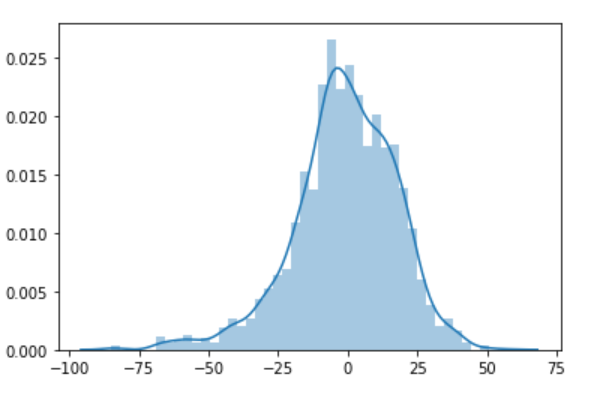
The MSE had even decreased for Lasso regressor.You can also observe this in the above plot.

1. KNN - GridSearchCV :

MAE: 14.090681688765967

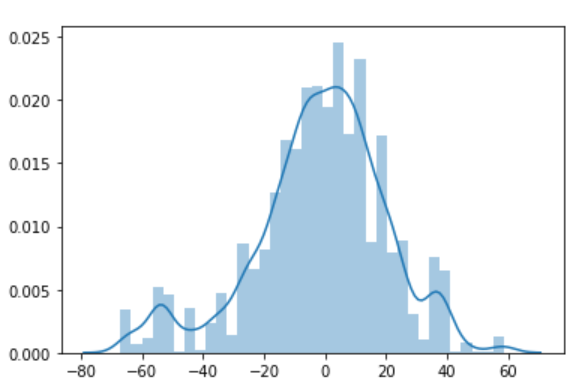
MSE: 339.61800361352414

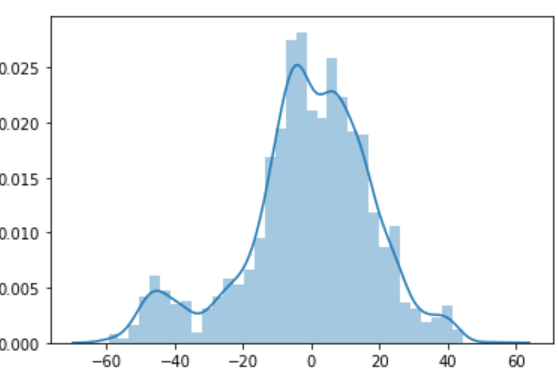
RMSE: 18.428727672129842



From the above stat and plot, its clear that KNN was not performed well enough.

1. Decision Tree Regression and Random Forest Regression:





You can observe in the code. Both the models are also failed to give good results.

**Conclusion :**

As per the experimentation the Lasso Regression model gave us the best results. That is MSE: 250.96.

Note: This Results can be improved if we further do more hyper parameter tuning.