We started by exploring the data and visualising the target variable with date, we found a clear pattern. Owing to this we created features from date such as day of week, day of year, day of month, month and year. Using these features with the given raw features we tried few models from linear to XGBoost. Among these XGBoost proved to be the best and gave us a public LB score of 133.xx.

After this we tried binning the continuous weather variables like “Average\_Breeze\_Speed”, “Direction\_Of\_Wind” and so on. Based on our intuition we also aggregated months into seasons since the footfall may directly be dependent on it. The visualisations also indicated a clear difference in the mean of footfall corresponding to different seasons. Adding to these we created range features for pressure, breeze speed etc. from given max and min values. These all features gave us an improved score of 123.xx. Tuning the model further improved the score to 118.xx. And this was the score on which we were stuck where nothing seemed to work further. Another set of features that we tried were grouped mean of footfall by “Park\_ID and binned direction of wind”, “Park\_ID and min moisture” etc., but they led to overfitting.

As nothing was now working we thought of diving deep into the data with more visualisations. We plotted nearly all variables with date and sensed some possible noise in them. To remove the noise we thought of smoothing the variables with their moving averages / rolling means with windows of 3, 7 and 30 days and this was something which gave us a huge jump in our score from 118 to 107.xx. We didn’t stop here and thought of another feature “percentage change” of weather variables from one day to next day since somehow we wanted to convey the algorithm the change in weather from a day to another. This worked too and our score further improved to 100.xx. We now played with different window for moving averages and checked how they reflected on CV while avoiding overfitting. Our final model had around 43 features which scored 96.xx and 86.xx on public and private LB respectively.

What we learned is feature engineering is one of the key elements to improve your score if done in the right way.