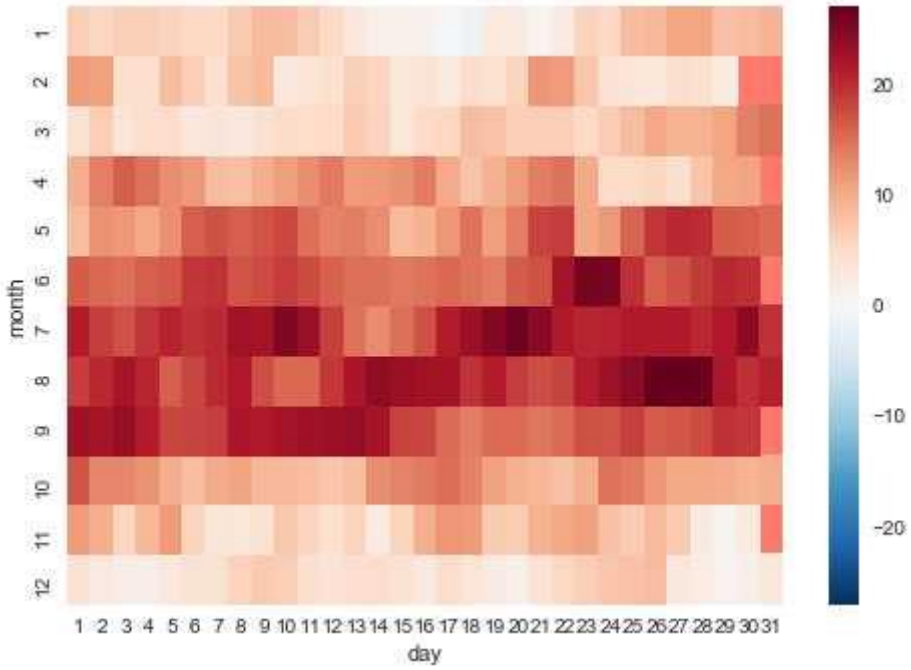
WEATHER FORECASTING IN BASEL SWITZERLAND

Data Science Crew



Gaurav. Joshi (MT2018035)
Manpreet. Singh. Tuteja (MT2018058)
Tushar. V. Bharadwaj (MT2018126)

DESCRIPTION

From the publicly available meteorological data of Basel Switzerland, we are performing Data Analysis on Weather Patterns, Precipitation, snowfall, and how these features have an affect on the mean temperature of the city.

By determining the strongly co-related factors, we predict the mean temperature based on humidity and barometer readings for that day.

ABOUT THE DATASET

We've obtained the data from Meteoblue, which is a meteorological service created at University of Basel. It provided over 20 years of weather data on an hourly and daily basis of Basel Switzerland and it is publicly available at https://www.meteoblue.com/en/.

We've taken the daily data from 1st January 2012 to 1st July 2017 instead of hourly data, the data collected from here are, Temperature, Humidity, Pressure, Precipitation, Snowfall, Cloud Cover, Sunshine Duration, Solar Radiations, Wind Speeds and Wind Gusts.

APPROACH TAKEN

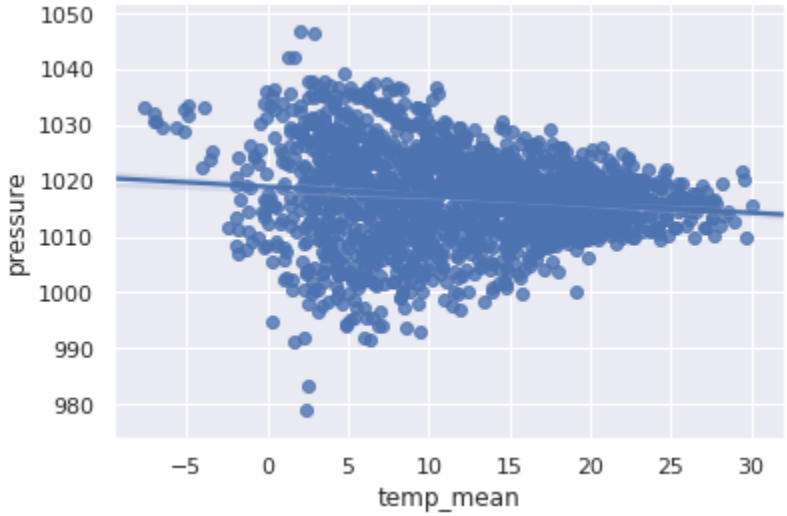
At first, we obtained some domain knowledge by reading on how different factors that affect the temperature of a place, and how it varies with respect to each other. After gathering some domain knowledge, we cross verified how much does each available feature play a role in affecting the temperature. With help of data visualization, we were able to pin point which all features had no co-relation with temperature, hence we chose not use them in our model. Then we created a co-relation matrix to see how much each of the selected features played a role to get better idea of how much weightage does each of the feature have. Thereafter, we proceeded to try with multiple regression models and tested which one of them had a good accuracy, and picked best few models and stacked them together to give a better model.

DATA ANALYSIS

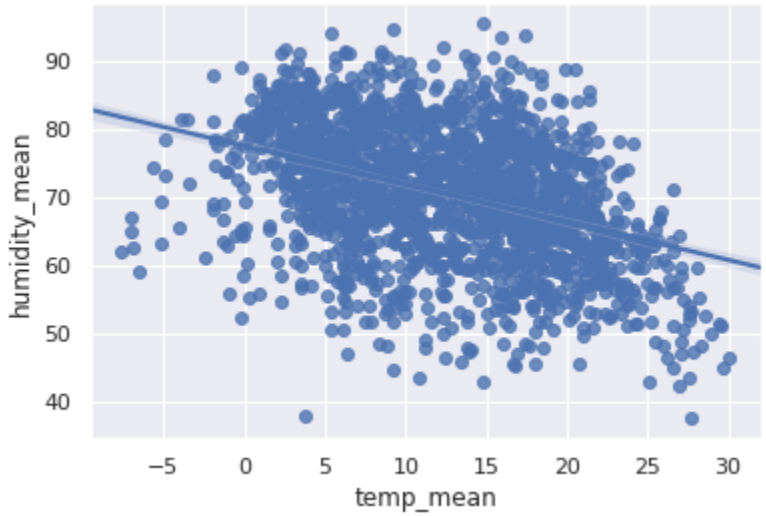
About 47 features were present in the meteorological dataset. To know how each of these features were having an effect on the mean temperature we plotted the scatter plot of these features with respect to temperature and determined the relevance of these for predicting temperature.

Some of these scatter plots are as under:

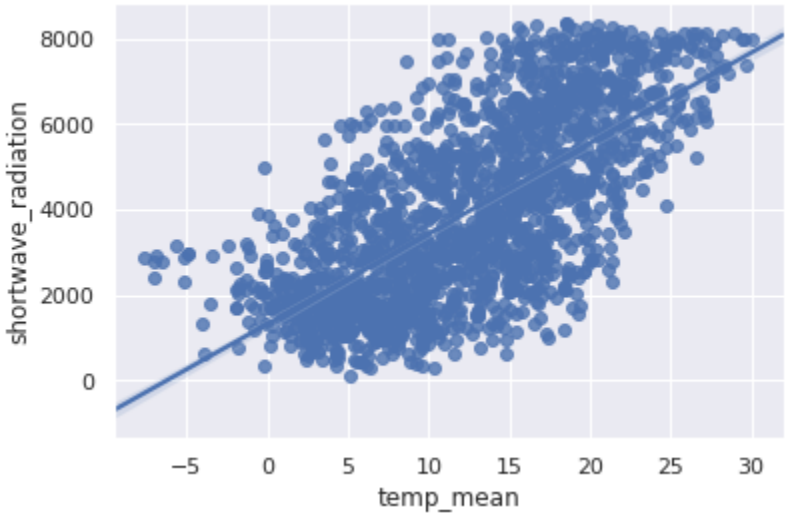
1) MEAN PRESSURE VS MEAN TEMPERATURE



2) MEAN HUMIDITY VS MEAN TEMPERATURE



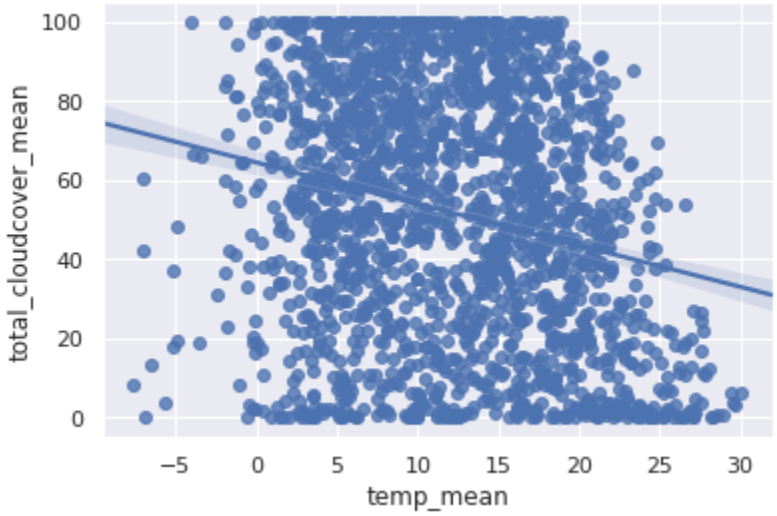
3) SHORTWAVE RADIATION VS MEAN TEMPERATURE



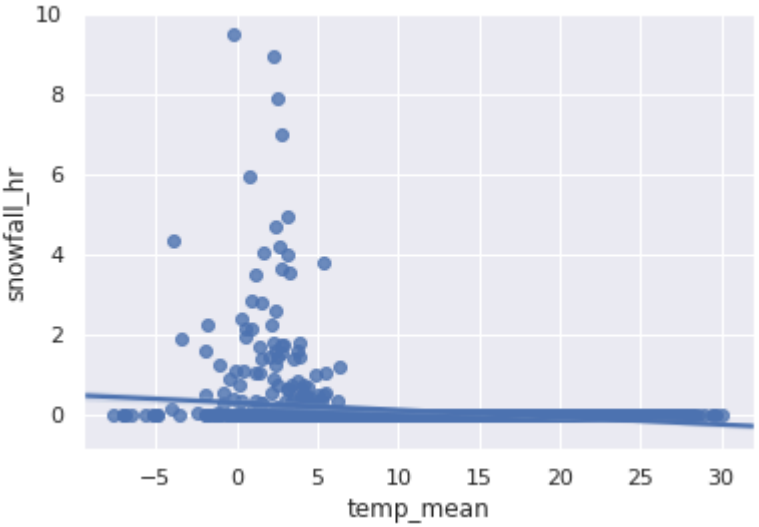
In the above scatter plots, it can be analysed that most of the data-points are coagulated within a given distance around the line hence they're important features. However, we also uncovered some of the features that had no relation with the mean temperature and would eventually add to decreasing the accuracy of the model if they were considered.

Some of them are as under:

1) TOTAL MEAN CLOUD COVER VS MEAN TEMPERATURE

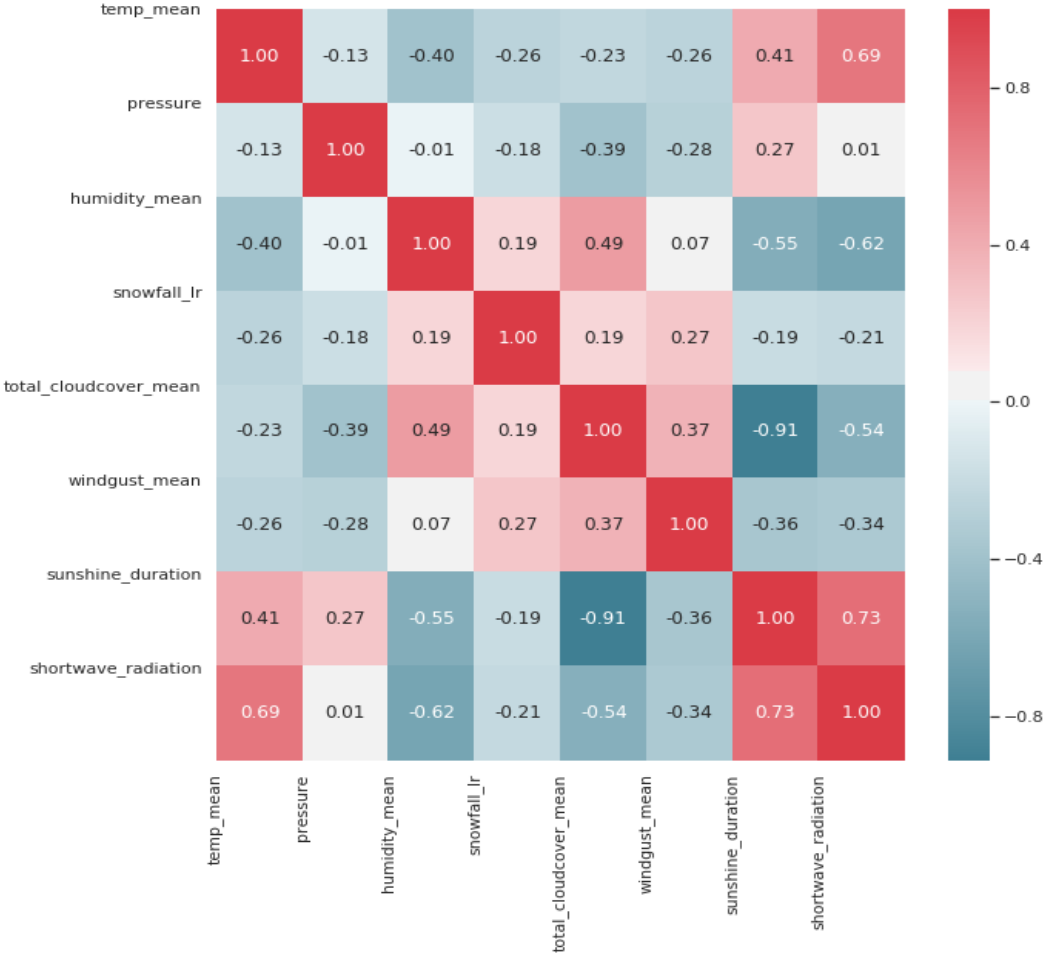


2) HIGH RESOLUTION SNOWFALL VS MEAN TEMPERATURE



CO-RELATION MATRIX

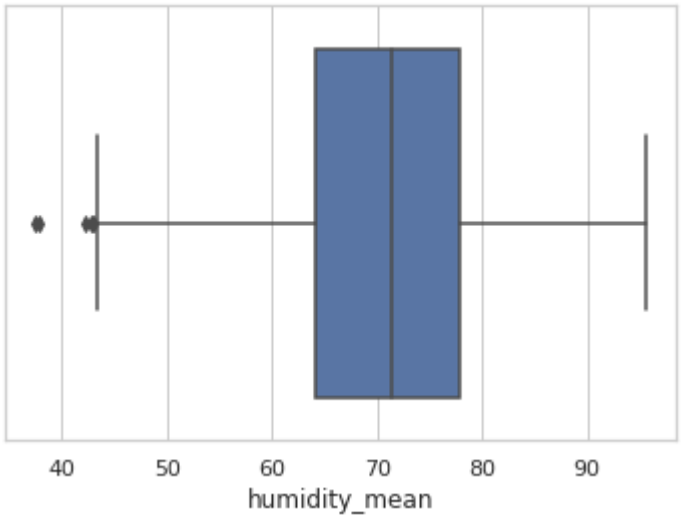
To get more insights about the effect of each of the features we analysed the data by plotting a heatmap of the correlation matrix which we annotated to find out how much exactly each of these features affect each other.
Correlation matrix for prominent features.



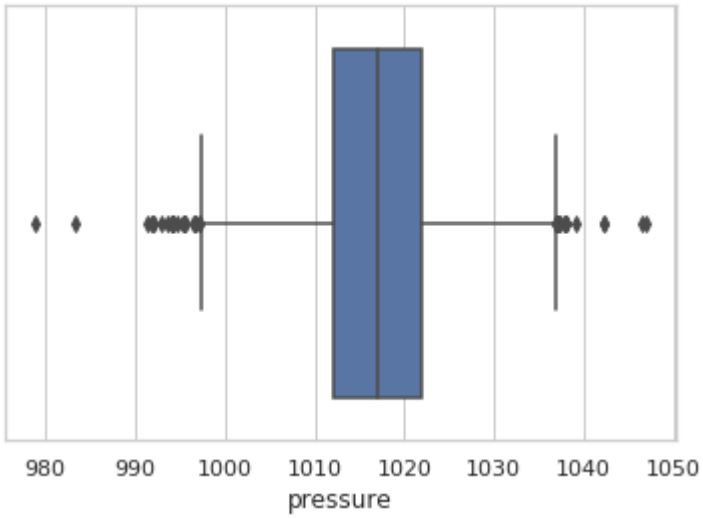
SCALING

Scaling can play a big role in differentiating between a good and bad model and it is necessary to visualize how these features range in terms of their values for this we used box plots.

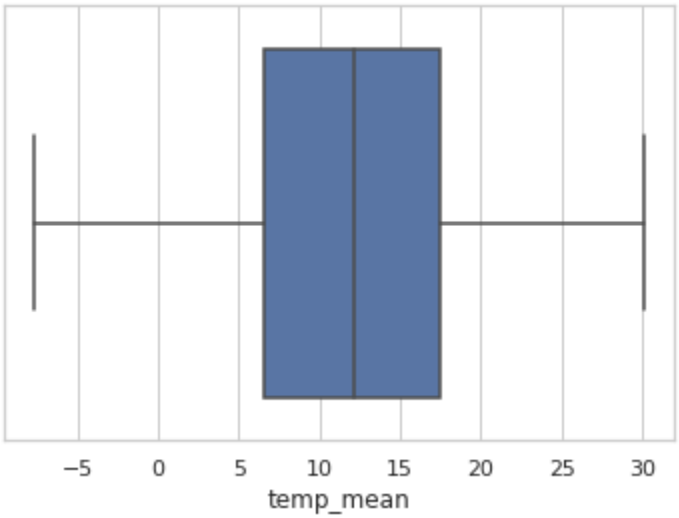
BOX PLOT FOR MEAN HUMIDITY



BOX PLOT FOR MEAN PRESSURE



BOX PLOT FOR MEAN TEMPERATURE



This was very important to determine what we need to scale.

FEATURE ENGINEERING & DATA CLEANING

Data Cleaning: The initial data provided by the meteoblue was poorly formatted and hence we had to convert their data into csv, and re-format names and fix the alignment so that we could it could actually be converted to a proper csv format that could be opened with excel.

Feature Engineering: From the data analysis phase, we were able to pick out the features that played no role in determining the temperature, such as the cloud-based features and days. Hence, we chose to remove all those features. Looking at boxplot graphs of Humidity and Pressure, we scaled them down to bring them both to same scale as other data. Using dew point min and max, we merged them both to dew point mean since both dew point min and max were partly linearly dependent.

MODEL TRAINING & MODEL BUILDING

Model Training

Multiple Models were Trained such as

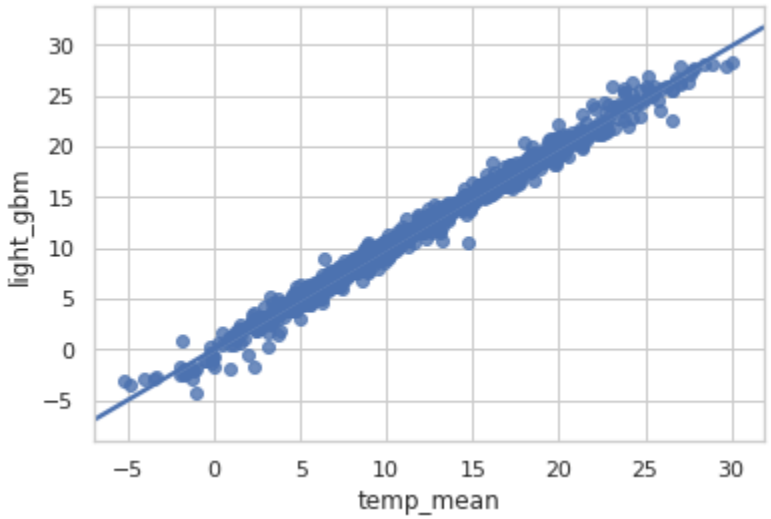
1) *Linear Regression:* - A Solo Linear Regression Model produced a RMSE of 0.53.
2) *Random forest Regressor:* - A Solo Random Forest Regressor Model produced a RMSE of 0.83, but after parameter tuning it reduced to 0.8.
3) *LGBM Model:* - A Solo LGBM Model produced a RMSE of 0.76.
4) *Support Vector Regressor:* - A Solo Support Vector Regressor Model produced a bad result of 6.96.

Model Building

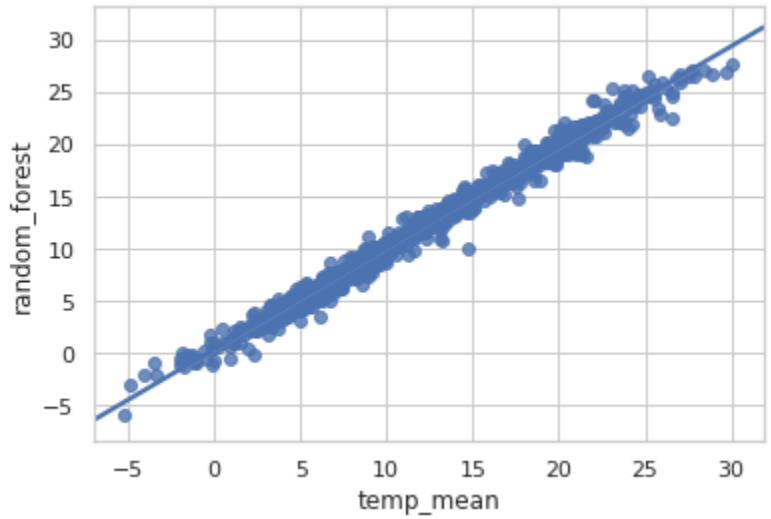
To further improve the possible accuracies, we performed stacking.

1) *Linear Regression + Light GBM Stacked on Linear Regression* – The combination provided us with an accuracy of 0.7 which was worse than either models alone
2) *Linear Regression + Random Forest Stacked on Linear Regression* – This combination provided us with an accuracy of 0.502, which is a good improvement over the previous single models.
3) *Linear Regression + Random Forest Stacked on XG Boost* – This combination provided us with a RMSE of 0.9, which again wasn't an improvement on previous scores.

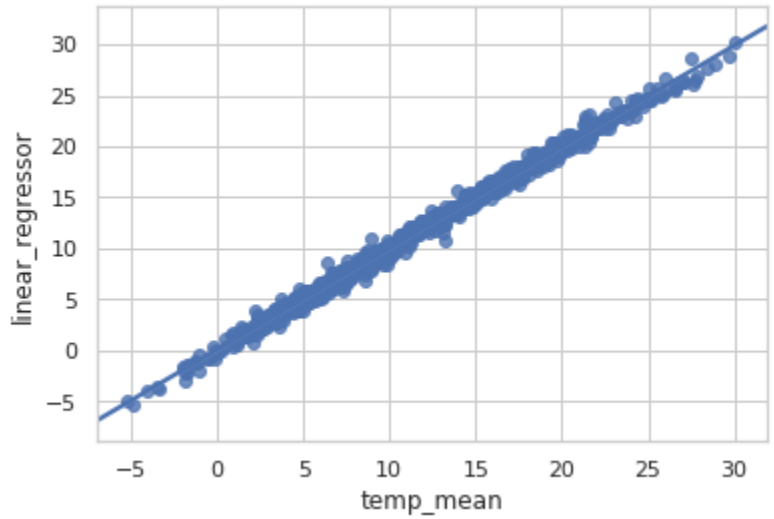
1) LIGHT GBM REGRESSION



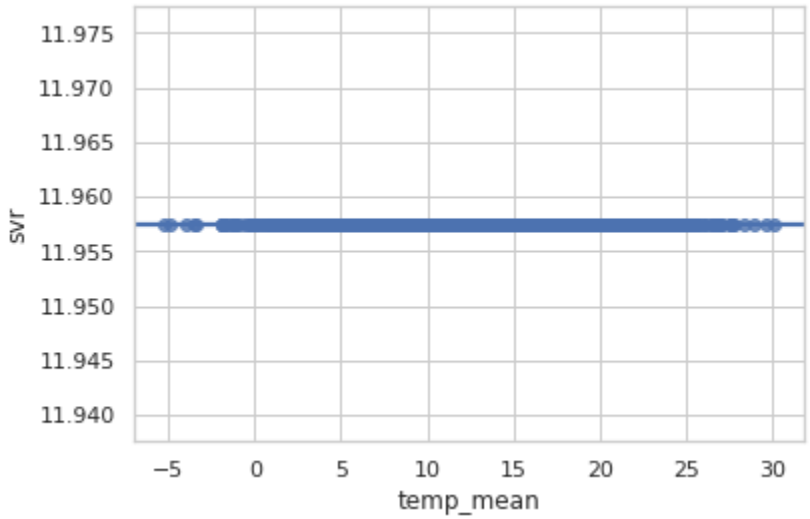
2) RANDOM FOREST REGRESSION



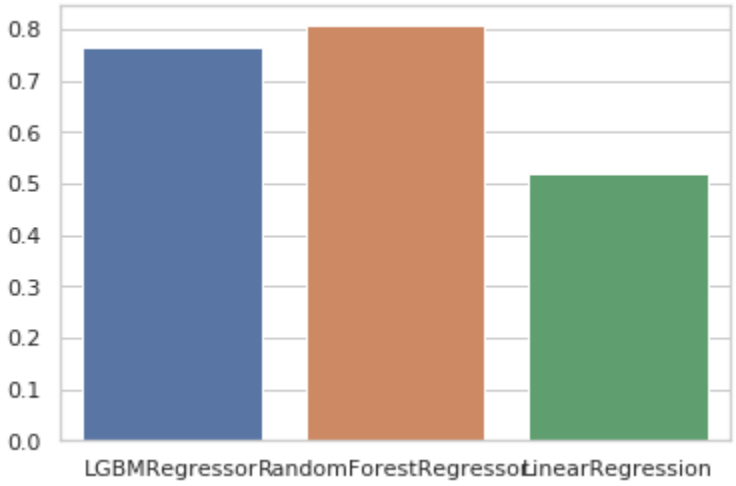
3) LINEAR REGRESSION



4) SUPPORT VECTOR REGRESSOR



RMSE BASED COMPARISION OF DIFFERENT MODELS



VALIDATION AND TEST METRICS

The models used a simple validation by splitting data into train and test split at 50:50 Ratio with 50% used for the first layer of stacking and the 50% used for the final model. We chose to use leave one out validation instead of K-Fold validation since our data set size is quite small, and using K-Fold validation ended up providing us quite high variance and leave one out validation was computationally less expensive allowing us to perform faster testing and training.

The test metrics we chose is RMSE, which represents the simple standard deviation of the predicted values and the observed values. Since, we wanted very low variance from the actual answer we preferred to choose RMSE, since it penalizes quite harshly for outliers and in weather forecast system our main goal is to keep variance as low as possible and avoid outliers.