Prediction of weather for a city for a given period

The weather in US is so variant across the year and across different cities that it is hard to know what weather one might run into in a new city . When people travel between different places in US for business or pleasure, they might be ill prepared for the weather they face if they are not aware of the weather in the destination city. When planning in advance, the historic weather data can provide them with the prediction of weather for a particular city.
Also, people planning for vacation can pick and choose a vacation place based on the prediction of the weather for a given period or based on the given window of the temperature.

Target audience

With this analysis, a weather forecasting company can provide the customers with prediction of weather with the help of the historic data they posses.The customer would know what kind of temperatures to expect in the destination city.

Data to be used

The dataset being used contains historical hourly temperatures for 30 US cities.
Data is acquired from kaggle, originally acquired from using Weather API on the OpenWeatherMap website .
This dataset consists csv file of aroung 45k rows of data with hourly temperature for 30 cities from the year 2012 - 2017.

Methodology

The data being predicted has label assigned to it in the dataset, so this falls under Supervised learning models to predict the temperature for future time period from the available historic temperature dataset. I have tested the data on 2 different models, Random Forest Regression and Gradient Boosting Regression. The metric being used to evaluate the performance of the models is R Squared score and Mean Absolute Error.

Deliverables

The project report with visualization of the data analysis and the source code will be uploaded on the public github repository.

Data Wrangling

The dataset consists of temperatures for each hour for the year 2012- 2017 for 28 US Cities and 2 cities from Canada. The dataset is inspected for invalid entries by checking for NANs and taking care of those by forwarding filling NAN values. Also, I had to delete few days worth of data since it contained NANs for few cities. Since the dataset has 5 years worth of data, deleting few days worth of data will be have a big impact on it. There are few outliers in the dataset, I am assuming they are valid entries and that they are big change in temperatures.

The temperatures are given in Kelvin unit. I have converted it to Farenheit as that is the unit of temperature that we are familiar with. The dataset contained the temperatures for different cities for each hour. I formatted the dataframe to contain temperature for each city at 12 noon. The dataset didn't need much of cleaning since it was a pretty clean dataset with not much wrangling needed.

The code for this can be found at https://github.com/githubnk/SB/blob/master/Capstone_Project/Milestone%20report%20Cap1.ipynb

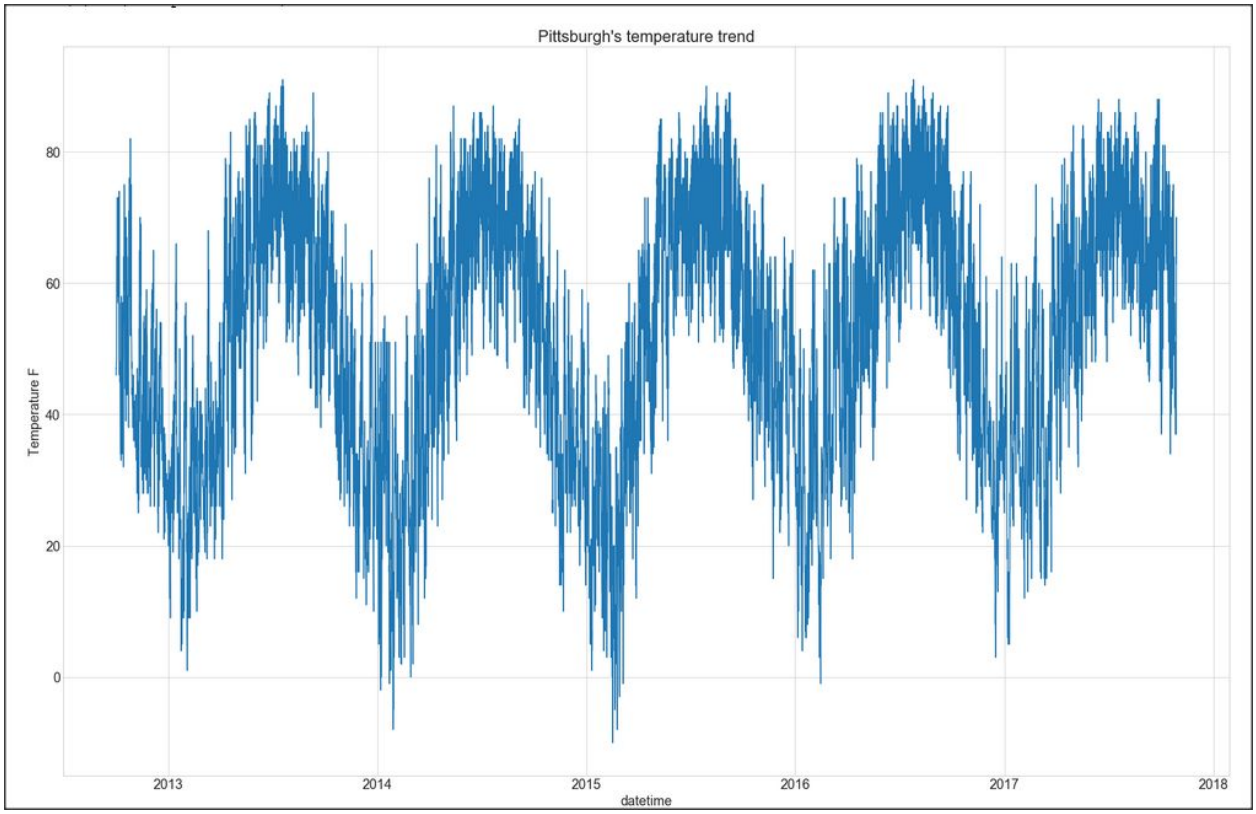
Exploratory Data Analysis

Next step was exploring the data to check if there were any trends, any cities which were correlated, exploration of outliers. I explored the dataset to see the trend of the weather in the cities, Vancouver, Pittsburgh and noted the winter months and summer months.

Exploring the temperature trend in Pittsburgh

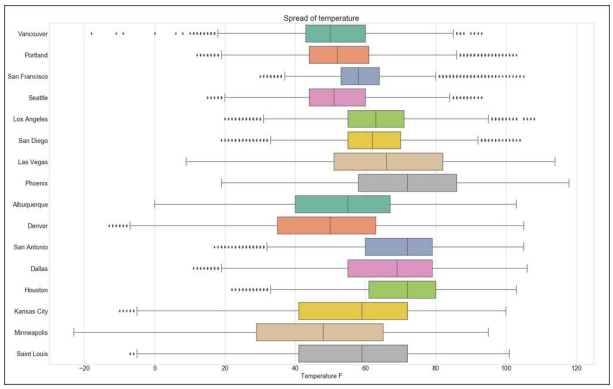
Pittsburgh's temperature trend as seen below has winter months from November to March, temperature starts going up in mid April , summer months typically being from May to

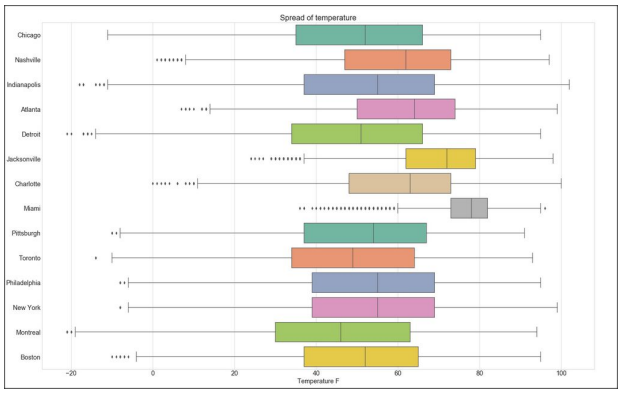
September. February of 2015 has been the coldest in the above data set. February 2015 temperatures have been the second lowest temperatures from 1948.



Spread of temperature

Also, checked the spread of temperature across all cities.



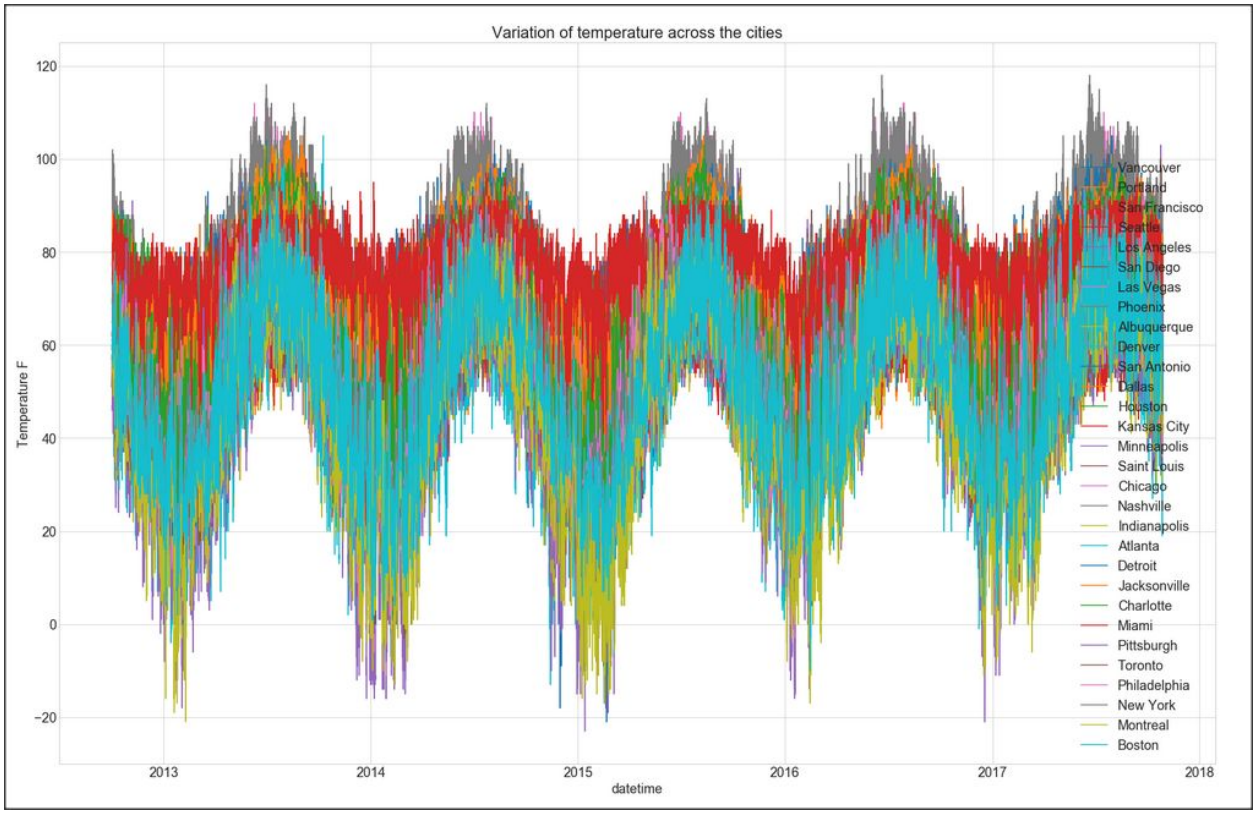


From above plots, we can figure out that the majority of the days the temperature in Vancouver is between 38F and 65F, Chicago is between 33F and 69F and so on. There are few outliers too in cities like Vancouver, Detroit, Toronto.

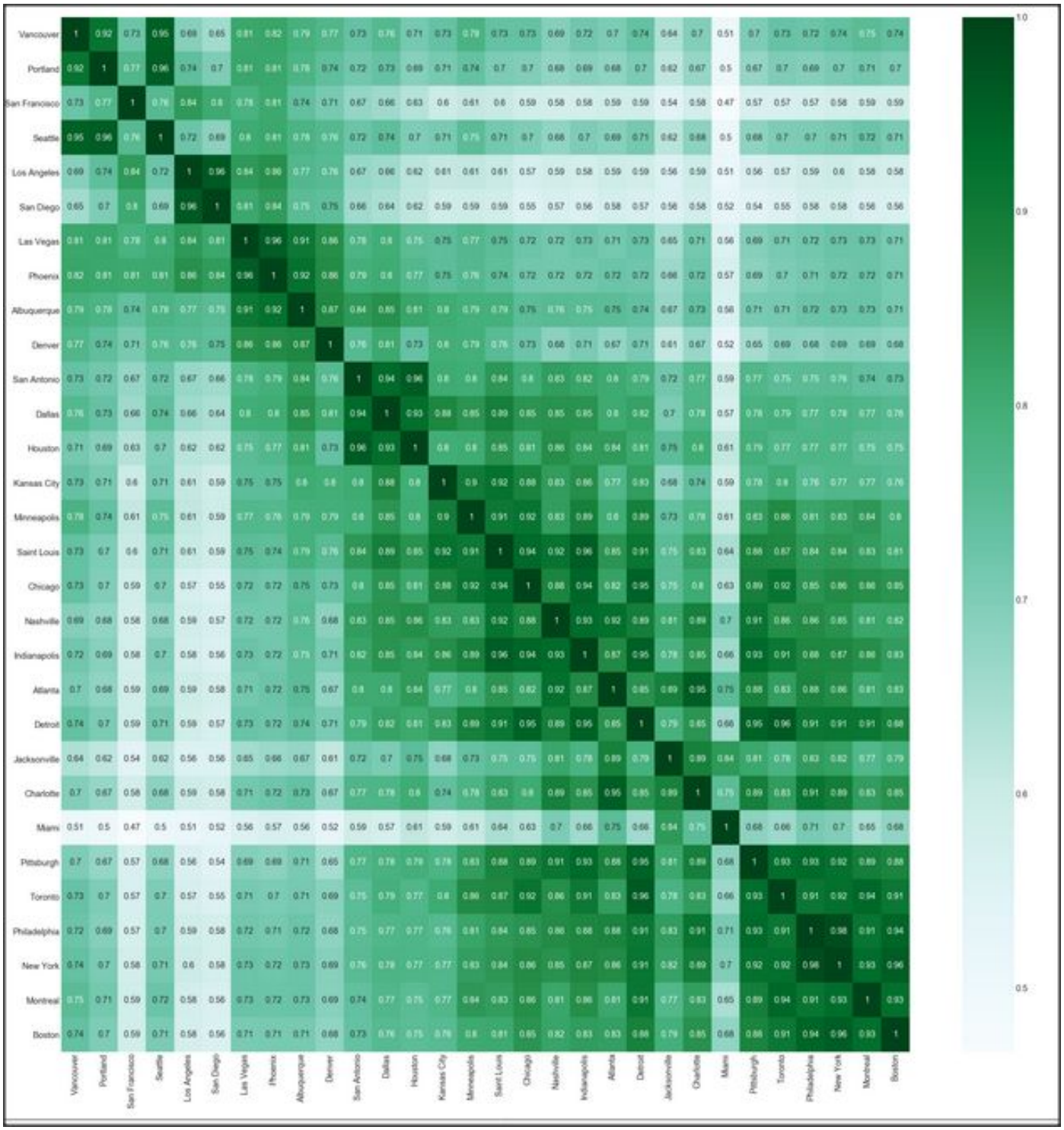
Top 3 warmest cities and top 3 coldest cities

The three warmest cities in the given dataset are Phoenix, Las Vegas and Dallas and it can be seen temperatures reaching max during the month of July.

The three coldest cities are Minneapolis, Montreal, Detriot. Coldest temperatures are seen in the month of January and December. Vancouver has knocked down Minneapolies to second coldest in the year 2014, which can be attributed to the cold front on Nov 30th with the temperature varying between -18F and -9F between Nov 30 2014 and Dec 02 2014



Exploring the correlation between the cities



None of the regions are negatively correlated which implies that the cities experience winter and summer around the same time. The level of correlation is different among the cities.

There are strongly related cities like

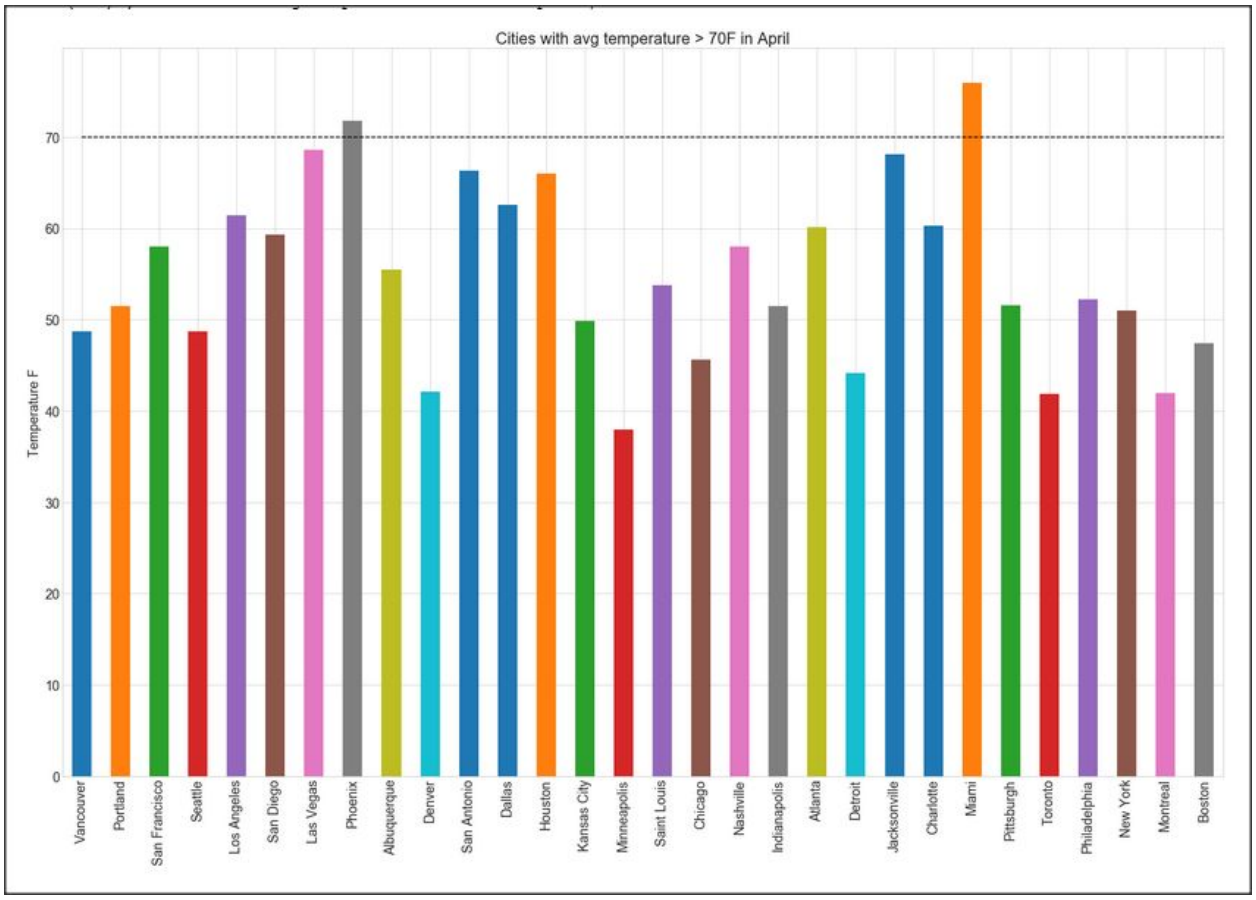
San Antonio, Dallas, Houston in the Southern region

Detriot, Pittsburgh, Toronta, Philadelphia, New York, Montreal, Boston in the NorthEast region

San Francisco, San Diego, Las Vegas, Phoenix, Albuquerque in Western Region

San Antonio, Dallas, Houston are strongly correlated with temperatures increasing and decreasing in the same way.
Portland, Seattle, Vancouver in the NorthWestern Region are strongly correlated

Which cities have Average 70F or more in the month of April, during the time of spring break?



From the graph below, we can see the cities which has average of 70F or more in the month of April - Phoenix, Miami

Modelling the data

I chose to model the data on Linear Regression, Random Forest Regression and Gradient Boosting Regression Models and check which model performs the best. The data set containing the temperatures of 30 cities is being modeled to predicted the weather of Vancouver which is part of the given data set. The data is wrangled and ready to be used for modeling. We will split the data into training set and train different models on it and test the model on the testing test. We will split the data set into 80% training and 20% testing data set.

For the times series data, appropriate models to choose would be time series models like ARIMA. But I want explore and learn about Linear Regression and so am trying out Linear Regression.

Performance Metrics

For Performance evaluation of the model, we will use the metric R^2(R squared) score and Mean Absolute Error(MAE).

R^2 is the statistical measure of how close the data aer to the fitted line.
The R^2 is defined as (1 - u/v) where

u is the residual sum of squares ((y_true - y_pred) ^ 2).sum() and
v is the total sum of squares ((y_true - y_true.mean()) ^ 2).sum().

R^2 score for better model will be close to 1. A model with R^2 score further away from 1 is considered bad model. R^2 is the percentage of variations in dependent variable which are explained by the independent variables present in the model.

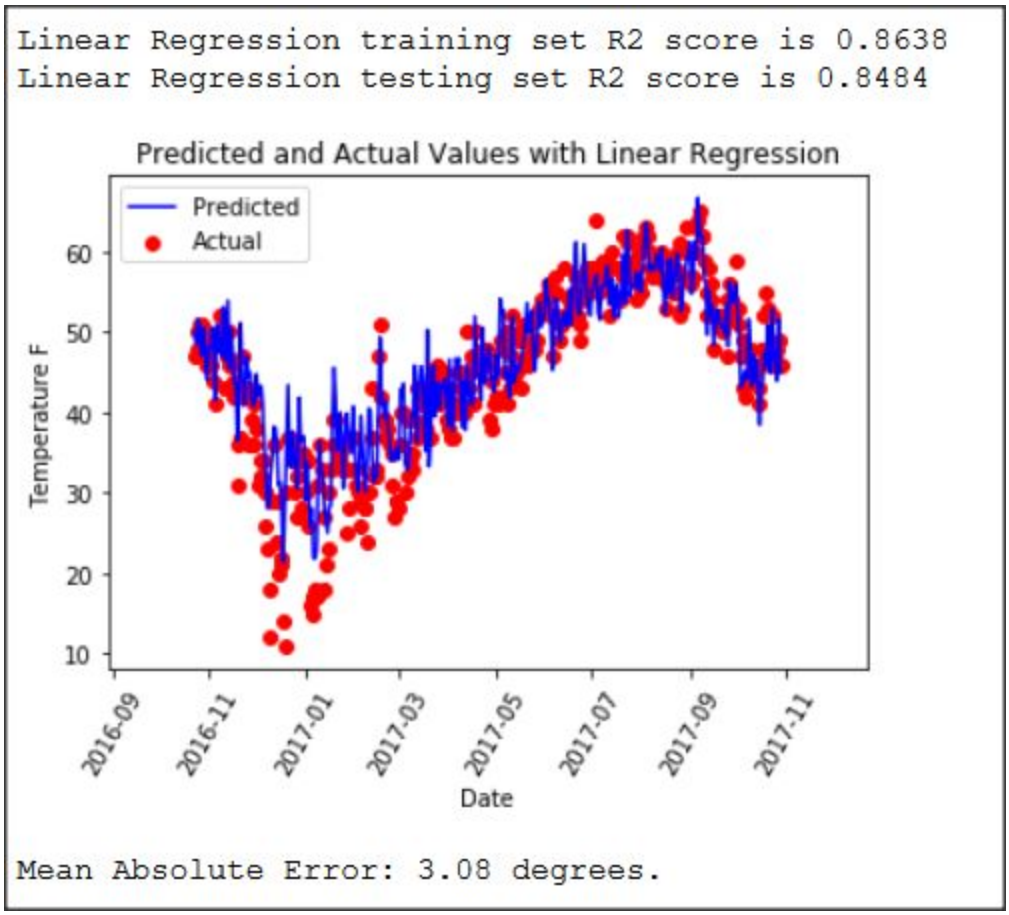
The Mean Absolute Error (MAE) is the sum of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions are. Large MAE suggests the model might be bad, small MAE suggests the model is good at predicting. MAE of 0 means it's a perfect predictor of the dependent variable output but it almost nevers happens.

What algorithm to chose

We are working with the data which is labeled and use that labeled data to predict the outcome which is called Supervised Learning. The data we are trying to predict,the temperature of a given city is continuous variable, hence Regression Models are to be picked. Also, the data not Linear. So, I want to model the data on Non Linear Models - Random Forest Regressor and Gradient Boosting Regressor. Just for curiosity, I tried to model on Linear Regression.

Linear regression

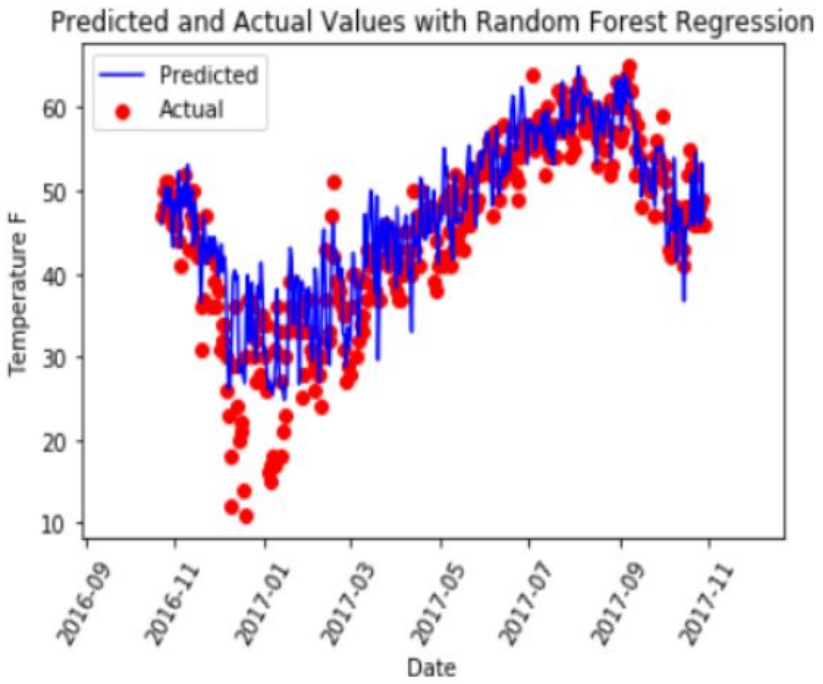
The results obtained when the data is modeled on Linear Regresion model is given below



Random Forest Regression

Random Forest Regression is a method of ensemble learning, that operates by constructing a multitude of decision trees at training time and outputting mean prediction of the individual trees.

```
Random Forest Regressor score on training set is  0.9785
Random Forest Regressor score on testing set is 0.8338
```



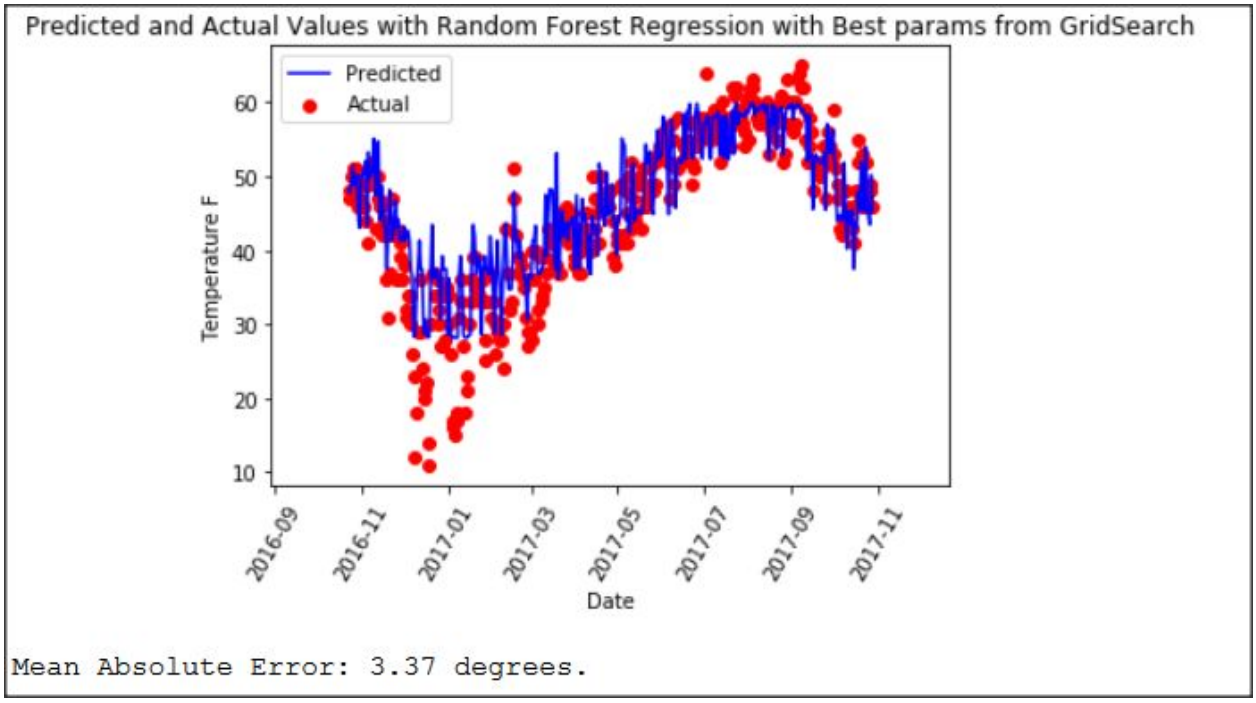

```
Mean Absolute Error: 3.27 degrees.
```

There is big difference in R^2 scores for training and testing data. This means that training data is being overfitted, the model is being trained very well by the training data but when the model is tried on new set of data, which can be testing data, it is not performing as good as on training set. We have to improve the model so that overfitting does not occur.
We can consider this as base model and it's performance metrics are measured by Mean Absolute Error and R^2 score. We will try to improve the performance by tuning the hyperparameters. This can be achieved through GridSearchCV where in a search will be performed for different set of parameters and returns the best set.

Random Forest with GridSearchCV, tuning the Hyperparameters

The best params obtained by GridSearchCV are

{'bootstrap': True, 'max_depth': 3, 'max_features': 30, 'min_samples_leaf': 5, 'min_samples_split': 8, 'n_estimators': 50}

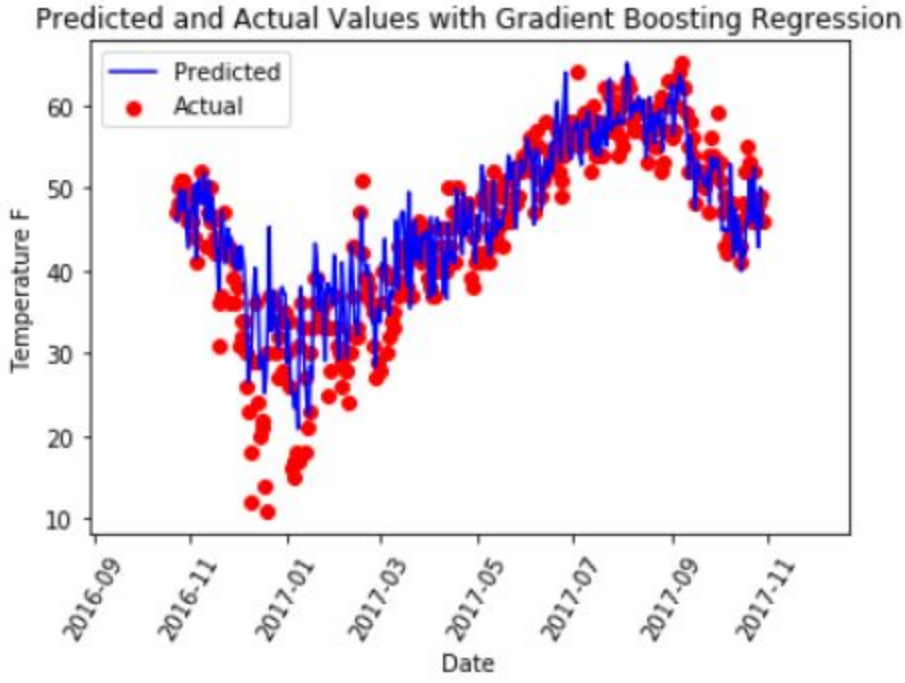


Compared to the base model, the difference in R^2 scores of training data and test data has reduced which means training data is not overfitted.

Gradient Boosting Regression

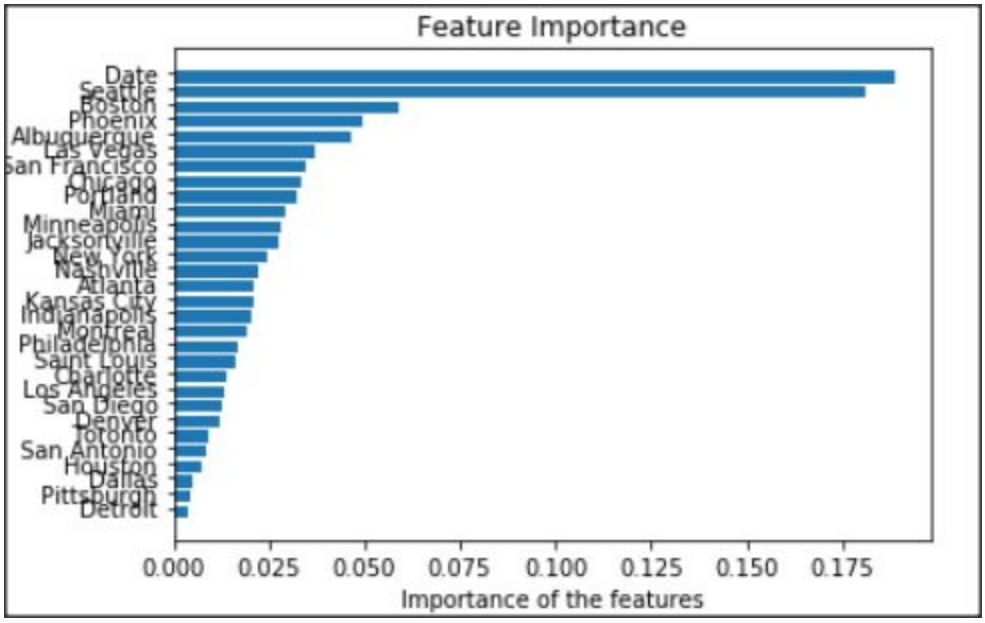
Gradient Boosting Regression uses the boosting technique, combining a number of weak learners to form a strong learner. Regression trees used as a base learner, each subsequent tree in series is built on the errors calculated by the previous tree.

```
Gradient boosting Regressor score on training set is  0.9421
Gradient boosting Regressor score on testing set is  0.8609
```

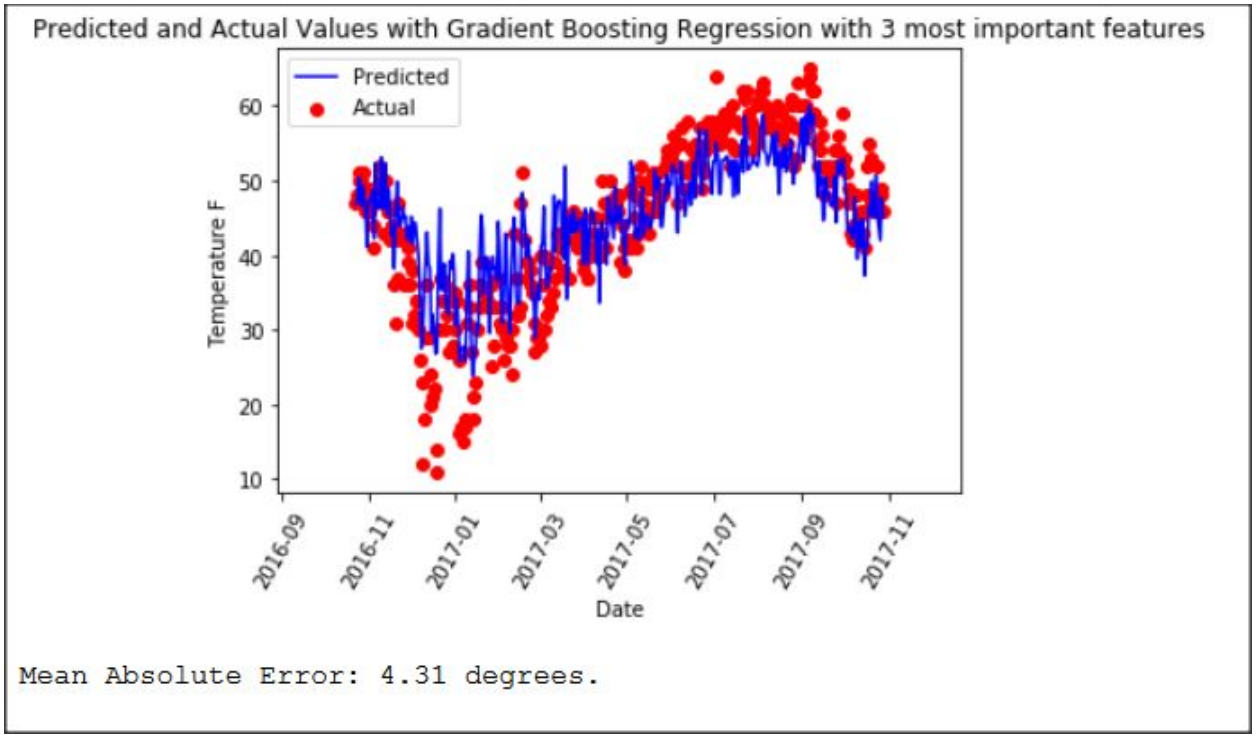



```
Mean Absolute Error: 2.94 degrees.
```

Gradient Boosting Regressor is performing better than Random Forest Regressor by having decrease in Mean Absolute Error but there is overfitting of data. We will check if it can be improved by considering only the important features in the feature vector. We obtain the top 3 important features from the feature vector of the model and use that as the X variable.



As is seen above, 'Date', 'Seattle', 'Boston' are the top features when the target variable is 'Vancouver'. So, we will setup the X with those values.



Model with top 3 important features is not performing as well as the base model. It goes to say that the other features are playing a role in predicting the dependent variable even though their importance is not much.

Conclusion

Considering all the models we tried, The Random Forest Regressor with best params seems to give the best results with Mean Absolute Error of 3.37% and good fit of training and test data.

Given above is the prediction for the city, Vancouver. Other city's prediction can be done similarly by setting up the target variable with required city and setting up the X variable.