Predict PM 2.5 Index and Design Web Application
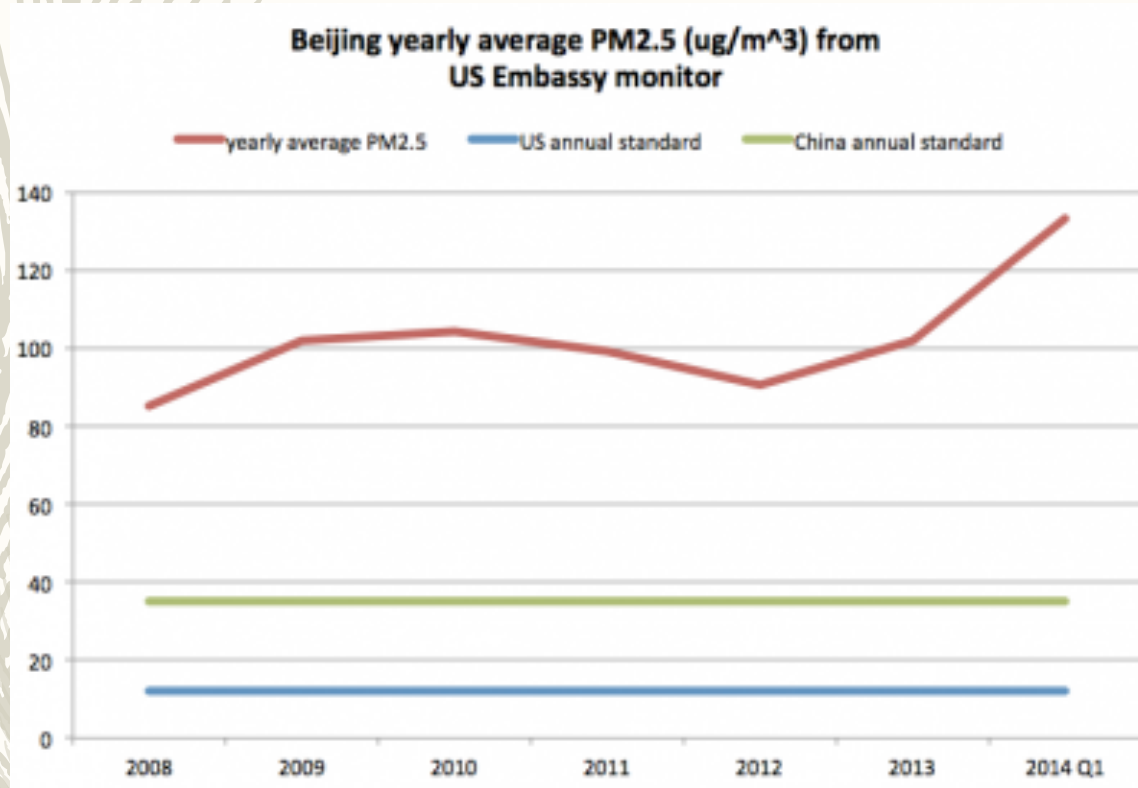
Ruofan Wang

Lu Yin

Shenghui Zhou

Air pollution in China

Beijing yearly average PM2.5 (ug/m^3) from US Embassy monitor

Introduction > Motivations and Goals > Data Visualization and Analysis > Web Design and Application > Conclusion

Motivation and Goal

– Observes the PM 2.5 data from the online resources and then fits the data into the several models to predict future PM 2.5 index. We choose the model based on the MSE and RMSE of the models

– Designs a web application and interface by using Python in order to let users to plan their activities wisely.
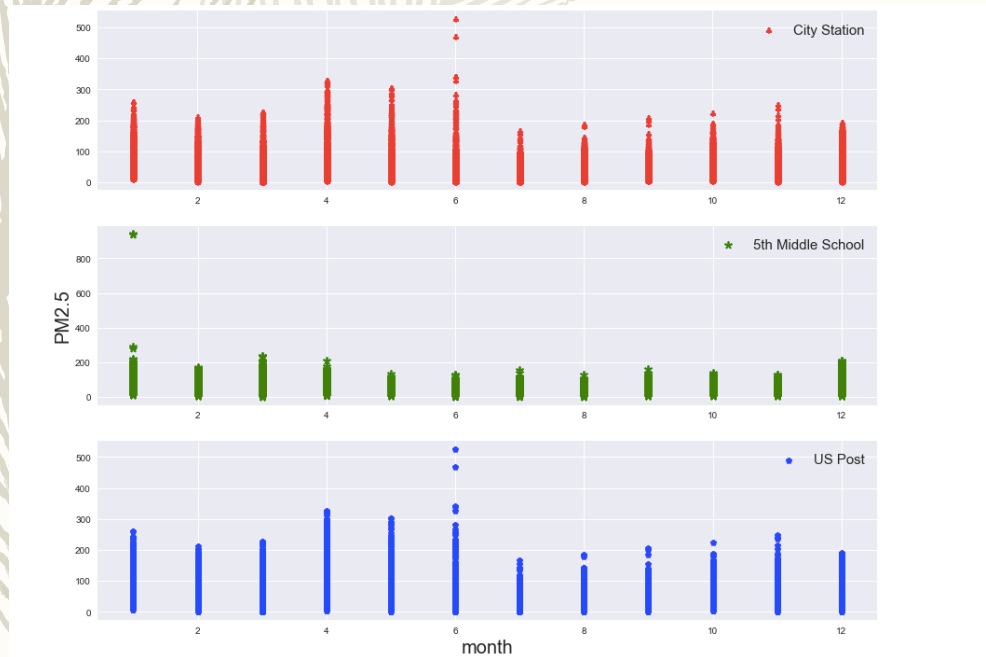
Data Cleaning and Preparation

- We initially plot three figures that showing the PM2.5 distribution in three stations, which are the city station,5th Middle School and US Post within the five years.

- Since the location of 5th Middle School shares the most popularity and data with less missing, we decide to make prediction that focus on the data collection from this particular area.

Data Visualization and Analysis

- Multilinear Regression

- Decision tree and Random Forest
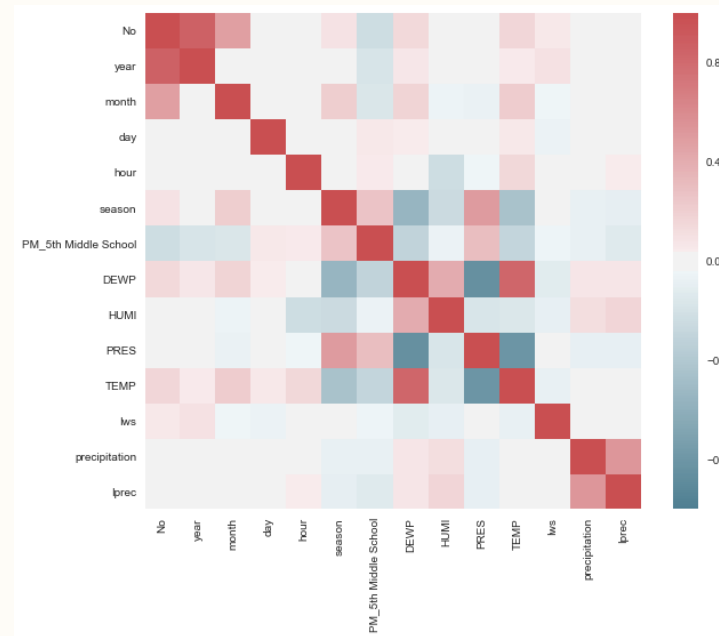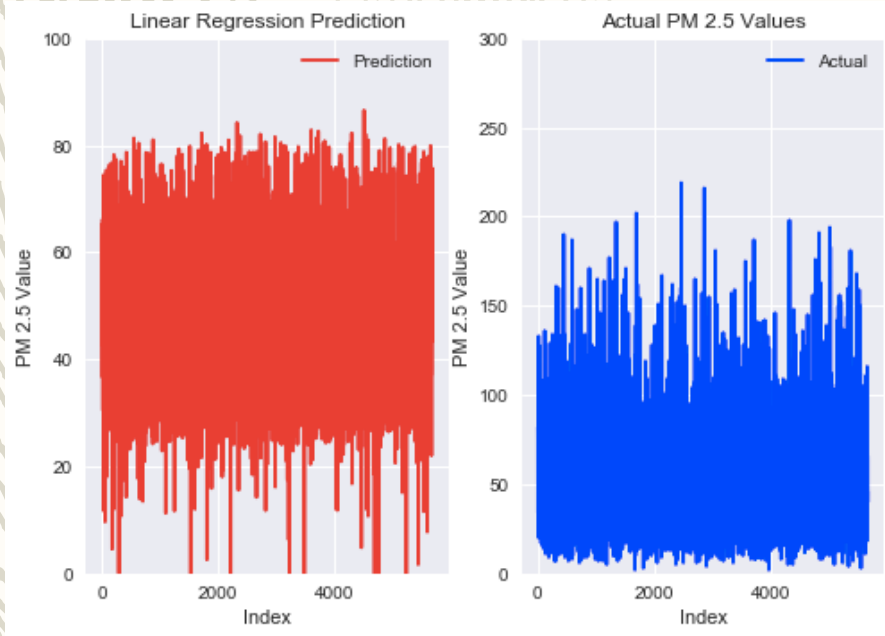
- XGBoost Regressor

- ARIMA model

Multilinear Regression

– We choose MLR model as our baseline model.

Tree-based models

– Decision Tree

– Random Forest

– XGBoost Regressor

ARIMA model
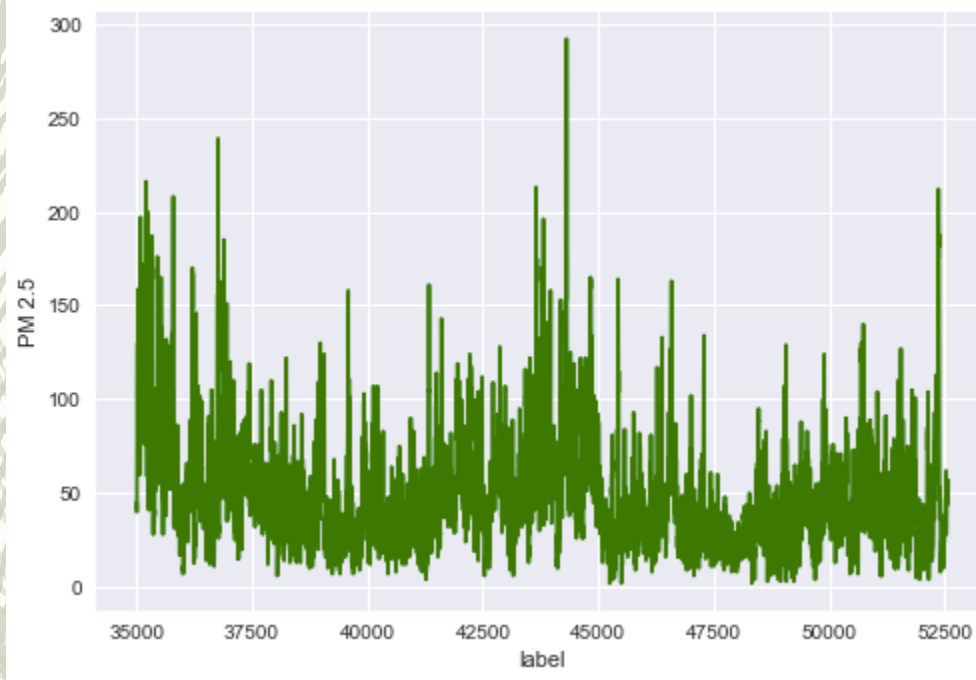
– From the ADF test result, we can see the p-value is 3e−223e−22, which is strong enough to reject the original hypothesis. Also, the value of this test is −11.9−11.9, which is smaller than the three threshold in the levels of1%,10%,5%1%,10%,5%. So we can assume that our data is stationary.

ARIMA model

– According to the AIC results, we select ARIMA(1,0,1)*(0,1,1,12) as our model.

```
==================================================================
              coef    std err        z      P>|z|    [0.025    0.975]
------------------------------------------------------------------
ar.L1       0.9386     0.002    534.112    0.000     0.935     0.942
ma.L1       0.0852     0.004     22.670    0.000     0.078     0.093
ma.S.L12   -1.0168     0.001   -690.111    0.000    -1.020    -1.014
sigma2     65.4380     0.355    184.155    0.000    64.742    66.135
==================================================================
```
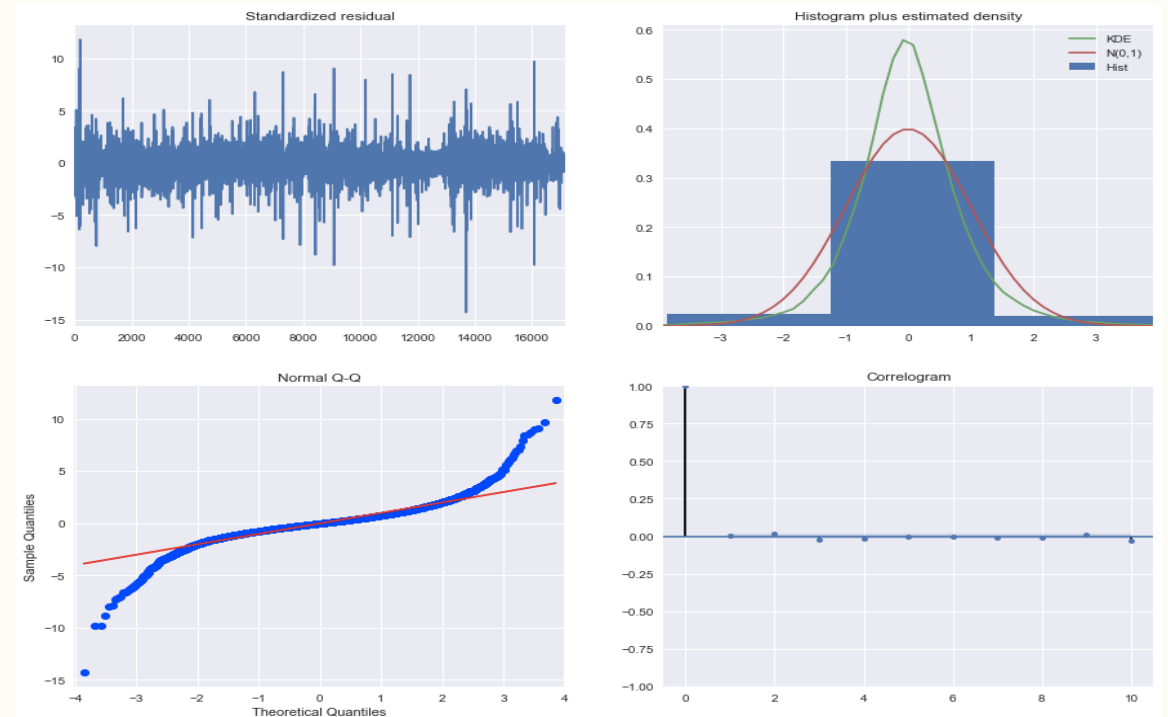

Introduction | Motivations and Goals | Data Visualization and Analysis | Web Design and Application | Conclusion

Web Design and Application

– As the final goal of this project, our team designs a web application to show the predictions of PM 2.5 to potential customers. More importantly, we implement the website with different labels to realize our business goals.

– Our team could combine our results with the real-world business problems and help people manage their indoor and outdoor activities accordingly .

Introduction › Motivations and Goals › Data Visualization and Analysis › Web Design and Application › Conclusion

Web Design and Application

– If the PM2.5 Index is higher than 100, it turns out to be red. Index higher than 100 means the outdoor activities could be harmful for human bodies. Therefore, the button would direct the users to those sellers who offer healthy products to clean air and absorb harmful PM2.5.

– On the other hand, the green button means index of PM 2.5 is lower than 100 and the outdoor activities are recommended. Then, this button will direct the users to workout applications and websites for outdoor activities.

PM2.5 In Guang Zhou

Tomorrow	2 Days Later	3 Days Later
0	159	116
2013-12-29	2013-12-30	2013-12-31
Outdoor	Indoor	Indoor

Copyright © 2017. NYU CDS DS-GA-1007.

| Introduction | Motivations and Goals | Data Visualization and Analysis | Web Design and Application | Conclusion |

Conclusion

– Our team explores from Multilinear regression to tree-based models and then use the ARIMA model to fit the time lagged data. Based on the evaluation of MSE for all those models, we choose the XGBoost regression as our the optimal model.

– Getting the good prediction result is not our final goal. As a group of ambitious graduate students, we try to make the business value out of our predictions. By designing the Website, we direct the users to different websites based on different PM 2.5 predictions and realize the business opportunity.