Nitrogen dioxide prediction with XGBoost and Regression models

(Stuttgart, Germany)

Ruprecht-Karls-Universität Heidelberg

Institute of Geography

Seminar: Geographic Applications of Machine Learning

Teacher: Benjamin Herfort, Lukas Winiwarter, Michael Schultz

Julian Käflein Anna Buch

Matriculation number: 2932816 Matriculation number: 3464541

j.kaeflein@stud.uni-heidelberg.de a.buch@stud.uni-heidelberg.de

# Introduction

**Aim:**

**Predict the particulate matter in the atmosphere close to frequently used main streets in the Stuttgart valley basin. How precise is our algorithm to the real measured concentration of nitrogen dioxide?**

**Introduction**

**Why important?**

Air pollution is currently a major task for municipality (Stadtverwaltung) of bigger cities in europe. Especially in rush hour the concentration is frequently overshot. Due to the breach of the EU policy (EU-Richtlinie 2008/50/EG) the Eu strives for proceedings against multiple german cities eg. Stuttgart, which would lead soon to monetary fines for the communes and the state. High values of air pollution affects also the human health in long and short-term, concerned are mainly residents who live or work close to such streets.

The overstepped air pollution values are highly spatial and temporal located, limited to small areas close to main streets and manly during rush hour. A good example is since the public discussion the well known Neckartor in Stuttgart. So a regional and dense grid of measuring stations is needed to generate a lot of high-quality input data to apply machine learning methods. The topographical position of Stuttgart within a valley basin leads in the combination of weather conditions to those overshot particulate matter concentration. In this context inversions, also called stationary temperature inversions, are common phenomenon in Stuttgart valley basin and are highly positive correlated to air pollution. So a closer look to those weather conditions and their parameters like wind direction and speed, precipitation and further meteorological parameters, would be interesting and part of the regression analysis.

# Data presentation

**Data description,**

**feature survey**

**metadata analysis (fitness for use)**

# Choice of appropriate Machine Learning Methods

**Explanation of Methods- which + why are chosen**

The following subsection gives a short overview of three selected machine learning algorithms, their operation and performance.

In this study two types of *multiple decision-trees models* are described; the more suitable model is later used in practise. The third model is a Regression model and was chosen with regard to the noise in the dataset and the assumption that none of the applied algorithms will include sufficiently all relevant input features. In this case an adequate prediction of the target wouldn’t be possible, so we decided to simplify the prediction to a binary classification problem, using Logistic Regression. The target of this regression algorithm is to classify if the predicted values are over or under the limit of 40 μg/m³.

The validation results of the Logistic Regression were compared with the validation of binary values derived from the XGBoost predicted target.

Data Preparation for Binary Classification (XGB ~ LogisticReg)

The target variable are the real measured hourly NO2 values. To compare the discrete prediction values of the XGBoost model with the binary predictions of the Logistic Model, the predicted target from XGBoost were categorized based on the NO2 limit. Those now binary organized target values derived from XGBoost and the ones from the Logistic Regression were validated and their results compared by using metrics like accurary, precision, recall and F1 score.

Data Preparation for validation of XGB ~ RidgeReg (discrete values)

Furthermore the discrete target variable of XGBoost was compared with the prediction based on a Ridge Regression Model. The performance of both algorithms was validated by error measurements. A forecast of both model predictions is shown in figure xxx.

**Random Forest**

While growing trees Random Forest uses a random subset of features to find the best feature of the subset to split the node. In comparison, other decision models use the best feature of the whole sample. Each tree grows parallel and separate from the other trees. In the end the decisions of each tree are used to generate the final decision. The random forest “..results in a greater tree diversity, which (once again) trades a higher bias for a lower variance, generally yielding an overall better model” (Géron 2019 , p. 199).

The best split of a branch can be determined by Gini-index or cross entropy (r2d3. website). Of consequence is that the properties on which a decision is based, is comprehensible (white box) and less complex than in other machine learning methods like neural networks. Still random forest models usually are not used for forecasts caused by the long-lasting training process when huge input datasets were used. Also the high amount of computational resources while creating parallel in time multiple trees, leads to the decision to use another more suitable model for NO2 prediction. The chosen model is faster than random forest and the tree are growing sequentially which less overloads the computational power.

**XGBoost – ensemble of Decision trees**

**Paper :** [**https://arxiv.org/abs/1603.02754**](https://arxiv.org/abs/1603.02754)Chen & Guestrin 2016

Buch . Géron 2019 Hands on ML

A very successful and widely used Gradient Boosting is XGBoost, an Extreme Gradient Boosting which is based on collections of decision trees. It was developed by Chen xxx year and is available by its corresponding python library. Benefits of XGBoost are:

* scalability in each processing steps/scenarios
* portability
* short running time

Like other Gradient Boosting methods it’s also possible to combine XGBoost with other cost functions (via the loss hyperparameter?-check) (Géron 2019, p. 210). Combinations with XGboost are often used successfully in ML competitions likewise the Netflix prize competition in 2009.

Similar to well-established decision tree models XGBoost is built up on boundary values, tree pruining and use, likewise random forest, column sampling to achieve a shorter run time and to prevent overfitting (Chen & Guestrin 2016, p. 9).

XGBoost is known for well performance on decision-tree-predictions based on large and sparse datasets. The small timesteps (hourly steps), which are used in this work, leads to a large input dataset of up to xxxxxx observations during the two-year study time (1year=8760 hours). Due to may devective input data with possible outliers, XGBoost seems to be a good choice for predictions of NO2 concentration.

* Good: highly scalable to larger datasets, optimized for extremely efficient computational performance, and handles sparse data with a novel approach.
* XGboost higher scalability (Chen & Guestrin 2016)

XGboost prevents overfitting via a regularized model (Chen & Guestrin 2016, S. 7). In the so called regularized objective the predictions from all (corresponding) leaves of each tree are calculated. These predictions are sum up (depending on their leaf weights) to the final prediction (Chen & Guestrin 2016, p. 2f)

XGBoost combines … from random forest and gradient boosting, resulting in lower prediction error. The Gradient Boosting takes smaller learning rate (slower teps) by doing the predictions sequentially.

* Boosting takes slower steps, making predictors sequentially instead of independently. ... By combining the advantages from both **random forest** and gradient boosting, **XGBoost** gave the a prediction error ten times lower **than** boosting or **random forest** in my case. (

https://liuyanguu.github.io/post/2018/07/09/extreme-gradient-boosting-xgboost-better-than-random-forest-or-gradient-boosting/#random-forest

Figure XGBoost.multi.trees.plot

***Softmax Regression***

*Instead of the closely related logistic regression the Softmax Regression can be use with more than two features/variables or classes. Based on multinomial Regression it is also called Multinomial Logistic or Maximum Entropy Classifier. Due to the property of Softmax Regression for multi-class classifications it is also commonly used in advanced machine learning algorithms like network sense. In this work Softmax Regression is used for predictions on how much each single feature (wind, rain..) influences the target variable, the hourly PM concentration.*

* *🡪 multiple calssi defined for different features 🡪 AIM: maximise predicions on all the clasisfiers , helps if input feature should belong to each of the classes in bild*
  + *🡪 Extense log reg to softmax reg*
* *One of the benefits of a simple model like softmax is that we can visualize the weights for each of the classes, and see what it prefers – check (https://medium.com/@awjuliani/simple-softmax-in-python-tutorial-d6b4c4ed5c16)*

**Logistic Regression**

This regression type is based on the operating mode of a linear regression, but instead of outputting a discrete model result like in a decision tree model or Ridge Regression, a Logistic Model generates a binary classification of its prediction.In this study the model calculates, if a prediction is above or lower than the nitrogen dioxide boundary of 40 μg/m³.

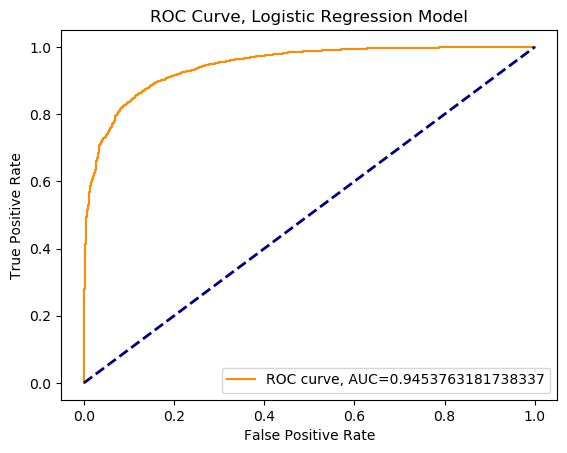
The steps how a model prediction is labeled to a particular class, are explained in the following. First the loss function of a logistic regression (logistic loss) calculates the weighted sum of the input features including the bias like a Linear Regression does. But instead of showing the model result directly, the estimated probability for each instance is displayed. The estimated probabilities (logistics) of the model results are a *Sigmoid Function*, which assigns to each instance one of both classes (Géron 2019, p. 144f).

The probability of the classification for each instance can be displayed via the sklearn function predict\_proba().

*This boundary value is a short term exposure value, the EU policy allows, beneath further exceptions, a 18 time overstep of this limit during a year. To simplify the calculations just the short term exposure value is used (*[*https://www.umweltbundesamt.de/themen/luft/luftschadstoffe-im-ueberblick/stickstoffoxide/stickstoffdioxid-gesundheitliche-bedeutung-von#grenzwerte-stickstoffdioxid*](https://www.umweltbundesamt.de/themen/luft/luftschadstoffe-im-ueberblick/stickstoffoxide/stickstoffdioxid-gesundheitliche-bedeutung-von#grenzwerte-stickstoffdioxid)

# Choice of hyperparameters

For a first look into the performance of the logistic model a *Receiver* *Operating Characteristic* curve (ROC) is plotted. Like expected the curve shows in the beginning a high recall ratio, while the error ratio stays low. With shifting the decision threshold the rate of recall to error ration changes more and more. The Area under the curve (AUC) is a parameter to determine the predictive accuracy of the model. A value of 94.5% is quite well.

****

**Standardization**

Another possibility to improve the validity of the model is to standardize the input features, to archive a unit variance of all features. Regularized models react sensitive to different scales of the input features. By Standardization the mean value is subtracted and divided by the standard deviation. To use the same transformation later on the testing set the StandardScaler was firstly used on the training dataset and with respect to use the features of the testing dataset (<https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler>)

* If you scale each feature to have the same standard deviation (i.e. divide each column of the model matrix by its standard deviation), then the magnitude (absolute value) of the coefficients is a direct measure of the importance of the corresponding features. See Schielzeth 2010 Methods in Ecology and Evolution "Simple means to improve the interpretability of regression coefficients" <https://doi.org/10.1111/j.2041-210X.2010.00012.x>.

(stack overflow: <https://stats.stackexchange.com/questions/423469/how-to-measure-the-impact-influence-of-a-feature-y-on-logistic-regression-mode> )

**PCA**  
The Principal Component Algorithm is a common method to prevent overfitting, extract noise and also to minimize large dataset by removing nonrelevant information. It keeps the relevant data and declassify the rest of the dataset as noise. The input features, except the target, are summarized and reduced to newly transformed principal components.

To find the most suitable number of components we set, instead of a concrete number, a ratio of variance of 95 % (p. 225). Those preserved components were used as new input features in the models of the Logistic- and Ridge Regression. Before the transformation into new feature space the training and testing sets were standardized, the target variables of both sets remained unchanged.

**Further hyperparameter tuning**

The regularization strength of the logistic regression model is controlled by the parameter C. It is the inverse of the regularization parameter, so a lower value of C induces a stronger regularization. In this case the regularization term will minimize the weights in the cost function. To find the best value for C a cross-validation estimator already implemented in the sklearn LogisticRegressionCV() was chosen. The best predictions could be archived by a low regularization of C = 2. Alternatively the C parameter an be chosen manually or by GridSearch()

To avoid overfitting of the Logistic Regression model some further parameters in sklearn LogisticRegression() function can be added. So the number of iterations was set as default to 100, to stop training

A *l2 regularization* term (Ridge Regression) was set to the cost function to shrink the maximum size of the feature coefficients. The regularization is used in combination with the solver ‘sag’ (*Stochastic Average Gradient descent*). It is employed for large datasets and quick computation, but the data needs to be scaled before. While the common *Lasso regularization* does a kind of feature selection while setting the coefficients of less important features to 0, the *l2 regularization* also penalize the coefficients of very large coefficients.

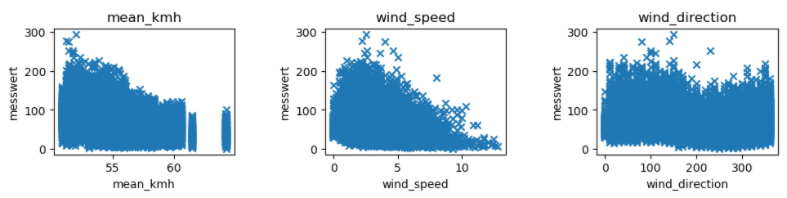
In general regularization techniques try to find the best balance between bias and variance of a dataset. Ridge Regression is usually applied for data with multicollinearity; a high correlation between multiple features. Additional the normalization of the loss function can also remove noise from the data, here the *l2 regularization* is used primary to minimize overfitting of the *logistic loss* (loss function of logistic regression). (*https://www.bogotobogo.com/python/scikit-learn/scikit-learn\_logistic\_regression.php*)

**Ridge Regression**

At last a regularized Linear Model with l2 regularization was chosen for comparison with the predicted discrete values of the XGBoost model. The scatterplots during data preparation show that multiple features *show a collinearity to each ot*her. Like Linear Regressions the least squares method is used, but least squares are minimized by the Rdige Regression. Like earlier in the logistic Regression the weigths are shrinked, so an overfit of the model can be prevent.

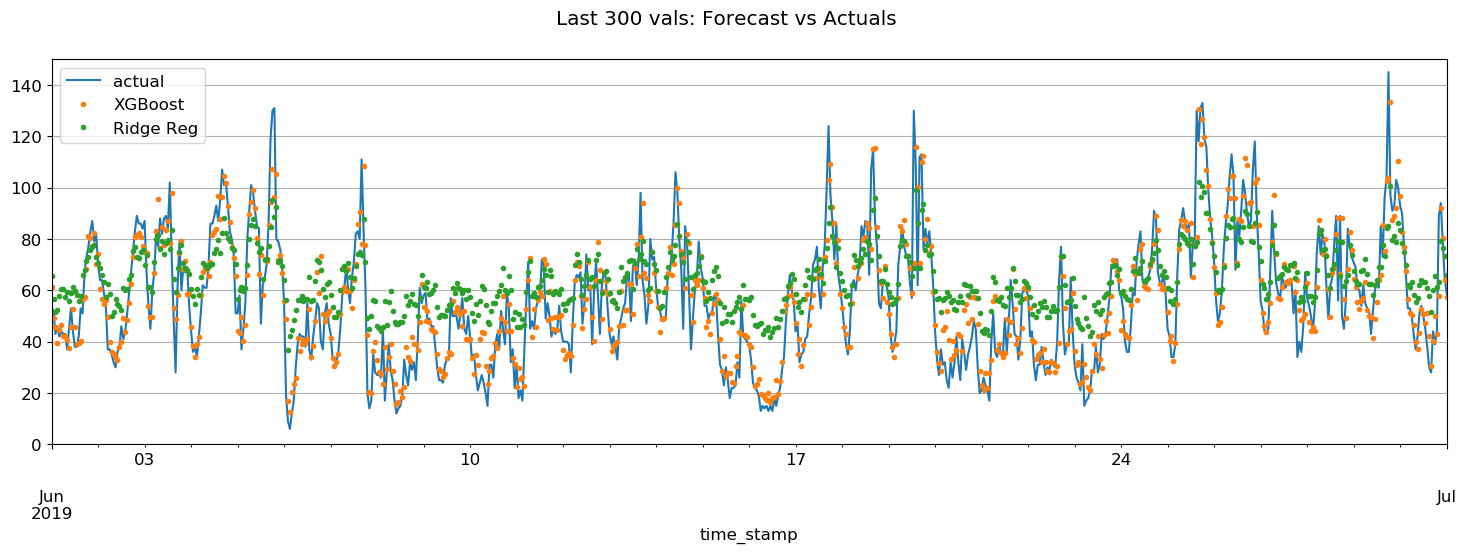
For multiple regression data with multicollinearity, variances are large-> so may far from true value

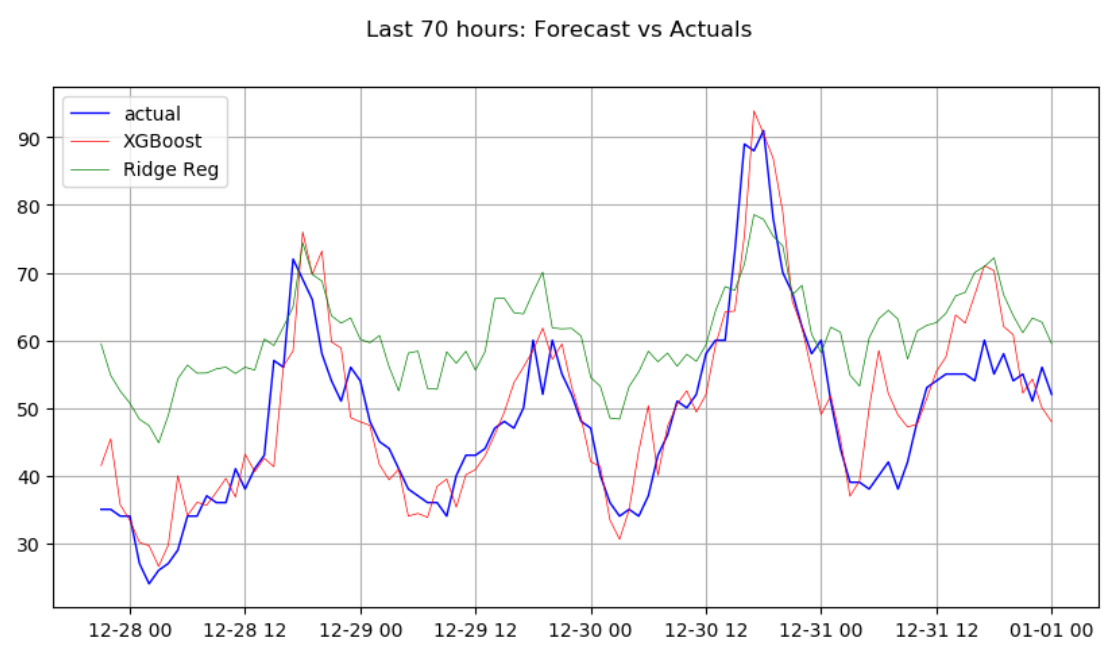
* Regularized linear model
* To reduce standard errors -> Adding degree of bias to regression estimates
* **Multikollinearität** ist ein Problem der [Regressionsanalyse](https://de.wikipedia.org/wiki/Regressionsanalyse) und liegt vor, wenn zwei oder mehr erklärende Variablen eine sehr starke [Korrelation](https://de.wikipedia.org/wiki/Korrelation) miteinander haben. Zum einen wird mit zunehmender Multikollinearität das Verfahren zur Schätzung der Regressionskoeffizienten instabil und Aussagen zur [Schätzung](https://de.wikipedia.org/wiki/Sch%C3%A4tzung) der Regressionskoeffizienten zunehmend ungenau (wiki)



# Survey of most relevant model results, their comparison and interpretation

For quantitative evaluation of the model performance several metrics were used as validation parameters. To evaluate the discrete predictions of hourly NO2 concentration the regression metrics Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R², Coefficient of determination) were applied; - accuracy, precision, recall and F1-Score for the classification task. A combination of the classification metrics is useful due to their limited explanatory power by individual view.





A first qualitative overview via the timeseries of NO2 concentration predicted by XGBoost and Ridge Regression indicates significantly a higher reliability of the XGBoost model (orange) to the actual measured data (blue), than the Ridge Regression Model (green) does. A closer look to a shorter 4-day-period (fig xxx) displays that peaks and minima are well explained by the XGBoost model. Also the Regression Model reproduces the curve progression, but it is less sensitive to lower and higher extreme values. The 4-day-timeseries leads to the assumption that the Regression Model frequently overestimates the reference values, this can be closer analysed by comparing the parameter values of MSE, MAE, R².

**Target reached?**

**Which method is the best**

All three models show similar parameter values with a F1 Score always around 90 %, this can be interpreted as a well performance of all three models. The parameter of F1 Score is more robust against an unequal amount of false positive to false negative that’s why it is here used for a rough estimation.

Tab xxx: Predictions of the measurement category, above or lower than the NO2 limit value of 40 u/m³.

|  |  |  |  |
| --- | --- | --- | --- |
|  | XGBoost | Logistic Regression | Ridge Regression |
| Accuracy | 88.8 % | 87.5 % | 84.2 % |
| Precision | 90.2 % | 88.5 % | 86.3 % |
| Recall | 93.2 % | 93.2 % | 90.5 % |
| F1 Score | 91.6 % | 90.8 % | 88.3 % |

The validation parameters should be seen in combination with their weights of the input features. *The feature importance of the logistic regression and the XGBoost model show both a strong overfitting through the high relevance of especially the first feature.?? ÜBERPRÜF QUELLE*

It is not possible directly to compare coefficients from different models, but between variations of a single model e.g. by changing one hyperparameter. In this case a dimensionality reduction was applied at the logistic regression model. The used PCA algorithm leads to a reduction from 287 to 91 features, by preserving a ratio of variance of 95 %. Performing the logistic regression after dimensionality reduction improved the principal components to coefficient values lower or equal to 0.90. In comparison without PCA the most important feature had a coefficient of 3.4 with a wide gap to less important features. The values of the validation parameters decreased marginally caused by the limitation of the features.

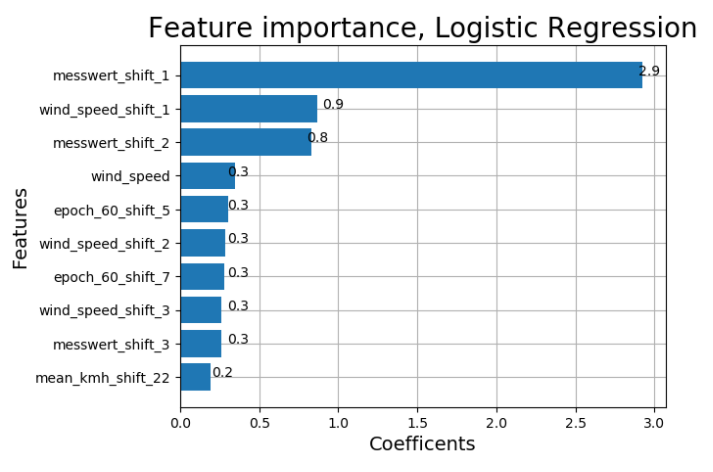


Fig : Coefficients of the ten most useful features of the logistic regressin model without using a PCA.

# Discussion of model result

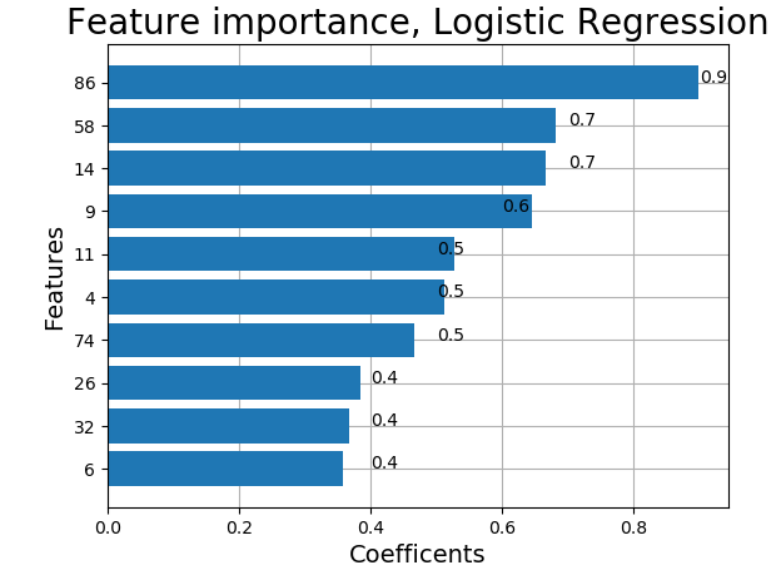
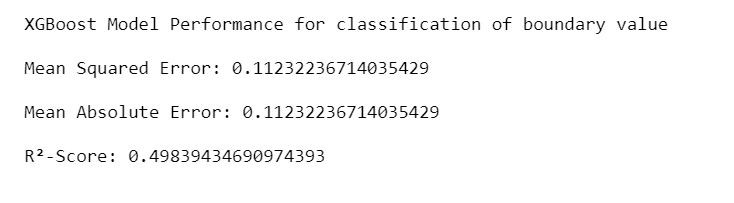
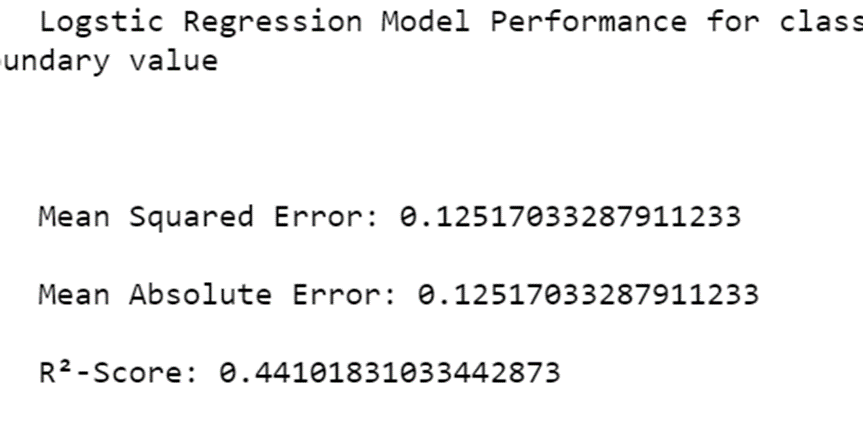


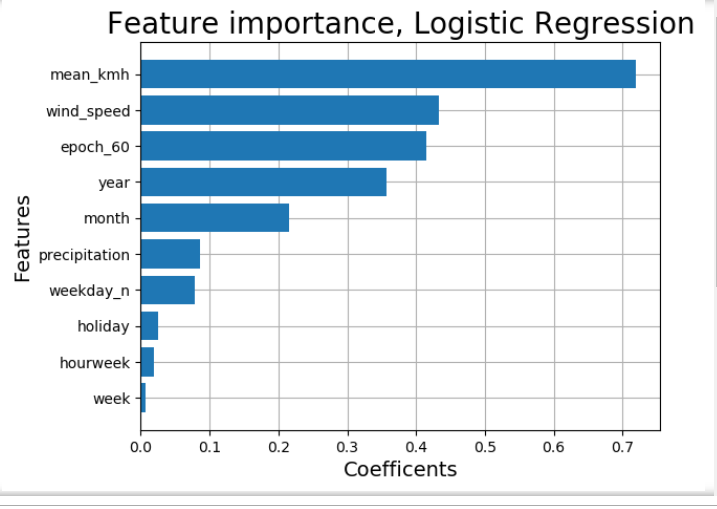
Fig : Coefficients of the ten most useful features of the logistic regressin model after using a PCA. Except the PCA all other hyperparameters and the dataset remained unchanged.

Ridge Reg= 0.95 ration

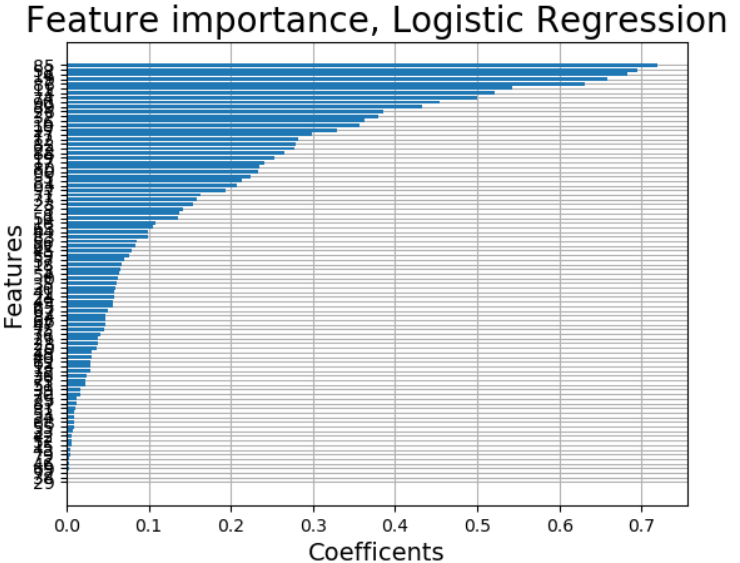
Tab xxx: Model predictions of hourly NO2 concentration.

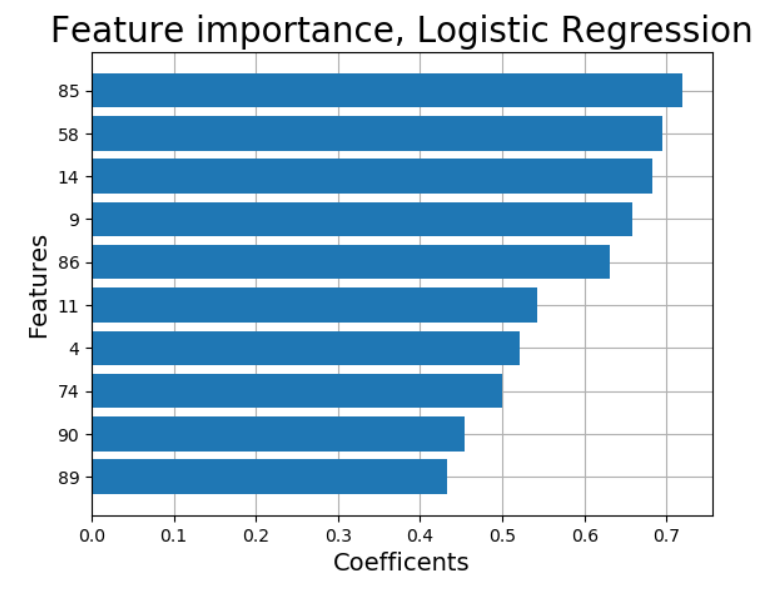
|  |  |  |
| --- | --- | --- |
|  | XGBoost | Ridge Regression |
| MSE | 32.2 | 164.3 |
| MAE | 4.1 | 9.9 |
| R²-Score | 0.82 | 0.7 |

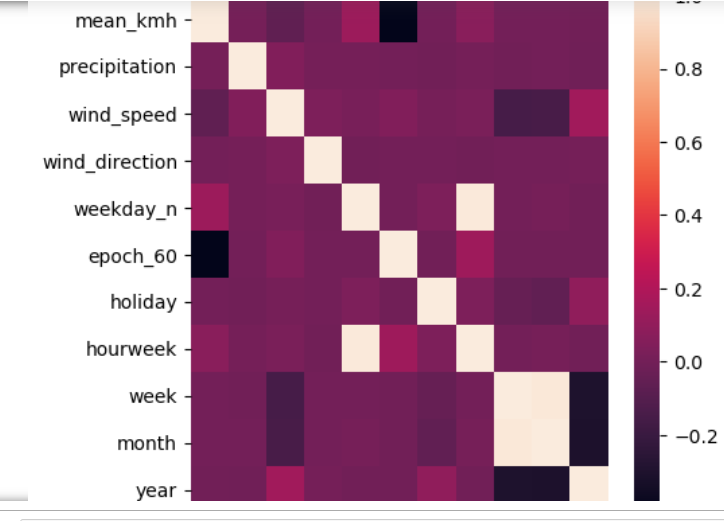


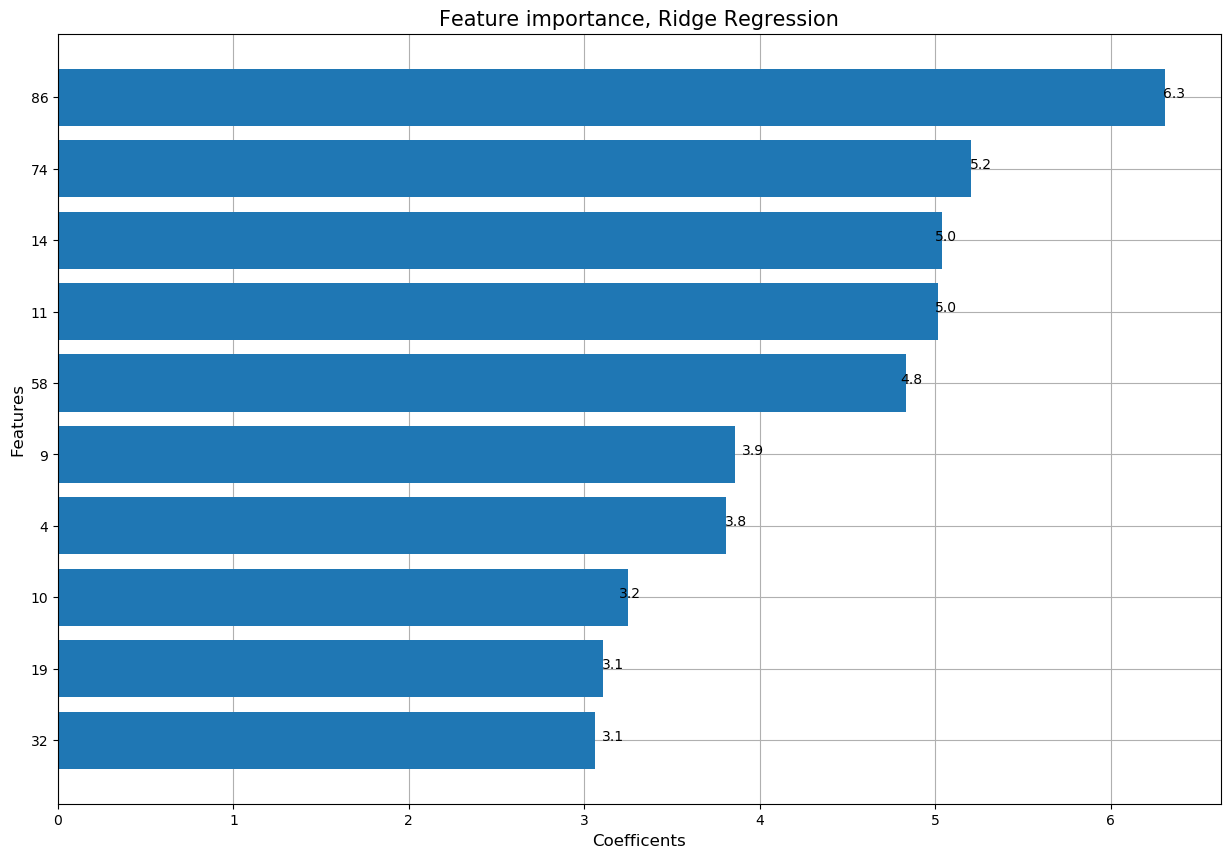
**PCA = 95 % ratio of variance that reserves**

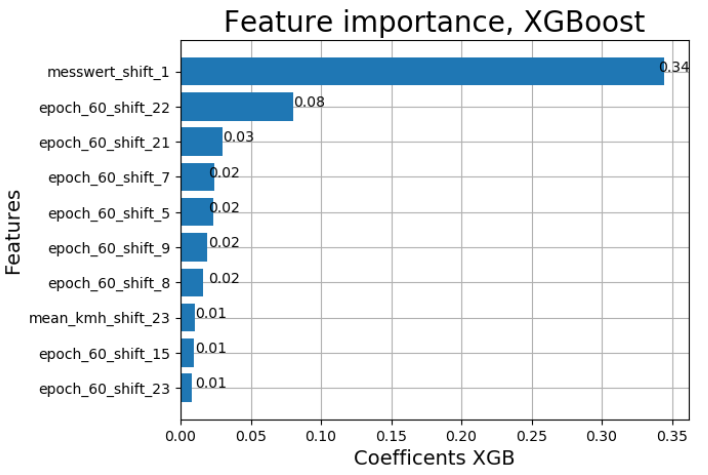
**2 Plots (all features, and top 10 features)**

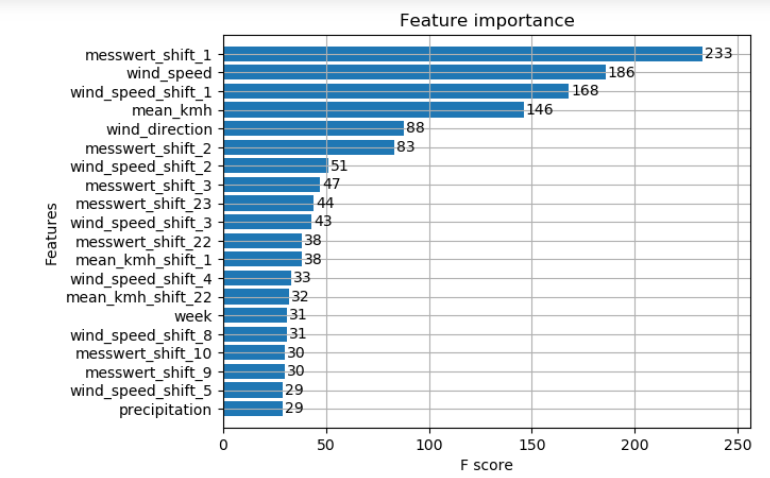


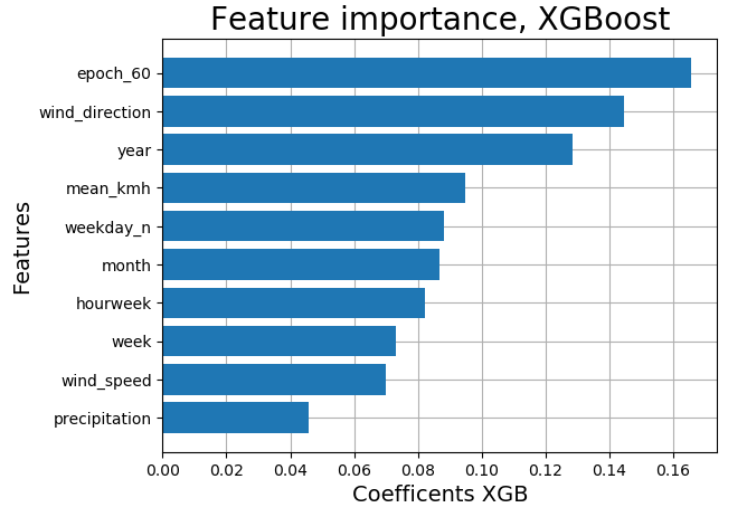


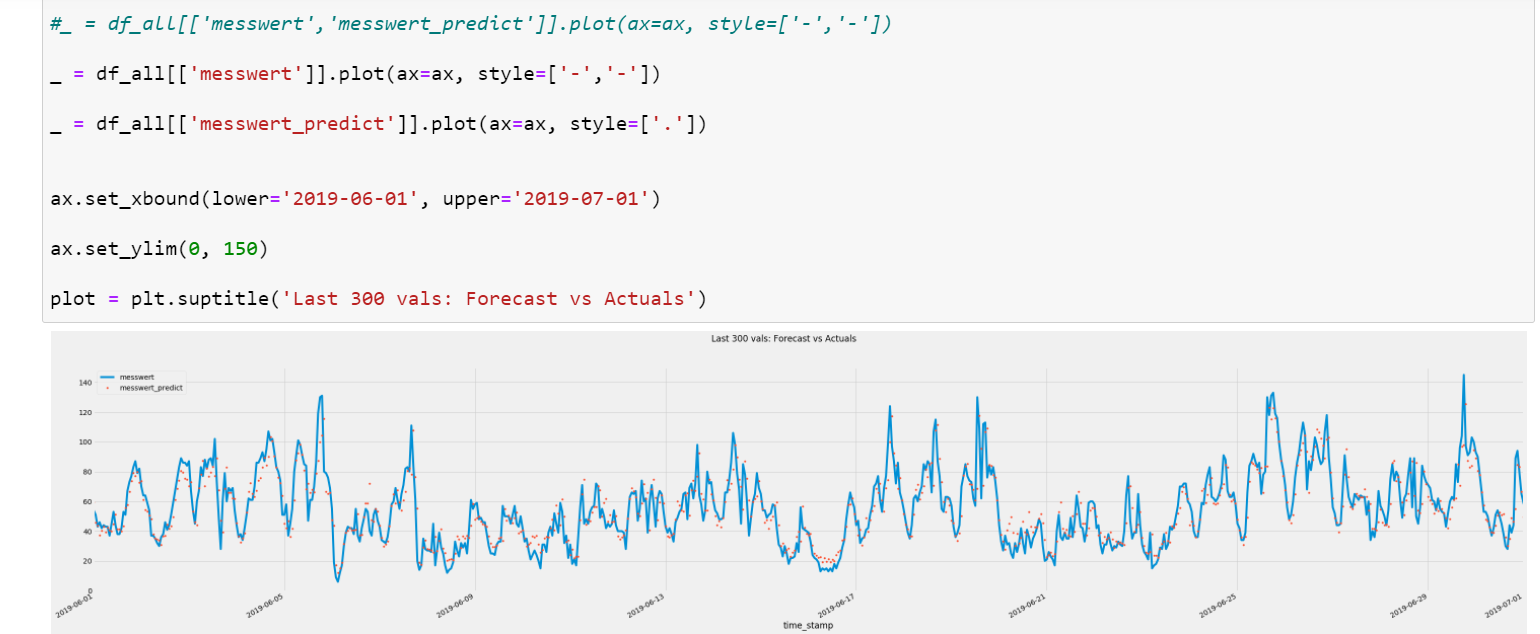


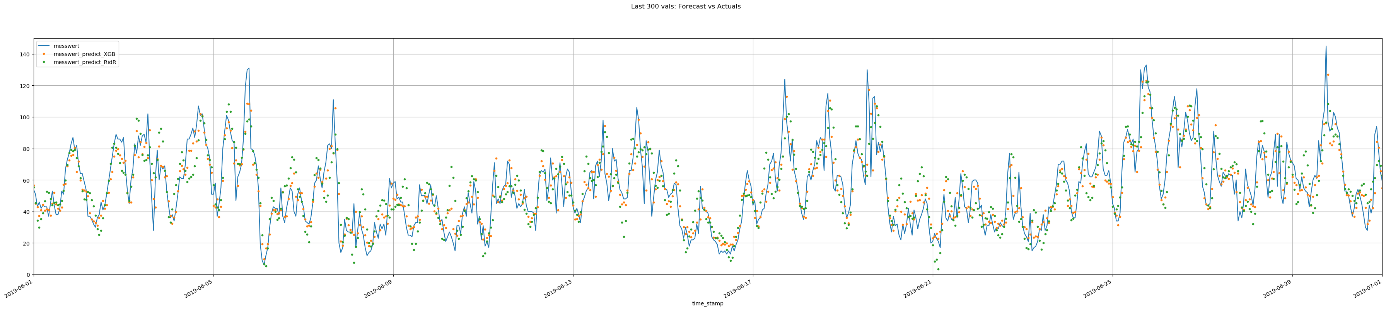


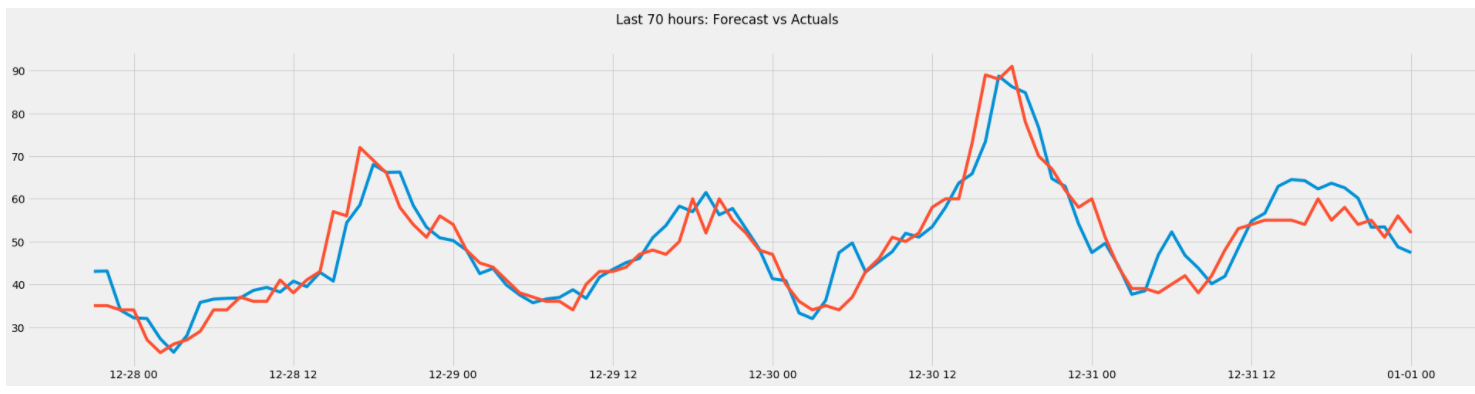




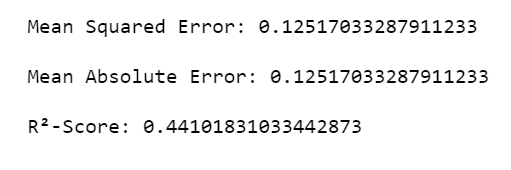
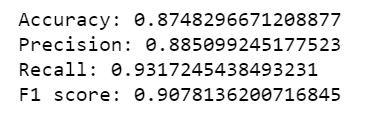


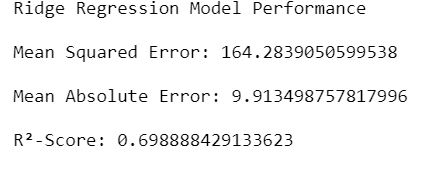






Logistic REG: (40 u/m³, + Wind.cols, standard+pca= None = best solution)



XGBoost : discrete NO2 values

