Predictive Analytics for Business Forecasting

PROJECT

Submitted To –

Prof. Pradeep Dadabada

Submitted By -

Group 3

Avinash Reddy - 2021PGP216

Ishan Kapse - 2021PGP095

Merry Ann Matthew - 2021PGP214

Pearl Makkar - 2021PGP108

Sarvesh Patkar - 2021PGP107

TABLE OF CONTENTS

Data Description

The dataset contains air quality data and AQI (Air Quality Index) at hourly level for Delhi city in India. It contains data assessed every hour from 2015 to 2020. There are several particles in the air that have a direct impact on air quality. The goal of our research is to determine what fraction of air particulates are significantly contributing to the city's elevated level of pollution. In addition, we are attempting to discover trends and seasonality in the data over time so that pre-emptive actions can be implemented to control pollution and maintain air quality.

The data has been made publicly available by the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of Government of India. They also have a real-time monitoring app: https://app.cpcbccr.com/AQI_India/

Columns present in the data are explained below

PM2.5 : Particulate Matter 2.5-micrometer in ug / m3

PM10 : Particulate Matter 10-micrometer in ug / m3

NO : Nitric Oxide in ug / m3

NO2 : Nitric Dioxide in ug / m3

NOx : Any Nitric x-oxide in ppb

NH3 : Ammonia in ug / m3

CO : Carbon Monoxide in ug / m3

SO2 : Sulphur Dioxide in ug / m3

O3 : Ozone in ug / m3

Benzene : Benzene in ug / m3

Toluene : Toluene in ug / m3

Xylene : Xylene in ug / m3

AQI : Air Quality Index

Introduction

A government commission on combating pollution has mandated a halt to all non-essential building activity in New Delhi and its surrounding districts as the air quality in the Indian capital continues to deteriorate.

The number of hospital visits is rising every year as a result of the nation capital's rising pollution levels. People have been blaming inhaling bad air quality for a variety of health problems for a number of years. Pollution has affected every age group, causing frequent complaints about respiratory conditions like a persistent cough and cold, asthma, and sinusitis, as opposed to a few years ago when only older people were vulnerable to respiratory problems. Additionally, those who have a chronic heart condition are now fighting for their lives due to additional health problems, frequently visiting the emergency room on days with high AQI (Air Quality Index).

The air quality had been in the "severe category" for multiple days in the month. The city's total air quality index (AQI) was around 320, and data from Delhi's "air quality early warning system" indicated that it will continue to deteriorate over the next several weeks. The air quality is projected to worsen but remain in the extremely bad category as per forecasts.

Monitoring Air Quality has become extremely important as the consequences are far severe. Through this exercise, we would be able to identify contaminated locations, the amount of pollution, and the air quality level with the use of the data gathered through air quality monitoring. The effectiveness of local air pollution management programmes might be assessed with the use of air quality monitoring.

Objective

Our ultimate goal is to create new and more efficient models that estimate pollution levels with high accuracy in a short span of time. Many current pollution forecasting systems rely solely on linear methodologies, which ignore nonlinear relationships in data. We hope to illustrate the possibility of applying nonlinear machine learning algorithms to improve pollution forecasts in this study. In the future, aggregation of model selection and time series analysis may be investigated for improved model efficiency and higher accuracy in less computational time.

ARMA

“in time series analysis to describe stationary time series”

What is it?

ARMA models cGiven a time series of data Xt, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA(p,q) model where p is the order of the AR part and q is the order of the MA part.

ARMA models can be estimated by using the Box–Jenkins method.

Applications

1. Application of the ARMA Model to Describe and Forecast the Flotation Feed Solids Flow Rate

Approach:

We imported the data of hourly AQI of Delhi, cleaned it, imputed missing values using backward fill.

We found the ‘p’ value of ARMA using the Partial Autocorrelation Function (pacf) and found the ‘q’ value using the Auto-Correlation Function (ACF). Using the ‘p’ and ‘q’ values decided, we generated the ARMA model such that the AIC score was also observed to be low. Using the model fitted, we generated the predictions and subsequently evaluated the model using RMSE, MAPE, and the Theil’s U Statistic values.

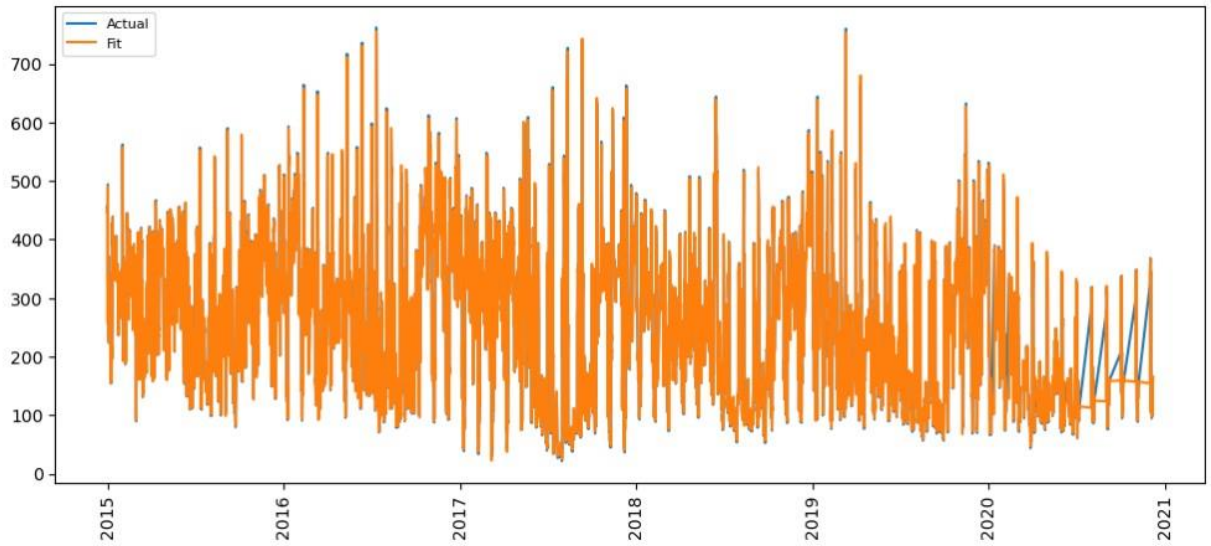
Metrics:

RMSE= 20.135896895461453

MAPE= 2.8856774281796236

Theil’s U = 0.03523106860556639

Graphical representation:



ARIMA

What is it?

The ARIMA model uses statistical analyses in combination with accurately collected historical data points to predict future trends and business needs. The ARIMA model is typically denoted with the parameters (p, d, q), which can be assigned different values to modify the model and apply it in different ways.

Applications

1. Forecasting the quantity of a good needed for the next time period based on historical data.
2. Forecasting sales and interpreting seasonal changes in sales
3. Estimating the impact of marketing events, new product launches

Approach:

We imported the data of hourly AQI of Delhi, cleaned it, imputed missing values using backward fill.

We then performed auto-arima using the pmdarima library, found the ARIMA model which has the lowest AIC score. We then performed ARIMA modeling using the coefficients obtained using the Auto-ARIMA model. Using the model fitted, we generated the predictions and subsequently evaluated the model using RMSE, MAPE, and the Theil's U Statistic values.

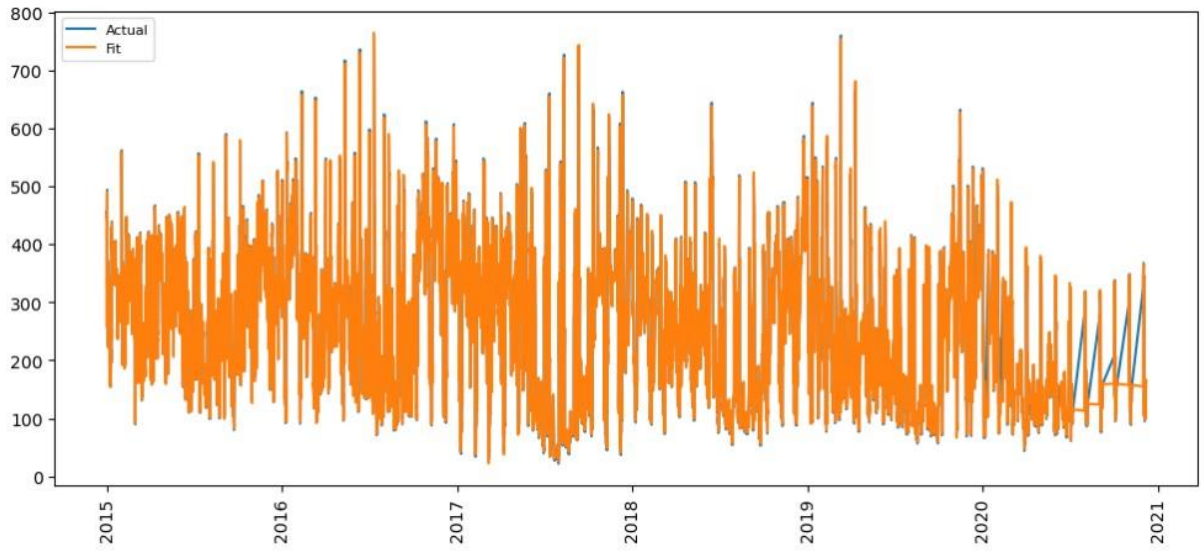
Metrics:

RMSE= 20.11480714837514

MAPE= 2.8983417661208386

Theil's U = 0.03519391420110315

Graphical representation:



SARIMA

What is it?

ARIMA and SARIMA are both algorithms for forecasting. ARIMA takes into account the past values (autoregressive, moving average) and predicts future values based on that. SARIMA similarly uses past values but also takes into account any seasonality patterns.

Applications

1. Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa
2. Application of SARIMA model to forecast Diwali sales

Approach:

We imported the data of hourly AQI of Delhi, cleaned it, imputed missing values using backward fill.

We then performed auto-arima using the pmdarima library, found the SARIMA model which has the lowest AIC score. We then performed SARIMA modeling using the coefficients obtained using the Auto-ARIMA model. Using the model fitted, we generated the predictions and subsequently evaluated the model using RMSE, MAPE, and the Theil's U Statistic values.

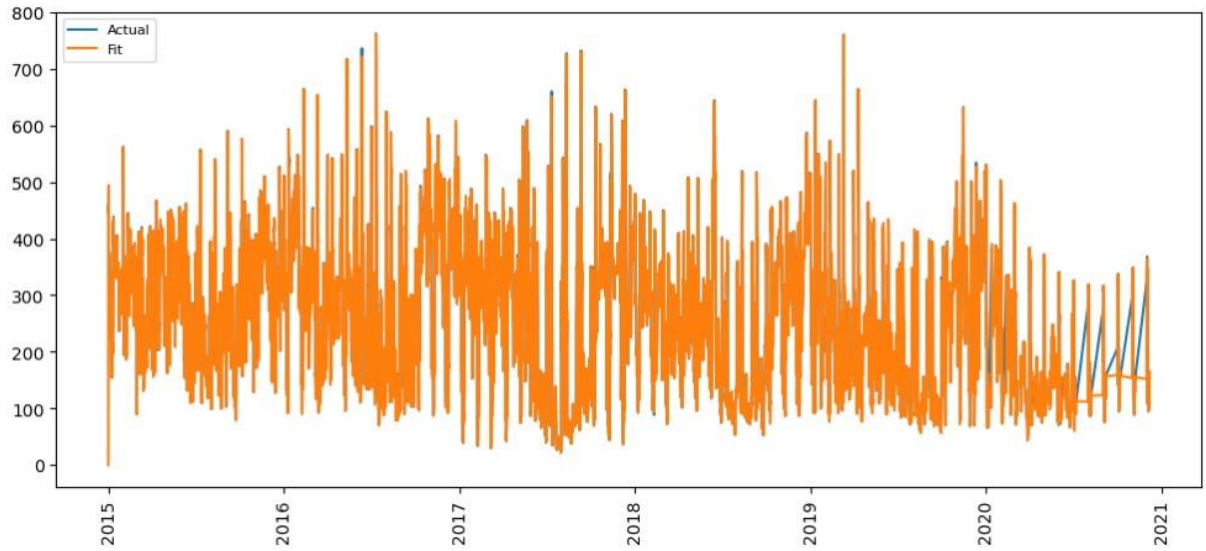
Metrics:

RMSE= 20.27952762884133

MAPE= 2.600944127581861

Theil's U = 0.03548276423654854

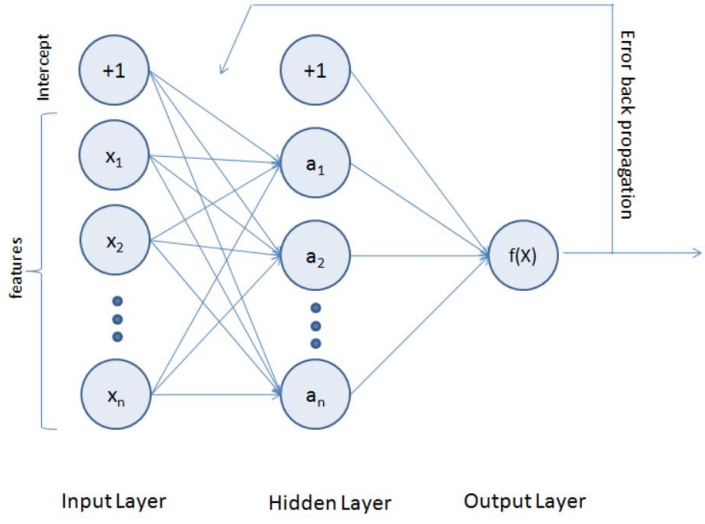
Graphical representation:



Multi-Layer Perceptron (MLP)

What it is:

Multilayer perceptron or feedforward neural networks are the simplest artificial neural network used in any real-world problem. Neural networks can be visualized as a series of connected layers that form a network connecting an observation's feature values at one end, and the target value (e.g., observation's class) at the other end.



Applications:

1. Airline Marketing Tactician
2. Backgammon
3. Data Compression – PCA
4. Driving – ALVINN
5. ECG Noise Filtering
6. Financial Prediction
7. Hand-written Character Recognition
8. Pattern Recognition/Computer Vision

Approach:

We imported the data of hourly AQI of Delhi, cleaned it, imputed missing values using backward fill, scaled the data, performed modeling using MLP Regressor which had activation function 'relu' and 10 hidden layers. Then we compared the model using the following performance metrics.

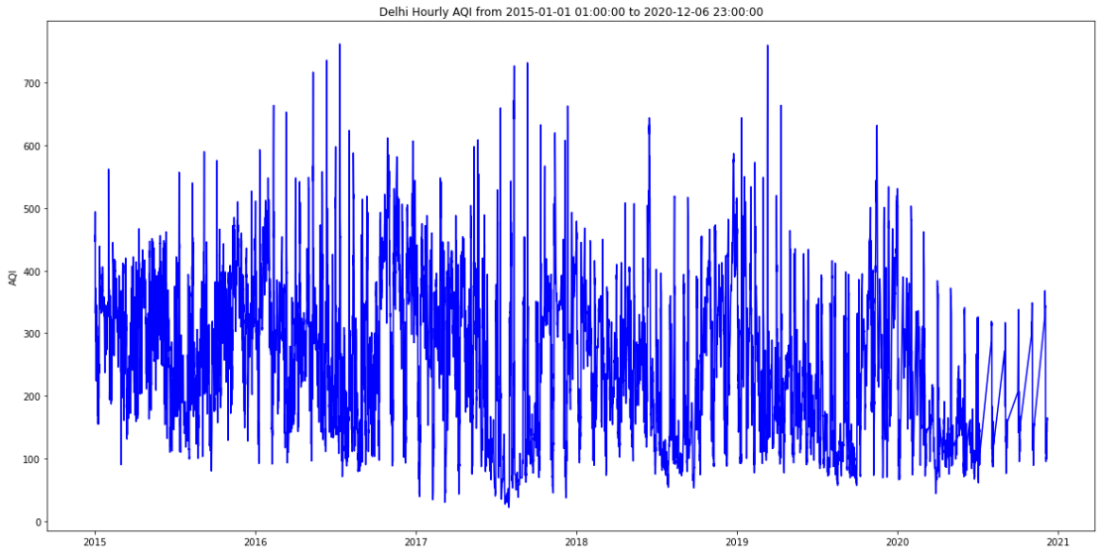
Metrics:

RMSE= 23.44147622521217

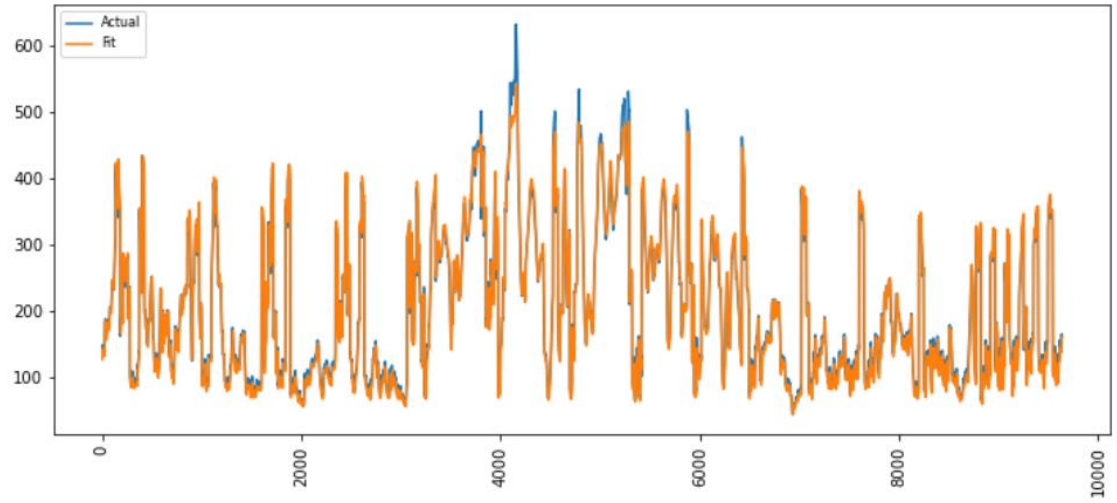
MAPE= 6.4455

Theil's U = 0.05035588303592504

Graphical representation:



The above graph shows the distribution of original data.



The above graph shows the performance of the model that is Actual vs Predicted values.

CNN

What is it?

It is used for deep learning and consists of three layers: a convolutional layer, a pooling layer and a fully connected (FC) layer. CNN employs parameter sharing that makes it less computationally intensive than other neural networks. This helps CNN to deal with scalability and overfitting issues.

Applications

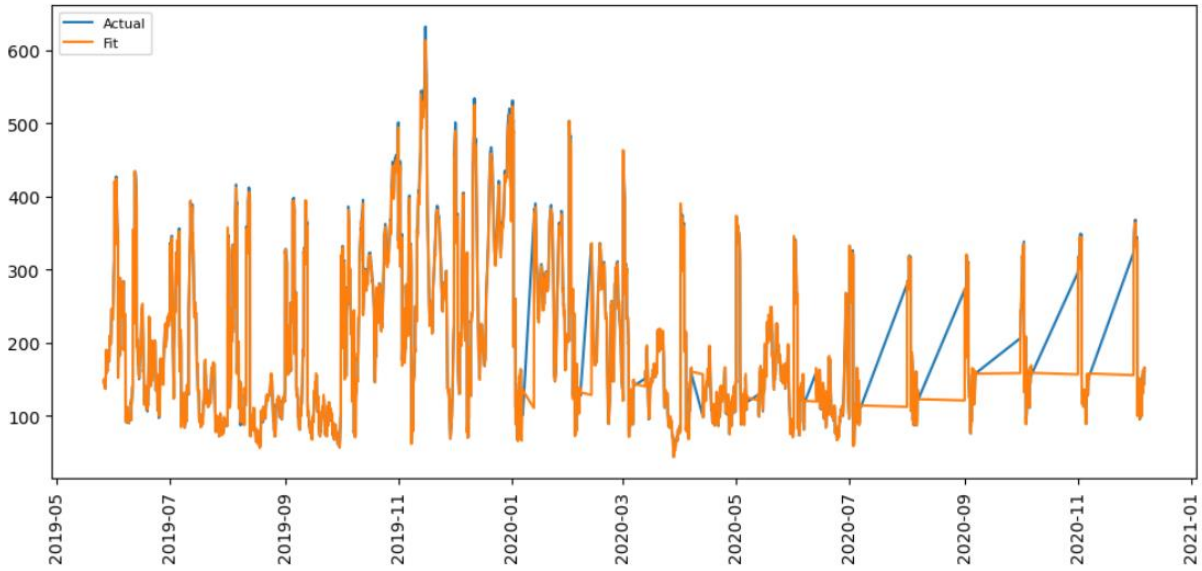
1. Image and pattern recognition
2. Speech recognition
3. Natural language processing
4. Video analysis

Metrics:

RMSE= 17.29905999777562

MAPE= 2.732814

Theil's U = 0.03721927110847361



Quantile Regression Neural Network

What it is

For the purpose of forecasting, the most extensively used neural network is single hidden-layer feedforward network. QRNN. It consists of m input neurons for predictors X1, X2, . . ., Xm,which are connected to n hidden neurons in a single hidden layer,which, in turn, are connected to one output neuron that yields predictand. The difference between traditional feedforward ANN and QRNN lies in the process of training.

Usually, the activation function at the output layer is a linear transfer function. The output can be calculated at each time t for each quantile individually.

Applications

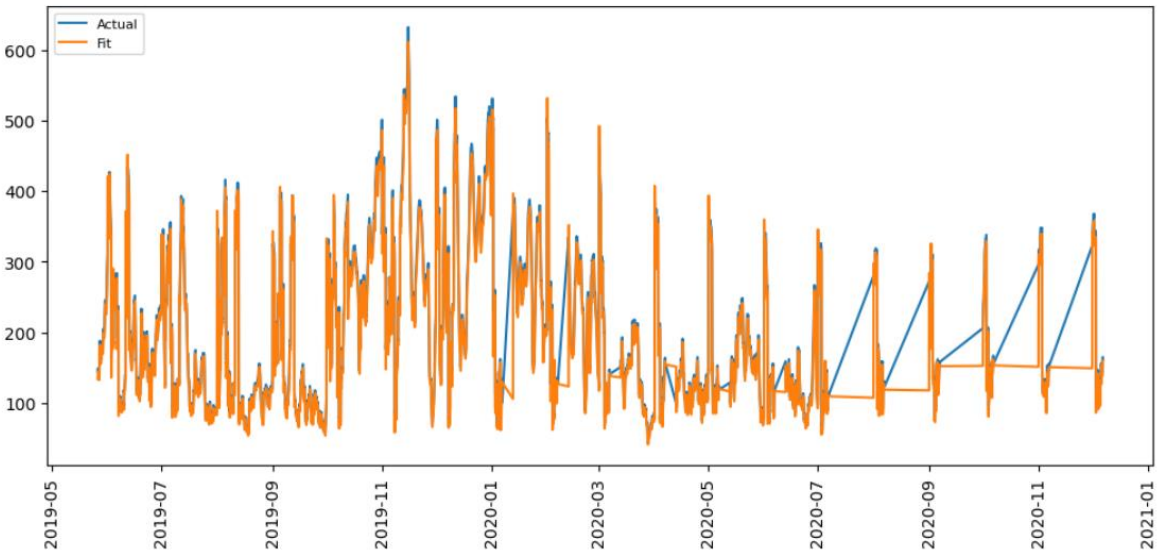
1. video classification
2. speech synthesis
3. natural language processing

Metrics:

RMSE - 19.64655650787658

MAPE - 4.7390237

Theil's U - 0.042714353844779084

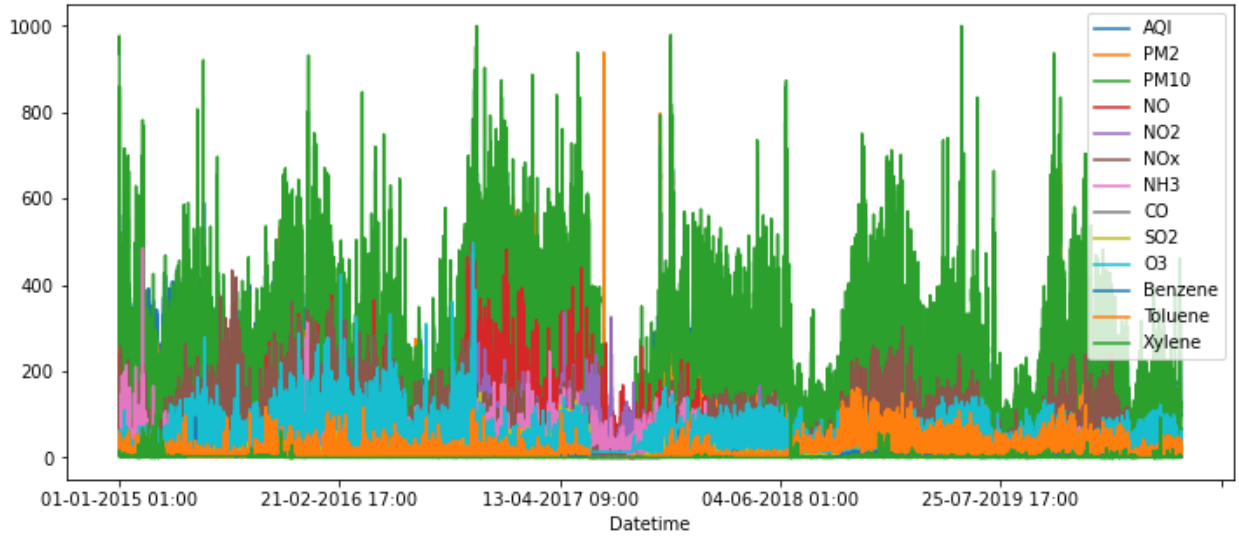


Multiple Linear Regression

What it is?

A statistical method known as multiple linear regression is used to forecast the result of a variable based on the values of two or more other variables. It is a development of linear regression and is occasionally just referred to as multiple regression. The variables we use to predict the value of the dependent variable are known as independent or explanatory variables, whereas the variable we want to predict is known as the dependent variable.

In our dataset, we had 12 columns which were deciding the target variable AQI (Air Quality Index)



From our analysis, coefficients are as below

Intercept	79.363059
PM2	0.538729
PM10	0.309618
NO	-0.283113
NO2	0.433355
NOx	0.102824
NH3	0.018457
CO	4.944888
SO2	0.535455
O3	0.368081
Benzene	-0.070521
Toluene	-0.432134
Xylene	0.213775

Through summary, we found that columns like NH3, Benzene and Xylene have p-value greater than 0.5 which means they have least significant impact on the final Air Quality Index (AQI) and can be removed from our final evaluation.

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.679
Model:	OLS	**Adj. R-squared:**	0.679
Method:	Least Squares	**F-statistic:**	6815.
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	0.00
Time:	21:20:25	**Log-Likelihood:**	-2.1818e+05
No. Observations:	38596	**AIC:**	4.364e+05
Df Residuals:	38583	**BIC:**	4.365e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	79.3631	1.102	72.045	0.000	77.204	81.522
PM2	0.5387	0.007	77.756	0.000	0.525	0.552
PM10	0.3096	0.005	68.012	0.000	0.301	0.319
NO	-0.2831	0.011	-26.141	0.000	-0.304	-0.262
NO2	0.4334	0.019	23.066	0.000	0.397	0.470
NOx	0.1028	0.011	9.455	0.000	0.082	0.124
NH3	0.0185	0.022	0.841	0.400	-0.025	0.061
CO	4.9449	0.132	37.423	0.000	4.686	5.204
SO2	0.5355	0.041	13.098	0.000	0.455	0.616
O3	0.3681	0.011	32.058	0.000	0.346	0.391
Benzene	-0.0705	0.190	-0.372	0.710	-0.442	0.301
Toluene	-0.4321	0.027	-16.226	0.000	-0.484	-0.380
Xylene	0.2138	0.120	1.783	0.075	-0.021	0.449

Metrics:

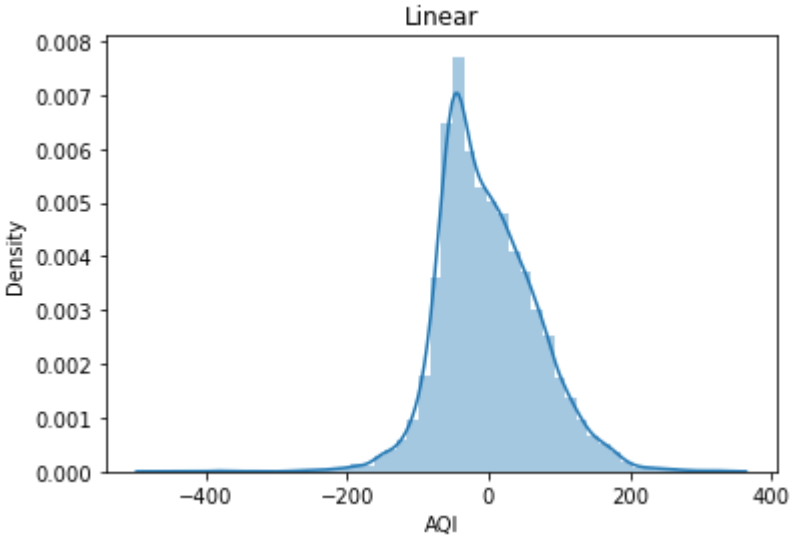
RMSE= 69.71093739842571

Rsquare= 0.6760592310414468

MAPE = 27.84446783262378

Thiel's u = 0.12283793629700637

Normality Test Results



NormaltestResult (statistic=448.2242366329083, pvalue=4.670288265961249e-98)

The null hypothesis is that the residual distribution is Normally distributed. Since the p-value < 0.05, we can reject the null. In other words, we can confidently say the residuals are **not Normally distributed.**

Breuschpagan test

['Lagrange multiplier statistic', 'p-value', 'f-value', 'f p-value']

(10717.83934491567, 0.0, 1236.1441823004163, 0.0)

The test is significant meaning the data violates the assumption of homoscedasticity, i.e. **heteroscedasticity is present in the data.** What to do? Either one can transform the variables to improve the model, or use a robust regression method that accounts for the heteroscedasticity.

Ridge Regression

"to eliminate multicollinearity in data models"

What it is:

When a data set exhibits multicollinearity or when there are more predictor variables than observations, ridge regression can be used to build a parsimonious model (correlations between predictor variables).

We applied Ridge model to find out the better fit for the model and obtained results are as follows:

Model Coefficients for input variables

([51.51950736, 44.18357374, -13.94422237, 12.76229998,

4.27247703, -0.17304691, 13.94314412, 5.41671076,

12.82115584, -0.14531681, -7.58047868, 0.81767544])

Metrics

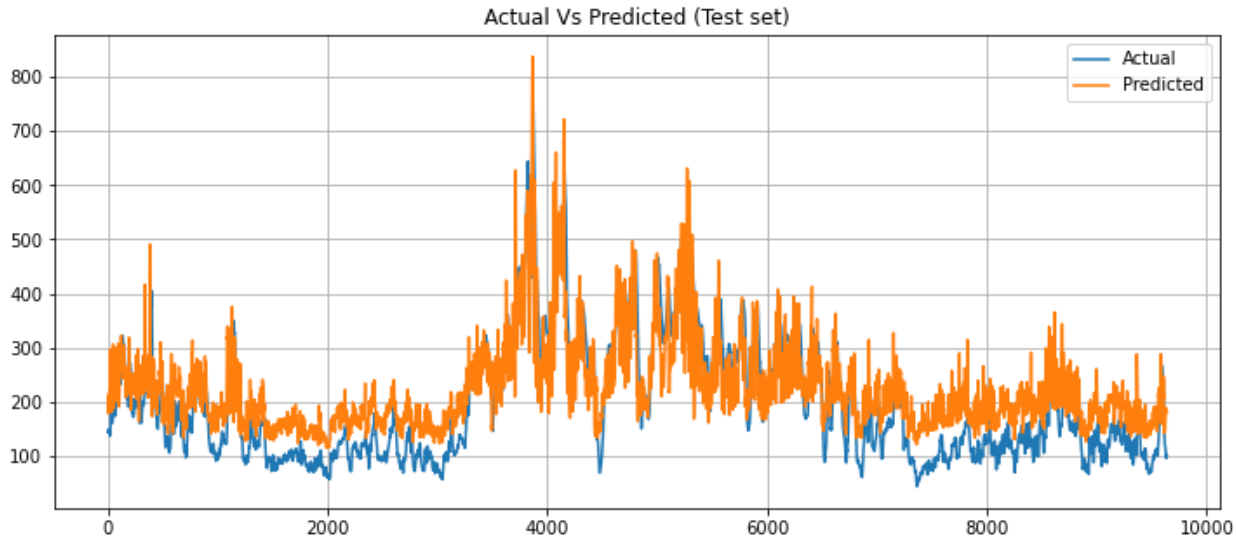
RMSE= 69.12897328409153

Rsquare= 0.6758039899007767

MAPE= 92.58068274949474

Thiel's u= 0.12283794304929037

Graph of Actual vs Predicted (Test Set)



Lasso Regression

Shrinkage is used in the linear regression method known as lasso regression. When data values shrink toward a middle value, such as the mean, this is called shrinkage. Simple, sparse models are encouraged by the lasso approach (i.e. models with fewer parameters). When models exhibit high levels of multicollinearity or when you want to automate specific steps in the model selection process, such as variable selection and parameter elimination, this particular type of regression is well suited.

Least Absolute Shrinkage and Selection Operator is referred to by the abbreviation "LASSO."

Metrics

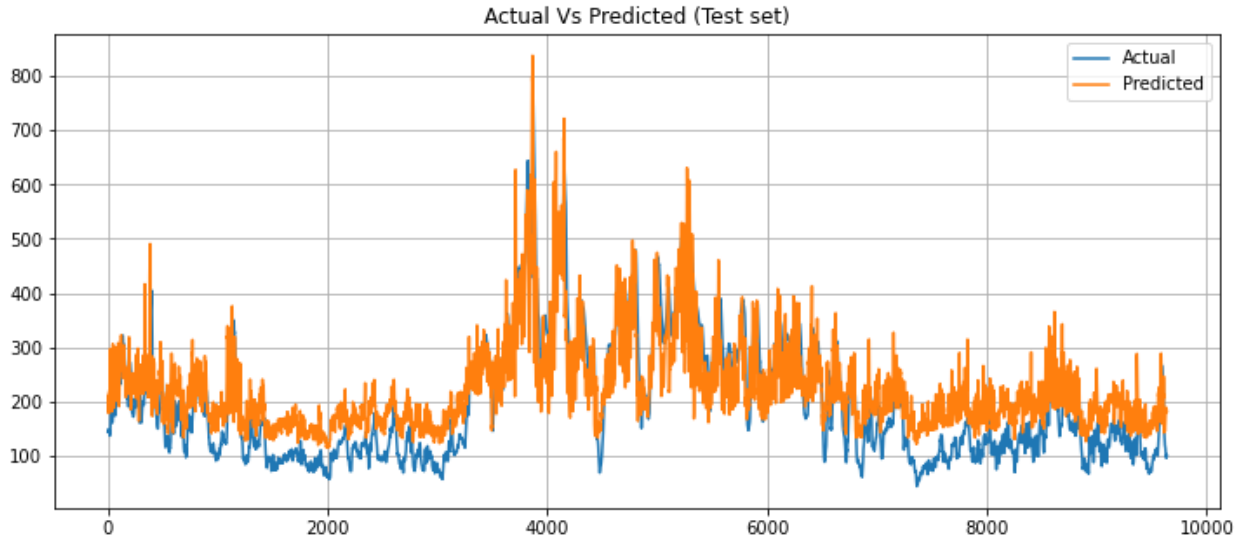
RMSE= 66.88509391669227

Rsquare= 0.661632783997421

MAPE = 39.490885302557686

Thiel's u= 0.14267166228150532

Graph of Actual vs Predicted (Test Set)



LSTM

What is it:

Long short-term memory (LSTM) is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network (RNN) can process not only single data points (such as images), but also entire sequences of data (such as speech or video).

Applications:

1. Unsegmented, connected handwriting recognition
2. Speech recognition
3. Machine translation
4. Robot control
5. Video games
6. Healthcare

Approach:

We imported the data of hourly AQI of Delhi, cleaned it, imputed missing values using backward fill.

We partitiond the data into Training and Testing sets and generated the LSTM model using 200 epochs. Using the model fitted, we generated the predictions and subsequently evaluated the model using RMSE, MAPE, and the Theil's U Statistic values.

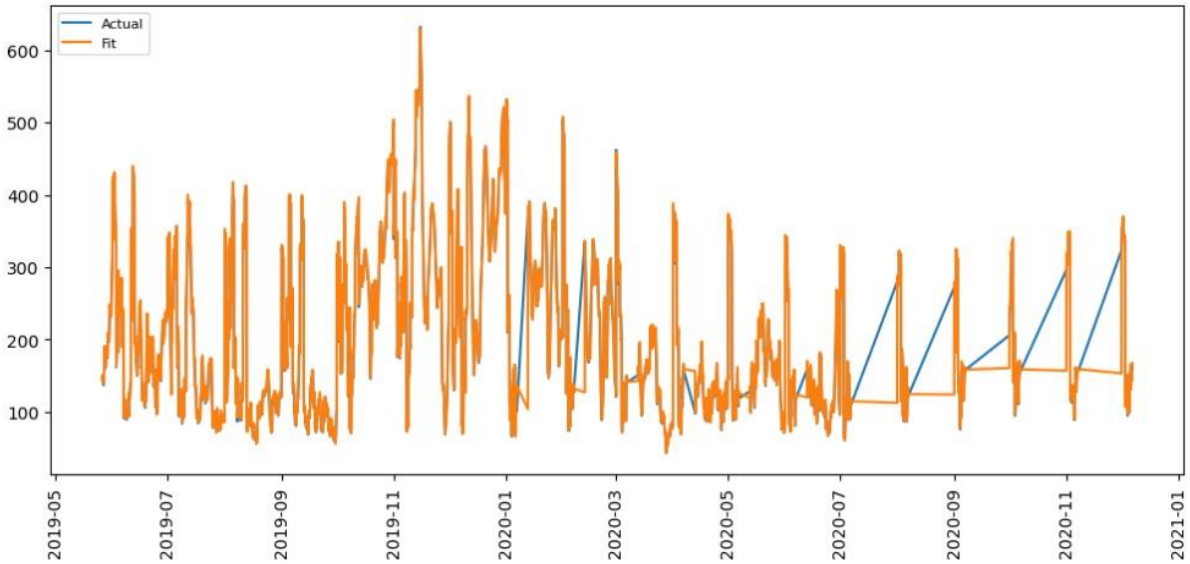
Metrics:

RMSE= 17.187847120074313

MAPE= 2.6115632

Theil's U = 0.03673530113732017

Graphical representation:



Decision Tree

What it is

Decision Trees are versatile Machine Learning algorithms that can perform both classification and regression tasks, and even multi output tasks. They are very powerful algorithms, capable of fitting complex datasets. Decision Trees are also the fundamental components of Random Forests which are among the most powerful Machine Learning algorithms available today. One of the many qualities of Decision Trees is that they require very little data preparation. In particular, they don't require feature scaling or centering at all. A decision tree consists of 3 types of nodes.

- Root node
- Branch node
- Leaf node (class label)

Applications

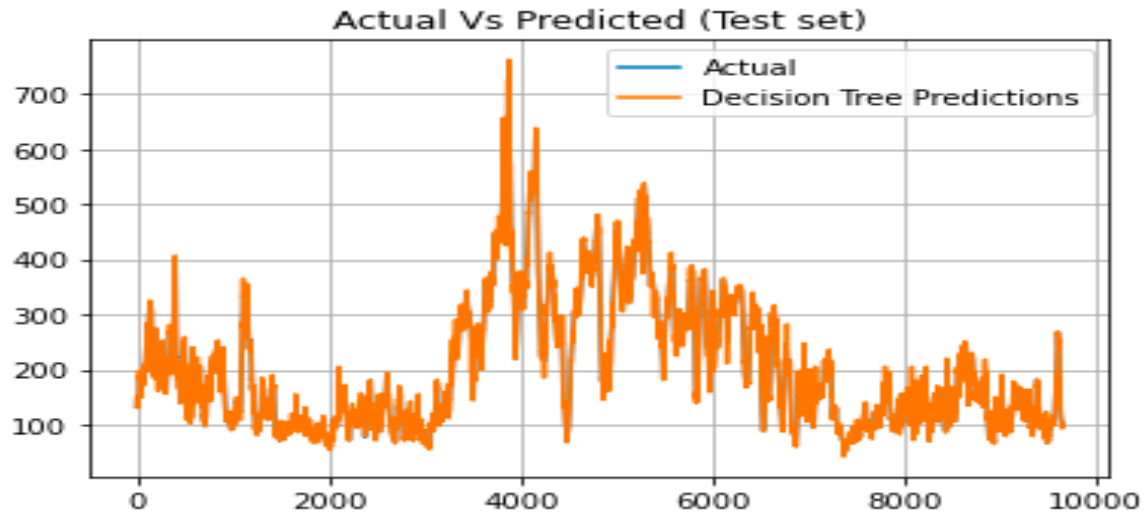
1. Buying something from any online shopping portal where we get several recommendations based on what we are buying.

RMSE= 6.572538539482582

MSE= 43.19826285298383

MAPE= 2.0156334454029796

Theil's= 1.422863884779911



Random Forest

What it is

A common problem with decision trees is that they tend to fit the training data too closely (i.e., overfitting). This has motivated the widespread use of an ensemble learning method called random forest. In a random forest, many decision trees are trained, but each tree only receives a bootstrapped sample of observations (i.e., a random sample of observations with replacement that matches the original number of observations) and each node only considers a subset of features when determining the best split.

Applications

Banking Industry

- Credit Card Fraud Detection
- Customer Segmentation
- Predicting Loan Defaults on LendingClub.com

Healthcare and Medicine

- Cardiovascular Disease Prediction
- Diabetes Prediction
- Breast Cancer Prediction

Stock Market

- Stock Market Prediction
- Stock Market Sentiment Analysis
- Bitcoin Price Detection

E-Commerce

- Product Recommendation
- Price Optimization
- Search Ranking

Metrics:

RMSE= 4.5834798879722065

MSE= 21.00828788344571

MAPE= 1.6805306544718421

Theil's= 0.9923446351539901

Where DT is Decision Tree and RF is Random Forest.

AdaBoost

What it is

In one form of boosting called AdaBoost, we iteratively train a series of weak models (most often a shallow decision tree, sometimes called a stump), each iteration giving higher priority to observations the previous model predicted incorrectly.The end result is an aggregated model where individual weak models focus on more difficult (from a prediction perspective) observations. In scikit-learn, we can implement AdaBoost using AdaBoostClassifier or AdaBoostRegressor. The most important parameters are base_estimator, n_estimators, and learning_rate.

Applications

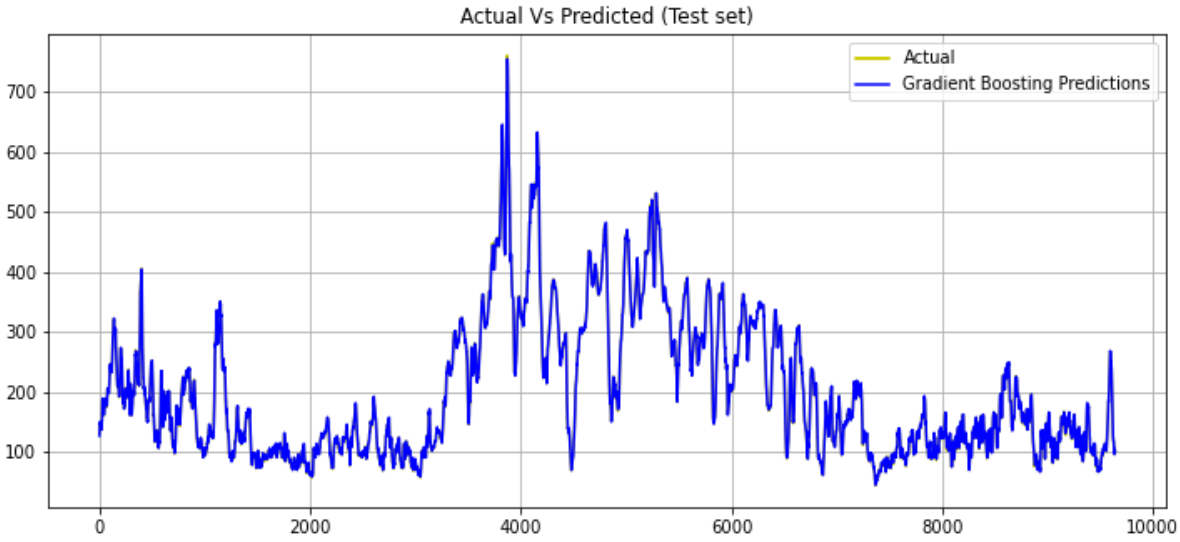
1. predicting customer churn
2. classifying the types of topics customers are talking/calling about.

AdaBoost RMSE= 34.05281036865356

AdaBoost MSE= 1159.5938940034794

AdaBoost MAPE= 17.710175758612902

AdaBoost Theil's= 7.336726280535233



Gradient Boosting

What it is

Due to the stage wise addictivity, the loss function can be represented in a form suitable for optimization. This gave birth to a class of generalized boosting algorithms known as generalized boosting algorithm (GBM). Gradient boosting is an example implementation of GBM and it can work with different loss functions such as regression, classification, risk modelling etc

Applications

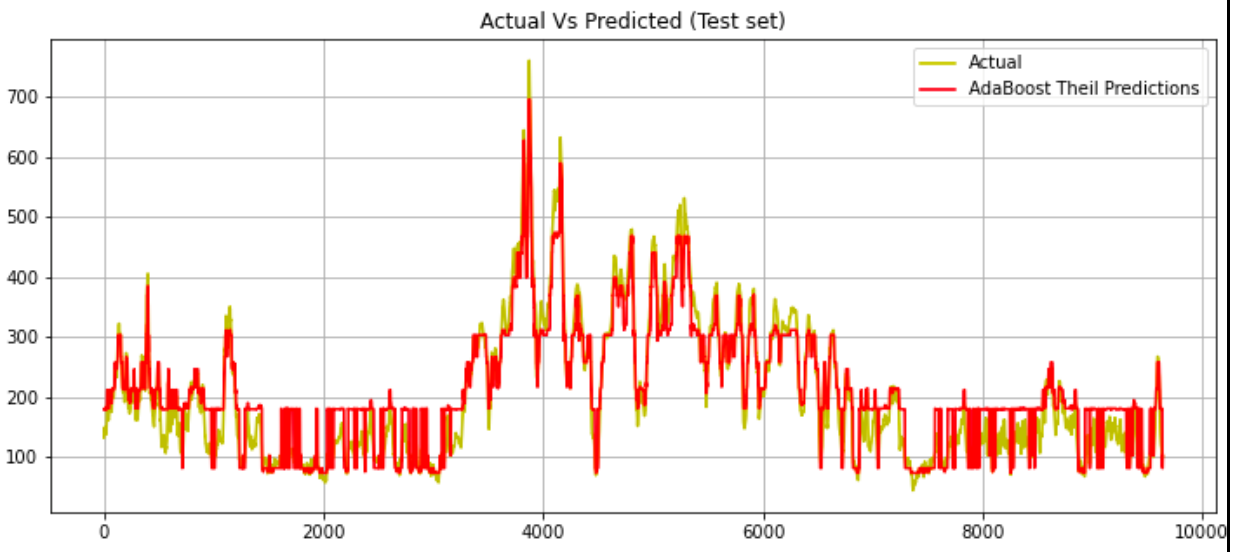
1. Gradient Boosting Algorithm is generally used when we want to decrease the Bias error.
2. Gradient Boosting Algorithm can be used in regression as well as classification problems.
3.

Gradient Boosting RMSE= 3.9935933952273963

Gradient Boosting MSE= 15.948788206403885

Gradient Boosting MAPE= 1.762239393102629

Gradient Boosting Theil's= 0.8649216544735704



Xtreme Gradient Boosting

What it is

Xgboost used a more regularized model formalization to control over-fitting, which gives it better performance.

Regularization: Standard GBM implementation has no regularization like XGBoost, therefore it also helps to reduce overfitting. In fact, XGBoost is also known as 'regularized boosting' technique.

Applications

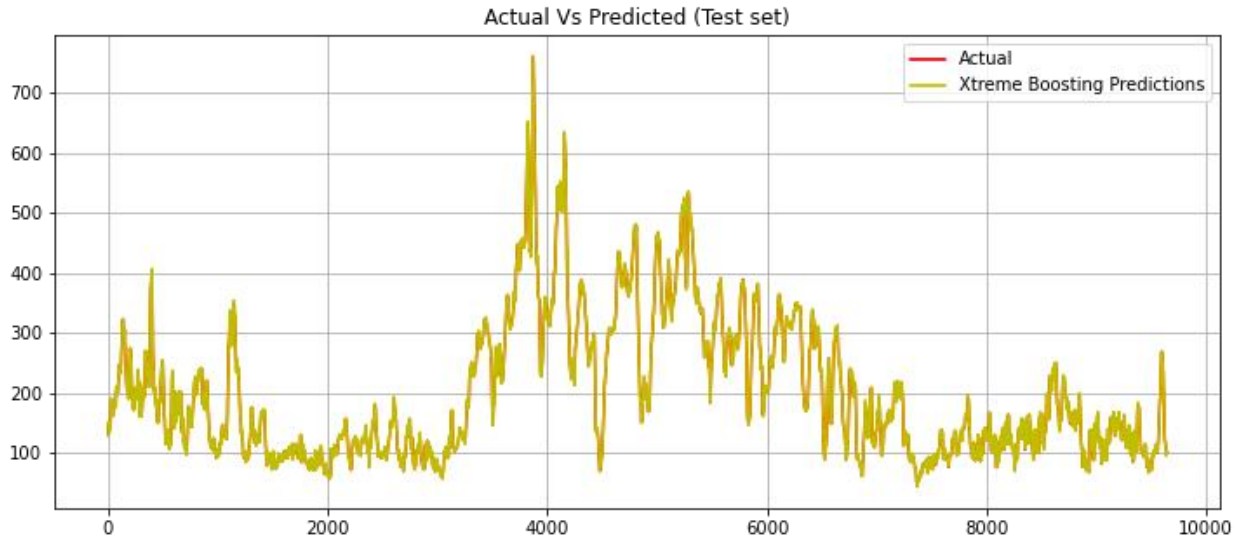
1. It provides parallel tree boosting
2. It is the leading machine learning library for regression, classification, and ranking problems.

Extreme Gradient Boosting RMSE= 3.182271261281875

Extreme Gradient Boosting MSE= 10.126850380380533

Extreme Gradient Boosting MAPE= 1.385762205720915

Extreme Gradient Boosting Theil's= 0.6891596812746276



SVM

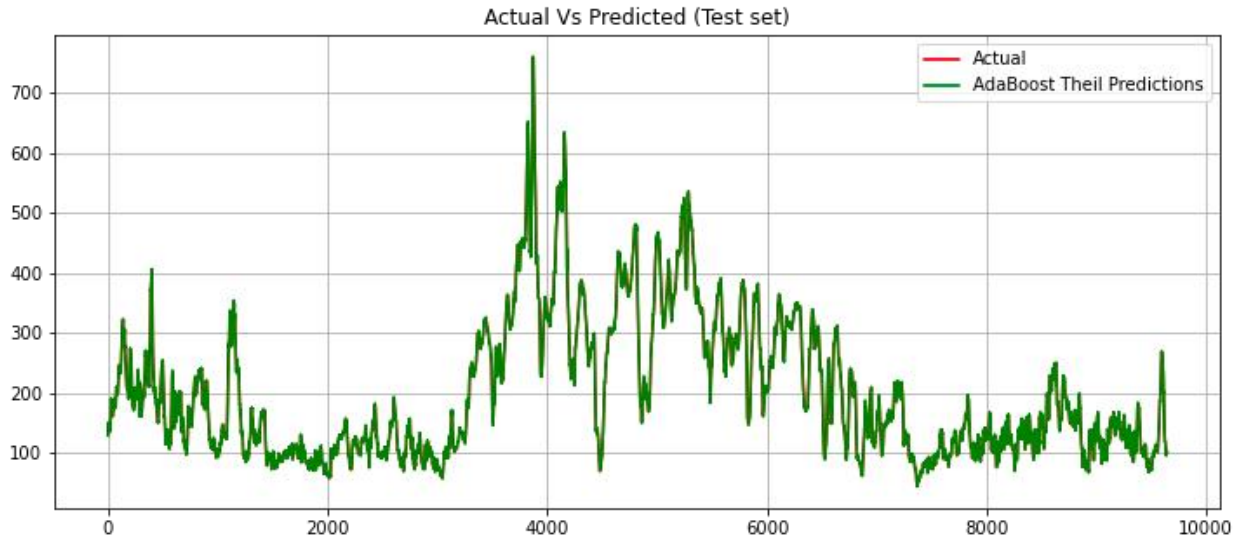
The key objective of SVM is to draw a hyperplane that separates the two classes optimally such that the margin is maximum between the hyperplane and the observations. The following figure illustrates that there is the possibility of different hyperplanes. However the objective of SVM is to find the one which gives us a high margin.

SVM RMSE= 10.173607771633113

SVM MSE= 103.50229509103367

SVM MAPE= 0.016323108592109597

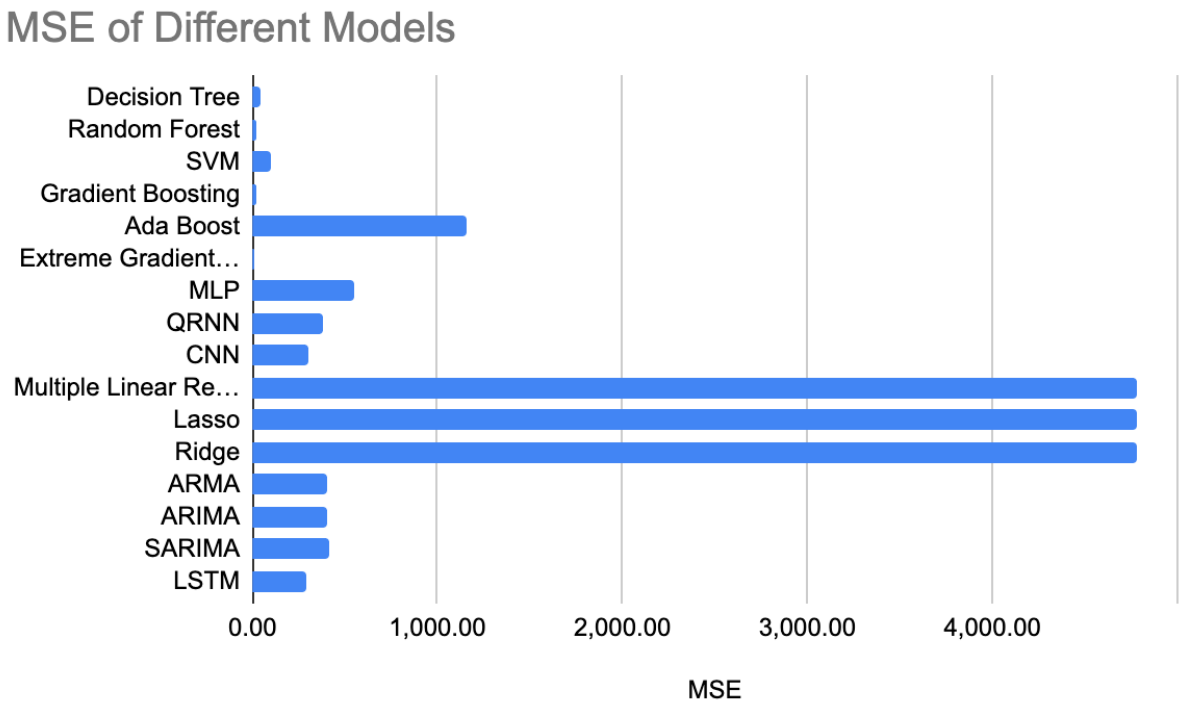
SVM Theil's= 2.209742133898767

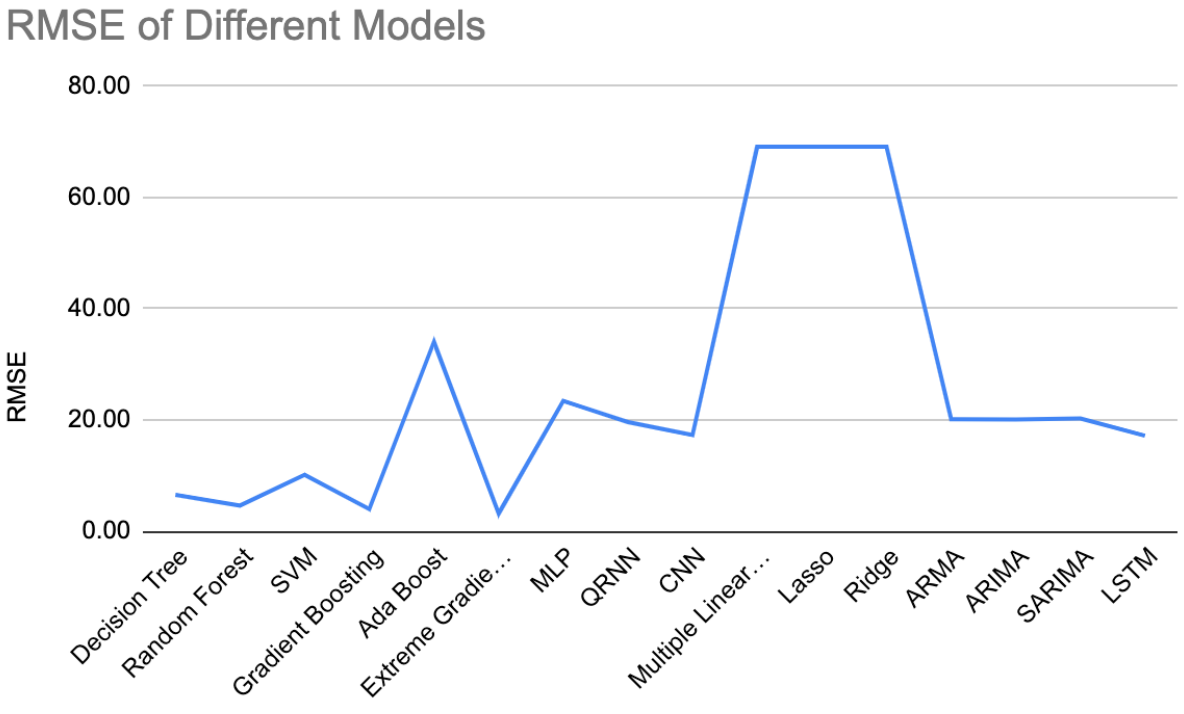


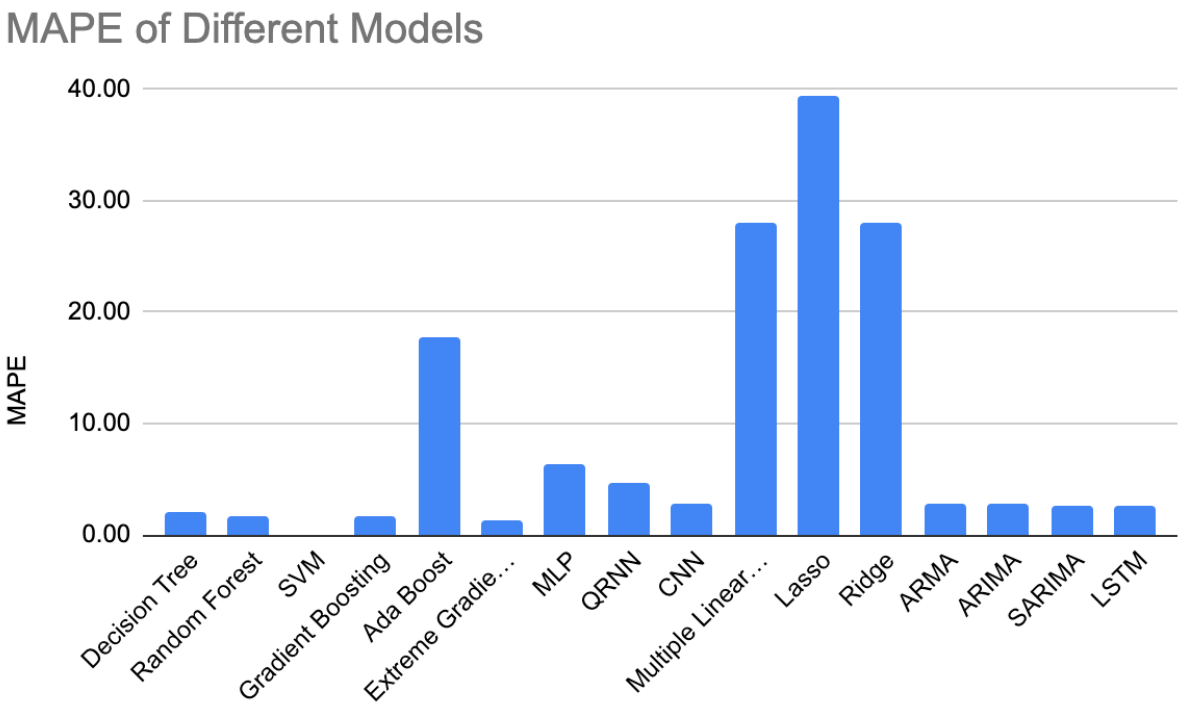
Results

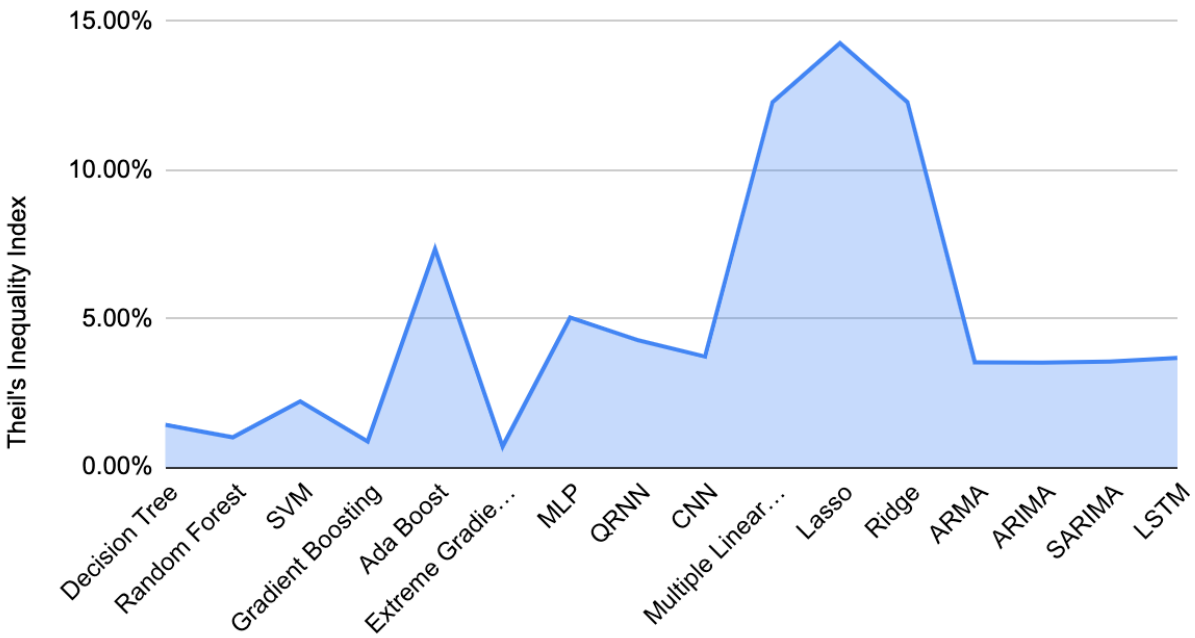
	MSE	RMSE	MAPE	Theil's Inequality Index
Decision Tree	43.20	6.57	2.02	1.42%
Random Forest	21.37	4.62	1.69	1.00%
SVM	103.50	10.17	0.02	2.21%
Gradient Boosting	15.95	3.99	1.76	0.86%
Ada Boost	1,159.59	34.05	17.71	7.34%
Extreme Gradient Boosting	10.13	3.18	1.39	0.69%
MLP	549.50	23.44	6.45	5.04%
QRNN	385.99	19.65	4.74	4.27%
CNN	299.26	17.30	2.73	3.72%
Multiple Linear Regression	4,778.82	69.13	28.08	12.28%
Lasso	4,779.19	69.13	39.49	14.27%
Ridge	4,778.81	69.13	28.08	12.28%
ARMA	405.45	20.14	2.89	3.52%

ARIMA	404.61	20.11	2.90	3.52%
SARIMA	411.26	20.28	2.60	3.55%
LSTM	295.42	17.19	2.61	3.67%






MAPE of Different Models
MAPE
40.00
30.00
20.00
10.00
0.00
Decision Tree
Random Forest
SVM
Gradient Boosting
Ada Boost
Extreme Gradie…
MLP
QRNN
CNN
Multiple Linear…
Lasso
Ridge
ARMA
ARIMA
SARIMA
LSTM


Theil's Inequality Index Comparision
Theil's Inequality Index
15.00%
10.00%
5.00%
0.00%
Decision Tree
Random Forest
SVM
Gradient Boosting
Ada Boost
Extreme Gradie…
MLP
QRNN
CNN
Multiple Linear…
Lasso
Ridge
ARMA
ARIMA
SARIMA
LSTM

From the results obtained above, we can see that **Extreme Gradient Boosting** has the least RMSE while regression models have high RMSE values

Managerial Implications

- Particulate Matter are minute particles of solid or liquid substances floating in the air. In 2020, 4.1 million fatalities worldwide from heart disease and stroke, lung cancer, chronic lung disease, and respiratory infections were caused in part by PM 2.5 exposure.Through analysis, we can determine when particulate matter exposure is likely to be at its highest and what safety precautions need to be taken to manage it.
- Additionally, we observed that even in Delhi with high pollution levels, certain timings remain more severe.
- We learned that Delhi's air quality standards continue to be in the "poor" range, and that serious measures and regulations are needed to reverse this trend.

Future Scope

The Commission for Air Quality Management should consider which particles are creating a direct serious impact on the Air quality index and which particles are deteriorating to health.

Industries that are increasing their pollution levels and endangering people's health ought to be strongly urged to do so.

By observing the trends and seasonality components of the data, CAQM can be more vigilant in their actions.

References

1. **https://www.imf.org/~/media/Files/Publications/WP/2017/wp17108.ashx**
2. **https://corporatefinanceinstitute.com/resources/data-science/ridge/**
3. **https://en.wikipedia.org/wiki/Autoregressive%E2%80%93moving-average_model**
4. **https://corporatefinanceinstitute.com/resources/data-science/autoregressive-integrated-moving-average-arima/**
5. **https://en.wikipedia.org/wiki/Long_short-term_memory#:~:text=March%202022**