Olabisi Balogun
January 21st, 2021

Capstone Project

Data Scientist Nanodegree

Project Overview

Charter schools were established in the United States to offer free quality education to K-12 students at no cost. It is seen as a better alternative to public K-12 schools. Charter schools are frequently audited based on their students' performance (Bulkley and Fisher, 2002). Improving students' academic performance is the aim of many charter schools in the United States. The project aims to track the performance of students that attend charter school in a specific subject like Math. The performance tracker gives the school leadership an insight of how students are performing in Math based on the percentage of students classified into 3 levels: Meeting Standards, Partially Meeting Standards, and Not Meeting Standards, after taking a test. The dataset for the project was gotten from **https://opendata.cityofnewyork.us**, New York City repository for publicly accessible dataset. The performance tracker will be based on a time series model – Vector Autoregression (VAR). VAR is a feature of the Python scikit.statsmodels package that was developed by McKinney, Perktold, Seabold in 2011.

Project Statement

The goal of the project is to design a web application to predict the future performance of students in Math from a specific school. The application predicts the percentage of students that would fall into each of the performance standards - Meeting Standards, Partially Meeting Standards, and Not Meeting Standards in the future.

A web application is designed with Flask. The web app displays the trend and forecast of the performance ratio in three specified Math learning standards for a selected school using Plotly

charts. The model that fuels the prediction is based on VAR, a statistical time series model from the statsmodel package in Python.

Metrics

According to this reference, (https://en.wikipedia.org/wiki/Forecasting#Forecasting_accuracy), there are several metrics that can be used to evaluate time series models. Several metrics like mean absoluter error (MAE), mean squared error (MSE), root mean squared error (RMSE) among other metrics. RMSE, MSE and MAE are efficient accuracy measures when the forecast error is on the same scale as the data. It is best for use when the dataset includes target variables of the same scale. This attribute makes RMSE and MAE a good choice because the target variables in the project's dataset are on the same scale. Besides, RMSE captures how much difference the forecast values (prediction) are from the original values in the dataset (test data). RMSE and MAE are suitable accuracy measure for regression models. RMSE and MAE give almost similar result, albeit MAE values are smaller. Using both together gives a boost of confidence in the reported error score.

Data Exploration

Understanding the Data:

The dataset- '2013-2019 Math Test Results Charter – School' was provided by the NYC Department of Education. It is the Math test results of students in charter schools in New York City. The data consists of 5267 observations and 18 features. The dataset has the scores of 191 charter schools in NYC from 2013 to 2019. To get the correct descriptive statistics of columns, there is a need for data type conversion. An error occurred during the data type conversion indicating the presence of literal string in the specified numeric columns. The handling of this

error was the first step in cleaning the dataset. A further investigation into the dataset revealed missing data for specific grades of schools. I made a decision to remove these observations because of the presence of an observation that accounts for the result of all grades in that specific year.

The original data set consist of the following features: 'Unnamed Column', 'DBN', 'School Name', 'Grade', 'Year', 'Category', 'Number Tested', 'Mean Scale Score', '# Level 1', '% Level 1', '# Level 2', '% Level 2', '# Level 3', '% Level 3', '# Level 4', '% Level 4', '# Level 3+4', '% Level 3+4'.

Initial Data Cleaning:

- Removed the 'Unnamed Column' because it was irrelevant
- Removed the 'Category' column because it contained same values for all the observations, which makes it irrelevant.
- Removed rows with literal string 's' in all the score columns
- Converted the features to appropriate data types

Further scrutinizing of the dataset revealed some redundant columns. The number of students in each level of performance and the percentage of students in each level of performance represent the same information. I dropped the number columns and kept the percentage column and also used the merged version of students in level 3 and 4 since both represented students meeting learning standards, although level 4 means 'meeting learning standards with distinction'.

The next step was to rename the columns to be more descriptive. With the help of the data dictionary that accompanied the dataset, I had the good understanding of what each column portrayed.

- '% Level 1' changed to 'Not_Meeting_Pct' (Percentage of students not meeting learning standards)
- '% Level 2' changed to 'Partially_Meeting_Pct' (Percentage of students partially meeting learning standards)
- '% Level 3+4' changed to 'Meeting_Pct' (Percentage of students meeting learning standards)

Data Visualization

The distribution of each performance level:

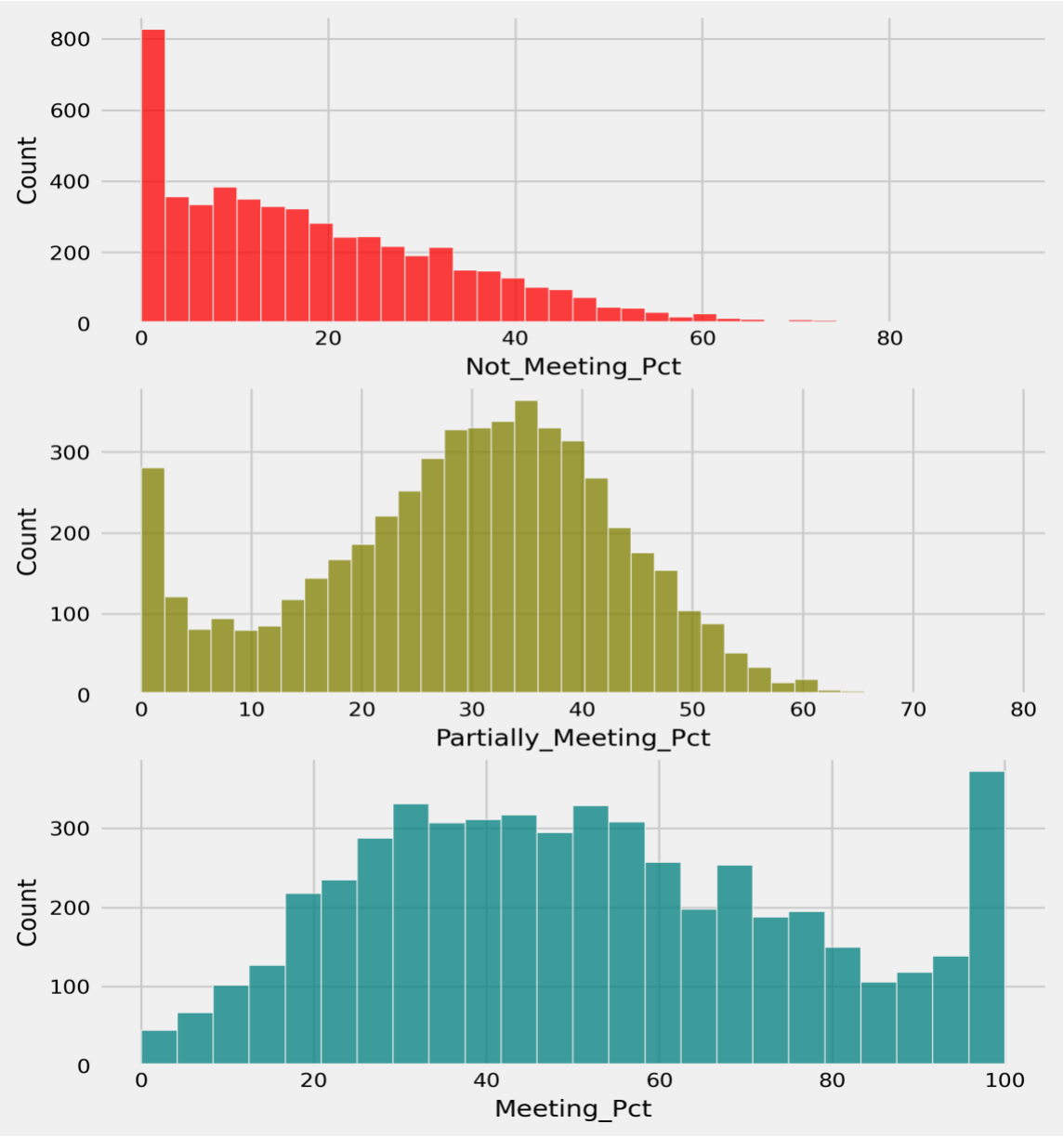

Figure 1: Distribution of the three performance levels

The distribution of the percentage of students not meeting learning standards in Math is right skewed; most schools have less than 60 percent of their students not meeting learning students, in fact a high fraction of the schools does not record any student in this category.

Olabisi Balogun
January 21st, 2021

The distribution of the students partially meeting learning standards is almost a normal distribution with an abnormal spike of at the left tail, which probably suggests significant number of schools have minimal percentage students in the partially meeting learning standards category.

The distribution of percentage of students in the meeting learning standards category is also almost a normal distribution with an abnormal spike at the right tail, which probably indicates many schools have high percentage of students doing really well in Math. This explains why there is a spike at the left tail of both the not meeting learning standards and partially meeting learning standards categories. A high percentage of students doing really well indicates a low percentage of students struggling.

The distribution of the number students taking the exam:

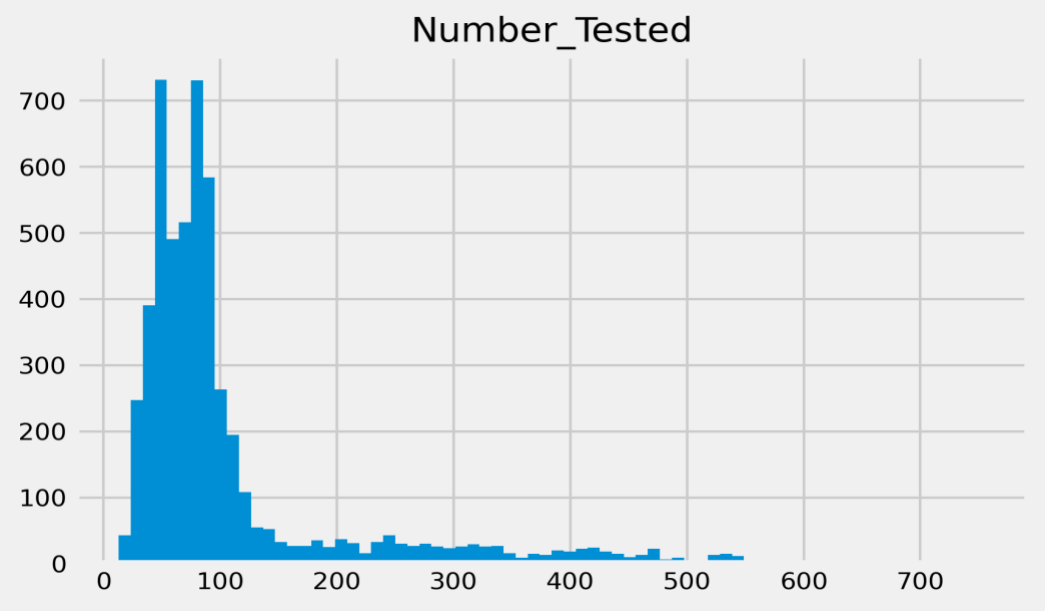

Figure 2: Distribution of number of students writing Math exam annually

This distribution is right skewed, the majority of schools have less than 100 students taking the Math exams at a particular time.

The distribution of the mean score of each exam:

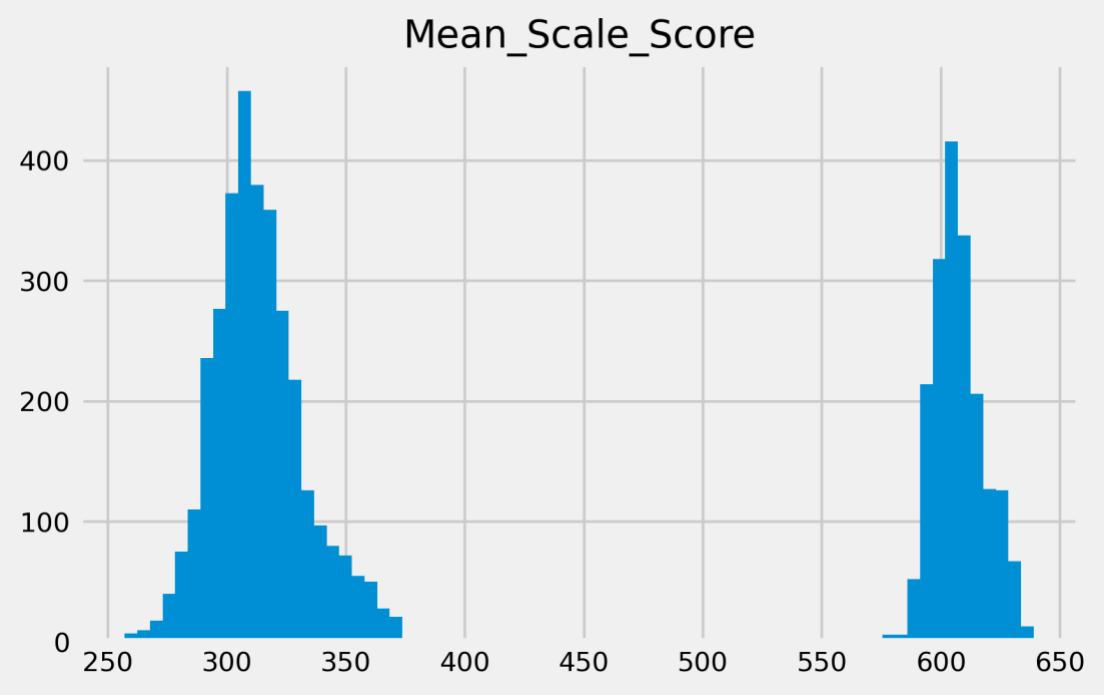

Figure 3: Distribution of calculated mean scores

This distribution depicts an interesting information. From the diagram, we can see that the schools can be categorized in two distinct groups. Schools that are really doing well and schools that are struggling.

It is worth looking at the behavior of the schools over the years.

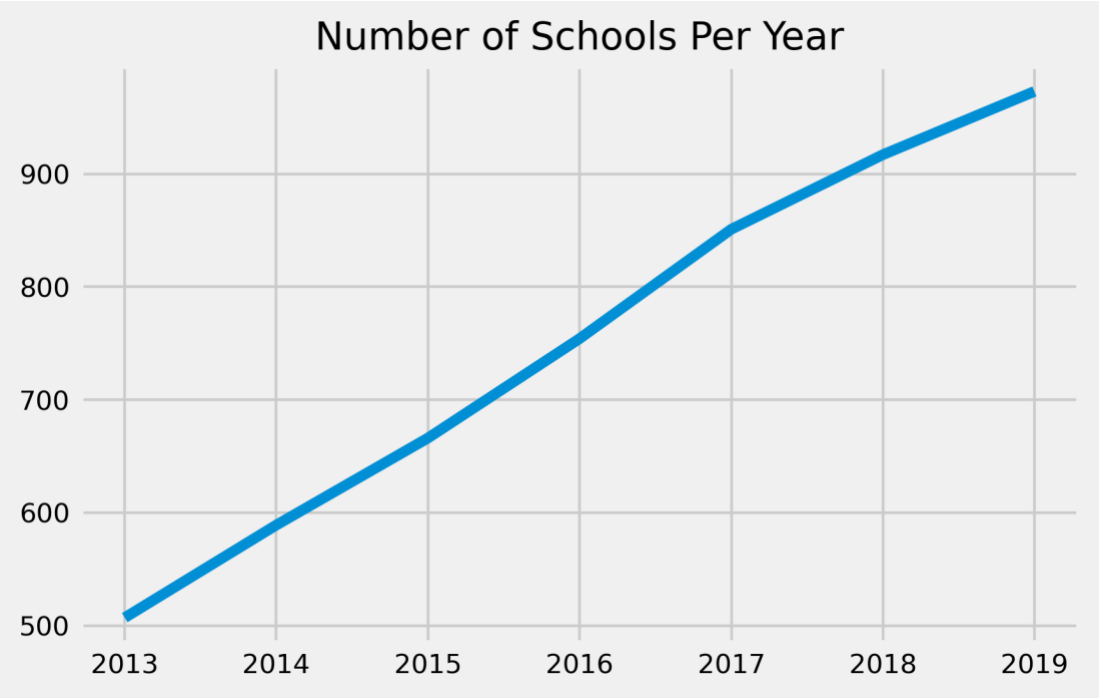

Figure 4: Trend of charter school

The number of charter schools in New York City has been on the rise. This could be probably due to their recorded successes or the dwindling belief in the public K-12 education system in New York City.

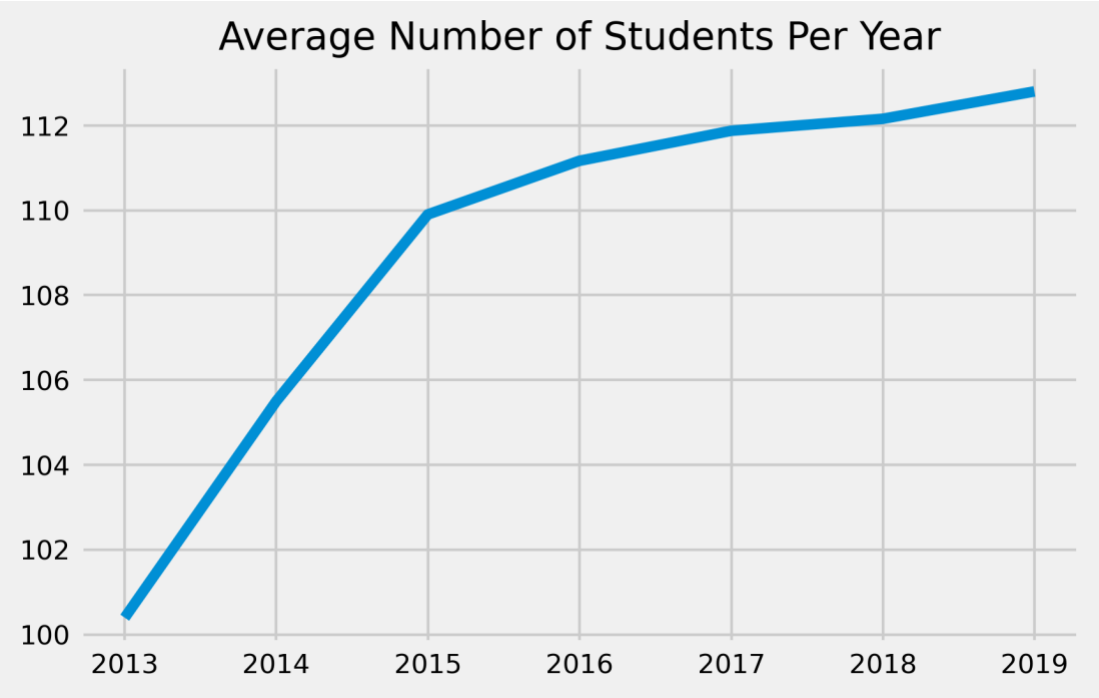

Figure 5: Average number of students in charter schools per year

It is expected that increase in the number of charter schools implies more students. This shows a recorded continuous increase in the number of students attending charter schools.

To check the correlation between variables:

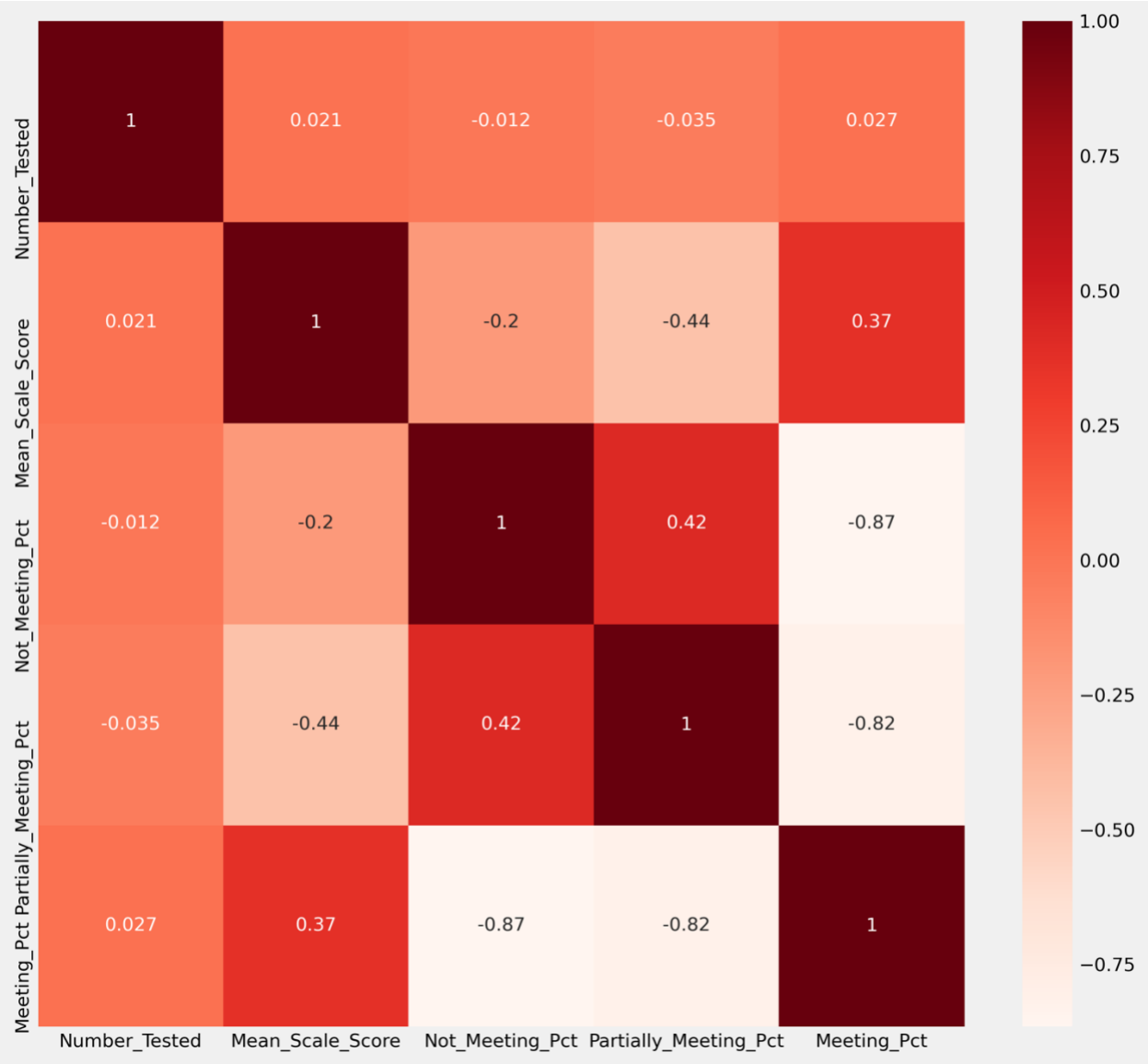

Figure 6: Correlation matrix of numeric features

The correlation matrix revealed the expected dependency in this scenario. As more schools have more students in the meeting learning standards category, they will have less percentage of students in the other two categories. This has also been shown by the histogram distribution above of the three features.

Olabisi Balogun
January 21st, 2021

Methodology

After much research into what algorithm would be suitable for the goal of this project- Predict Performance Levels of Charter Schools in NYC. The first algorithm that came up was the Multioutput Regression but then there is the 'time' component which is essential to be captured by the model. Finally, I concluded that the problem could be classified as a Multivariate Multi-step Time Series Forecasting. Data Preparation has to be in line with the algorithm to be used.

I looked into (Autoregression) AR models many of which are used for univariate. Although, literature suggest they could still be used to predict multi-output. I wanted a more efficient algorithm. Further research revealed Vector Auto Regression (VAR) model. After a good study on this algorithm and problems it has been used to solve, I resolved to using it for the problem at hand.

VAR takes into consideration the previous occurrence of variables (lag variables) and their dependencies with other features in the dataset when building the model. Here, only the target variables are needed in the dataset.

Implementation

Testing for autocorrelation of each target variable is one important test for using VAR using the lag_plot function from the Pandas.

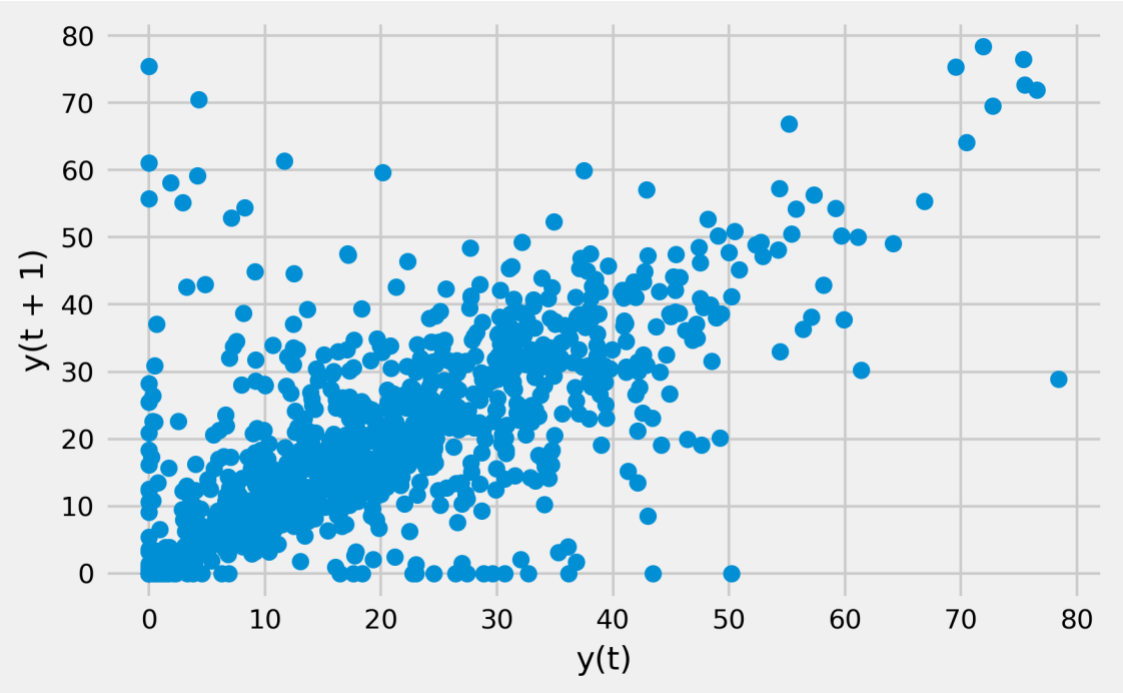

Figure 7a: Autocorrelation plot for 'Not_Meeting_Pct' feature

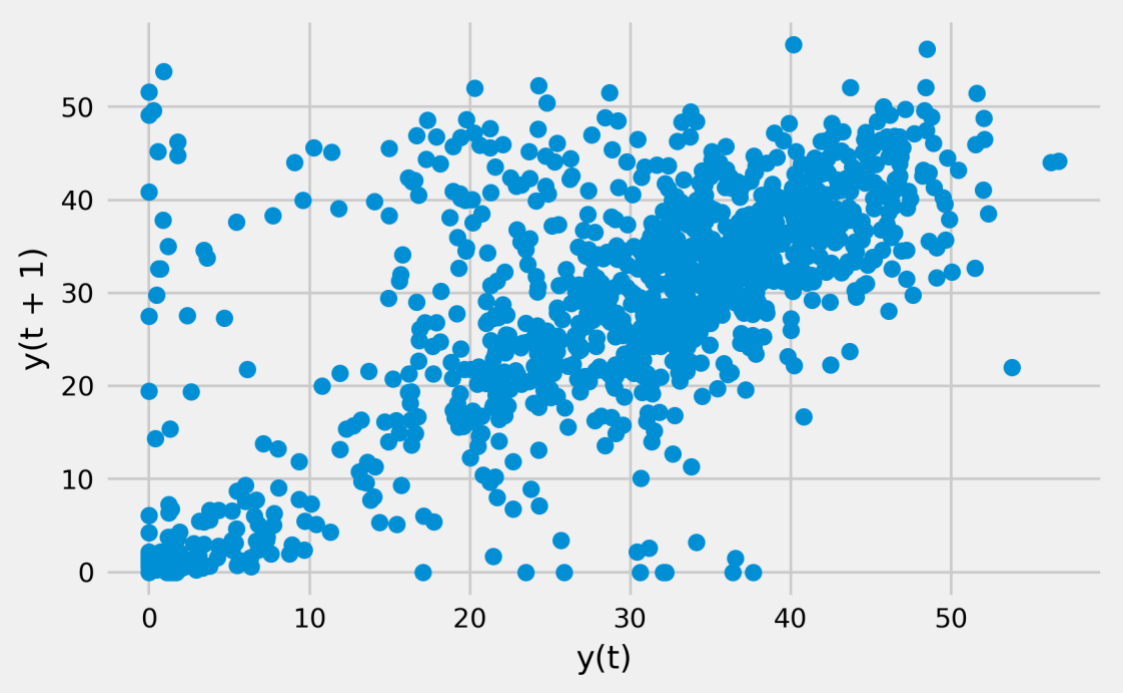

Figure 7b: Autocorrelation plot for 'Partially_Meeting_Pct' feature

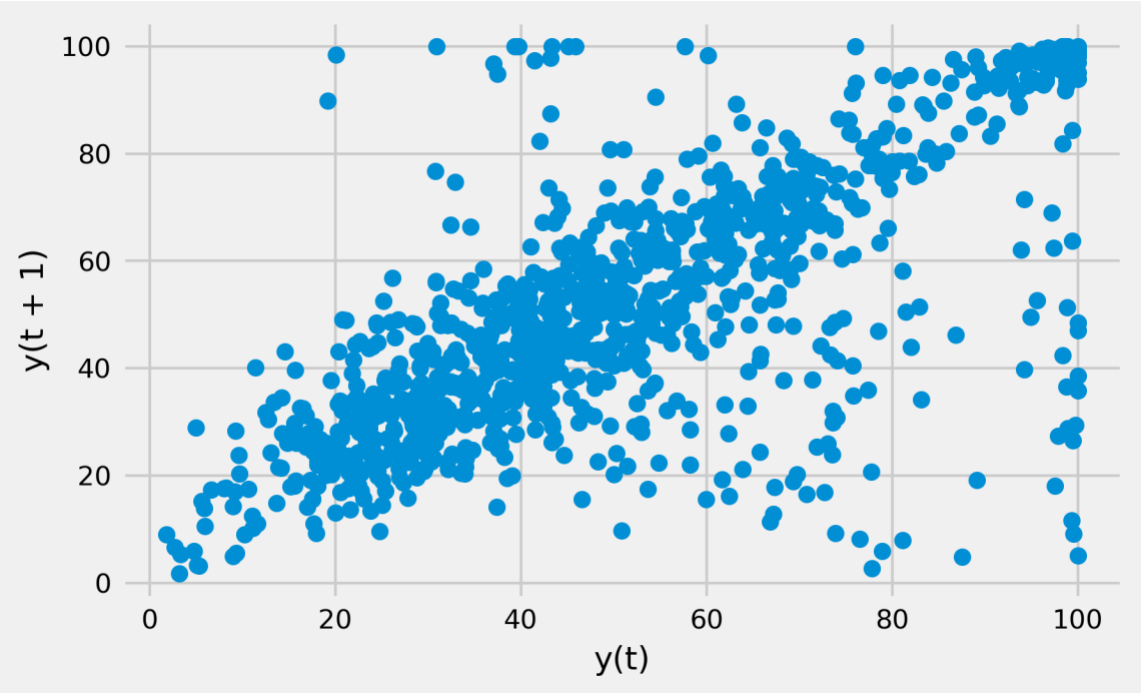

Figure 7c: Autocorrelation plot for 'Meeting_Pct' feature

The autocorrelation plots show there is a high correlation between previous values of each variables and their present values. A characteristic that shows VAR could work well with the dataset.

Data Preparation

A unique characteristic to the model building is building the model for each school because the school can be classified as 'different sites. The dataset was further grouped by school and date ('Year') feature was set as datetime index. There were scores for different grades from a school in a year, so I resampled the data to get the aggregate data per year for a school with multiple entries (from different grades). To train a model, I get the data for the school in context, split into training and test, then train the model on the training data for that school.

Olabisi Balogun
January 21st, 2021

Coding Complications

- During the training process, an error reported training data contains constant value. VAR algorithm do not work well with data that has not shown significant variability over time. I tried to introduce noise data with minimal effect to gain variability in the data, but it did not resolve the issue. I opted for the removal of schools whose training data has not shown variability over time.

- The Math scores were collected from multiple schools, so the prediction needs to be per school. For that I planned to create a model for each school, but some schools do not have enough data (schools with only a year data) to split into train and test for time series modeling so I removed them from the dataset.

Before building the model, to make sure each column data is stationary (stationarity is necessary for a time series model), I applied differencing to the training dataset. After building the models for the schools, I carried out Durbin Statistic test to check if the models captured the pattern in each training set of specific school data. Durbin Waston Statistic is test used to check if the time series model has captured the pattern in the dataset efficiently.

Olabisi Balogun
January 21st, 2021

school_id	not_meeting	partially_me	meeting_pct
84K333	1.32	0.93	1.22
84K355	0.71	0.62	1.17
84K356	1.17	0.7	1
84K357	1.76	1.11	2
84K358	1	1.07	1.95
84K360	1.91	2	1.16
84K362	2.15	0.63	1.67
84K367	0.52	0.28	1
84K379	0.45	0.43	0.58
84K406	0	0.08	1
84K508	1.07	1.2	1.12
84K517	1	0.79	1
84K536	2.95	1	2.95
84K538	0.06	2.72	2.41
84K593	0.36	1.11	1.8
84K608	1.03	1.59	1.19
84K626	1.16	1.06	0.85
84K648	0.78	1	1.17
84K649	1	1.15	0.68
84K651	2.62	2.77	2.8
84K652	1.88	1.89	1.78
84K701	0.05	1.58	0.13
84K702	0.38	1	1.46
84K704	0.39	2.4	0.44
84K707	1.06	1.47	0.43
84K710	1	1.42	1.44
84K711	0.7	0.72	0.31
84K712	0.93	1.54	1.56
84K724	0.2	0.53	1.47
84K730	0.77	1.05	1.12
84K731	1.38	0.75	1
84K737	2.23	1.15	1.67
84K740	1.47	0.74	1.39
84K742	0.71	1.91	1.18
84K744	0.58	2.59	0.28
84K746	1.71	1.23	1
84K757	1.66	1	0.53
84K758	1.85	1.98	1.8
84K769	2.64	2.68	2.63
84K774	0.16	0.36	0.49
84K775	2.26	1.75	1.35
84K777	1.38	2.07	1.83
84K780	2.21	0.35	0.62
84K782	0.84	0.93	1

84K785	1.24	2.44	1.18
84K791	2.59	2.6	1.53
84K792	2.58	0.98	1.26
84K793	0.48	1.73	1.95
84M065	1.55	1.27	1.41
84M068	0.56	2.67	0.16
84M202	1.32	1.98	1.98
84M279	0.41	0.4	0.84
84M284	0.83	1.35	2.34
84M320	0.78	1	2.62
84M329	2.02	1.47	1
84M330	1	0.17	1
84M335	0.8	0.45	0.89
84M336	2.71	1.9	2.02
84M341	0.67	1	0.91
84M350	1.18	1.1	1.06
84M351	2.14	1	1
84M353	0.57	1.26	1.6
84M382	0.68	1	2.5
84M384	1.11	0.54	2
84M385	1.86	0.36	1.28
84M386	1.95	1	2
84M388	2.33	0.86	1.73
84M430	1.36	0.75	0.86
84M478	2.82	0.16	0.14
84M481	0.17	1	0.89
84M482	1.77	1.7	1.39
84M483	0.81	0.98	1.01
84M518	0.4	0.2	0.53
84M523	1	1.89	1.8
84M702	2.64	2.5	2.22
84M704	0.68	0.81	1.16
84M705	1.39	1.77	1.76
84M708	1.26	1.09	0.79
84M709	1.42	1	1
84M726	0.03	1	0.36
84M861	1.16	1	1.16
84Q083	0.92	1.21	1.09
84Q170	1.8	1.69	1.6
84Q298	0	1	1
84Q304	1.12	0.99	1.38
84Q321	0.73	0.49	1
84Q340	1.82	1.94	1.26
84Q341	1.58	1.53	1.27
84Q342	1.53	1.55	1.02
84Q359	0.53	0.72	0.8
84Q704	0.99	1	0.84
84Q705	1.49	1.3	2.21
84Q706	1.62	0.1	2.18
84R067	1.74	1.89	2.12
84R073	2.04	1.57	0.89
84X133	0	0.47	0.2
84X165	0.14	1.32	2.15
84X177	2.78	0.87	0.76
84X185	2.67	1.05	1.58
84X255	2.75	1.82	0.66
84X309	1.3	0.17	1.64
84X345	2.99	1.93	1.29
84X346	0.62	1.7	1
84X378	0.4	0.89	0.78
84X389	0.32	1.35	0.76
84X394	0.73	0.95	0.8
84X398	0.87	0.65	1.26
84X407	2.65	2.31	2.7
84X419	1.66	2	0.62
84X422	2.93	1	0.89
84X461	1.27	0.68	0.61
84X482	1.11	0.63	1.31
84X487	2.77	1.86	2.8
84X488	1.33	1	1
84X491	0.08	2.5	1.4
84X493	1.15	0.1	0.06
84X494	0.96	2.83	1
84X496	1.68	1.44	1.98
84X538	1.95	1.98	1
84X554	1.47	1.99	1.6
84X703	0.33	0.48	1
84X704	0.76	0.7	1
84X705	1.3	0.81	1.78
84X706	0.87	0.66	1.08
84X717	2.6	2.6	2
84X718	1.03	1.26	1.12
84X730	0.74	1.36	1.46

Figure 8: Result of Durbin-Watson Statistic

The criteria for Durban Statistic states: value closer to 2 implies no serial correlation (model explained pattern in data well), value closer to 0 means positive correlation, and value closer to 4 means negative correlation. The result is quite mixed, some schools' model captured the time series pattern very well, while other schools' model did not. A keen observation of schools with 'perfect' model is that those schools had significant percentage in each level of performance in the training data. This again points out to how VAR model appreciates variability in the variables.

Forecast vs Actual

Since there are 127 models, one for each school. I picked three schools randomly to show their model prediction versus the actual values in the test data.

Schools with id: 84K333, 84M068, 84Q340 are chosen to plot their forecast and original values.

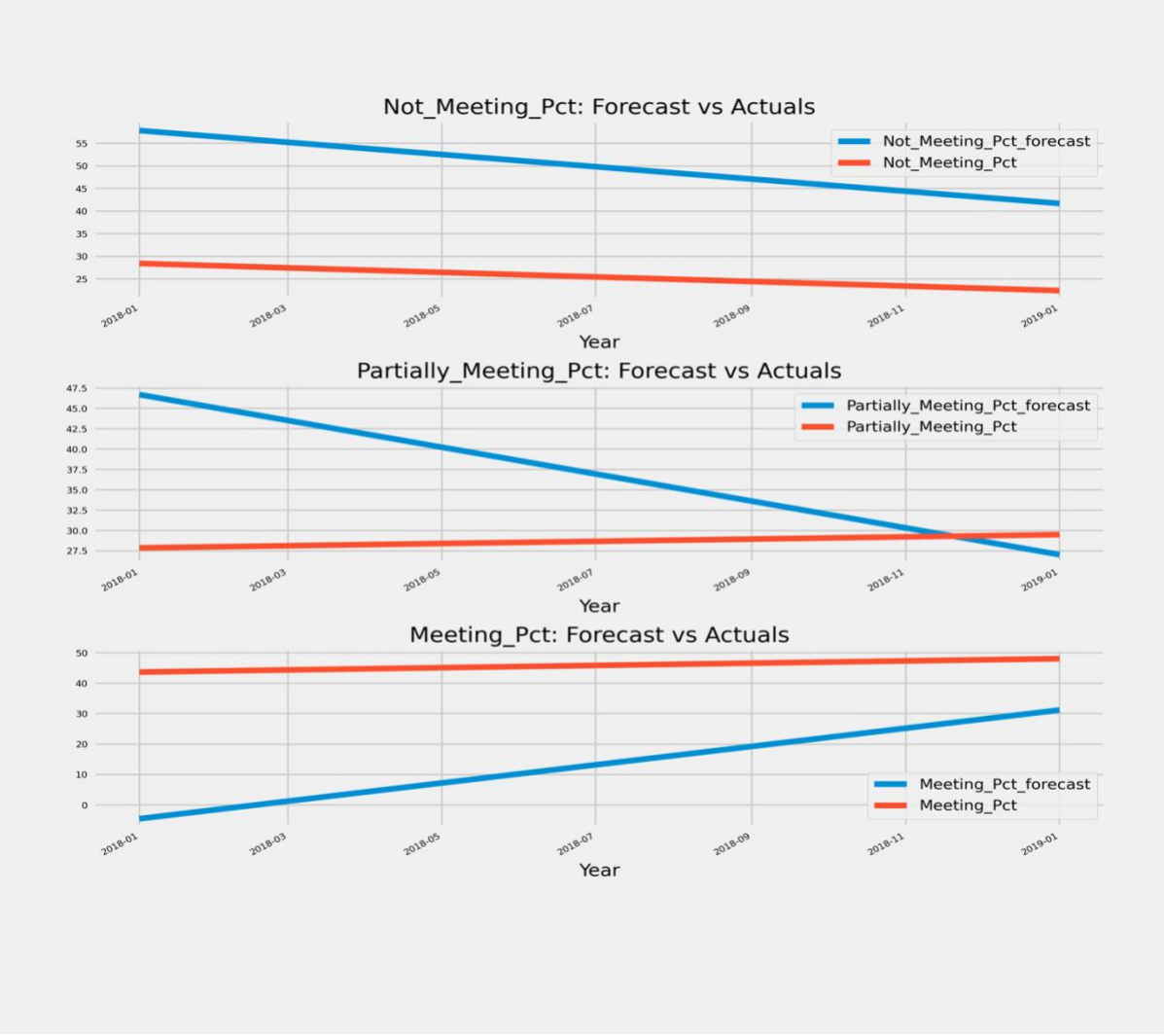

Figure 9a: Forecast vs Actuals: School id: 84K333

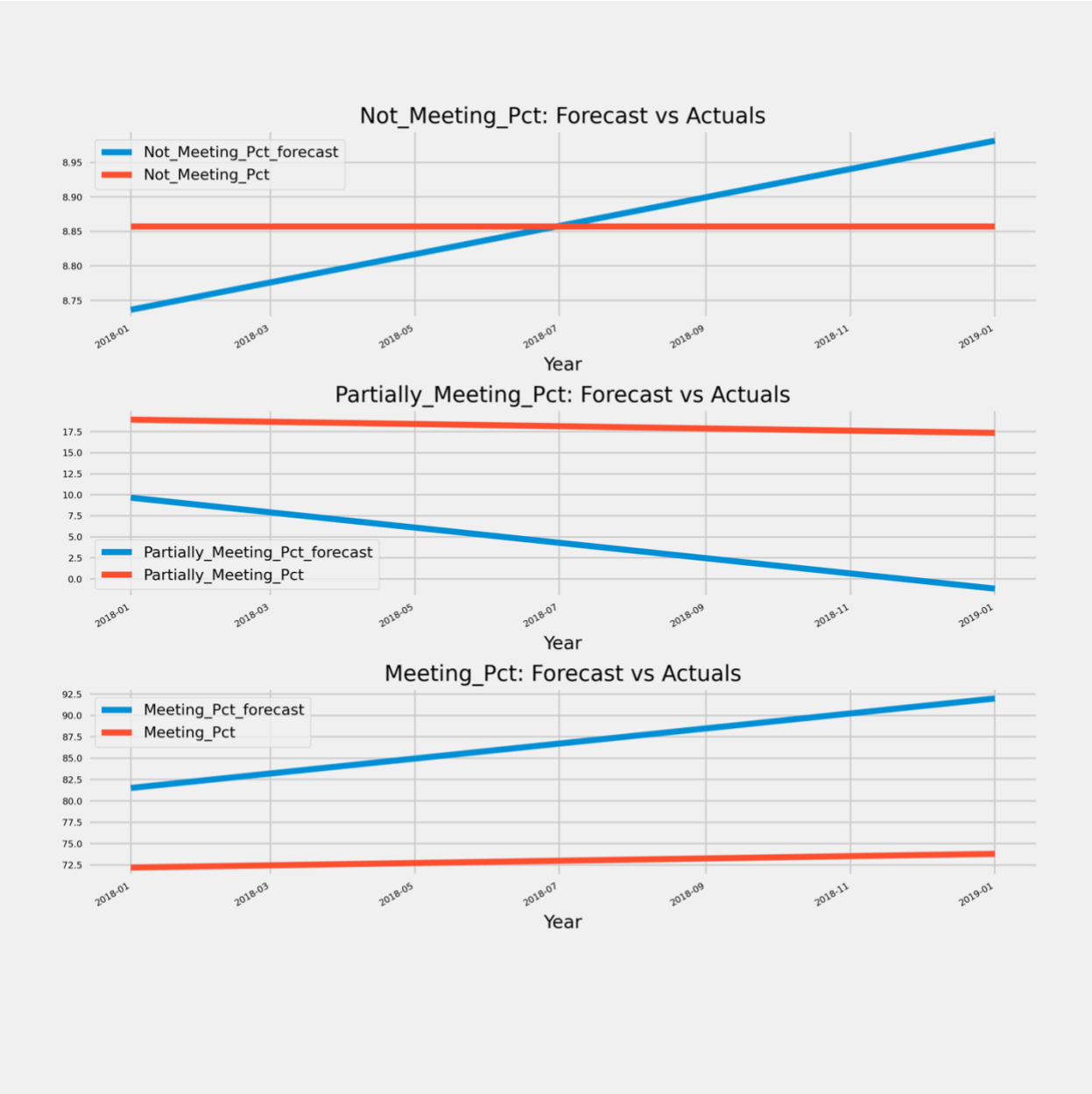

Figure 9b: Forecast vs Actuals: School id: 84M068

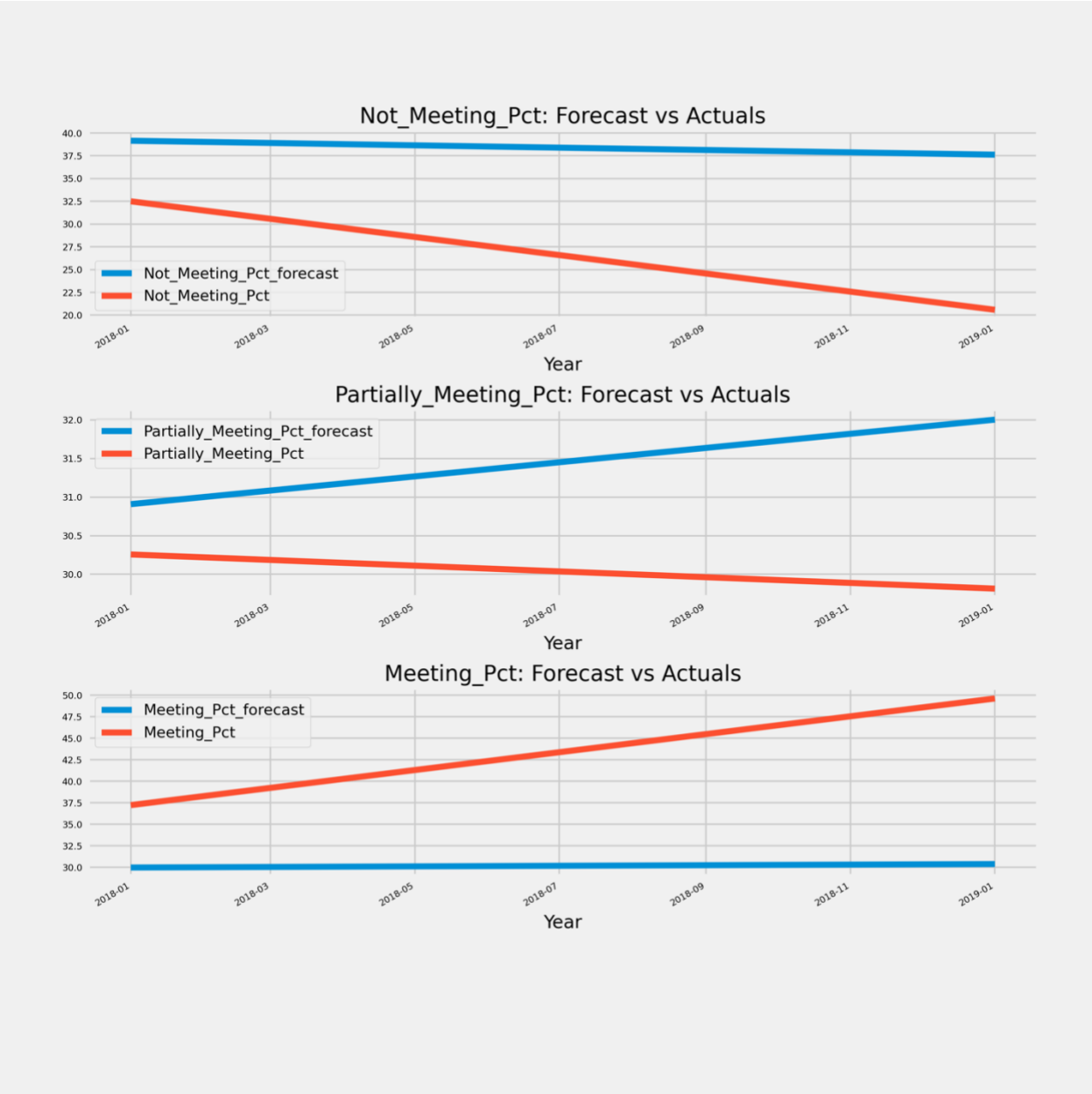

Figure 9c: Forecast vs Actuals: School id: 84Q340

Olabisi Balogun
January 21st, 2021

The model accuracy was evaluated using the rmse and mae. For the 3 schools above the model accuracy is given below. The rmse and mae for all the model can be gotten when you run train_model.py

```
Model Accuracy for school with id 84K333
Forecast Accuracy of: Not_Meeting_Pct
mae :  24.3261

rmse :  24.846

Forecast Accuracy of: Partially_Meeting_Pct
mae :  24.3261

rmse :  24.846

Forecast Accuracy of: Meeting_Pct
mae :  32.5108
rmse :  36.0781



Model Accuracy for school with id 84M068
Forecast Accuracy of: Not_Meeting_Pct
mae :  0.1226

rmse :  0.1226

Forecast Accuracy of: Partially_Meeting_Pct
mae :  0.1226

rmse :  0.1226

Forecast Accuracy of: Meeting_Pct
mae :  13.7478
rmse :  14.4459



Model Accuracy for school with id 84Q340
Forecast Accuracy of: Not_Meeting_Pct
mae :  11.8459

rmse :  12.9339

Forecast Accuracy of: Partially_Meeting_Pct
mae :  11.8459

rmse :  12.9339
```

```
Forecast Accuracy of: Meeting_Pct
mae :  13.2566
rmse :  14.5502
```

Refinement

To improve the accuracy of a VAR model, increasing the lag order is a one approach. The first round of training used lag order of 1. So I tried lag order of 2 and there was a reduction in the rmse of the models presented above.

```
Model Accuracy for school with id 84K333
Forecast Accuracy of: Not_Meeting_Pct
mae :  13.3544

rmse :  13.817

Forecast Accuracy of: Partially_Meeting_Pct
mae :  13.3544

rmse :  13.817

Forecast Accuracy of: Meeting_Pct
mae :  21.4626



Model Accuracy for school with id 84M068
Forecast Accuracy of: Not_Meeting_Pct
mae :  3.799

rmse :  3.9643

Forecast Accuracy of: Partially_Meeting_Pct
mae :  3.799

rmse :  3.9643

Forecast Accuracy of: Meeting_Pct
mae :  11.4461
rmse :  11.9243



Model Accuracy for school with id 84Q340
Forecast Accuracy of: Not_Meeting_Pct
mae :  3.9812
```

```
rmse :  4.4448

Forecast Accuracy of: Partially_Meeting_Pct
mae :  3.9812

rmse :  4.4448

Forecast Accuracy of: Meeting_Pct
mae :  3.6737
rmse :  4.4967
```

Model Evaluation and Validation

The model accuracy score improves with the number of lags. The optimal lag order for training the models is 2 as any number above that would not produce useful result. The plots of forecast vs actuals values of the target variables presented above showed linearity, so for I decided there was really no need for differencing. For future forecasting, I used the whole dataset for each school and forecasted future values. The forecast is plotted as a visualization on the web application, which can be run from *charter.py* file on the terminal.

Justification

At the initial stage of the project, I considered using Multioutput Regression model. But after researching and gaining more understanding of my project scope, I realized the presence of time and the prediction based on time. VAR is a time series model, and it takes the dependency of past events and other variables as a feed to the occurrence of the next timeline. It captures the concentration of time in the project. VAR is one of the few time series algorithms that can predict multivariate target variables. The introduction of noise data was very helpful in making the model robust to withstand constancy in the training data

Olabisi Balogun
January 21st, 2021

Conclusion

The aim of the project is to forecast the trend of three performance levels in Math in NYC Charter Schools- Not Meeting Learning Standards, Partially Meeting Learning Standards, and Meeting Learning Standards. To achieve this, I built a web application for displaying the trend and future prediction of schools in the three learning standards. A school is selected, then three charts, each representing one learning standard is displayed. The model does a pretty good job at forecasting as the forecast lines depict a reasonable continuation to the observed trend. Schools with higher level in the meeting standards chart have corresponding lower levels in the other two categories and vice versa. The application gives a holistic overview of how students in NYC charter schools are doing in Math. It could also serve as an alert for poor performance.

This project could help educators and superintendent of schools initiate early intervention in schools with high percentage of students falling behind. School improvement could be measured and schools exceling could serve as a learning model for other schools in the district. The complete code for this project- data cleaning, analysis, model training, and evaluation could be found at: https://github.com/Olabisi-Balogun/School_Performance_Prediction

Olabisi Balogun
January 21st, 2021

Improvement:

- Getting more data for each school would improve the accuracy of the models' forecasting.
- Trying another multivariate time series model- Vector Autoregression Moving-Average (VARMA), an extension of VAR, to see if better forecasting and higher model accuracy can be achieved.
- While I was researching a model that fits the problem, I came across the use of deep learning methods in multivariate forecasting. My knowledge for deep learning is currently limited so this is an area I would like to explore with the problem. Using appropriate deep learning algorithms to carry out the forecasting.

Olabisi Balogun
January 21st, 2021

References:

Bulkley, K., & Fisler, J. (2003). A decade of charter schools: From theory to practice. *Educational Policy*, *17*(3), 317-342.

McKinney, W., Perktold, J., & Seabold, S. (2011). Time series analysis in Python with statsmodels. *Jarrodmillman Com*, 96-102.

How to Develop Multivariate Multi-Step Time Series Forecasting Models for Air Pollution https://machinelearningmastery.com/how-to-develop-machine-learning-models-for-multivariate-multi-step-air-pollution-time-series-forecasting/

A Multivariate Time Series Guide to Forecasting and Modeling (with Python codes) https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/

An End-to-End Project on Time Series Analysis and Forecasting with Python https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/

Olabisi Balogun
January 21st, 2021

VAR Prediction on Coronavirus (Italy) https://www.kaggle.com/sagivmal/var-prediction-on-coronavirus-italy