Using Transfer Learning to Address Sustainability Issues

Vaibhav Gupta, Ally Racho, Divya Rustagi

Pennsylvania State University – DS440: Capstone Project

ABSTRACT

The team was assigned with the task of using transfer learning to address sustainability issues. Air quality has become more evident in recent years due to climate change and pollution becoming monitored more. The EPA tracks many cities' air quality index (AQI) and keeps historic records of the data. There are data insufficient cities however that the EPA does not keep information on. To overcome the issue of data insufficient cities and the lack of AQI data, we proposed a neural network which learns on data connected to AQI for a source city – Manhattan, NY – and then transfers the knowledge to target cities – Chicago, IL and Philadelphia, PA – and predicts the AQI measurements for a given week and year between 2018 – 2022. It was found that normalized RMSE for the Chicago model was 0.224 and the Philadelphia model was 0.200. Given these results, our models are predicting relatively accurately. We also noted that the COVID-19 pandemic may have been a factor in the accuracy of our model for years 2020 and 2021. A Streamlit web application was created to demonstrate the model's findings and allow for user-interaction with the model.

OBJECTIVES

1. Create an efficient and effective transfer learning neural network with factors connected to air quality to predict the weekly AQI values of target cities
2. Accurately predict the weekly AQI values of the target cities
3. Deploy the model to a website which includes different functionalities and user-input

CHALLENGES

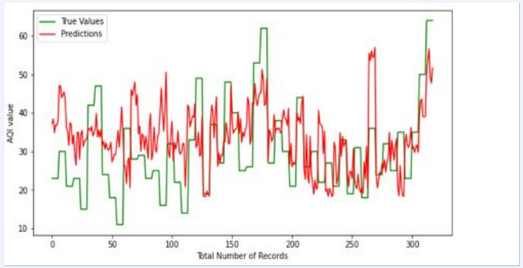
1. Some target cities were too data insufficient thus the model was not able to predict the AQI values accurately.
2. Huge datasets caused issues with RAM and time-efficency.
3. New methods and technology had to be learned and understood prior to beginning the project.

METHODS

The New York dataset was first split into train and test sets. We used 2018-2020 as our training data and 2021 for the testing. It was then scaled using Standard Scalar. Three models were used – RandomForestRegressor, SVM and Custom Neural Network. The metrics used and given scores for each are presented below with the best outcome of the test data bolded.

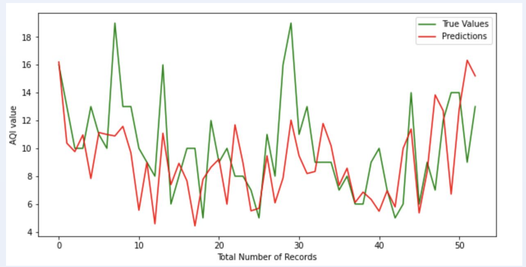
Random Forest Regressor	R2	3.9
SVM	R2	-18.8
Custom Neural Network	**MSE**	**1.1399**

The transfer learning model was created using the neural network as it produced the best results. Below is an image of the true values (green) to predicted values (red) of our transfer learning model tested on 2021 NYC data. The normalized RMSE for the model was found to be 0.201.



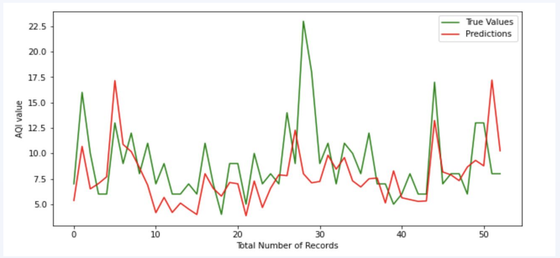
From this model, we transferred the knowledge learned to our target cities of Chicago and Philadelphia. Below are the plots of true values to predicted values found for each.

Chicago plot of true to predicted values for the year 2021:

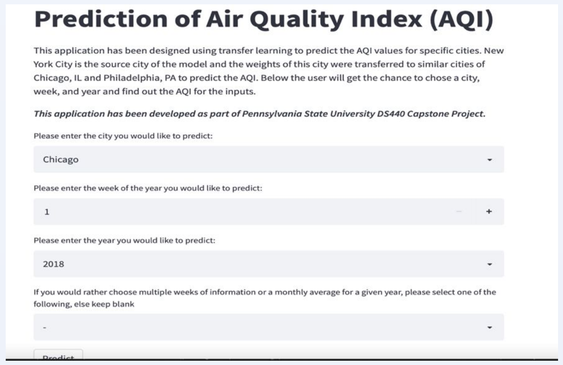


METHODS CONT.

Philadelphia plot of true to predicted values for the year 2021:



A Streamlit app was then developed to allow for user-interaction with the model and allow them to decide which city, week, and year they would like to predict the AQI value for. Other functionality is included as well. Below is a screenshot of the dashboard where users are can make certain inputs and then find out the predicted values along with geocoordinates of where the AQI was found, what the predicted and real AQI values actually mean, plot of real to predicted values for selected year, and features used in the modeling .



RESULTS

The normalized RMSE found for the target cities was calculated. For Chicago, the model was found to be 0.247 and for Philadelphia it was 0.200. Generally, an normalized RMSE found between 0.2-0.5 indicates the model is predicting relatively accurately. We also noted that COVID-19 may have been a factor with the models ran on the synthetic data used for the web application. For example, we observed our model predicted higher AQI values for later weeks in Philadelphia in 2020.

RESULTS CONT.

We believe our model's predictions are higher due to what the model was trained on and appear inaccurate because of the COVID-19 pandemic and it's aid in improving air quality. Including more data of pre/post pandemic to train the model on may be beneficial in improving the overall accuracy in the future.

CONCLUSIONS

We were successful in our tasks and were able to create a neural network based on New York City data which predicts the air quality of a specific week and year and transfers it to the cities of Philadelphia and Chicago. We realized later the transfer of the model to Hazleton would not be feasible as there is no documented air quality index history for the city, causing the model to not be runnable due to the extremity of the data insufficiency and inability to verify any predictions. Future work in adjusting and optimizing the model to be able to handle these types of cities would be vital. We plan to also see if the model can still be enhanced and predict values even more accurately. Additionally, we successfully created a front-end dashboard which allows for user input of the city, week, and year and outputs the AQI measurement calculated from the model. Having this web application allows us to expand our audience to others who are interested in air quality information and would like an easy to use, comprehensive application. The website also completes the project and ties all the information together in one spot.

REFERENCES

Gupta, V., Racho, A., Rustagi, D. (2022) DS440-Transfer-Learning-Address-Sustainability-Issues [Source code] https://github.com/vbgupta/DS440-Transfer-Learning-Address-Sustainability-Issues