Video Transcript for Group Project - Developing Machine Learning Models for Predicting Atmospheric Emissions

Intro slide: By Alexandra Settle, Devin van Rooyen, Hammed Arowosegbe, Yao Kwadzo

Project Introduction (45 seconds)

For this project, we - the team - have been engaged as contract data scientists by Athana Data Science Services which is a small company specialising in the provision of data science consultancy services to public and private sector organisations. We were provided with datasets from the London Datastore and expected to perform an in-depth analysis of atmospheric emissions from 2013. The aim was to decide upon 3 recommended machine learning algorithms to model to predict future atmospheric emissions in London and specify any predictive insights gathered from the data.

This video exhibits the steps taken to determine the machine learning algorithms, to develop the machine learning models and to predict insights. The developers used Jupyter notebook and Scikit-Learn Python machine learning frameworks and its components to code and execute the machine learning algorithms. Lastly, an assessment on performance was conducted of each model using a machine learning matrix and this video report explains differences we found in each model's performance.

Understanding the process (1 min 10 sec)

Looking at machine learning algorithms, there is no one solution or one approach that fits all. There are several factors that can affect your decision to choose a machine learning algorithm. The type and kind of data we have plays a key role in deciding which algorithm to use. Some algorithms can work with smaller sample sets while others require multiple samples. The team went through the following steps to ensure the most accurate machine learning models were chosen:

1. Data cleansing
2. Data augmentation
3. Problem categorisation

To ensure the data was easily interpretable, data processing was conducted to ensure the data set was scrubbed appropriately for machine learning. The team conducted a review of the data for accuracy, consistency, uniformity, completeness and validity and found duplicated or missing values in numerous columns and rows. We'll show an example of the data cleansing later in this video.

Next, the team looked at augmenting the data. Data augmentation looks to increase the amount of data by adding modified copies of existing data sets. This acts as a regulariser and assists in managing the overfitting of data.

Lastly, our developers worked on problem categorisation. Problem categorisation arranges the data into classes, or categories, to understand what type of problem is occurring to help identify trends. We'll show some examples of this later in the video.

GitHub Repository (40 seconds)

Source control was an important element to this assignment as it required active collaboration from 4 individuals. For this purpose, a GitHub repository was used to collaborate on the code, testing, and technical documentation.

The repository contains a license - Apache 2.0 -, a README file with instructions on the project premise and getting started instructions. Within the repository there is also a source and documentation directory. The source directory contains the Jupyter files with data preprocessing and the models chosen, as well as the CSV file with the data set. The documentation folder contains information on contributing to the repository.

Data Processing (45 seconds)

As mentioned in the introduction, the first step in this project was to review the data from the London Datastore and perform a data quality assessment. This assessment is a scientific and statistical evaluation of the data to determine whether the quality of data is suitable to train effectively and efficiently.

As you can see from the example video, to prepare our dataset for machine learning work, an emphasis was placed on editing the data type format, removing irrelevant and remapping certain columns, and data encoding. Columns with number entries were properly formatted to floating point numbers, column labels were renamed for clarity, and other columns were removed. To predict the target variable, the columns were assessed based on their data quality fit. For example, pollutant versus the type of pollutant emitted from the vehicles. And finally, data was encoded accordingly. Entries were converted to be numeric feature points for machine learning work.

What are the predictive insights (20 seconds)

Predictive insights analyse data and learn predictions about future outcomes and performance of a given topic. After cleansing the database, an initial correlation matrix shows a clear relationship between the total emissions and the type of vehicle. The correlation matrix looks like this [matrix].

By aggregating the data by borough and pollutant, the input variables were identified as borough, pollutant, road length, and road traffic stress, and the output variables as total emissions.

Developing the ML models (1 min 30 sec)

Next, we defined the machine learning algorithms we wanted to use for our project based on the predictive insights.

Ordinary least squares, or OLS, is a method for estimating the unknown parameters in a linear regression model. It works by minimising the sum of squared vertical distances - dependent variable Y and independent variable X - between the observed responses in the dataset and the responses predicted by the linear approximation. We chose to use ordinary least squares as our first model because we wanted a simplier technique to model a linear relationship between our input variables - Borough, Pollutant, Road Length - and a continuous numerical output variable - total emissions.

Next, we selected Gradient Boosting Regression, an ensemble technique, because we want to handle multicollinearity and non-linear relationships. Multicollinearity is when several independent variables in a model are correlated and thus produce less reliable statistical inferences. Gradient boosting regression can also be used for classification and regression. In gradient boosting, each predictor corrects its predecessor’s error. It works by finding nonlinear relationships between the model target and features.

Lastly, we picked support vector machine, or SVM, to capture more complex relationships between our datapoints without having to perform difficult transformations on our own. Support vector machines are a set of supervised learning methods used for classification, regression and outliers detection. The objective of the support vector machine algorithms is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

Performance assessment of the ML models (30 sec)

There are various ways to evaluate a machine learning model's performance. Performance metrics are a part of every machine learning pipeline. The first metric we used is the the determination coefficient of R squared. R squared is the most important index for verifying the accuracy of the predicted result of a regression algorithm, and it has a range of 0 and 1. It gives a clue of how well the trend of the model results is able to track the trend of actual data with a normalised value. The larger the value, the more accurate the predictions. R2 value of 1 would mean that the regression model makes predictions without any error.

The second metric we used is the Mean Absolute Error (MAE). This metric evaluates the absolute error between predicted value and actual value. Where n is the number of samples, yi is the actual value (real Total Emissions), and yˆi is the model predicted value, which is the predicted Total Emissions of the ith sample. The smaller the MAE value the more accurate the predictions.

Our conclusions showed that even though SVM has slightly lower Mean Absolute Error than GBR, overall, GBR seems to be produce the most accurate model.

Summary (1 min)

The goal of ordinary least squares was to provide insights into cause–effect relationships. One of the challenges that the team faced was choosing the wrong independent feature. Even though most of our features were suitable, the model was confused by poorly selected features (for example, road length) that added a high signal-to-noise ratio to the learning process.

Support vector machines was initially the preferred method as it yielded notable accuracy with less computational power. However, two identified limitations when training and testing the data was the fact that support vector machines did not handle large data sets well, and as such the training run did not run very well.

Gradient Boosting Regression is a relatively easy-to-use algorithm. Notwithstanding its resilience, it is very sensitive to outliers because each classifier is indulged to correct the inaccuracies in the predecessors. The team noted the limitation and was fortunate enough to not have encountered such challenges. However, of the three methods, the team noted that gradient boosting regression performed better than ordinary least squares and supported vector machines.

Image References

Commentary, G. (2020). Governor, legislators helping to ensure equity to computer science education in California. CalMatters. [online] 2 Apr. Available at: https://calmatters.org/commentary/my-turn/2020/04/governor-legislators-helping-to-ensure-equity-to-computer-science-education-in-california/ [Accessed 1 Aug. 2022].

Opiyo, T.C. and B. (2021). Best Practices When Training Machine Learning Models. [online] Contract Engineering, Product Design & Development Company - Cardinal Peak. Available at: https://www.cardinalpeak.com/blog/best-practices-when-training-machine-learning-models [Accessed 1 Aug. 2022].

Ravindran, B. and Ghose, A. (n.d.). Interpretable models | Robert Bosch Center for Data Science and Artificial Intelligence. [online] rbcdsai.iitm.ac.in. Available at: https://rbcdsai.iitm.ac.in/blogs_tags/interpretable-models/ [Accessed 1 Aug. 2022].