**CONTEXT**

Bike riding has been a very important section of urban transport due to its ability to contribute to its fast, sustainable, flexible and cost-efficient transportation. While bicycle rides contribute to healthy and active lifestyle, there is a sudden push from the local government authorities, that boosts its citizens to commute using bikes. This has been implemented using various types of policy measures including the construction and improvement of bicycle lanes and proper bike parking facilities. While boosting bike sharing in cities being one of the major improvements over the last decade, the introduction of e-scooters and e-bikes is another factor that increases the attractiveness of such plans. However, in order to understand the bike flows graphs, there is a need to focus on external factors like current bike flows and what factors influence rides and riders such as weather and safety of such bike rides. While there has been a recent spike in bicycle sharing during the pandemic because of obvious reasons, this analysis aims to look into the prediction of bicycle ride sharing based on multiple factors for the bike ride sharing scheme by the Transport for London (TFL) in the city of London.

**DATA**

The dataset is obtained from the Transport for London (TFL) website. The data was exported in an .xlsx file which had multiple datasets compiled into one. The data had daily usage data, monthly usage data, yearly usage data and percentage of usage variation for bikes of the TFL (often known as the Santander Scheme) operating in the City of London. Even though there were other sets of data present, the daily usage data was considered for the analysis. The reason for using this is because it lets us have the large number data points which is the fundamental for any data science analysis.

Another dataset has been considered for this analysis – the power generation dataset. The data was downloaded from the website of British Electricity Generation and it was used to see if Bike Rides from the TFL schemes are affected by power generation in the UK. The UK holidays and historical weather data were also used as features for the prediction. For the historical weather data, the data was taken from a website [www.worldweatheronline.com/](http://www.worldweatheronline.com/) using the API from the website. The data license for the API was purchased for personal usage and has been used here for obtaining the weather data. The historical data weather data for the dates in the analysis has been saved as ‘london.csv’ file and have been attached in the submission.

**DATA IMPORTING AND CLEANING**

Data Cleaning refers to the process of correcting bits and parts of data to ensure we achieve high data-integrity. For this analysis, multiple steps were undertaken, such as importing the raw csv files of all the relevant datasets and then filtering them on the basis of the features that were chosen to be important for the analysis. The details of the importing and the cleaning process has been mentioned in the jupyter notebook.

**EXPLORATORY DATA ANALYSIS**

One of the most important things for any machine learning analysis is the Exploratory Data Analysis. EDA refers to the most critical process of machine learning – to gauge the data, perform initial investigations, discover trends, spot anomalies and to check patterns using statistical summaries and graphical representations. Similarly for this data, an intensive EDA was performed to gain insights before getting our hands dirty with it.

All the EDA that has been used has been enumerated below with their respective observations and the business perspective associated with it:

**Analysis 1:** Mean rides over the years

**Observation:** While there has been a steady increase in the number of bike rides for the first few years, the number of rides has almost reached a saturation point over the recent years. The average number of hires for the year 2021 has been low because of the fact that the analysis uses data from the first couple of months of the same year.

**Business Perspective:** TFL bike is a bike sharing scheme which uses a dock for bike parking.

While the usage has reached a saturation point, an increase in number of bike dock stations might increase the number of bike rides.

**Analysis 2:** Mean rides over the months.

**Observation:** Most rides are during the summer months which is very intuitive.

**Business Perspective:** Bike maintenance and bike relocation should be given priority during these months.

**Analysis 3**:  Mean rides based on weekdays and weekends

**Observation:** Average rides are higher during the weekdays.

**Business Perspective:** Bike maintenance could be done on the weekends because of lesser usage.

**Analysis 4**:  Mean rides based on days of the week.

**Observation:** Tuesdays, Wednesdays and Thursdays have the highest mean usage of bikes.

**Business Perspective:** A further analysis of where these rides originated and ended, could provide more insight of usage patterns and could let analysts identify areas of more usage on certain days of the week.

**Analysis 5:** Bike rides over the month compared on the basis of Weekdays and Weekends

**Observation:** Difference of bike hires on weekdays and weekends over the months of the year varies.

**Business Perspective:** A further analysis of where these rides originated and ended, could provide more insight of usage patterns and could let analysts identify areas of more usage on certain days of the week.

**Analysis 6:** Distribution of Rides

**Observation:** Daily rides crowd cluster around 20000 to 35000 per day.

**Business Perspective:** Provides an insight on the daily revenue of the company

**Analysis 7:** Bike rides over the week compared on the basis of the months of the year

**Observation:** Fridays, Saturdays and Sundays have lower rides irrespective of the month of the year.

**Business Perspective**: A further analysis, decomposing the data on more granular pattern, like the hourly usage could provide more insight on the usage patterns**.**

**MACHINE LEARNING**

Now that the data cleaning and EDA has been done, the next stage would be using the data to train and test a machine learning model. From a high level, it’s evident that bike rides depend on multiple factors, such as weather, time, day of the week, month, etc. The primary motivation for this machine learning analysis is to predict the number of bike rides, primarily using the TFL data and then using other datasets such as the holiday dataset and historical weather dataset to see if the data are correlated and if they could be used as features for prediction. We are particularly interested to see if the bike prediction depends on the power generation dataset.

**ONE HOT ENCODING**

One Hot Encoding is generally used to deal with categorical variables in the data. The word encoding refers to representing each piece of data so that a machine learning algorithm can understand. The year feature has been one-hot encoded because it’s value ranges between 2010 and 2021. As the numerical value of the year feature increases every year (almost after every 365 rows), it should be noted that this might affect the machine learning model. To ensure that every single entry in the ‘year’ feature has equal weight on the machine learning model, we had to encode it.

**MODELING**

**DATASETS**: Machine Learning has been used in the next section to predict the number of bike rides. 4 different cases have been considered as per the analysis demand. The datasets used for the prediction are as follows:

* TFL data
* TFL data merged with weather data and holiday data
* TFL data merged with power generation data and holiday data
* TFL data merged with weather data, power generation data and holiday data

**ALGORITHM**: While the objective of the modelling exercise is to predict bike hires, we will be using regression-based analysis techniques. 5 different machine learning models have been used for this analysis:

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor
* ADA-Boost Regressor
* Neural Network

Finally, ARIMA and SEASONAL ARIMA (SARIMA) has been used for the prediction of Bike Rides as part of the time-series model. ARIMA stands for Auto-Regressive Integrated Moving Average while SARIMA is the acronym for Seasonal Auto-Regressive Integrated Moving Average. ARIMA and SARIMA are well identified methods used for time series prediction such as this bike sharing problem which has index as the date.

While other algorithms also could have been used for this analysis, time constraint has been a prime reason for applying these 5 regressions and 2 time-series analysis techniques.

**TRAIN AND TEST**

The data has been split into train and test datasets with 60% data used for training and 40% used for testing purpose.

**METRICS**

The evaluation metrics of the model are enumerated at the end of training and testing each machine learning model. The metrics considered for this analysis are Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and Mean Absolute Percentage Error. The reason for choosing these is because they are industry standard metrics used to evaluate performance of models. Finally, a scatter plot has also been shown as a metric for each machine learning model which lets the viewer recognise the efficacy of the model. The closeness of the scatter between the test value vs the predicted value and *y=x* line shows how efficient the model is (*ref* : jupyter notebook).

**RESULTS**

The results of different models have been compiled with the metrics chosen and have been elucidated below. The models were used on 4 different datasets –

1. TFL data
2. TFL data, weather data and holiday data
3. TFL data, power generation data and holiday data
4. TFL data, power generation data, weather data and holiday data.

From the results below, it can be seen that the ADA Boost performs best among all the regression techniques used for the predicting the number of bike rides for the cases (**a**) – (**d**). This is because Adaptive Boosting algorithm is an ensemble technique which combines multiple weak models into a strong model. ADA Boost sequentially grows decision trees as weak learners and punishes incorrectly predicted samples by assigning higher weight to them after each round of prediction. One of the major disadvantages of this ADA Boost is that it takes a large computational time.

For the sake of brevity, the results from the ADA-Boost model for the datasets (**a**) to (**d**) would be focused on.

1. **Model with only TFL Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **MAPE** |
| Linear Regression | 6448.5 | 66507327.4 | 8155.2 | 31.5 |
| Decision Tree | 4069.3 | 30027910.1 | 5479.7 | 21.2 |
| Random Forest | 4000.8 | 29681972.9 | 5448.1 | 21.0 |
| **ADA Boost** | **4045.3** | **27243208.1** | **5219.5** | **20.2** |
| Neural Network | 4338.6 | 34352519.6 | 5861.1 | 22.6 |

**Discussion**: While using only the TFL data in case (**a**), we can see that the model could not be generalised well and hence a percentage error of 20.2 could be seen on daily prediction of bike rides. This is because, this dataset only uses features like day of the week, month and year. While these features are very important for predicting the number of bike rides, this dataset doesn’t include the most important factors that affects the bike ride – weather and holidays.

**b. Model with only TFL Data, Weather Data and Holiday Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **MAPE** |
| Linear Regression | 3553.3 | 24499071.4 | 4949.6 | 19.1 |
| Decision Tree | 4136.5 | 29763984.5 | 5455.6 | 21.1 |
| Random Forest | 3254.8 | 18612487.5 | 4314.2 | 16.7 |
| **ADA Boost** | **3303.6** | **18012943.5** | **4244.1** | **16.4** |
| Neural Network | 3669.1 | 25677296.8 | 5067.2 | 19.6 |

**Discussion**: For case (b), it can be seen that as the weather data was added into the dataset along with the holiday data, there has been a considerable jump in system accuracy. The errors decreased by approximately 4%, validating the need of weather data for the bike prediction problems.

**c. Model with only TFL Data, Power Generation Data and Holiday Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **MAPE** |
| Linear Regression | 5493.2 | 51081604.7 | 7147.1 | 27.6 |
| Decision Tree | 4455.9 | 35599724.2 | 5966.5 | 23.1 |
| Random Forest | 4090.9 | 31501625.3 | 5612.6 | 21.7 |
| **ADA Boost** | **4199.3** | **30303117.2** | **5504.8** | **21.3** |
| Neural Network | 7595.1 | 95879133.7 | 9791.7 | 37.9 |

For case (c), as similar to case (a), running the model without the weather features reduces the accuracy. In this model, the weather data was removed and instead the power generation dataset was included in the model to specifically answer the question - “**Can national electrical power generation help estimate how many bikes are hired?**” While its very intuitive to assert that a country’s power generation numbers might not affect its bike prediction numbers, we did a heatmap correlation to weigh the correlation of the feature set. While most of the features were not correlated to the number of bike rides, few of them had a correlation (very slim) and was considered for modelling. From the results, it is evident that the features were definitely correlated by chance as including them did decrease the overall performance accuracy of the model.

**d. Model with only TFL, Power Generation Data, Weather and Holiday Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **MAPE** |
| Linear Regression | 3535.4 | 24212469.6 | 4920.6 | 19.0 |
| Decision Tree | 3888.5 | 28374324.4 | 5326.7 | 20.6 |
| Random Forest | 3137.3 | 17907989.9 | 4231.7 | 16.3 |
| **ADA Boost** | **3213.0** | **17504361.8** | **4183.8** | **16.1** |
| Neural Network | 4535.0 | 33273671.3 | 5768.3 | 22.3 |

The model which includes all the features – i.e., TFL data, Power Generation Data, Weather Data and Holiday data gave the best results. While more features do increase the accuracy of the model in some cases, sometimes it introduces complexities like parallelism and multi-collinearity in the high dimensions. The main rationale behind using the heat map correlation was to identify such features that were needed for this analysis. It’s evident from the results that using all the datasets- i.e, the TFL dataset, the Power Generation dataset, the Weather dataset and the Holidays dataset, the best machine learning model could be developed which has the least mean absolute percentage error.

**e. Time Series Modelling**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **MAPE** |
| ARIMA | 10373.6 | 169166593.9 | 13006.4 | 50.3 |
| **SARIMA** | **8915.2** | **123952656.3** | **11133.4** | **43.1** |

One of the Machine Learning techniques that deals with time series data is time-series modelling. As the name suggest, time series modelling involves working with data that are based on time (hours, minutes, days, months etc) to device meaningful conclusions on trends and pattern of the data. Time series modelling is almost the first go for of any machine learning model where there are serially correlated data as in this example of bike ride sharing. For the time-series modelling, the most common machine learning technique - ARIMA and SARIMA have been used.

From the results above, it can be seen that Seasonal ARIMA model performed better than ARIMA model. This is because the bike sharing model has a seasonality factor which is taken care by the SARIMA algorithm. The ARIMA and SARIMA parameters are both automatically chosen by the auto-arima function of scikit-learn library of python pandas which choses the order of the p, d and q parameters by minimizing the ‘**AIC**’ criteria of the model. While the machine learning analysis using time-series shows a considerable error between the test data and the predicted data, it has to be kept in mind that time series data does takes care of complex factors like stationarity which are not considered by other models.

**BUSINESS IMPLICATIONS**

The analysis focus on the prediction of bike rides based on historical data of bike usage by the TFL bikes in London along with other exogenous factors. This main aim of such analysis is to exploit Machine Learning technologies that will drive the future of such bike sharing schemes and, in doing so, create an entirely novel data function to offer bike scheme operators.

It is well known fact that the principal operating costs of bike share costs are bike redistribution and bike maintenance. Given that demand for bikes in a city varies hugely based on external factors such as time, day of week, weather, travel disruption, special events, etc., it has proven extremely difficult to map and manage demand and supply of bikes till date. The main rationale behind the prediction of bike rides is to address these. From the bike ride data, the nexus between machine learning and ride data would focus on developing algorithms, analysing trends and develop models that would stream predictions for demand of bikes, address bike availability schedules and predict maintenance.

**FUTURE SCOPE**

As future steps for this analysis, there are a number of things that could be done to improve the predictions of the model. While hyper-parameter tuning for the regression algorithms might increase the model efficiency and reduce the errors, including granular data, like the start hour might increase the model efficacy. Fetching bike ride data based on the journey start and stop locations helps to identify travel patterns and might address the problem of bike relocation better. Another method of improving performance is to add weather data as exogenous variables to the time-series model which might decrease the prediction error and has been kept as the future steps for this exercise.