Abstract :

Bike-sharing programs have received increasing attention in the recent years due to their positive impact on the environment coupled with the awareness created by the environmental activists. In addition to the environmental advantages, cycling aids in improving public health, increasing cycling population and contributing to transit use. With the increasing number of bike-sharing service providers around the world, a real-time demand-supply tracking system which brings together the service providers and bike enthusiasts is implemented. Though there are several bike-sharing applications readily available for the users, currently there are no apps for the bike-sharing owners to track the demand-supply for bikes in real-time for any given station. We implemented a real-time demand-supply system that provides to the bike owner information regarding the demand-supply for bikes in any particular station at any particular time through a dashboard and also does forecasting of the bike demand. The Bike Demand forecasting uses the following classifiers: random forest, gradient boosting and linear regression, to predict the number of bikes in demand at a particular station in real-time. Results from the above-mentioned methods are compared and the method with the best results is chosen to supplement information provided by the dashboard to show the demand. For all three of our datasets, XGBoost on the dataset with all stations outperformed all the other models and hence XGBoost on trips data was chosen for the web application. Bike share providers will know the demand for any particular station which will enable them to fetch bikes from stations using the Web UI. Bike shortages due to uneven bike distribution are avoided with high accuracy due to fast prediction.

An important question in planning and designing bike-sharing services is to support the user's travel demand by allocating bikes at the stations in an efficient and reliable manner which may require accurate short-time demand prediction. This study focuses on the short-term forecasting, 15 min ahead, of the shared bikes demand in Montreal using a machine learning approach.

Keywords :

- *Business Understanding*
- *Data summary*
- *Data Collection*
- *Data Wrangling*
- *Feature Engineering*
- *EDA*
- *Model Building*
- *Conclusion on Bike Sharing Demand Prediction*
- *Impliction To Business*

Introduction :

Public bike-sharing systems were suggested at the beginning of the millennium but have been gaining momentum only in the last decade. The main premise of implementing bike-sharing systems is to promote sustainable mobility in urban areas.

They offer a convenient and easy-to-use service for residents for short-distance trips. Moreover, they are capable to improve first/last mile connection to other travel modes, reduce traffic congestion and energy consumption and decrease environmental impacts of daily travels. Furthermore, communities who organize a bike-sharing program increase physical activity and encourage remarkable health benefits to the users.

Bike-sharing system as a sustainable and affordable travel mode is not without its challenges for both users, e.g., perceived lack of safety, and operators. In this context, bike repositioning or rebalancing has been recognized as an important operational challenge. Bike demand is basically non-stationary, meaning that it varies over time and space. The fluctuating demand may cause the uneven distribution of bikes across different stations where some stations may be totally saturated while concurrently others may lack bikes. This can be related to the “tidal flows” of bike-sharing trips, with certain areas in the city encountering a deficit of bike availability. For instance, during the morning rush hour, residential areas generate a high number of commuting trips towards the areas of employment. This could possibly lead to the problem of insufficient bikes in those areas in that period.

The service quality has positive and significant effect to increase the bike-sharing popularity, attract more users, and improve the overall economic performance of bike-sharing Sensors. The bike imbalance issue may cause reduced service reliability, user dissatisfaction and decrease attraction and user engagement in bike-sharing program which may fail to meet the expectations of sustainable transport system implementation. Therefore, it is of great importance to understand and predict travel demand to support planning and day-to-day operation of bike-sharing systems. Accurate and reliable trip demand predictions across the city over different times of day allow system operators to better plan bike redistribution and fleet rebalancing. Hence, application of advanced predictive models has recently received a lot of research interest, as revealed by a recent literature review by Albuquerque et al. on Machine Learning techniques’ contributions applied to bike-sharing systems to improve urban mobility.

Like most bike-sharing systems, the one in Montreal faces the operational challenge of redistributing bikes across the stations to meet travel demand. While optimization algorithms can support such operation, they must rely on relevant travel forecasting demand able to anticipate where bikes will be required in the short-term. This is typically the missing component since bike-sharing demand combines both regular and irregular use patterns and fluctuates according to various events. Moreover, the recent COVID-19 pandemic drastically impacted all features of daily travel for all transport modes. It has become even more challenging to anticipate plausible travel behaviors at all forecasting horizons with the high uncertainty related to post-COVID activity systems. In this context, proposing tools able to forecast pickups while accounting for changing spatial-temporal patterns is critical for network operators that must ensure good fit between provided services (shared bike availability) and evolving demand. The focus of this study is on the short-term prediction, in a 15 min horizon of bike-sharing usage in Montreal. The timespan is selected in such a way that it accommodates for predicting the bike demand during the COVID-19 pandemic using past data. More specifically, this study aims to compare performance of different models fitted to time-series data when the period under analysis includes important disruptions as the one faced during the COVID-19 pandemics. Firstly, this study employs Louvain method to identify and cluster communities in the bike-sharing network to account for the interactions between stations and be less volatile than station-based modeling. Secondly, the study introduces data structure preparation where historical demand, feature engineering, weather conditions, and temporal variables are incorporated. Thirdly, the study employed deep neural networks for short-term travel demand prediction in a bike-sharing system. In this regard, a hybrid model composed of convolutional neural network (CNN) and long short-term memory (LSTM) is suggested to forecast the time series changes of bike usage. Finally, to examine the effectiveness of the proposed structure, the study compares the performance of some competitors using mean absolute error (MAE) and root mean squared error (RMSE) as evaluation measures.

Problem Statement:

Maximize: The availability of bikes to the customer.

Minimize: Minimise the time of waiting to get a bike on rent.

The main goal of the project is to: Finding factors and cause those influence shortage of bike and time delay of availing bike on rent. Using the data provided, this paper aims to analyse the data to determine what variables are correlated with customer churn, if any. Hourly count of bike for rent will also be predicted.

Data Description:

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. Data which is collected from a rented bike provider company form Seoul to get analysed, involves usage details of customers from. The data was taken from rented bike Provider Company. It has 8760 rows and 14 columns. Most columns related to hourly bike count for rent. Other column was indicative of weather condition affecting bike count per hour.

Dataset Preparation:

The bike sharing demand prediction dataset from rented bike provider company from Seoul contains 14 features and 8760 observations of a complete year I.e. from 1.12.2017 to 31.11.2018. Below Table shows the data features.

Data-set description:

Feature Name	**T**
Date : year-month-day	D
Rented Bike Count	I
Hour	I
Temperature(**°C)**	F
Humidity (%)	I
Wind speed (m/s)	F
Visibility (10m)	I
Dew Point temperature (**°C)**	F
Solar Radiation (MJ/m2)	F
Rainfall (mm)	F
Snowfall(cm)	F
Seasons	C
Holiday	C
Functioning day	C

Feature Breakdown:

Date: *The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, we need to convert into date-time format.*

Rented Bike Count: *Number of rented bikes per hour which our dependent variable and we need to predict that*

Hour: The hour of the day, starting from 0-23 it's in a digital time format

Temperature (°C): *Temperature of the weather in Celsius and it varies from -17*°C to 39.4°C*.*

Humidity (%): Availability of Humidity *in the air during the booking and ranges from 0 to 98%.*

Wind speed (m/s): Speed of the wind while booking and ranges from 0 to 7.4m/s.

Visibility (10m): Visibility to the eyes during driving in “m” and ranges from 27m to 2000m.

Dew point temperature (°C):*Temperature At the beginning of the day* and it ranges from -30.6°C to 27.2°C.

Solar Radiation (MJ/m2): Sun contribution or solar radiation during ride booking which varies from 0 to 3.5 MJ/m2.

Rainfall (mm): The amount of rainfall during bike booking which ranges from 0 to 35mm.

Snowfall (cm): Amount of snowing in cm during the booking in cm and ranges from 0 to 8.8 cm.

Seasons: Seasons of the year and total there are 4 distinct seasons I.e. summer, autumn, spring and winter.

Holiday: If the day is holiday period or not and there are 2 types of data that is holiday and no holiday

Functioning Day: If the day is a Functioning Day or not and it contains object data type yes and no.

Exploratory Data Analysis :

Exploratory data analysis is an statistical way of understanding the data which is usually done in a visual way.The graphs plotted in explotary data analysis are for better understanding of data to the analyst. For the current data set exploratory data analysis is done as follows.

Since we have to predict the number of bikes that will be rented, the best way to begin is with the variable to predict, "count". We can stratify the "count" distribution as boxplots for the categorical variables, and draw the "count" and numeric variables in another plot.

Correlation Among Variables:

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour.

Graphical Representation Of The Results:

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analysed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

Algorithms

1. Linear regression.

Linear regression is the most widely used and simplest method to predict demand in various contexts. Due to its simplicity and straightforward economic intuition in explaining the relationship between predictors and the outcome, we use linear regression as a benchmark against which other more advanced models are compared for their predictive power. The linear regression model is given as

y=β0+∑inβixi+εy=β0+∑inβixi+ε

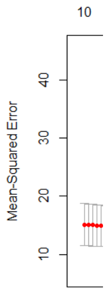
where βiβi is the coefficient of feature xixi, β0β0 is the constant, and εε is the random error.

2. Lasso Regression:

Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.

The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|.$$



3. Ridge Regression:

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

we have concluded that we would like to decrease the model complexity, that is the number of predictors. We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response. Removing predictors from the model can be seen as settings their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way. This way, we decrease model complexity while keeping all variables in the model. This, basically, is what Ridge Regression does.

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda|$$

4.DECISION TREE:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be *"learned"* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called *recursive partitioning*. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.

5. Random Forest.

Random forest is a almighty tool which ensembles decision trees and bagging .The base learner of random forests is a binary tree constructed by recursive partitioning (RPART) and then developed using classification and regression trees. Binary splits of the parent node of a random forest splits data into two children’s nodes and increases homogeneity in children nodes compared to parent nodes. Note that a random forest does not split tree nodes based on all variables; instead, it chooses random variable subsets as candidates to find the optimal split at every node of every tree. Then the information from the nn trees is aggregated for classification and prediction. Random forests also provide the importance of each feature by accumulated Gini gains of all splits in all trees representing the variable discrimination ability:

imporj=1#trees∑v∈xjGain(xj,v)imporj= 1#trees∑v∈xjGain(xj,v)

where Gain(xj,v)Gain(xj,v) is the gain of the Gini index of feature xjxj combined with node.

6. Gradient Boosting:

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term ‘gradient’ is related here.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc.

Conclusion

While machine learning models are completely data driven, we have attempted to incorporate social economic variables in the models to predict bike sharing demand. Despite the fact that these variables are barely useful in explaining and predicting short-term bike demand because they are constant, they did reveal demand differences between docking stations that are characterized by different social economic conditions. The roles that these variables play are to reveal population and economic activity that may differ across districts where bike docking stations are located. In this regard, bike sharing demand at the station level could perhaps be divided into basic demand, which is determined by social economic factors and induced demand, which changes with weather, pollution as well as a wide range of features that vary in the short term or even instantaneously. We advanced studies conducted by V E et al. [32] and E and Cho [14] in predicting bike demand in Seoul in the sense that they only addressed the induced demand for bike sharing on a daily basis.

The best model is the random forest model in our study, and the most important features are precipitation, the number of Covid-19 cases, the level of O3, heat index, and the level of PM10. The most important categories of features for the random forest model are Covid-19 outbreak, followed by air pollution and weather. Almost all social economic features are the least important, however they played a role in enhancing the performance of the models. The SVM is also an acceptable model. The features in the categories of weather, Covid-19 outbreak and traffic accidents have highest average weights. These results indicate that weather features such as precipitation, temperature, heat index, wind chill temperature as well as Covid-19 outbreak have huge impacts on bike sharing demand in Seoul. Further research can focus on many other potential features that influence bike sharing demand and many other machine learning algorithms such as Multilayer Perception Model.

References

1. Akın M (2015) A novel approach to model selection in tourism demand modeling. Tour Manage 48:64–72. https://doi.org/10.1016/j.tourman.2014.11.004
2. Bi J-W, Han T-Y, Li H (2020) International tourism demand forecasting with machine learning models: the power of the number of lagged inputs. Tour Econ. https://doi.org/10.1177/1354816620976954
3. Chen X, Ishwaran H (2012) Random forests for genomic data analysis. Genomics 99:323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

4. Glantz SA, Slinker BK (1990) Primer of applied regression and analysis of variance. McGraw-Hill, Health Professions Division
5. Claveria O, Monte E, Torra S (2018) Modelling tourism demand to Spain with machine learning techniques. The impact of forecast horizon on model selection. Revista de Economia Aplicada, **24**(72):109–132