MATH6185 Operational Research and Data Science
Case Study 1

Predicting London Bike Usage Using Machine Learning

Supervisor: Dr Xiang Song
Email: xiang.song@port.ac.uk

2023/24

Mathematical Sciences
University of Southampton

Aims of the case study

In 2007, London was suffering from major traffic congestion and its consequences, such as high levels of pollution and slow journey times. In 2010 it followed Paris and Brussels in introducing a cycle hire scheme, which enabled cyclists to hire a bike from one of London's hundreds of docking stations (761 as of January 2016). The take-up was immediate and encouraging, with a million journeys in the first 10 weeks. (https://www.centreforpublicimpact.org/case-study/londons-cycle-hire-scheme)

The appearance of the bike-sharing scheme raises a lot of advantages such as reducing greenhouse gas emissions, alleviating traffic congestion, especially implicitly increasing exercise and enhancing health. However, the management of shared bicycles scattered everywhere in the city has become a serious problem. Placing enough bicycles at a certain time in high-demand places can maximize the utilisation of bicycles and improve the convenience of people. So, the forecast for bike-sharing demand is quite necessary to improve the distribution of bicycles which ensures enough bicycles for the public all the time. (Gu and Lin, 2023).

In this case study, we will delve into various data science methodologies to address the challenges posed by bike-sharing management. Firstly, we will conduct customer behaviour analysis to scrutinize rental duration and frequency, aiming to uncover user preferences and habits. Next, employing multivariate analysis, we will explore correlations among different variables such as trip durations, station popularity, and time of day, with the objective of revealing underlying usage patterns. Furthermore, geospatial analysis will be employed to map trip start and end locations, enabling the identification of popular routes and areas with significant cycling activity. Finally, machine learning techniques will be utilised to forecast bike-sharing demand, providing valuable insights for effective resource allocation and management within the bike-sharing system.

Note: You must, where you write your project:

Acknowledge TfL as the source of the Information by including the following attribution statement 'Powered by TfL Open Data'

Instructions

- Please read the instructions below.
- Completed work should be submitted via Blackboard before 15:59 on Thursday, 25 July 2024. The deadline is strict and penalties for late work will be applied in accordance with the University's late work policy.
- Your submission should include:
    - a written report (handwritten reports are not accepted). This should be submitted via the Turnitin link in the 'Assignments' tab ('MATH6185 Case Study
- Submission').
    - the Python code of your calculations. This should be submitted via the file upload link in the 'Assignments' tab ('MATH6185 Python Code Submission').
- The case study must be carried out and written up independently (see University's Academic Integrity Guidance). The marks are distributed as follows:
    - 60 marks for the answers. All the answers must be presented in your written report.

    - 30 marks for the presentation of your answers: layout, plots, interpretation, appropriate use of references and academic style of writing. Careful explanation and clear presentation of results and Python code are important.
    - 10 marks for originality, expressing your own ideas.
- In your submission please attach the following two files:
- The report in a file called report-ID.pdf, where ID is your student ID number;
- The Python code called code-ID.XXX, where ID is your student ID number.

What should be covered in your report?

1. In this first question we try to analyse urban mobility patterns, station performance, and cycling preferences among London's diverse population.
   (a) Data Collection and Preparation: The data sets 351JourneyDataExtract02Jan2023-08Jan2023.csv to 384JourneyDataExtract15Nov2023-30Nov2023.csv were sourced directly from the Transport for London's official website, which provides open data to encourage public use and analysis. More details and related datasets can be found at Transport for London (TfL) (https://cycling.data.tfl.gov.uk/).

   The dataset includes the following variables for each ride:
   - Number: A unique identifier for each trip (Trip ID).
   - Start Date: The date and time when the trip began.
   - Start Station Number: The identifier for the starting station.
   - Start Station: The name of the starting station.
   - End Date: The date and time when the trip ended.
   - End Station Number: The identifier for the ending station.
   - End Station: The name of the ending station.
   - Bike Number: A unique identifier for the bicycle used.
   - Bike Model: The model of the bicycle used.
   - Total Duration: The total time duration of the trip (in a human-readable format).
   - Total Duration (ms): The total time duration of the trip in milliseconds.

   Students are expected to identify sources of data related to London bike usage in **month x in 2023** and explore the data to gain insights into the patterns and relationships, where x is the final digit of the student ID. For example, if student ID is 1234567. Then, $x = 7$. If $x = 0$, then revise $x = 0$ to $x = 10$.

   Hint: Consider using "import pandas as pd" and "pd.concat()" in Python to retrieve the data for the entire month.

   (b) Customer Behaviour Analysis: Analyse the duration and frequency of rentals to understand user preferences and habits.

   Hint: Consider using "pd.to_timedelta()" in Python to convert "Total Duration" column to timedelta. Consider using "import seaborn as sns" and "sns.histplot()" to plot histogram of trip durations.

   (c) Geospatial Analysis: Map the start and end locations of trips to identify popular routes and areas with high cycling traffic.

   Hint: Consider using "Data['Start station'].value_counts().head(5)" to group data by start station and count the number of trips.

   (d) Multivariate Analysis: Explore relationships between different variables, such as trip durations, station popularity, and time of day, to uncover underlying patterns in bike usage.

   Hint: Consider using "Data['Start date'].dt.hour()", "Data['Start date'].dt.day_name()" and "Data['Start date'].dt.day" to extract hour, day of the week, and day from the 'Start date'.

   Consider using "Data.pivot_table()" and "sns.heatmap(pivot_table)" to show the heatmap of Hourly Rentals by Day of the Week.

(e) Documentation and Reporting:

- Document the entire process, including data collection and pre-processing.
- Present the key functions used in Python.
- Present the results clearly with visualisations such as plots, tables, and charts.
- Discuss the findings and conclude with recommendations for improving bike usage in London based on the insights gained from the analyses in sections (b), (c), and (d).

2. In the second question, we aim to develop a predictive London bike usage model using machine learning techniques in Python software.

(a) Data Collection and Preparation:

- Identify sources of data related to London bike usage **at a specific station (chosen by students) in 2023**.
  https://cycling.data.tfl.gov.uk/
  Hint: Visit https://www.visualcrossing.com/weather/weather-data-services and search for London to find historical weather information.
- Explore the data to gain insights into the patterns and relationships.
  Hint: Consider using "groupby('day').size()" to analyse the daily trends within the month.

(b) Feature Engineering:

- Select relevant features that might influence bike usage, such as, time of day, day of the week, holidays, weather, etc.
- Engineer new features if necessary, such as aggregating data over different time periods or creating interaction terms.
- Conduct correlative tests and other statistical analyses to understand the relationships and distributions within the data.
  Hint: Consider using "corr()" and "heatmap()" functions.
- Encode categorical variables and normalize numerical variables as needed.

(c) Model Selection and Training:

- Choose suitable machine learning algorithms for building the predictive model, such as regression, decision trees, random forests, or neural networks.
- Consider hybrid techniques and demonstrate initiative by exploring combinations of different methods. For example, using clustering before regression.
- Split the data into training and testing sets to evaluate model performance.
- Train multiple models with different algorithms and hyperparameters.
- Evaluate each model using appropriate evaluation metrics (e.g., RMSE, MAE, R^2 score).

(d) Model Evaluation and Optimisation:

- Compare the performance of different models and select the best-performing one based on evaluation metrics.
- Refine hyperparameters utilising methods such as grid search or random search. Whenever feasible, offer a mathematical rationale for the selection of parameters.

(f) Documentation and Reporting:

- Document the entire process, including data collection, pre-processing, feature engineering, model selection, training, evaluation, and optimisation.
- Provide detailed explanations of the chosen techniques and algorithms.
- Present the key functions used in Python.
- Present the results clearly with visualisations such as plots, tables, and charts.
- Discuss the implications of the findings and any limitations of the models.
- Conclude with recommendations for improving bike usage in London based on the insights gained from the predictive model.

Reference

1. Colin Cameron, A. and Windmeijer, F. A. G., “An R-squared measure of goodness of fit for some common nonlinear regression models,” J. Econometrics, 77(2), 329–342(1997). https://doi.org/10.1016/S0304-4076(96)01818-0 Google Scholar
2. DecisionBrain. "Bike Sharing." Available at: https://decisionbrain.com/bike-sharing/ (Accessed: 22 May 2024).
3. Ding, C., Wang, D., Ma, X. and Li, H., “Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees,” Sustain., 8(11), (2016). Google Scholar
4. Eberly, L. E., [Topics in Biostatistics], Humana Press, Totowa, 165–187(2007).Eberly, L. E., [Topics in Biostatistics], Humana Press, Totowa, 165–187(2007).
5. Eren, E. and Uz, V. E., “A review on bike-sharing: The factors affecting bike-sharing demand,” Sustain. Cities Soc., 54, (2020). https://doi.org/10.1016/j.scs.2019.101882 Google Scholar
6. Feng, Y. and Wang, S., “A forecast for bicycle rental demand based on random forests and multiple linear regression,” Proc. IEEE/ACIS International Conference on Computer and Information Science, ICIS, 101–105(2017).
7. Feng, Y. and Wang, S., “A forecast for bicycle rental demand based on random forests and multiple linear regression,” Proc. IEEE/ACIS International Conference on Computer and Information Science, ICIS, 101–105(2017).
8. Feasibility study for a central London cycle hire scheme, Final report, November 2008, Transport for London
9. Gu, K.Y., Lin, Y., Prediction for bike-sharing demand in London using multiple linear regression and random forest. Proceedings Volume 12803, Fifth International Conference on Artificial Intelligence and Computer Science (AICS 2023); 128031I (2023) https://doi.org/10.1117/12.3009514
10. London Datastore. "Number of Bicycle Hires," created July 2010.
11. Marill MD, K. A., “Advanced statistics: Linear regression, Part II: Multiple linear regression,” Acad. Emerg. Med., 11(1), 94–102(2008). Google Scholar
12. Raschka, S. and Mirjalili, V., [Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2], Packt Publishing, Brimingham, (2019).
13. Raschka, S. and Mirjalili, V., [Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2], Packt Publishing, Brimingham, (2019).
14. Transport for London. (August 2023). TfL Cycle Hire Trip Data. Retrieved [Date Retrieved], from https://tfl.gov.uk/info-for/open-data-users/our-open-data.
15. Transport for London. Feasibility study for a central London cycle hire scheme, Final report, November 2008.
16. Xu, Y., “Research and implementation of improved random forest algorithm based on Spark,” Proc. IEEE International Conference on Big Data Analysis, ICBDA, 499–503(2017).Xu, Y., “Research and implementation of improved random forest algorithm based on Spark,” Proc. IEEE International Conference on Big Data Analysis, ICBDA, 499–503(2017).

17. Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J. and Li, Z., “Deep multi-view spatial-temporal network for taxi demand prediction,” Proc. AAAI, 32(1), (2018). Google Scholar
18. Zhai, H., Cui, L., Nie, Y., Xu, X. and Zhang, W., “A comprehensive comparative analysis of the basic theory of the short term bus passenger flow prediction,” Symmetry, 10(9), (2018). https://doi.org/10.3390/sym10090369 Google Scholar