**Machine Learning Project Proposal**

**DTSC 691: Applied Data Science**

**Seydou Toure**

**Forecasting Bike Rental Demand:**

**A Machine Learning Approach**

**Background**

Bicycle-sharing programs have emerged as a modern evolution of conventional bike rental services, offering an automated approach to the entire process encompassing membership registration, bicycle rental, and return. Users can conveniently pick up a bicycle from one location and return it to a different one. Presently, there are approximately 500 such programs globally, boasting a fleet of over 500,000 bicycles. These systems have garnered significant attention due to their substantial impact on urban transportation, environmental sustainability, and public health.

Beyond their practical urban applications, bicycle-sharing systems present intriguing opportunities for research due to the unique nature of the data they generate. Unlike other forms of public transportation such as buses or subways, these systems meticulously log each trip’s duration, starting point, and destination. This capability transforms bicycle-sharing networks into a de facto virtual sensor grid, poised to capture urban mobility trends. Consequently, it is anticipated that these systems could play a crucial role in identifying and understanding major urban events through continuous data monitoring.

**Project Overview**

The Bike Sharing Demand dataset typically encompasses a variety of features such as Datetime, weather conditions, temperature, humidity, and wind speed and categorical variables such as season, holiday, and working day, alongside the count of bikes rented per hour or day. The main objective of this machine learning project is to predict bike rental demand, aiding in better resource allocation and management for bike-sharing programs.

Through this project, one can practice various data science skills ranging from data cleaning and preprocessing, exploratory data analysis, feature engineering, model selection, training, and evaluation, to interpreting model results. The goal is to create a robust predictive model that can help optimize bike stock and distribution, ultimately enhancing bike-sharing services' efficiency and user experience.

The results will provide valuable insights for optimizing bike-sharing operations and contribute to urban planning and traffic management. Ethical considerations, particularly around data privacy and bias, are crucial to ensure responsible and fair use of the models developed.

The emphasis on exploratory data analysis aids in uncovering trends, correlations, and distribution of the target variable. Ensuring robust model evaluation through cross-validation and performance metrics is crucial, alongside interpreting model predictions and understanding feature importance.

**Project Description**

**Data Set:**

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect rental behaviors. The core data set is related to the two-year historical log corresponding to the years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available at [http://capitalbikeshare.com/system-data](#_top).

The data was aggregated on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information is extracted from <http://www.freemeteo.com>.

**Files**

- Readme.txt

- hour.csv: bike sharing counts aggregated on an hourly basis. Records: 17379 hours

- day.csv: bike sharing counts aggregated daily. Records: 731 days

**Project Objective and Scope**

The main objective of the Bike Sharing demand prediction project is to develop a machine learning model capable of accurately forecasting the number of bike rentals at different times of the day, given a set of features such as weather conditions, time, day of the week, and season. The aim is to enable Bike Sharing companies to optimize their operations, ensuring that bikes are available when and where they are needed, thereby improving customer satisfaction and operational efficiency.

**Data Description**

Both hour.csv and day.csv have the following field, except hr which is not available in day.csv.

**instant:** record index.

**dteday:** date.

**season:** season (1: springer, 2: summer, 3: fall, 4: winter).

**yr:** year (0: 2011, 1:2012).

**mnth:** month (1 to 12).

**hr:** hour (0 to 23).

**holiday**: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule).

**weekday:** day of the week.

**workingday:** if the day is neither weekend nor holiday is 1, otherwise is 0.

**weathersit:** 1- Clear, Few clouds, Partly cloudy, Partly cloudy

2- Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3- Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain +

Scattered clouds

4- Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

**temp:** Normalized temperature in Celsius. The values are divided to 41 (max).

**atemp:** Normalized feeling temperature in Celsius. The values are divided to 50 (max).

**hum:** Normalized humidity. The values are divided to 100 (max).

**windspeed:** Normalized wind speed. The values are divided to 67 (max).

**casual:** count of casual users.

**registered:** count of registered users.

**cnt:** count of total rental bikes including both casual and registered.

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) for a Bike Sharing Demand dataset typically involves visualizing and statistically analyzing the data to uncover patterns, relationships, and insights that can inform subsequent modeling decisions. By conducting thorough EDA, valuable insights can be gained, helping to guide the subsequent stages of feature engineering, model selection, and model evaluation in the machine learning project.

**Data Preparation and Cleaning**

Data preparation and cleaning are crucial to ensure the Bike Sharing Demand dataset is ready for machine learning modeling. The process starts with handling missing values, either by imputing them using statistical methods or removing the affected rows or columns. Any outliers or anomalies detected during exploratory data analysis need to be addressed, as they can distort the model's predictions. This can involve transforming the data or removing the outliers altogether.

Categorical variables such as season, weather conditions, and day of the week need encoding to convert them into a numerical format that machine learning models can interpret. Techniques like one-hot encoding or label encoding are commonly used for this purpose. Temporal features like datetime require the extraction of useful components such as hour, day, and month, creating new features that can provide additional context to the model.

Normalization or standardization of numerical features ensures that all features contribute equally to the model’s performance, preventing features with larger scales from dominating the model’s behavior. The dataset may also benefit from feature engineering, creating new features that can help improve model performance. This could involve interaction terms, polynomial features, or domain-specific features derived from existing data. Finally, it is essential to split the dataset into training and testing sets, ensuring an unbiased evaluation of the model's performance. The split can be done randomly or considering the time-series nature of the data, ensuring that future data is not used to predict past or present bike demand.

By diligently preparing and cleaning the data, the machine learning model will have a solid foundation to learn from, potentially leading to more accurate and reliable predictions of bike sharing demand.

**Model Training**

In the model training phase for the Bike Sharing Demand dataset, various machine learning algorithms are applied to learn from the prepared and cleaned data. Commonly used models include linear regression for its simplicity and interpretability, decision trees, and ensemble methods like random forests and gradient boosting machines for their ability to capture complex non-linear relationships.

The dataset is divided into training and validation sets, ensuring that the model's performance can be evaluated on unseen data, providing a more accurate estimate of its real-world performance. Cross-validation techniques, particularly time-series cross-validation, are often employed due to the temporal nature of the data, ensuring that the model is robust and performs well across different time periods.

Hyperparameter tuning is conducted to find the optimal settings for each model, enhancing their predictive capabilities. This can be done using techniques like grid search, random search, or more advanced methods like Bayesian optimization. The models are then evaluated using appropriate performance metrics. For regression tasks like bike sharing demand prediction, common metrics include Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

By the end of the model training phase, the best-performing model(s) are selected based on their performance metrics and interpretability, ready to be validated on the test set or deployed for making predictions on new data.

**Model Evaluation**

Model evaluation for the Bike Sharing Demand dataset involves assessing the performance of the trained machine-learning models to ensure they make accurate and reliable predictions. This is done using a separate test set, which has not been seen by the models during training, to simulate real-world conditions and provide an unbiased performance estimate. Performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared are calculated to quantify the models' prediction errors and assess their goodness of fit. For a demand prediction task, lower values of MAE and RMSE, and a higher R-squared value are desirable, indicating more accurate and reliable predictions.

Residual plots and prediction error analyses are conducted to identify any systematic patterns in the models' errors, helping to uncover areas where the models may be underperforming or biased.Finally, the models' interpretability is considered, ensuring that their predictions and decision-making processes can be understood and trusted by stakeholders, which is crucial for deployment and adoption in real-world settings.

Through rigorous model evaluation, the best-performing and most reliable model is selected, ready for deployment or further refinement, ensuring it meets the project's objectives and provides accurate demand predictions for bike-sharing services.

**User-Interface Integration**

Integrating the bike-sharing demand prediction model into a user interface (UI) facilitates easy access and interpretation of the predictions by end-users, which could be operations managers, planners, or other stakeholders. This involves creating a frontend that communicates with a backend server where the machine learning model is hosted. The UI typically includes input fields for users to enter or select relevant features such as date, time, weather conditions, and other pertinent information. These inputs are then sent to the backend server where the machine learning model processes the data and returns the predicted bike demand.

The front-end displays the prediction results in an easily understandable format, possibly including visualizations like charts or graphs that show demand trends over time. It may also provide additional information, such as confidence intervals or recommendations for bike redistribution. Error handling and validation checks are implemented to ensure that the data input by the user is valid and within expected ranges, providing prompts or error messages as necessary to guide the user.

The integration ensures real-time or near-real-time predictions, allowing for quick and informed decision-making. The UI is designed to be user-friendly, ensuring that users do not need a deep understanding of the machine learning model to make use of it. Security measures are put in place to protect both the user’s input data and the model’s predictions, ensuring that sensitive information is not exposed.

By integrating the bike-sharing demand prediction model into a user interface, the model’s insights are made accessible and actionable to end-users, enhancing decision-making processes and operational efficiency for bike-sharing services.

**Capstone complexity**

The capstone complexity of the Bike Sharing Demand machine learning project stems from the need to carefully handle time-series data, address a variety of external factors, select, and tune appropriate machine learning models, integrate the model into a user-friendly interface, and ensure the robustness, reliability, interpretability, and ethical soundness of the entire system.

**Software**

For a Bike Sharing Demand machine learning project, a variety of software tools and libraries are utilized to handle different stages of the project:

**Programming Languages**: Python or R are commonly used for their extensive libraries and community support in data science and machine learning.

**Data Cleaning and Preprocessing**: Libraries such as Pandas assist in cleaning and manipulating the dataset. It provides functionalities to handle missing values, outliers, and transforming variables.

**Data Visualization**: Matplotlib and Seaborn are utilized for creating visualizations to understand the data distribution, detect anomalies, and identify patterns during exploratory data analysis.

**Machine Learning Libraries**: Scikit-learn offers a wide variety of algorithms for model training and evaluation. XGBoost or LightGBM can be used for more advanced ensemble models. TensorFlow can be considered if deep learning models are needed.

**Model Evaluation**: Scikit-learn also provides functions to calculate performance metrics, perform cross-validation, and conduct other evaluation tasks.

**Hyperparameter Tuning**: Libraries like RandomizedSearchCV or GridSearchCV in Scikit-learn help in tuning the model to find the optimal set of hyperparameters.

**Development Environment**: Jupyter Notebooks will be used as it provides an interactive coding environment, making it easier to write, debug, and visualize code.

**Web Framework for UI**: For integrating the model into a user interface, web frameworks like Flask can be used to set up a backend, and a frontend server.

By leveraging these software tools and libraries, we can effectively handle the various aspects of the Bike Sharing Demand machine learning project, from data analysis and modeling to deployment and user interface integration.

**Project Completion Plan**

**Week 1: Project Initialization and Data Collection**

During the first week of the Bike Sharing Demand machine learning project, we will focus on laying a solid foundation by clearly defining the project's objectives and establishing success criteria to guide subsequent activities. The collection of the Bike Sharing Demand dataset, along with any other relevant data that might be required for comprehensive analysis, is a priority. This initial stage is crucial for aligning expectations, preparing resources, and ensuring that the project is set on the right track from the outset.

**Week 2: Data Cleaning and Preprocessing**

In Week 2, the focus is on data cleaning and preprocessing. We will conduct a comprehensive initial assessment to understand the quality and completeness of the data, ensuring that any deficiencies are promptly addressed. Attention is given to identifying and rectifying any issues with missing values, outliers, and anomalies that could potentially skew the analysis. To lay a strong foundation for the subsequent stages of the project, we will also engage in data type conversions and begin initial feature engineering efforts, setting the stage for more advanced data manipulation and analysis in the following weeks.

**Week 3: Exploratory Data Analysis (EDA)**

During Week 3, dedicated to Exploratory Data Analysis (EDA), the project delves into a comprehensive analysis. The focus lies on understanding the distribution of the target variable, which represents bike demand, along with other key features that are crucial to the prediction task. We will investigate any evident temporal trends, seasonality, and recurring patterns in bike usage to grasp how demand fluctuates over time. To solidify these insights, various relationships between bike demand and other features are explored using a combination of visualizations and statistical analyses, aiming to uncover potential correlations, trends, and anomalies that could inform subsequent stages of the machine learning project.

**Week 4: Advanced Data Preprocessing and Feature Engineering**

In Week 4, the focus is on advanced data preprocessing and feature engineering, building upon the insights gained from the exploratory data analysis (EDA) conducted in the previous week. We will work on creating new features that could enhance the model’s predictive power, carefully encoding categorical variables to ensure they are appropriately handled by machine learning algorithms. Numerical features undergo scaling to standardize their ranges and prevent features with larger scales from disproportionately influencing the model. Considering the time-series nature of the bike-sharing demand data, the dataset is meticulously split into training, and test sets, ensuring that temporal patterns are preserved, and future data is not mistakenly used to predict past or present demand, which is crucial for maintaining the integrity of the model evaluation process.

**Week 5: Model Training**

During Week 5 of the Bike Sharing Demand machine learning project, the focus is on model training, encompassing the selection and training of a variety of machine learning models ranging from simpler to more complex structures. This phase involves meticulous hyperparameter tuning to enhance the models' performance and optimize their predictions. To ensure the models accurately capture the underlying patterns in the data, they are evaluated using appropriate performance metrics and validation techniques, providing a comprehensive understanding of their efficacy and areas for potential improvement.

**Week 6: Model Evaluation and Selection**

Week 6 focuses on Model Evaluation and Selection, where all the previously trained models undergo a rigorous evaluation process using the reserved test set to ensure their performance is robust and reliable. The assessment involves a deep dive into residual plots and various error metrics, which helps in pinpointing areas that might need further tuning or adjustment. The ultimate goal of this phase is to compare the models and select the one that not only excels in terms of predictive accuracy and aligns with the performance metrics but also meets the specific business requirements and objectives of the project. This careful selection ensures that the model chosen is well-suited for deployment, providing reliable and useful predictions for bike sharing demand.

**Week 7: Integration and Finalization**

In Week 7, the focus shifts to integrating the chosen machine learning model with a user-centric interface, facilitating easy access and interpretation of the bike demand predictions for end-users. Rigorous testing is conducted to verify the system’s functionality and reliability, ensuring a seamless user experience. Comprehensive documentation is created to detail the methodologies employed, insights derived, challenges faced, and recommendations for future enhancements. The project findings and the operational system are then presented to stakeholders, providing them with a thorough understanding of the project’s outcomes and potential impact. Finally, considerations and plans are made for the deployment and ongoing maintenance of the machine learning model, ensuring its continued accuracy and relevance in predicting bike-sharing demand.

**Presentation Plan**

The presentation for the Bike Sharing Demand machine learning project systematically covers the project from inception to completion, emphasizing its importance for operational efficiency and user satisfaction. It begins with an introduction to the project’s goals, followed by a detailed look at the dataset, highlighting any unique challenges faced during data preparation and the strategies employed to address them. The key insights from the exploratory data analysis are shared using visual aids, enhancing the audience’s understanding of demand patterns and feature relationships. The model training phase is then dissected, explaining the selection of various machine learning models, the hyperparameter tuning process, and the evaluation metrics used.

A comprehensive analysis of the model evaluation and selection process is presented, detailing each model’s performance, and justifying the final choice. This section also discusses the model’s accuracy, reliability, and potential limitations. The integration of the model with a user-friendly interface is demonstrated, showcasing how end-users interact with the system and access the bike-sharing demand predictions.

The presentation rounds off with a summary of the main findings, the project’s impact, lessons learned, challenges faced, and recommendations for future initiatives. Throughout, the presentation will maintain a clear and concise style, utilizing visuals and demonstrations to engage the audience and enhance their understanding of the project’s achievements and implications.

# References

Fanaee-T, H., & Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 2(2-3), 113-127.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Brownlee, J. (2016). Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End. Machine Learning Mastery.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc.

Aggarwal, C. C. (2016). Neural networks and deep learning: A textbook. Springer.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.

McKinney, W. (2012). Python for data analysis. O'Reilly Media, Inc.

Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics Press.

[http://capitalbikeshare.com/system-data](#_top).

<http://www.freemeteo.com>.

<https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset>

<https://ictactjournals.in/paper/IJCT_Vol_14_Iss_3_Paper_6_2998_3004.pdf>