Python Group Assignment

Assignment Content

Top of Form

*The administration of Washington D.C wants to make a deeper analysis of the usage of the bike-sharing service present in the city in order to build a predictor model that helps the public transport department anticipate better the provisioning of bikes in the city. For these purposes, some data is available for the years 2011 and 2012.*

**Practice (6 points)**

**Main goals**

As part of the team of the hired consultancy firm, your work is to **build an interactive, insightful and complete web application about the bike-sharing service in the city**for the head of transportation services of the local government. As part of the requirements, there are two big goals:

* 1. ***The customer wants a deep analysis of the bike-sharing service. He wants to know how the citizens are using the service in order to know if some changes must be made to optimize costs or provide a better service.***
  2. ***The customer also wants a predictive model able to predict the total number of bicycle users on an hourly basis. They believe this will help them in the optimization of bike provisioning and will optimize the costs incurred from the outsourcing transportation company.***

**Grading criteria**

**Data Analysis (descriptive analytics) (1.5 points)**

* 1. Ensuring data quality (correctness, consistency, missing values, outliers...). Done
  2. Plotting clear and meaningful figures. In process
  3. Giving insights on what seems relevant for prediction and what does not.
  4. Discussion on missing values and outliers: No missing values.
  5. Treatment of text and date features. Yes, categorization already done + date time features
  6. Generation of extra features (e.g., season, yes/no holiday, hours of daylight, combinations of features, quantization/binarization, polynomial features) To do.
  7. Provide useful explanations about the decisions taken and the results obtained

**Machine Learning (predictive analytics) (1.5 points)**

* 1. Choosing the best model to solve the problem (linear and non-linear).

Minimum try LinearRegression and then others

* 1. Split data correctly to train and test your models

Done

* 1. Tuning model parameters with validation

In process/Done

* 1. Plotting predictions vs. reality for additional insights

TO DO. (plot residuals).

* 1. Deep analysis of results and insights given the features

Does not overfit, error for train and test stay similar. Feat importance, explain what happened to humidity.

* 1. Provide explanations about the decisions taken (chosen model, hyperparameter settings, evaluation metric employed, etc.)

TO DO: RMSE/MAE, better for explainability in terms of units. Explain why neg\_rmse

* 1. **NOTES**:
  2. Limit the libraries to those seen in class: sklearn, xgboost, lightgbm, catboost, pycaret, pandas, etc.

**Interactive report with Streamlit (2 points)**

* 1. **Requirement:** Use Streamlit (https://streamlit.io/)
  2. Include all knowledge from the previous points in a unified tool that comprises all knowledge about data and satisfies the requirements of the customer.
  3. Include a simulator, that, for a given model, it's able to provide estimations on demand of the number of bicycle users given the input features.
  4. Use separate sections to tell your story with cohesion. Missings, creation of new variables, exploratory data analysis, modeling, and recommendations about the actions to take by the results.
  5. Include Plotly charts to show your insights but also to show the analysis of the results of your models.
  6. Include some kind of interactivity to your graphs by including text boxes, sliders, and any kind of input type element in Streamlit.
  7. **NOTES**:
  8. To include graphics, data, and the results of the model you can export all elements you need from the notebook (in CSV, json, or any other format) and then read it again in the dashboard script.

These grading criteria contain just some points to take into account. The final grading will take into account more things like programming style, cleanliness, the correct use of pandas and other libraries (without redundant code), and the creativity of the group to provide an original solution that solves the business problem. Use your creativity and your already-acquired knowledge of engineering, machine learning, plotting, and programming to build a sophisticated tool for the local government department.

**Presentation (4 points)**

The exposition has a limit of **5 minutes per group** (excluding questions). You are free to choose the number of presenters, everyone doesn't need to talk. Teams running out of time will receive a grade penalty.

The presentation should be oriented to the business and the technical side, with the support of your interactive visualization tool. You can use slides to introduce and explain your work, but the interactive visualization tool must be shown.

After that, expect questions about the decisions taken in the project.

All members of the group should be ready to answer questions, as well as those that didn't present.

**Grading criteria**

The grading of the presentation will be made through **peer review.**At the end of the presentation session, all groups must deliberate and give an honest, and objective score to the rest of the presentations. The final grade will be the average value of all the scores.

* 1. Quality of explanations (you understand what the provider is trying to explain)
  2. Quality of analysis. All analyses made must contain useful information for the business. The provider must be able to extract the value from data and explain it.
  3. Look & feel. The tool and graphics presented look good.

***Please, provide honest and objective scores based on your perception of the work of your peers. Avoid previously agreed grades. This is a funny and motivating way to shine and learn from others.***

**Submission**

* 1. Jupyter Notebook with all code developed for this assignment. This includes all exploratory data analysis and machine learning models. Name: ***Group<groupname>\_Notebook\_Assignment2.ipynb***
  2. Python script that launches the streamlit dashboard. Name: ***Group<groupname>\_StreamlitSrc\_Assigment2.py***
  3. This script must be ready to be executed with the command ***streamlit run <yourscriptname>***without errors.

Bottom of Form

*The administration of Washington D.C wants to make a deeper analysis of the usage of the bike-sharing service present in the city in order to build a predictor model that helps the public transport department anticipate better the provisioning of bikes in the city. For these purposes, some data is available for the years 2011 and 2012.*

As part of the team of the hired consultancy firm, your work is to **build an interactive, insightful and complete web application about the bike-sharing service in the city**for the head of transportation services of the local government. As part of the requirements, there are two big goals:

* 1. ***The customer wants a deep analysis of the bike-sharing service. He wants to know how the citizens are using the service in order to know if some changes must be made to optimize costs or provide a better service.***
  2. ***The customer also wants a predictive model able to predict the total number of bicycle users on an hourly basis. They believe this will help them in the optimization of bike provisioning and will optimize the costs incurred from the outsourcing transportation company.***

Ladies and Gentlemen,

Welcome to today's presentation where we delve into the insightful analysis of the bike-sharing service in collaboration between the Paella Consulting Group and the esteemed townhall of Washington D.C. Our focus lies on predicting usage patterns to enhance the provisioning of bikes in the city, ensuring optimal service for our citizens.

Let me begin by commending the robustness of the data ingestion system provided. In our extensive analysis, we found the dataset to be remarkably clean:

* No missing values were encountered, ensuring data integrity.
* Relevant outliers, which often skew analyses, were notably absent.
* The consistency observed across various variables and measures reflects the accuracy of the dataset, aligning seamlessly with real-world expectations.

Moving on to user behavior, our analysis reveals distinct seasonal trends. [Display stacked area chart]. The data exhibits a clear weekly pattern, with peaks coinciding with labor days, characterized by the distinctive 'M shape' daily usage. Conversely, weekends and holidays witness a notable decrease in bike usage. Notably, there's a significant disparity between regular users and casual users, with casual users comprising approximately 20-25% of total users between 2011 and 2012.

[Present evolution of percentage for casual users]

Given these insights, we could therefore argue that if the city of Washington aims to target each group independently with different strategies, it would be convenient to forecast the bike demand for each group independently. This will have deep implications when it comes to creating a ML model as we will see.

Before delving into our predictive modelling approach, let's explore some key observations from our data.

* Factors such as [humidity] and [wind speed] do not exhibit clear linear correlations with total users. For the latter, higher wind speeds on the right-hand side of the plot generally lower the bikes usage.
* Temperature and feeling temperature have clear correlation, therefore only the first is necessary. [temp and atemp].
* We could elaborate it also plays a key-role the weather situation, [Boxplot] where we see that as the weather worsens, the usage of bikes also decreases and shrinkages.

With the overall perspective on the data analysis, we can show you now our predicting tool that will allow you to know for each user’s group how many bikes will be needed.