Data Analytics

Micro Project – Classifying Building Functions

Introduction:

You're employed at the city of Hamburg as data scientists. Because of a new tax-law the tax office would like to predict the usage of new buildings around Hamburg grouped by the different citys, so they can plan to get a more accurate picture on how much money from housing taxes the city of Hamburg and the districts will receive. You have been provided with the "df_hamburg.parquet"-File by the building office. This dataset includes data of buildings located in the Hamburg rural area and around the city, together with additional information on the single buildings. Your goal is to predict the building_type value for each building grouped by the city its located in.

Tasks:

Your project should include the following points and should also be included in your project presentation:

- Data Understanding and Cleanup: Describe your dataset and give an insight into the data you're working with. This should also include visualization and giving an insight into its most important features.
- Choosing a model: Choose a model for your specific problem. This should include a definition of the prediction problem. Compare different models and explain which model you have chosen and why it's the best suited one for your project. Evaluate the performance of the model as well on a city with good data quality. Also predict values for a city with low data quality and show the differences in performance.
- Evaluation: Talk about the experiences you have collected during your project. What were your biggest challenges, how did you overcome them and talk about your biggest lesson learned during the project.
- Creating a presentation: Your presentation should include and represent the most important aspect of the tasks named above. It should also include a brief description of the single steps that you've taken to achieve your goals.

(Note: This project is meant to give you a head on approach on data analytics. Also creative solutions such as feature engineering, the collection of additional data, etc. are welcome and will have a positive impact on the amount of bonus points you will earn from this project.)

Additional information:

A .parquet file is a kind of data frame like a .csv or pandas file and is usually used inside a apache environment.

A basic introduction on how to work with .parquet files and how to convert them into dataframes in python can be found here:

https://www.youtube.com/watch?v=KLFadWdomyI