Projects for Machine Learning

Students must complete a project, which will count for 40% of the total grade (10% pitch, 30% submitted project).

Detailed instructions

Choosing your project

You can work in teams of at most 3 people, and each team must elect a “captain”.
The team’s captain must send an email to gitaliano@luiss.it and dtorre@luiss.it with the subject [PROJECT22/23 - Machine Learning MSc] and cc the components of the team.
The email should contain:

First Project Preference
Name Surname student id of member 1 (“captain”)
Name Surname student id of member 2
Name Surname student id of member 3
Second Project Preference

Each captain (and only the captain) must send the email by April 3, 2023. If you are not included in an email sent by that date, you will be assigned to a project and to a team by the instructor.

Deadline

You must submit your project by May 17, 2023. This is a firm deadline,

What to submit for your project.

Each group must submit via mail to gitaliano@luiss.it and dtorre@luiss.it, the URL of a GitHub repository. The repository's name should be the student id of the “captain”.

The repository should contain:

1. a “README.md” file with the following information:
    a. Title and Team members
    b. **Introduction** – Briefly describe the project.
    c. **Methods** – Describe your proposed ideas (e.g., features, algorithm(s), training overview, design choices, etc.) and your environment so that:
        i. A reader can understand why you made your design decisions and the reasons behind any other choice related to the project
        ii. A reader should be able to recreate your environment (e.g., conda list, conda env export, etc.)
        iii. It may help to include a figure illustrating your ideas, e.g., a flowchart illustrating the steps in your machine learning system(s)
    d. **Experimental Design** – Describe any experiments conducted to validate the target contribution(s) of the project. Indicate the main purpose of each experiment, in particular:

i. The main purpose: 1-2 sentence high-level explanation
ii. Baseline(s): describe the method(s) that you used to compare your work to
iii. Evaluation Metrics(s): which ones did you use and why?

e. **Results** – Describe the following:
i. Main finding(s): report your final results and what you might conclude from your work
ii. Include at least one placeholder figure and/or table for communicating your findings
iii. All the figures containing results should be generated from the code.

f. **Conclusions** – List some concluding remarks. In particular:
i. Summarize in one paragraph the take-away point from your work. o Include one paragraph to explain what questions may not be fully answered by your work as well as natural next steps for this direction of future work

2. Single notebook called “main.ipynb” with ALL the code used for the project. The notebook must have the following characteristics:
   a. Text and code cells must alternate from start to finish. The text cell above must describe the contents of the code below and its output so that a reader can easily follow up on your implementation. In particular:
      i. You must explain what you will do and why you chose to do so.
      ii. You must explain the outputs of the cell (if any) with particular attention to describing figures such that a reader already knows what he is going to see.
3. An additional folder named “images” contains the figures displayed in the “README.md”.

Academic Integrity

You must write the code by yourself. Any abuse or violation will be taken into account during the evaluation. Any code that, for some (nonsensical) reason, is not written by yourself must be referenced (with a link to the original code).

Copying the projects from other teams is also strictly forbidden. Your code will be validated by anti-plagiarism software. In the unlikely event of two projects being very similar, we will follow the Netflix Prize rules: only the first project published on GitHub will get the grade, and the other will get nothing.

Datasets

All datasets are available in the zipped format and can be accessed on the learn.luiss.it platform by navigating to the Datasets folder in the Project section.

1) Customer Segmentation

A large corporation is seeking to develop a targeted email campaign for its Brazilian subsidiary and has collected data on its customers. As a member of the management team with expertise in machine learning, your task is to segment customers using the RFM (recency, frequency, monetary value) strategy. The dataset includes information on orders, customers, sellers, payments, products, and geolocation data. The goal of this project is to identify the optimal number of segments and assign each user to one of them to facilitate targeted marketing campaigns. By leveraging customer segmentation, the store can create personalized and relevant marketing messages that result in increased sales and customer loyalty.

Variables description:

- order_id: unique order identifier
- customer_id: the key to the orders dataset. Each order has a unique customer_id
- customer_unique_id: the unique identifier of a customer.
- customer_city: customer city name
- customer_state: customer state
- order_item_id: sequential number identifying the number of items included in the same order.
- product_id: product unique identifier
- price: item price
- freight_value: item freight value item (if an order has more than one item, the freight value is split between items)
- payment_type: method of payment chosen by the customer.
- payment_installments: number of installments chosen by the customer.
- payment_value: transaction value.
- order_status: the order status (delivered, shipped, etc).
- order_purchase_timestamp: purchase timestamp.
- order_approved_at: purchase approval timestamp.
- order_delivered_carrier_date: order posting timestamp. When it was handled by the logistic partner.
- order_delivered_customer_date: actual order delivery date to the customer.
- order_estimated_delivery_date: the estimated delivery date informed to the customer at the purchase moment.
- shipping_limit_date: seller shipping limit date for handling the order over to the logistic partner
- product_category_name: root product category, in Portuguese.
- product_category_name_english: root category of product, in English
- product_name_lenght: number of characters extracted from the product name.
- product_description_lenght: number of characters extracted from the product description.
- seller_id: seller unique identifier
- seller_city: seller city name
- seller_state: seller state

Specific Task Requirements

- Perform an Exploratory data analysis (EDA) with visualization using the entire dataset. Discuss correlations and how the data is distributed. In particular try to answer to these questions:
    - Looking the *price* do you think the dataset is balanced?
    - Looking the *customer_city* distribution do you think the dataset is balanced?
- Preprocess the dataset (remove duplicates, impute NaNs, encode categorical features with one hot encoding, check for anomalies, not necessarily in this order).
- Pick one technique to perform market segmentation based **on each of these** scores:
    - **Recency value:** time since a customer's last purchase.
    - **Frequency value:** refers to the number of times a customer has made a purchase.
    - **Monetary value:** refers to the total amount a customer has spent purchasing products
- Identify the proper number of clusters, and evaluate different options.
- Describe the properties of the clusters you have identified.
- Describe the properties of the customers belonging to each cluster

2) Popularity score of music tracks

A large company that provides music streaming services is trying to understand which factors contribute to the popularity of the songs they host on their platform. You have been hired to perform an exploratory data analysis (EDA) on this dataset provided by the music intelligence department (MID) to gain insights into the characteristics of popular songs on the platform and to identify patterns and trends that can inform recommendations for future music releases. The dataset contains information about songs including their artists, album names, track names, popularity, duration, and various audio features such as danceability, energy, and instrumentalness.

By analyzing this data, we can gain insights into what makes a song popular and what types of songs are likely to be enjoyed by different types of users. This can help the company to build more accurate and effective recommendation models, which in turn can improve user engagement and retention on the platform.

Variables description

- track_id: A unique identifier for each track.
- artists: The artists who performed the track. A single track can have multiple artists, separated by a comma.
- album_name: The name of the album that the track appears on.
- track_name: The name of the track.
- popularity: The popularity score of the track, ranging from 0 to 100.
- duration_ms: The duration of the track in milliseconds.
- explicit: A binary value indicating whether the track contains explicit lyrics.
- danceability: A score indicating how danceable the track is, ranging from 0 to 1.
- energy: A score indicating the energy level of the track, ranging from 0 to 1.
- key: The key that the track is in (e.g., C, D, E, etc.).
- loudness: The loudness of the track in decibels (dB).
- mode: The mode of the track (major or minor).
- speechiness: A score indicating how much speech-like content is in the track, ranging from 0 to 1.
- acousticness: A score indicating how acoustic the track is, ranging from 0 to 1.
- instrumentalness: A score indicating how instrumental the track is, ranging from 0 to 1.
- liveness: A score indicating the presence of an audience in the recording, ranging from 0 to 1.
- valence: A score indicating the positivity of the track, ranging from 0 to 1.
- tempo: The tempo of the track in beats per minute (BPM).
- time_signature: The time signature of the track (e.g., 4/4, 3/4, etc.).
- track_genre: The genre of the track (if available).

Specific Task Requirements:

- Perform an Exploratory data analysis (EDA) with visualization using the entire dataset. Discuss correlations and how the data is distributed. Notice that listening to one or more songs is part of the EDA. Try to answer to these questions:
    - Do you think the metrics provided by the MID are objectively measurable? Why or why not.
    - Do you think that the track_genre is something objectively measurable? Why or why not.
- Conduct descriptive statistics and visualizations to explore the correlations between other different features of the songs.
- Define another DataFrame without the track_genre column. Use K-means to group songs into different categories based on their audio features. In particular pick K = the number of unique genres available from the original dataset. Now compare the clustering result with the label given by the track_genre. Is this technique good for inferring the genre of a track? Discuss why.
- Build predictive models using machine learning algorithms to predict the popularity of songs based on their audio features.
- Evaluate the performance of the models using metrics such as accuracy, precision, recall, and F1 score.
- Interpret the results of the analysis and provide insights and recommendations for music producers and artists based on your findings.

3) Asteroids classification

The most famous space program agency in the observable universe provided you with this dataset which contains information on Near-Earth Objects (NEOs) they have been detected by Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE) mission. Your task is to analyze the dataset to gain insights into the characteristics of asteroids that could potentially impact Earth and to analyze this data and identify patterns and trends that can inform recommendations for monitoring and preventing potentially hazardous asteroid impacts.

Dataset

- Neo Reference ID: A unique identifier for each NEO.
- Name: The name of the NEO (if available).
- Absolute Magnitude: The brightness of the NEO as seen from a distance of one astronomical unit (AU).
- Est Dia in KM(min): The estimated minimum diameter of the NEO in kilometers.
- Est Dia in KM(max): The estimated maximum diameter of the NEO in kilometers.
- Est Dia in M(min): The estimated minimum diameter of the NEO in meters.
- Est Dia in M(max): The estimated maximum diameter of the NEO in meters.
- Est Dia in Miles(min): The estimated minimum diameter of the NEO in miles.
- Est Dia in Miles(max): The estimated maximum diameter of the NEO in miles.
- Est Dia in Feet(min): The estimated minimum diameter of the NEO in feet.
- Est Dia in Feet(max): The estimated maximum diameter of the NEO in feet.
- Close Approach Date: The date of the NEO's closest approach to Earth.
- Epoch Date Close Approach: The date and time of the NEO's closest approach to Earth in epoch format.
- Relative Velocity km per sec: The velocity of the NEO relative to Earth in kilometers per second.
- Relative Velocity km per hr: The velocity of the NEO relative to Earth in kilometers per hour.
- Miles per hour: The velocity of the NEO relative to Earth in miles per hour.
- Miss Dist.(Astronomical): The minimum distance between the NEO's orbit and Earth's orbit, measured in astronomical units (AU).
- Miss Dist.(lunar): The minimum distance between the NEO and the Moon's orbit, measured in lunar distances (LD).
- Miss Dist.(kilometers): The minimum distance between the NEO and Earth in kilometers.
- Miss Dist.(miles): The minimum distance between the NEO and Earth in miles.
- Orbiting Body: The celestial body that the NEO orbits.
- Orbit ID: A unique identifier for the NEO's orbit.
- Orbit Determination Date: The date of the NEO's orbit determination.
- Orbit Uncertainity: A measure of the uncertainty in the NEO's orbit.
- Minimum Orbit Intersection: The minimum distance between the NEO's orbit and Earth's orbit.
- Jupiter Tisserand Invariant: A measure of the NEO's relationship to Jupiter's orbit.
- Epoch Osculation: The date and time of the NEO's orbit determination in epoch format.
- Eccentricity: A measure of how elliptical the NEO's orbit is.

- Semi Major Axis: The distance from the center of the NEO's orbit to its farthest point (the semi-major axis).
- Inclination: The angle between the NEO's orbit and the ecliptic plane.
- Asc Node Longitude: The longitude of the ascending node of the NEO's orbit.
- Orbital Period: The time it takes for the NEO to complete one orbit around its orbiting body.
- Perihelion Distance: The closest distance between the NEO and the Sun.
- Perihelion Arg: The angle between the perihelion (closest approach to the Sun) and the ascending node of the NEO's orbit.
- Aphelion Dist: The farthest distance between the NEO and the Sun.
- Perihelion Time: The date and time of the NEO's closest approach to the Sun in epoch format.
- Mean Anomaly: The position of the NEO in its orbit relative to its mean position.
- Mean Motion: The speed at which the NEO moves in its orbit.
- Equinox: the reference frame for the position of the asteroid.
- Hazardous: binary variable indicating whether the asteroid is classified as "potentially hazardous" according to the criteria set by NASA. An asteroid is considered potentially hazardous if its orbit comes close enough to Earth's orbit that it could potentially collide with Earth in the future.

Specific Task Requirements

- Perform an Exploratory data analysis (EDA) with visualization using the entire dataset. Discuss correlations and how the data is distributed. Notice that a quick online search of these objects is part of the EDA.
- Conduct descriptive statistics and visualizations to explore the relationships between different features of the asteroids, such as the correlation between size and speed. Try to answer to these questions:
    - How many hazardous NEO are present in this dataset? Is it balanced?
    - Which are the most correlated features? Explain why.
- Build predictive models using machine learning algorithms to predict the likelihood of an asteroid impact based on the features measured by the National Aeronautics and Space Administration.
- Evaluate the performance of the models using metrics such as accuracy, precision, recall, and F1 score.
- Interpret the results of the analysis and provide insights and recommendations for monitoring and preventing potentially hazardous asteroid impacts based on your findings.

4) Air transportation fare prediction

You've been hired by one of the top earning travel companies. They provided you this dataset that can be used to predict flight prices for different airlines and routes, which can be helpful for customers who are looking to book flights at the best possible prices.The goal of this project is to predict the prices of flights given this information, which can be helpful for both customers and airlines in understanding the factors that affect flight prices and making informed decisions about booking flights.

Dataset

- airline: The name of the airline operating the flight.
- flight: The flight number.
- source_city: The city of departure.
- departure_time: The time of departure.
- stops: The number of stops on the flight.
- arrival_time: The time of arrival at the destination.
- destination_city: The city of arrival.
- class: The class of the ticket (e.g. economy, business, first).
- duration: The duration of the flight in hours and minutes.
- days_left: The number of days between the date of data collection and the date of departure.
- price: The price of the ticket in the local currency

Specific Task Requirements

- Preprocess and clean the data, handling missing values and outliers as necessary.
- Perform an Exploratory data analysis (EDA) with visualization using the entire dataset. Discuss correlations and how the data is distributed. Notice that searching for yourself the price of one of these routes can be part of EDA. Try to answer to these questions:
    - What is the metric that correlates the most to price?
    - Explain why.
- Select appropriate features for your model, and transform the data as necessary to prepare it for modeling.
- Split the data into training and testing sets.
- Train and evaluate at least 3 regression models, using appropriate metrics to assess their performance.
- Tune the hyperparameters of your best-performing model to achieve the best possible performance.
- Interpret the results of your analysis and provide insights and recommendations based on your findings.