DATA SCIENCE CERTIFICATION EXAM

UpX Academy

1 Nov 2020 Maximum Duration: 5 hrs.

INSTRUCTIONS

Note: This question paper consists of five sections: Section A, Section B, Section C, Section D and Section E. Please attempt all the relevant sections (according to course opted) to get certified.

Each one of the problems has its corresponding dataset present in the zip folder. You can follow the given guidelines as a rough path for solving the problems. Feel free to employ any algorithm you wish. Try different ones and choose the best one.

You are required to submit the code(.ipynb notebook) , the error of the model (also mention the metric you are using to evaluate your model) and a small write-up of the approach you've taken, mentioning why you chose a particular algorithm and how you proceeded.

Please write your answers in a MS-Word doc file, create a zip file containing your doc and codes. Send the zip file to dssupport@upxacademy.com .

All the submissions (code, report, etc.) will undergo a plagiarism test and copied papers will be subjected to disqualification.

SECTION A – Data Analytics with Python

Duration: 1 hour

Each question carries 100 marks. You can choose **1 out of 2 tasks** in this section. To clear this section, you're required to score at least 70% marks in the task you choose.

Task 1: Global Space Launches

This dataset contains data about every Space launch from 1957 to the Modern-day.
The attributes are given below:

Company Name - Name of Organization responsible for launch
Location - Name of launch location
Details - Specification of each rocket
Status Rocket - Showing if a rocket is currently in use
Rocket- Cost in Millions of rocket used
Status Mission- One of 4 categorical elements showing the the result of the launch
Country of Launch - the country where the launch took place
Company's Country of Origin - the country that the organization is from
Private or State Run -the organizations category think SpaceX for private and NASA for State
DateTime - Date and Time
Year - year of launch
Month - Month of launch
Day - Day of launch
Date - Date of launch
Time - Time of launch

Objective

Perform Exploratory Data analysis on the dataset to understand the following:

1. Illustrate with appropriate plot the countries wherein successful launches have taken place over the years.
2. Which are the major players in the private sector of the space industry?
3. Which organization is the most collaborative and works the most with other countries?Illustrate with plots
4. Analyse at what time of the day do the launches usually fail.
5. Explore with plots the budget allocated to GSLV and PSLV launches from ISRO. Also assess the success rate of launches conducted by ISRO for GSLV and PSLV spaceships via plots through the years.

You can come up with more insights on the data. Please use **Python** Programming.

Task 2: International Football Analysis

This dataset contains women's international football results throughout the years.

The dataset file has 9 columns:

- date - date of the match
- home_team - the name of the home team
- away_team - the name of the away team
- home_score - full-time home team score including extra time, not including penalty-shootouts
- away_score - full-time away team score including extra time, not including penalty-shootouts
- tournament - the name of the tournament
- city - the name of the city/town/administrative unit where the match was played
- country - the name of the country where the match was played
- neutral - TRUE/FALSE column indicating whether the match was played at a neutral venue

Objective

1. What is the distribution of matches played by countries through the years? Illustrate via appropriate plots
2. Which countries host the most matches where they themselves are not participating in?
3. What is the distribution of goals scored by top five winning teams in the FIFA World cup through the years? Illustrate via appropriate plots
4. Analyse via plots whether there was advantage for the home teams during the Olympic games through the years
5. Analyse the Indian football team performance over the years in AFC Championship, and Asian Cup qualifications

You can come up with more insights on the data. Please use **Python** Programming.

SECTION B – Machine Learning

Duration: 1 hour

Each question carries 100 marks. You can choose **1 out of 2 tasks** in this section. The grading will be done out of 2 tasks. To clear this section, you're required to get at least 70% marks in the task you choose.

Task 1: Apartment Real estate analysis

This dataset is extracted from prominent real estate portals. The targeted housing market for this project is Dubai, United Arab Emirates. Predict the price of an apartment from the following attributes:

- id-property id
- neighborhood-neighborhood name
- latitude-location data
- longitude-location data
- price-market price
- size_in_sqft-covered area of apartment
- price_per_sqft-price per square feet for the apartment
- no_of_bedrooms-number of bedrooms in apartment
- no_of_bathrooms-number of bathrooms in apartment
- quality-quality based on number of amenities. Contains category labels Ultra, High, Medium, and Low

The rest of the attributes are either True or False

- maid_room
- unfurnished
- balcony
- barbecue_area
- built_in_wardrobes
- central_ac
- childrens_play_area
- childrens_pool
- concierge
- covered_parking
- kitchen_appliances
- lobby_in_building
- maid_service
- networked
- pets_allowed
- private_garden
- private_gym
- private_jacuzzi
- private_pool
- security
- shared_gym
- shared_pool
- shared_spa
- study
- vastu_compliant
- view_of_landmark
- view_of_water
- walk_in_closet

Objective

Predict the apartment price using numerical data.

Guidelines

1. Explore and prepare the data.

2. Create training and testing data for the model
3. Train and test the model using any three regression algorithms or more and compare rmse error.
4. Show or visualize the output.

Task 2: Atmospheric particle classification

The dataset is Monte Carlo generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. The task is to classify the high energy particles to discriminate statistically those caused by primary gammas (signal), from the images of hadronic showers initiated by cosmic rays in the upper atmosphere (background) The attributes of the dataset are given below:

1. fLength: - major axis of ellipse [mm]
2. fWidth: - minor axis of ellipse [mm]
3. fSize: - 10-log of sum of content of all pixels [in #phot]
4. fConc: - ratio of sum of two highest pixels over fSize [ratio]
5. fConc1: - ratio of highest pixel over fSize [ratio]
6. fAsym: - distance from highest pixel to center, projected onto major axis [mm]
7. fM3Long: - 3rd root of third moment along major axis [mm]
8. fM3Trans: - 3rd root of third moment along minor axis [mm]
9. fAlpha: - angle of major axis with vector to origin [deg]
10. fDist: - distance from origin to center of ellipse [mm]
11. class: g,h - gamma (signal), hadron (background)

Objective

Based on the given features, classify the particle as belonging to either gamma or hadron based on various attributes.

Guidelines

1. Explore and prepare the data
2. Create training and testing data for the model
3. Train and test the model using any three or more classification algorithms and compare accuracies between all.
4. Show or visualize the output.

SECTION C – Advance Machine Learning

Duration: 1 hour

Each question carries 100 marks. You can choose **1 out of 2 tasks** in this section. To clear this section, you're required to score at least 70% marks in the task you choose.

Task 1: Star classification

Star Classification uses the spectral data of stars to categorize them into different categories. The modern stellar classification system is known as the Morgan–Keenan (MK) classification system. It uses the old HR classification system to categorize stars with their chromaticity and uses Roman numerals to categorize the star's size.

In this Dataset, we will be using Absolute Magnitude and B-V Color Index to Identify Giants and Dwarfs.

The attributes of the dataset are given below:

1. Vmag-Visual Apparent Magnitude of the Star
2. Plx-Distance Between the Star and the Earth
3. e_Plx-Standard Error of Plx
4. B-V-B-V color index.
5. SpType-Spectral type
6. Amag-Absolute Magnitude of the Star
7. TargetClass-Whether the Star is Dwarf (0) or Giant (1)

Objective

The classification goal is to predict whether the star is dwarf or giant.
Please apply any two boosting techniques learned in the module to solve this business problem.

Task 2: Predict impact of air quality on mortality rates

The goal is to predict mortality rates (number of deaths per 100,000 people) for each English region using daily means of ozone (O3), nitrogen dioxide (NO2), PM10 (particulate matter with diameter less than or equal to 10 micrometers), PM25 (2.5 micrometers or less) and Temperature..

The description of data are as follows:

1. Id - a unique id
2. region - an identifier of a region in England
3. O3 mean - ozone, daily average computed for a particular region
4. PM10 mean - particulate matter 10 micrometers or less in diameter, daily average
5. PM25 mean - particulate matter 2.5 micrometers or less in diameter, daily average
6. NO2 mean - nitrogen dioxide, daily average
7. Temperature mean - Temperature at 2 m, daily average
8. mortality rate - number of deaths per 100000 people.

Objective

The objective is to predict the mortality rate using advanced machine learning techniques. Please apply the stacking technique learned in the module to solve this problem. Please implement in python notebook

SECTION D – Deep Learning with NLP

Duration: 1 hour

Each question carries 100 marks. You can choose **1 out of 2 tasks** in this section. The grading will be done out of 2 tasks. To clear this section, you're required to get at least 70% marks in the task you choose.

Task 1: Disease Prediction

This dataset contains information about diseases. It has 132 parameters on which 42 different types of diseases can be predicted.

The file has 133 columns. 132 of these columns are symptoms that a person experiences and the last column is the prognosis. The symptoms do not require addition description These symptoms are mapped to 42 diseases you can classify these set of symptoms to.

Objective

Classify symptoms into the 42 diseases.

Guidelines

1. Explore and prepare the data
2. Create training and testing data for the model
3. Use a deep neural network and proper activation function to classify into particular disease.
4. Test the model
5. Show or visualize the output.

Task 2: Classification of news based on headlines

This is a text classification task.
This dataset contains headlines from The Onion articles and real "Onion-like" news articles from the subreddit. Your goal is to classify the news into The Onion articles which are labeled 1 and the NotTheOnion articles which are labeled 0.

Objective

Perform text classification of the given dataset.

Guidelines

1. Prepare the data
2. Create training and testing data for the model
3. Train the model in Keras using NLP models covered in Deep Learning track .
4. Test the model

SECTION E (Tableau)

Duration: 1 hour

This section is a shortened format of the Tableau Desktop 10 qualified associate exam
Each question carries 20 points, in order to pass this section, you need to score 75%. Answer all questions.

Please use a separate word document to type in your answers. Be concise in your replies.
Please install Tableau 10 Desktop edition or Tableau Public edition, you may download the free trials from the Tableau homepage through providing your email

For submitting graphs you have to save all the worksheets on the tableau public and send the tableau public profile link to dssupport@upxacademy.com.

Data set for this section: New York city Airbnb

The dataset attributes description is provided below:

id: A unique number identifying an Airbnb listing.

name: name of the listing

host_id: A unique number identifying an Airbnb host.

host_name: Name of the host

neighborhood_group: Region of the Area

neighborhood: Area

room_type: One of “Entire home/apt”, “Private room”, or “Shared room”.

latitude: latitude co-ordinates

longitude: longitude co-ordinates

price: The price (in $US) for a night stay. In early surveys, there may be some values that were recorded by month.

minimum_nights: The minimum stay for a visit, as posted by the host.

number_of_reviews: The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of reviews can be used to estimate the number of visits. Note that such an estimate will not be reliable for an individual listing (especially as reviews occasionally vanish from the site), but over a city as a whole it should be a useful metric of traffic.

last_review: Latest review

reviews_per_month: The number of reviews that a listing has received per month.

host_listings_count: The number of listings for a particular host.

availability_365: The number of days for which a particular host is available in a year.

Questions

1. Find the Top 20 Neighborhood Areas having high count of No of Reviews and give your inference
2. Using LOD calculations find which Neighborhood has a high Average price with respect to Neighborhood Group and give your inference.
3. a) Create Pareto Chart for Price, Availability-365 and Neighborhood Group
   b) Create Waterfall Chart for Price vs Neighborhood Group
4. Plot a Tree Map for Host Name vs Price w.r.t No of Reviews
5. Plot a Bubble Chart to see which Room_Type has high Minimum Nights count
6. Visualize a map of Neighborhood Groups having high amount of No of Reviews

Create an interactive dashboard by making use of the above sheets.