DATA SCIENCE CERTIFICATION EXAM

UpX Academy

6 Dec 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 fully executed 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 Jupyter notebook/Colaboratory notebook, 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: Explore Sneaker Data

This dataset contains data about sneaker sales from between 9 Jan 2017 and 13 Feb 2019. There are 99,956 total sales in the data set. The sample consists of U.S. sales only.

The attributes are given below:

Order Date - Order Date of the sneaker
Brand - Brand of sneaker
Sneaker Name - Name of the sneaker
Sale Price - Sale price of the sneaker
Retail Price - Retail price of the sneaker
Release Date - Release date of the sneaker
Shoe Size - Shoe size of the sneaker
Buyer Region - Buyer region of the sneaker

Objective

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

1. Illustrate with appropriate plot the top 5 regions where the maximum sales have taken place over the years in terms of numbers.
2. What shoe sizes and brands are most popular?
3. Which shoes have the worst profit margins? Illustrate with plots
4. Analyse if the shoe sales have any impact during the days of December 20th to January 10th over the years. Illustrate with plots
5. Explore with plots what factors affect the profit margins of Adidas Yeezy Boost shoe sales in California.

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

Task 2: Explore residents in Singapore based on various factors

This dataset contains information about different ethnic groups living in Singapore dated from 1957 to 2018 along with their age group, gender, and population count..

The attributes are given below:

year - Year
level_1 - Describes ethnic/gender based groups
level_2 - Age groups
value - Population count

Objective

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

1. Identify the largest Ethnic group in Singapore. Illustrate their average population growth over the years through appropriate plots and what proportion of the total population do they constitute
2. Which are the largest age groups residing in Singapore? Explain using appropriate plots.
3. Explore population trends for ethnic group ‘Malays’ for both female and male genders. Illustrate with plots
4. Extract total number of ethnic groups residing in Singapore age wise.
5. Plot an year wise analysis of population for total residents of the age group 15 years to 39 years. Plot multiple plots in the same chart.

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: Mental Exhaustion analysis

According to an anonymous survey, about 450 million people live with mental disorders that can be one of the primary causes of poor health and disability worldwide. These days when the world is suffering from a pandemic situation, it becomes really hard to maintain mental fitness. Predict the burn rate from the following attributes:

- Employee ID: The unique ID allocated for each employee
- Date of Joining: The date-time when the employee has joined the organization
- Gender: The gender of the employee (Male/Female)
- Company Type: The type of company where the employee is working (Service/Product)
- WFH Setup Available: Is the work from home facility available for the employee (Yes/No)
- Designation: The designation of the employee of work in the organization. (In the range of [0.0, 5.0] bigger is higher designation)
- Resource Allocation: The amount of resource allocated to the employee to work, ie. number of working hours. (In the range of [1.0, 10.0] higher means more resource)
- Mental Fatigue Score: The level of fatigue mentally the employee is facing (where 0.0 means no fatigue and 10.0 means complete fatigue)

- Burn Rate: The value we need to predict for each employee telling the rate of Burn out while working (in the range of [0.0, 1.0] where the higher the value more is the burn out)

Objective

Predict the Burn Rate.

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: Asteroid classification

The dataset contains details about asteroids classified as potentially hazardous by NASA. The attributes of the dataset are given below:

1. Epoch (TDB) - Osculating epoch of the elements given as the modified Julian date (Julian date - 2400000.5) TDB
2. a (AU) - Semi-major axis of the orbit in AU
3. e - Eccentricity of the orbit
4. i (deg) - Inclination of the orbit with respect to the ecliptic plane and the equinox of J2000 (J2000-Ecliptic) in degrees
5. w (deg) - Argument of perihelion (J2000-Ecliptic) in degrees
6. Node (deg) - Longitude of the ascending node (J2000-Ecliptic) in degrees
7. M (deg) - Mean anomoly at epoch in degrees
8. q (AU) - Perihelion distance of the orbit in AU
9. Q (AU) - Aphelion distance of the orbit in AU
10. P (yr) - Orbital period in Julian years
11. H (mag) - Absolute V-magnitude
12. MOID (AU) - Minimum orbit intersection distance (the minimum distance between the osculating orbits of the NEO and the Earth)
13. ref - Orbital solution reference
14. class - Object classification: NEA="Near-Earth Asteroid", AMO="Amor", APO="Apollo", ATE="Aten", or IEO="Interior Earth Object".

(AU)-- Astronomical distance Unit: 1.0 AU is about 1.5x10^8 km (roughly the average distance between the Earth and the Sun).

Objective

Based on the given features, classify the asteroid as belonging to either of the classes NEA, AMO, APO, ATE, or IEO 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: Employee Promotion.

In this Dataset, we will be evaluating whether an employee will get a promotion or not based on various attributes .

The attributes of the dataset are given below:

1. employee_id
2. department
3. region
4. education
5. gender
6. recruitment_channel
7. no_of_trainings

8. age
9. previous_year_rating
10. length_of_service
11. KPIs_met >80%
12. awards_won?
13. avg_training_score
14. is_promoted

Objective

The classification goal is to predict whether the employee will get promoted or not.
Please apply any two boosting techniques learned in the module to solve this business problem.

Task 2: Engineering Graduate Salary prediction

Every year on an average 1.5 million students get their degree in engineering, but due to lack of skill required to perform technical jobs less than 20 percent get employment in their core domain.

A relevant question is what determines the salary and the jobs these engineers are offered right after graduation. Various factors such as college grades, candidate skills, the proximity of the college to industrial hubs, the specialization one have, market conditions for specific industries determine this. On the basis of these various factors, your objective is to determine the salary of an engineering graduate in India.

The description of data are as follows:

1. ID: A unique ID to identify a candidate
2. Salary: Annual CTC offered to the candidate (in INR)
3. Gender: Candidate's gender
4. DOB: Date of birth of the candidate
5. 10percentage: Overall marks obtained in grade 10 examinations
6. 10board: The school board whose curriculum the candidate followed in grade 10
7. 12graduation: Year of graduation - senior year high school
8. 12percentage: Overall marks obtained in grade 12 examinations
9. 12board: The school board whose curriculum the candidate followed
10. CollegeID: Unique ID identifying the university/college which the candidate attended for her/his undergraduate
11. CollegeTier: Each college has been annotated as 1 or 2. The annotations have been computed from the average AMCAT scores obtained by the students in the college/university. Colleges with an average score above a threshold are tagged as 1 and others as 2.
12. Degree: Degree obtained/pursued by the candidate

13. Specialization: Specialization pursued by the candidate
14. CollegeGPA: Aggregate GPA at graduation
15. CollegeCityID: A unique ID to identify the city in which the college is located in.
16. CollegeCityTier: The tier of the city in which the college is located in. This is annotated based on the population of the cities.
17. CollegeState: Name of the state in which the college is located
18. GraduationYear: Year of graduation (Bachelor's degree)
19. English: Scores in AMCAT English section
20. Logical: Score in AMCAT Logical ability section
21. Quant: Score in AMCAT's Quantitative ability section
22. Domain: Scores in AMCAT's domain module
23. ComputerProgramming: Score in AMCAT's Computer programming section
24. ElectronicsAndSemicon: Score in AMCAT's Electronics & Semiconductor Engineering section
25. ComputerScience: Score in AMCAT's Computer Science section
26. MechanicalEngg: Score in AMCAT's Mechanical Engineering section
27. ElectricalEngg: Score in AMCAT's Electrical Engineering section
28. TelecomEngg: Score in AMCAT's Telecommunication Engineering section
29. CivilEngg: Score in AMCAT's Civil Engineering section
30. conscientiousness: Scores in one of the sections of AMCAT's personality test
31. agreeableness: Scores in one of the sections of AMCAT's personality test
32. extraversion: Scores in one of the sections of AMCAT's personality test
33. nueroticism: Scores in one of the sections of AMCAT's personality test
34. openess*to*experience: Scores in one of the sections of AMCAT's personality test

Please note AMCAT is a job portal.

Objective

The objective is to predict the salary 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: Web Page Phishing Detection

Phishing is a type of cybercrime in which a target or targets are contacted by email, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords. The information is then used to access important accounts of the victims and it mostly results in identity theft and financial loss.
The provided dataset includes 11430 URLs with 87 extracted features. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs.

Objective

Classify websites into either phishing or legitimate based on various attributes.

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 a phishing or legitimate webpage.
4. Test the model
5. Show or visualize the output.

Task 2: Analysis of Sarcastic and Ironic statements

This is a text classification task.

Irony in language is when a statement is produced with one meaning but the intended meaning is exactly the opposite. The file contains 2 columns:

- tweet: The text of the tweet
- class: The respective class to which the tweet belongs. There are 4 classes -:
    - Irony
    - Sarcasm
    - Regular
    - Figurative (both irony and sarcasm)

This dataset contains **67997** comments, which have been labeled as irony,sarcasm,regular and figurative by human annotators.
Your goal is to classify each sentence into one of the 4 classes.

Objective

Perform classification of the given text into one of the 4 classes.

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: Data Breaches 2018 updated (.csv format)

Questions

1. Which sector has the maximum number of records stolen(make use of LOD calculations to find avg no of records stolen)
2. Which year has the maximum records lost and from which sector.
3. Visualize the methods by which the maximum records leaked.
4. By making use of filters on method of leak, find which entity records are lost in maximum numbers from all the methods of leak.
5. By making use of a tree map, find which sector has the maximum number of records stolen and which source name provides that information.

Create an interactive dashboard by making use of above sheets.