Data Science Bootcamp

Capstone Project 3

Data Science Academy

Project Description

This capstone project includes 2 separate tasks.

Please, submit your solutions to homework@dsa.az by August 18th, 23:59 2021.

1. Data Cleaning

This task aims at sharpening your python skills and testing your ability to conduct the advanced data cleaning techniques in real life. In order to successfully complete this task, you need to provide step-by-step implementation that leads to the results. Make sure that you also mark the correct answers for the multiple-choice questions.

Submission deliverables (task1.zip):

1) Jupyter Notebook (*.ipynb): Clearly showing the implementation details and the cell outputs
2) HTML file: HTML version of the Jupyter Notebook
3) requirements.txt: Clearly show the packages you have used for this task

Note: All code needs to be commented!

Please provide the python version used and OS environment details. Questions are equally weighted.

2. McKinsey&Company Prohack

This is an anonymized dataset provided by the one of the most prestigious consulting companies around the world that operate in many countries including Azerbaijan. In this task, you need to predict the target value ("y") given all the other features. At the end of this task, you will be ready for your first data science job as this task encomposes all of the necessary phases a top-notch data scientist in the finance industry passes through to successfully build an in-house prediction model. This task is composed of 2 components:

1) Prediction (Regression)
2) Optimization

You must show your authentic work in all of the following sections:
1.- Data cleaning
2.- Data Preprocessing
3.- EDA (Exploratory Data Analysis)
4.- Feature Engineering and Feature
Selection5 - Modeling
6 - Performance Evaluation
7 - Optimization

Submission deliverables (task2.zip):
1) Jupyter Notebook (*.ipynb): Clearly showing the implementation details and the cell outputs
2) HTML file: HTML version of the Jupyter Notebook
3) requirements.txt: Clearly show the packages you have used for this task
4) submission.csv file as it is described in the assignment details

Note: All code needs to be commented! Please provide the python version used and OS environment details. Prediction phase is worth **80%** and the optimization phase is worth **20%** of the final grade. A starter kernel is provided. Solution to this task will be released after December 9th, 2020.

1 - Data Cleaning

Electricity Contract Selection

INTRODUCTION

Just over a year ago you had a smart electricity meter installed at your home. This particular model of smart meter reports your usage to your electricity provider every hour. Your electricity contract is due for renewal and being a top financial analyst you decide to identify which contract will minimize your annual electricity cost.

There are 3 different types of electricity contracts available to you:

- **No Flex:** The cost per kWh of electricity is constant for the entire year
- **Monthly Flex:** The cost per kWh of electricity fluctuates depending on the month
- **Hourly Flex:** The cost per kWh of electricity fluctuates based on the time of day

To decide which contract is optimal, you decide to compare the costs under each of the contracts assuming your usage remains exactly the same as last year. You've contacted your current electricity supplier to obtain your electricity usage history for the last year. Unfortunately, they provided it to you in a poorly structured and unsorted format, so you will need to clean it up prior to doing your analysis.
In the data the hour field identifies when the hour starts, so 8AM would be from 8:00AM until 8:59AM.

SUPPORTING DATA

You've placed your usage history and the specifics of each of the electricity contracts into the supplied workbook. Your usage history is located on the "Usage" sheet while the specifics about each of the electricity contracts

QUESTIONS

Question 1

What is your average hourly electricity usage?

a. 0.641kWh

b. 0.782kWh

c. 0.884kWh

d. 0.937kWh

Question 2

What is your average electricity usage perhour in February?

a. 0.760kWh

b. 0.784kWh

c. 0.808kWh

d. 0.834kWh

Question 3

Which day of the week has the highestaverage usage?

a. Sunday

b. Monday

c. Tuesday

d. Wednesday

Question 4

What is the highest amount of electricityused in a continuous 4 hour period?

a. 17.237kWh

b. 17.327kWh

c. 17.422kWh

d. 17.487kWh

Question 5

Based on your historic electricity usage, what would your annual cost of electricitybe under the "Monthly Flex" contract?

a. $1350.73

b. $1420.06

c. $1450.26

d. $1493.77

Question 6

Based on your historic electricity usage, which of the three contracts would pro- duce the lowest annual cost?

a. The No Flex plan

b. The Monthly Flex plan

c. The Hourly Flex plan

d. Impossible to Determine

2 - McKinsey&Company Prohack

Achieve Singularity

"Beeep…Beeeep….Beeeep… Hooomans*, are you there?..."

This very strange transmission is coming from your narrowband radio signal receiver, pointed towards one of the farthest away galaxies. It's early morning, you are sitting in your radio observatory high in the mountains.

For the last 10 years you've been a Chief Data Scientist in one of the best astrophysics research teams in the world. You are enjoying a quiet time with a cup of coffee and reviewing the data reports from last night, when this strange sound arrived. You almost spill your coffee in surprise. "Am I dreaming?" is your first thought as you move closer towards the speaker and listen…

"Beep…Beeeep….Beeeep… To all Hooomans who can hear us – we need your help"
You lean closer and grab a notebook and a pencil – you don't really trust computers when it comes to such important tasks as taking notes from a radio transmission. You start recording everything that the strange voice from light years away is saying.

"… We need serious Data Science help and we know you Hooomans are the best at it…. We are an intergalactic species which have almost achieved singularity and the highest possible levels of development. We travel fast through space and explore other galaxies"

"The only essence that we consume is energy, measured in DSML units…Our populace is widespread and we live across many different star clusters and galaxies. What we need now is to optimize our well-being across all those galaxies… We have a lot of data but our computers and methods are too weak – we urgently need your data science knowledge to help us

Only **two steps** prevent us from achieving singularity:

1) To understand what makes us better off.

Our elders used the composite index to measure our well-being performance, but this knowledge has disappeared in the sands of time.
Use our data and train your model to predict this index with the highest possible level of certainty.

2) To achieve the highest possible level of well-being through optimized allocation of additional energy

We have discovered the star of an unusually high energy of 50000 zillion DSML.

We have agreed between ourselves that:
- no one galaxy will consume more than 100 zillion DSML
- at least 10% of the total energy will be consumed by galaxies in need with existence expectancy index below 0.7.

Think of our galaxies as your "countries" (or how you call them??) and our population as citizens. We have similar healthcare and wellbeing characteristic as you, Hooomans"
"We are sending all the data to you right now. Let the data be with you, Hoomans… … …"

Transmission suddenly ends. You put your notebook and pencil away and start thinking. Youreally want to help this species optimize their well-being. You open up Python and upload the dataset from the narrowband radio signal receiver. It will be another great day at the observatory today.

Description Data Received

The solutions are evaluated on two criteria: **predicted future Index values** and **allocated energy from a newly discovered star**

1) Index predictions are evaluated using RMSE metric

2) Energy allocation is also evaluated using RMSE metric and has a set of known factors thatneed to be taken into account.

Every galaxy has a certain limited potential for improvement in the index described by the following function:

Potential for increase in the Index = -np.log(Index+0.01)+3

Likely index increase dependent on potential for improvement and on extra energy availability is described by the following function:

Likely increase in the Index = extra energy * Potential for increase in the Index **2 / 1000

There are also several constraints:

- In total there are 50000 zillion DSML available for allocation and no galaxy at a point in time
- no galaxy should be allocated more than 100 zillion DSML or less than 0 zillion DSML.
- Galaxies with low existence expectancy index below 0.7 should be allocated at least 10% of thetotal energy available in the foreseeable future

3) Performance is based on a combined scaled metric:

80% prediction task RMSE + 20% optimization task RMSE * lambda

where lambda is a normalizing factor

4) The submission should be in the following format:

Variable	Description
index	Unique index from the test dataset in the ascending order
pred	Prediction for the index on interest
opt_pred	Optimal energy allocation

info@datascienceacademy.az
Aşıq Alı 24, Nərimanov rayonu