Homework 2

Data Analysis and Machine Learning with Python

Problem description:

There are four datasets and the corresponding questions. Each dataset is 50 points. Please choose **two** datasets for Homework 2, and **please tell me why you selected these two dataset**. Each question requires an explanation of your thoughts and actions, along with relevant models, evidence or charts if possible. Additionally, not all questions have standard answers.

If you have time and interests, you can also answer all questions with four datasets. This action will help you earn more points for Homework 2.

Note: same as Homework 1, you need to prepare a report recording all ideas, steps, processes, results, and so on of your answers for the questions.

Dataset 1 descriptions (maintenance_prediction.csv):

A company operates a fleet of devices that transmit daily sensor readings. They aim to develop a predictive maintenance solution to anticipate when maintenance should be conducted. This approach offers potential cost savings compared to routine or time-based preventive maintenance, as tasks are only performed when necessary.

The objective is to construct a predictive model using machine learning to predict the probability of device failure. When building this model, it's crucial to minimize false positives and false negatives. The target variable to predict is labeled "failure," with binary values of 0 for non-failure and 1 for failure.

Column descriptions
date: the data received day
device: device ID
failure: device can work or not, 0: workable/non-failure, 1: failure
metric1~9: features

Q1. How many unique device IDs are there in this dataset?

Q2. You are asked to do data analysis. What will you find?

Q3. You are asked to build a prediction model. Which kind of machine learning will be used and why? Supervised learning or Unsupervised learning? Regression, classification, or clustering? Which model will you use?

Q4. Can you find the important features, informative features, or coefficient values? (Note: the answer will depend on your selected machine learning model.)

Q5. There are two data from two devices, please predict the corresponding failure values.

	metric								
device	1	2	3	4	5	6	7	8	9
AA	127175526	4109.434	3.90566	54.63208	15.46226	258303.5	30.62264	30.62264	23.08491
BB	4527376	0	0	0	3	24	0	0	0

Dataset 2 descriptions (Insurance_dataset.csv and Insurance_validation.csv):

Context:

Insurance companies offering life, health, and property and casualty insurance are utilizing machine learning (ML) to enhance customer service, detect fraud, and improve operational efficiency. The data provided by an insurance company, which is not shared with other companies, is leveraged to benefit from ML. This particular company provides health insurance to its customers. A model can be developed to predict whether policyholders (customers) from the past year will also be interested in the vehicle insurance provided by the company.

An insurance policy is an agreement in which a company agrees to provide compensation for specified loss, damage, illness, or death in exchange for the payment of a specified premium. The premium is the amount of money that the customer must pay regularly to the insurance company for this guarantee.

For instance, you might pay an annual premium of USD $5,000 for a health insurance cover of USD $200,000 so that if you fall ill and need to be hospitalized in that year, the insurance company will cover the cost of hospitalization up to USD $200,000. The concept of probabilities comes into play here. For example, out of 100 customers paying a USD $5,000 premium each year, only a few (say 2-3) might get hospitalized that year, not everyone. This way, everyone shares the risk of everyone else.

Similarly, there is vehicle insurance where customers pay a premium each year to the insurance company so that in the event of an unfortunate accident involving the vehicle, the insurance company will provide compensation (called 'sum assured') to the customer.

Content:

Developing a model to predict whether a customer would be interested in vehicle insurance is highly beneficial for the company. With this prediction, the company can plan its communication strategy to reach out to those customers and optimize its business model and revenue.

We have information about:
- Demographics (Gender, Age, Driving_License, Region Code, Previously_Insured),
- Vehicles (Vehicle Age, Vehicle Damage),
- Policy (Annual Premium, Policy Sales Channel, Vintage)
- Target (Response)

Please use Insurance_dataset.csv to be the training and testing datasets.

Q1. Which kind of data selection method will you use to split csv data to training and testing datasets? sequential or random? WHY?

Q2. In class, we learned many model evaluation methods, such as confusion matrix, accuracy score, precision score, recall score, and so on. In addition to the confusion matrix and accuracy score, which must be used in the Q3 and Q4, if you

were to choose two evaluation metrics/scores, which two would you choose? Why?

Q3. Please use eight classification models taught in the class and find their own best parameters' settings.

Q4. Which prediction model is the best between these eight classification models? WHY?

Q5. Insurance_validation.csv is a validation dataset. Please use your best prediction model to get the "Response" and output the results to a csv file. The format of the csv file is the following sample.

id,	Response
57782,	1
286811,	0
…	…
420570,	0

Dataset 3 descriptions (feedback_sentiment.csv):

This dataset comprises customer sentiments expressed across various sources such as social media, review platforms, testimonials, and more. It includes text, sentiment (positive or negative), source of the sentiment, date/time of the sentiment, user ID, location, and confidence score. The sentiments reflect customers' opinions and experiences with products, services, movies, music, books, restaurants, websites, customer support, and more.

You can perform the following tasks on the dataset:

Data cleaning

Sentiment analysis

Statistical analysis

Data visualization

Text preprocessing

Topic modeling

Feature engineering

Machine learning modeling

Data integration

Data aggregation

Note: there is only one column in the csv file. You can use the separator argument within Pandas to read the data.

Q1. How many columns are there in this csv file and what are these columns’ names?

Q2. Do the following steps and show the results.

- Clean data
- Dop nan
- Convert data and time to datatime
- Create new features - month, day, and hour
- Create new features again - Total Words, Total Chars, and Total Words After Transformation of "Text," where Total Words After Transformation means *The natural logarithm of the word count of the "Text"*.

Q3. Do the following visualizations, eg., histplot, displot, barplot, kdeplot, etc., and write your findings.

1) by Sentiment 'Positive' and 'Negative'
2) by 'Source' and 'Sentiment'
3) by 'Location' and 'Sentiment'
4) by 'Confidence Score' and 'Sentiment'
5) by 'Month' and 'Sentiment'
6) by 'Day' and 'Sentiment'
7) by 'hour' and 'Sentiment'
8) by 'Total Words' and 'Sentiment'
9) by 'Total Chars' and 'Sentiment'
10) Wordcloud by Sentiment = Negative
11) Wordcloud by Sentiment = Positive
12) by Top 25 Negative Words
13) by Top 25 Positive Words

Q4. Build eight classification models taught in the class and find their own best parameters' settings with the dataset, where "sentiment" is set as the target.

Q5. Show the confusion matrix and classification report of your eight models, compare these models and write your findings.

Dataset 4 descriptions (store_sale_prediction.zip):

Context:
Forecasts aren’t just for meteorologists. Governments forecast economic growth. Scientists attempt to predict the future population. And businesses forecast product demand—a common task of professional data scientists. Forecasts are especially relevant to brick-and-mortar grocery stores, which must dance delicately with how much inventory to buy. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leading to lost revenue and upset customers. More accurate forecasting, thanks to machine learning, could help ensure retailers please customers by having just enough of the right products at the right time.
Current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing.

Potential Impact:
If successful, you'll have flexed some new skills in a real world example. For grocery stores, more accurate forecasting can decrease food waste related to overstocking and improve customer satisfaction. The results of this ongoing competition, over time, might even ensure your local store has exactly what you need the next time you shop.

Dataset Description:
You will predict sales for the thousands of product families sold

at Favorita stores located in Ecuador. The dataset.csv includes dates, store and product information, whether that item was being promoted, as well as the sales numbers. Additional files include supplementary information that may be useful in building your models.

File Descriptions and Data Field Information

- dataset.csv: The data, comprising a time series of features store_nbr, family, and onpromotion as well as the target sales.
    – store_nbr: identifies the store at which the products are sold.
    – family: identifies the type of product sold.
    – sales: gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).
    – onpromotion: gives the total number of items in a product family that were being promoted at a store at a given date.

- validations.csv: The validation data, having the same features as the data in the dataset. You will predict the target sales for the dates in this file. The dates in the test data are for the 15 days after the last date in the training data.

- stores.csv: Store metadata, including city, state, type, and cluster.
    – cluster is a grouping of similar stores.

- oil.csv: Daily oil price. Includes values during both the dataset.csv and validations.csv timeframes. (Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices.)

- holidays_events.csv: Holidays and Events, with metadata
    – NOTE: Pay special attention to the transferred column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to pay back the Bridge.
    – Additional holidays are days added to a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Additional Notes:

- Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

Q. Please predict the results of the data in validations.csv, output the results to a csv file, and record and explain all the steps/processes.

The format of the csv file is the following sample.

id,	sales
3000888,	0.0
3000889,	0.0
…	…