Homework 3

Data Analysis and Machine Learning with Python

Problem description:

There are two datasets and the corresponding questions. Each dataset is 50 points. If you use AI tool(s) to help this assignment, **please tell me why you selected that/those AI tool(s)** and **attach screenshots or a webpage of your conversation with the AI tool(s)**. 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.

Note: same as Homework 1 and 2, you need to prepare a report recording all ideas, steps, processes, results, and so on of your answers for the questions.

Dataset 1 descriptions (HW3-1.csv):

Dataset description:
- Number of samples: 5500
- Number of features: 1500
- Label: Binary class (0 or 1)

Objective:
This assignment is designed to deepen your understanding of different feature evaluation techniques—namely informative features, important features, and permutation importance—and how various machine learning models handle feature relevance and interpretation.

You will answer questions based on your analysis using eight classifiers:
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVC)
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Decision Tree
- Random Forest
- XGBoost

You are expected to use Python (e.g., scikit-learn, xgboost) to support your answers with empirical evidence when applicable.

I. General Concept Questions

(2%) Q1. What is the key difference between informative features and important features?

(2%) Q2. In what scenario might a feature be informative but not important to a model?

(2%) Q3. What could cause permutation importance scores to vary across different random seeds?

(2%) Q4. How can highly correlated features affect permutation importance?

II. Dataset Specific Questions

(4%) Q5. In this dataset, which features do you think are important? Do they have equal importance, or is there a hierarchy among them?

(4%) Q6. According to Q5, how would you validate whether the redundant features still contribute meaningfully to a model?

(4%) Q7. Is there any noise in the dataset? If yes, what is the effect of noise on informative features, important features, and permutation importance, respectively?

(4%) Q8. If you trained a model without using highly informative features, important features, or permutation importance, what would you expect to happen to model accuracy?

III. Model-Specific Questions

(16%) Q9. Use bar charts of informative features, important features, and permutation importance to compare eight classification models.

> Suggested ranking (easiest to hardest to interpret feature importance):
>
> 1. Decision Tree
> 2. Logistic Regression

3. Random Forest
4. XGBoost
5. GaussianNB
6. MultinomialNB
7. SVC (linear easier than RBF)
8. K-NN

(10%) Q10. For each model, are the top mutual information features also top permutation features? Why or why not?

Dataset 2 descriptions (HW3-2.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 training data 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
1. 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.

2. 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 validation data are for the 15 days after the last date in the dataset.

3. stores.csv

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

4. 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.)

5. 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).

6. sample_submission.csv (for you saving the results and submitting to NTU Cool)

- A sample submission file in the correct format.

** 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.

(50%) Question:
Please use the best-performing model—whether tuned manually or automatically—to predict the results for the data in validations.csv. Output your predictions to a CSV file following the format of sample_submission.csv. Be sure to record and clearly explain all the steps and processes you performed, including any preprocessing, parameter tuning, model selection, and evaluation.

The format of the sample_submission.csv file is as follows.

id,	sales
3000888,	0.0
3000889,	0.0
…	…