Machine Learning – Assignment 2

Due Date: 31.5.23

The Data

This dataset is a relational set of files that describes customers' orders over time, from grocery delivery company named "ShufersalML".

Your goal is to predict which products will be in a user's next order, based on its past orders.

The dataset contains a sample of over 3 million grocery orders from more than 200,000 users. For each user, between 4 and 100 of their orders are provided, along with the sequence of products purchased in each order, the week and hour of day the order was placed, and a relative measure of time between orders.

The dataset is anonymized and contains several files that are associated with each entity, such as customers, products, orders, aisles, and departments:

- **aisles.csv**
- **departments.csv**
- **products.csv** - provides information on each product, such as its name, aisle ID, and department ID.
- **order_products__*.csv** files - specify which products were purchased in each order.
    - **order_products__prior.csv** - contains past orders of customers. This file should be used for feature engineering, which involves creating new features from the raw data that can be used to train your model.
    - **order_products__train_test.csv** – contains orders and products that should be used for train and test the model.
- **target.csv** – contains the labels for each order and product combination of the train and test samples.

- **orders.csv** - indicates to which set (prior, train, test) an order belongs, and extra details about the orders.

Section A (Data Exploration and Visualization) *10 pts*

Explore the data using tables, visualizations, and other relevant methods.

- Plots should have an informative main title, axis labels and a legend.
- For each plot or table, provide a short description of **key observations**. Make sure to only include content which would be **meaningful** for a "ShufersalML" manager.
- The visualizations should be detailed and cover all relevant aspects of the data.
- The visualizations should highlight any interesting patterns or trends that can be observed in the data.
- The goal of this section is to get insights on the data which may or may not be relevant for the following sections.

Section B (Data Pre-processing) *30 pts*

Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections.

- Perform feature engineering on the data, including three specific features provided by the project guidelines and at least six additional features of your own choosing.

  Explain why you chose these features and how they may improve model performance.

  Note: For this task you are required to use only the **prior samples** (across all data sets)!

  Create the following features:

  - Weekday Purchase Probability - This feature calculates the probability of a customer buying a certain product during a specific day of the week (calculate for each product).
  - Product Purchase Frequency – This feature represents the number of times a customer has purchased a certain product in the past.
  - Days Since Last Order - provides information on the time elapsed since the customer last bought a particular product in its past orders.
  - At least six other features of your own choosing.

- Apply at least one type of imputation (if needed), one of transformations, and one of exclusion (i.e., feature selection).
- Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it's needed.

Section C (Future Order Prediction) *25 pts*

Use at least **three** different machine learning models to predict the future order of each customer, according to the target.csv.

Predict the value of column "Was_In_Order" and summarize the prediction results into a list contain the products in the future order (the final prediction).

- When training a model to predict future orders, you can use all the data of prior samples as features, **except** for any features that don't provide relevant information for predicting future orders. (However, you can still use these irrelevant features to engineer new features that can be used in the model).
- The implementation must include parameter tuning.
- Report a suitable measure to evaluate the performance of each model and compare the results.

- Present the models' results in a plot.

Section D (Clustering) *25 pts*

- Apply at least **two** clustering algorithms on the prior data to cluster the different customers.

  Make sure that before clustering the customers you created features that describe their buying behavior (create additional features if needed).
- Use parameter tuning based on the algorithms you selected.
- Identify the most important features that contribute to the differences between the clusters. Discuss your findings and find a way to demonstrate **visually** what similarities the clusters may have.
- Use a method (of your own choice) to estimate the quality of the clusters you created with each clustering algorithm. Visualize the results according to the method you selected.
- Use at least **one** clustering algorithm on the prior data to cluster the different products. Discuss about the differences between the clusters. Make sure to have relevant features and add them if needed.

Presentation *10 pts*

Create a short presentation (no more than 6 slides) that includes interesting findings of your choice. 3-4 presentations will be chosen to be presented in front of the class. The goal is to learn from other students' work.

Section E (Clustering and Dimensions Reduction - Bonus) *10 pts*

- Reduce dimensions of the **customers** data (from section D) using PCA algorithm.

- Show which principal components explain the majority of the variance in data, using a plot. Identify the features that are most strongly represented in each component.
- Use the top principal components to perform clustering on the customers, using the same clustering algorithms as before.
- Visualize the clusters before and after PCA. Compare the results to the clustering performed without PCA. Are there any differences in the clusters obtained? If so, try to explain why these differences exist, and discuss how does using PCA affects the results of clustering.

Section F (Chi-Square test - Bonus) *10 pts*

Perform a Chi-Square test on the **train** dataset to determine the relationship between the "Reordered" feature and the binary label "Was_In_Order".
The Chi-Square test is a statistical method used to test the independence of two events and is commonly used in feature selection for classification tasks (we can determine whether to exclude a feature if it is independent from the label).
This test helps us determine whether the " Reordered" feature should be selected as a predictor for model training.
Define a null hypothesis (that the two variables are independent) and accept or reject the null hypothesis based on the Chi-Square value, with 95% confidence that alpha (level of significance) equals to 0.05.

Section G (Performance - Bonus) *5 pts*

Machine learning models that outperformed other students' models for the future orders predictions may get additional points as long as the non-standard methodology to obtain superior results is also explained.

- In order to get the bonus points, you may want to apply multiple performance measures to ensure that we can compare your

performance on an equal basis to other projects, and that you did not sacrifice performance in a specific measure to outperform in another.

Submission

- The assignment should be submitted in pairs (only one submission).
- You are required to submit two files including sections A-F. One in **.ipynb** format and one in **.html**. Both files should also include the program's outputs. In addition, you are required to upload a **pdf** file of the presentation you prepared.
- The files' names should be of the form: **ML_HW2_#ID1_#ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to one week. Later submissions will not be accepted.