**Data Mining and Visualization (83676)**

**Final Project**

**Due: 11/7/24**

**General**

In this project you will apply in Python the concepts you have learned in the course. The main tasks are to analyze the given dataset, apply preprocess techniques and get a prediction using classification methods.

The submission should include three files:

* A Python notebook file (.ipynb) that has the full code and with all the results displayed in the same order as in the report.
* A report (word or PDF) with the project description and explanations of the process and the results.
* A CSV file with predictions for the attached test dataset.

The grade evaluation takes into consideration:

* Techniques used: Did you select appropriate and diverse techniques and justify why?
* The process you followed: Is it correct (given the techniques you used), did you describe it well?
* Interpretation of results: Did you correctly understand and interpret the results you obtained?
* Quality of writeup: Did you present your work well, in an understandable and usable manner?
* The classification model quality: The test set predictions will be ranked based on accuracy relative to the class, this part will constitute 10 points of the grade (those ranked at the bottom will receive 0 points).

**Problem Description**

Hotels often face challenges with reservation management due to unexpected cancellations. These cancellations can lead to revenue loss and inefficient resource allocation. Accurately predicting reservation cancellations is essential for hotels to optimize occupancy rates and maximize profitability. In this project, students will analyze a dataset of hotel reservations to develop a model that can predict which reservations are likely to be canceled. This will enable hotels to implement more effective reservation strategies and improve overall operational efficiency.

**Data Attributes**

The data contains booking details, customer information, and reservation specifics.

* ID: reservation id
* n\_adults: number of adults
* 'n\_less\_12': number of children aged less than 12
* 'n\_more\_12': number of children aged more than 12
* 'weekend\_nights': number of weekend nights
* 'week\_nights': number of week nights
* 'board\_type'
* 'booked\_tour': indicates whether a tour was included in the reservation
* 'room\_type'
* 'lead\_time': number of days between the reservation date and the arrival date
* 'purchase\_type'
* 'repeated': indicates whether the reservation is a repeat reservation
* 'n\_p\_cacellation': number of previous reservations that were canceled by the customer prior to the current reservation
* 'n\_p\_not\_cacellation': number of previous reservations not canceled by the customer prior to the current reservation
* 'price'
* 'n\_requests': number of special requests made by the guest
* 'date': date of the reservation
* 'is\_canceled': target value, 0 – not canceled, 1 – canceled

**Instructions:**

You are given two datasets: ‘hotels\_train’ and ‘hotels\_test’. You will initially explore and analyze the training dataset using statistical methods and visualization tools, and perform preprocess. Afterward, you should train a classification model on the preprocessed dataset and make a prediction for the attached testing dataset.

You should implement the following sections and add more necessary actions, analytics, and visualizations to enrich your work. All the analytics should be displayed in the notebook, the report should include explanations with examples.

* Show the data information, e.g., types of attributes, the attributes values etc.
* Show the data statistics, e.g., distribution, skewness, median and more.
* Show and explain attributes correlations.
* Show and explain visualizations that present interesting insights from the data, e.g., identify relations, trends, the effect of an attribute on the target variable etc.
* Data cleaning - check for each one of the problems and take care of them properly, e.g., missing values, inconsistency etc.
* If necessary, add and/or delete attributes.
* Data transformation - apply the appropriate methods to the required attributes, e.g., normalization, discretization etc.
* Implement the pre-process on the attached test dataset and describe what adjustments you made.
* Split the data to train and validation set and use cross-validation method.
* Choose and explain the most appropriate evaluation metrics for this problem.
* Train at least three different classifiers and present evaluation metrics to compare between them and describe the results. Include at least one new classifier that wasn’t covered in class and explain it briefly.
* Perform hyperparameters tuning for each of the models you train. Describe how you chose the hyperparameters to calibrate and their range.
* Choose the best model, in terms of statistical significance, to apply to the attached test data set to generate predicted classification. Save the results in a CSV file.
* Incorporate visualizations whenever possible.

\*Students will lose points from their final mark on the assignment if they use internet published analysis of the dataset in the project.

**Good Luck!**