CP322 Machine Learning - Assignment 1
Due Date: Sep 30, 2024 at 11:59 PM

About Submission

When writing and submitting your assignments follow these requirements:

- You are expected to submit a single Notebook file for this assignment.
- Extensions for assignment are only granted for medical reasons with a doctor's note. Assignments submitted within 48 hours after the deadline will have their grade reduced by 50%. Submissions beyond 48 hours post-deadline cannot be accepted and will receive a grade of 0.
- Your assignment should be submitted online through the MyLearningSpace website. Email submission is not accepted.
- Please document your program carefully.

Before You Start

Setting-up Your Software Environment

This part simply requires you to setup the programming environment that we will be using for the remainder of the course. As stated in class, you may install the required software in your personal computer, in which case we suggest you carefully ensure that everything you install is up-to-date, which will help avoid compatibility issues when your assignments are graded.

Programming assignments will require the use of Python as well as additional Python packages. Most of the relevant software is a part of the SciPy [1]stack, a collection of Python-based open source software for mathematics, science, and engineering (which includes Python, NumPy, the SciPy library, Matplotlib, pandas, IPython, and scikit-learn). The Anaconda Python Distribution[2] is a free distribution for the SciPy stack that supports Linux, Mac, and Windows. Ensure that your machine has the following software installed:

- Python (An interactive, object-oriented, extensible programming language.)
- NumPy (A Python package for scientific computing.)
- SciPy (A Python package for mathematics, science, and engineering.)

[1] https://www.scipy.org/
[2] https://www.anaconda.com/distribution/

- Matplotlib (A Python package for 2D plotting.)
- pandas (A Python package for high-performance, easy-to-use data structures and data analysis tools.)
- IPython (An architecture for interactive computing with Python.)
- scikit-learn (A Python package for machine learning.)

Create Your First Notebook

To create a new IPython Notebook, you simply need to open the terminal *Jupiter Notebook*, and this will bring up the IPython web interface from which you can select *New Notebook*. Once you are finished, rename and save your assignment, and this will generate an .ipynb file.

Objective

The Hotel Booking Cancellation Prediction Dataset is designed for data scientists and machine learning enthusiasts interested in predicting hotel booking cancellations. This real-world dataset includes various details about hotel bookings, such as customer demographics, stay specifics, booking features, and pricing information.

The main goal of this assignment is to apply machine learning techniques to predict the likelihood of hotel booking cancellations. You will engage in data exploration to understand key influences on cancellations, preprocess data, implement decision tree classifiers, and evaluate their models' effectiveness. Through this assignment, you will gain practical experience in managing and analyzing data, building predictive models, and evaluating their performance in a real-world context.

1 Data Exploration (9 points)

1. Load the dataset and display the first 10 rows. How many features does the dataset contain?
2. Provide a statistical summary for the numerical features of the dataset. Which feature has the highest mean?
3. Check the dataset for any missing values. Are there any missing values across the dataset? If yes, list the columns with missing values and their counts.
4. Plot histograms for three numerical features ('number of adults', 'average price', 'lead time') to examine their distribution.
5. Use boxplots to examine whether there are outliers in the 'average price' feature. Explain your observations.
6. Compute the correlation matrix and create a heatmap with 'seaborn' [3] for numerical features.
7. Create a new feature 'total nights' as the sum of 'number of weekend nights' and 'number of week nights'. Show the first 5 rows of your merging result.

[3] `https://seaborn.pydata.org/generated/seaborn.heatmap.html`

2 Decision Tree (6 points)

1. Train a Decision Tree classifier on the 'booking status' using features 'number of adults', 'average price', 'total nights' (continuous features), as well as 'type of meal' and 'room type' (categorical features). Convert the categorical features into numerical format using one-hot encoding. Split the data into an 80-20 train-test set Split the data into 80-20 train-test sets.

2. Evaluate the Decision Tree model using accuracy and display the confusion matrix.

3. Determine which feature is the most important for making predictions. Display the importance of each feature.