

MSBA 315 Assignment 3: Hotel Booking Cancellation Prediction

The goal of this assignment is to **develop a machine learning system that can predict whether a hotel reservation will be cancelled or not**. Towards this goal, you must select the “*best*” *preprocessing techniques, features, model, and model parameters*. **Apply 5-fold cross-validation** to train and evaluate your models using the area under the ROC curve (AUC).

1. Exploratory Data Analysis (EDA) [10 pts]

Report the most relevant insights that might be useful to improve the AUC performance of the predictive model.

2. Baseline System [10 pts]

Start by creating a **simple** baseline system based on **logistic regression**. Select some features (which you think have a prediction power, might be based on your EDA), train and **cross-validate** your model.

- Report the **average AUC** with the **standard deviation**.

3. Preprocessing [20 pts]

Conduct the appropriate preprocessing techniques toward improving the average AUC performance of your baseline system (using the same five fold CV), using for instance:

- Replacing missing values
- Removing outliers
- Encoding categorical variables (large number of categories might require binning/grouping)
- Scaling: Normalization/standardization

Report:

- Your best/selected preprocessing techniques
- The average AUC with the standard deviation of your baseline model (after preprocessing).

4. Feature Selection [20 pts]

The goal of this step is to *simplify the system by dropping less important features* for the selected model without significantly *decreasing the AUC value*. To find the more (or less) important features:

- Apply the recursive feature elimination with **cross-validation** ([RFECV](#)) to select the number of features
- Based on the above two feature selection techniques, what are the top **five** features (most powerful predictors)? Does it make sense?
- Report the percentage reduction in feature size and the average AUC with the standard deviation for the selected feature subset using the same baseline model

5. Model Optimization and Selection [20 pts]

Based on the subset of features and the preprocessing techniques selected in the previous step, train and **optimize the parameters** of KNN and SVM (with an RBF Kernel – gamma and C) using 5-fold cross-validation.

For each model, report:

- The average AUC with the standard deviation of the best model parameters
- Plot the (five) ROC curves of the final best model

6. Model Optimization and Selection [10 pts]

- Which preprocessing techniques, features, and model do you select for deployment (operation)?
- What recommendations do you provide?
- If your boss give you an extra month to further improve this project, what would you do?

MSBA 315 Assignment 3: Hotel Booking Cancellation Prediction

Notebook Organization and Code Structure [5 pts]

- Good notebook structure, organization, comments, and coding style are expected (similar to the labs and also consult the **Rubric** for project's code on Moodle)

Deliverable [5 pts]:

Two files, one notebook (also *exported* as html) divided into the above 6 sections (and subsections of your choice) and named:

- **File 1:** uid_firstname_lastname_assign3.ipynb (wk47_wael_khreich_assign3.**ipynb**) [1 pt]
- **File 2:** uid_firstname_lastname_assign3.html (wk47_wael_khreich_assign3.**html**) [1 pt]
 - You can follow this [link](#) or this [link](#) to convert **ipynb** -> **html** from colab
- **Do not include irrelevant experiments and outputs** in your notebook, only the final and relevant code/results that are sufficient to answer the question (the more concise your notebook the better). Describe (in writing) additional experiments if needed. [3 pts]

Dataset:

Use the dataset included in this folder: "hotel_booking.csv". It is a subset of the dataset posted on Kaggle (the only difference is the "reservation_status" column is dropped). You can check the data description on Kaggle:

<https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>

The **target** value is: "**is_canceled**"

- Value indicating if the booking was canceled (1) or not (0)