MSBA 315 Assignment 5: Who is going to buy?

The goal is to develop a machine learning system that can accurately predict whether a user will buy a product, based on session events collected from users on an e-commerce website. Toward this goal, you must select the *"best" preprocessing techniques, features, model, and model parameters*. Use the ROC curve for visualization and the area under the ROC curve (AUC) for evaluation. Start by splitting your training data into 70% for training and 30% for validation and always report the AUC score on the training set (train) and the validation set (valid) to observe and avoid overfitting or underfitting.

Part I [10 pts]

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

a) [5%] Visualize the training data, identify outliers and missing values, analyze relationships, etc.
b) [5%] Summarize the **most relevant findings and insights** gained from your analysis.

2. Baseline System [10%]

To begin with, you can build a simple baseline system using logistic regression (LogReg) using basic preprocessing and a subset of the features (which you believe may have a strong predictive power based on your intuition and EDA). For instance, you could consider dropping time-related features and any rows with missing values. You can then use the remaining numerical and categorical features to train your model.

a) Plot the ROC curves and report the AUC values on train and valid.

3. Improved System [40%]

You can now **iteratively** refine your preprocessing techniques and features (e.g., by including all features, engineering new ones, and then selecting the most powerful ones) to improve the performance of your model on the validation set.

a) [15%] Report and **discuss** the preprocessing techniques that helped improve the model performance.
b) [20%] Report and **discuss** the most powerful features that helped improve the model performance.
c) [5%] Plot the ROC curves and report the AUC values on the training and validation of the best preprocessing techniques and features.

4. Model Optimization and Selection [30%]

Based on the subset of features and the preprocessing techniques selected in the previous steps, train and optimize the hyperparameters of **LogReg**, **KNN,** and **Naïve Bayes** (*Gaussian* and *Multinomial*) to further improve the AUC on valid. For each model, report:

a) [20%] Report (in a pandas dataframe) the best AUC values for each optimized model on train and valid.
b) [5%] Discuss which model would choose for operations.
c) [5%] Discuss any issues you faced during training or testing and what you did to overcome these issues.
d) [5%] *Extra credit* if you can make ***SVM*** compete with the other classifiers and discuss how.

Notebook Organization and Code Structure [10%]

- [80%] Assessed based on readability, structure, modularity, and efficiency as described in the **Rubric** for the project's code available on Moodle.
- [20%] Your notebook should run with **no error** (after modifying the path of the dataset)

Deliverable:

- Submit your notebook to moodle under assignment 5
- Notebook Name: **uid_firstname_lastname_sectX.assign5-1.ipynb**
  (wk47_wael_khreich_sect2_assign5-1.ipynb)

MSBA 315 Assignment 5: Who is going to buy?

Part II [5 pts]

After submitting the notebook of Part I to Moodle, Your next objective is to participate in a Kaggle competition to achieve the highest AUC performance possible using the same training set as in Part I. You have the freedom to choose and test any classifier or technique that you prefer. Your grade will be based on two factors:

a) [60%] your rank relative to your classmates in your section
b) [40%] your final notebook's quality based on the Rubric for the project's code available on Moodle.

Deliverable:

- Submit your notebook to moodle under assignment 5
- Notebook Name: **uid_firstname_lastname_sectX.assign5-2.ipynb (wk47_wael_khreich_sect2_assign5-2.ipynb)**

Data Description:

This dataset contains session events collected from users on an e-commerce website. A user could have one or multiple sessions. While browsing the site, and based on the previous event, the user might get a discount offer as an incentive to purchase (there are 4 types of offers as described below). Each row in the dataset represents the information recorded in the last event of a user in a session (before leaving or buying).

Feature Mapping and Description

col1	user_id	Unique user identifier
col2	session_id	Unique session identifier
col3	session_start_time	When a user landed on the website
col4	session_expiry_time	When the session is considered as expired (if the same user refreshes the browser he/she will be assigned a new session_id)
col5	event_time	The time of the current event in a session (from which the data is extracted)
col6	event_time_zone	The time zone of the current event
col7	event_type	The type of event that was done by the user (loading a page, changing the cart...)
col8	offer_decline_count	Number of times the user declined the offer received
col9	user_status	New Customer (NC) or Old Customer (OC)
col10	cart_quantity	Number of items in the user's cart
col11	cart_total	Total $ amount of the user's cart
col12	last_offer_type	The type of the last offer received (in the previous event of the session)
		'F': Fixed offer; get a $5 discount if you spend $50 or more
		'P': Discount varies depending on the product price a user is looking at
		'S': Discount varies based on the cart $ amount (ask the user to buy more things to get the offer)
		'C': Discount varies based on the cart $ amount (ask the user to checkout now to get the offer)
col13	last_reward_value	The discount in $ that a user will get after meeting the minimum spend (specified in col14)
col14	last_spend_value	The $ amount that a user should spend to collect the reward (specified in col13)
col15	offer_display_count	Number of offers received in the session (up to the current event)
col16	user_screen_size	The pixel resolution of the device's screen that the user is utilizing
col17	offer_acceptance_state	Can be on the following:
		ACCEPTED: if the user accepted the offer during the session (the offer will be fixed after acceptance)
		DECLINED: if the user explicitly declined the offer (clicked on the "no thanks" button)
		IGNORED: if the user did not act on the offer and kept browsing
col18	**converted**	Your **target** variable: 1 if a user purchased something and 0 otherwise