DAV 5400 Project 4 (Week 13) (100 Points)

Predictive Analysis using scikit-learn

<u>You may work in small groups of no more than three (3) people for this project. **</u>**

As data analysts, we're often tasked with taking data in one form and transforming it for easier downstream analysis. In this Project, you'll use what you've learned in the course to prepare data for predictive analysis and then construct a predictive model using tools available within the **`scikit-learn`** library.

Many industries are interested in broadening their relationships with existing customers. To that end, companies will often attempt to sell additional products to their existing customers. For example, if you have a checking account with a particular bank, they will likely try to also sell you an auto loan or a mortgage, etc. You've been tasked by a large company with the development of a model that can predict whether or not a given existing customer is likely to purchase an additional product from the company. The company plans to use the output of such a model in an attempt to improve its customer retention and sales practices.

The data set you will be working with is comprised of more than 14,000 observations of 1 response/dependent variable (which indicates whether or not the new product was purchased) and 11 explanatory/independent variables. The company gathered data about customers to whom they offered the new product. You are given information about whether they did or did not sign up for the new product, together with some customer information and information about their buying behavior of two other products (Product A and Product B). A data dictionary for the dataset is provided below.

Attribute	**Description**
ID	Unique customer identifier
TARGET	Indicator of customer buying the new product (N = no, Y = yes)
Loyalty	Customer loyalty level, from low to high (0 to 3), 99 = unclassified
Age	Customer age in years
City	Unique code per city (where the customer resides)
Age_p	Age of customer's partner in years
LOR_m	Length of customer's relationship with company (in months)
Prod_A	Customer previously bought Product A (0=no, 1=yes)
Type_A	Type of product A
Turnover_A	Amount of money customer spent on Product A
Prod_B	Customer previously bought Product B (0=no, 1=yes)
Type_B	Type of product B
Turnover_B	Amount of money customer spent on Product B

In this Project, you will use **`scikit-learn`** to answer the question:

"***Which attribute or attributes are the best predictors of whether a customer will purchase an additional product?"***

The work you will need to do for this Project can be separated into two distinct phases: all of the work required for **Phase I** can be completed using the Python and Pandas skills you developed through Week 10 of this class (i.e., without any prior **`scikit-learn`** knowledge), while Phase II will require the use of **`scikit-learn`** to assess the predictive qualities of the attributes contained within your DataFrame.

Phase I: **Data Acquisition, Data Preparation & Exploratory Data Analysis (60 Points)**

- Construct a professional-quality introductory narrative that explains the purpose and objectives of your work.
- Load the provided dataset to you Github repository and read it from your Github repository into Jupyter Notebook.
- Study the dataset and the associated description of the data (i.e. "data dictionary").
- Create a pandas DataFrame **with a subset of the columns in the dataset**. You should include the **TARGET** column, which indicates **whether or not a customer purchased an additional product**, the **Age, Type_A, Type_B, lor_M** columns, and **at least two other columns** of your choosing.
- Convert the **TARGET** indicators in the first column to digits: for example, the "N" might become 0 and "Y" might become 1.
- Perform exploratory data analysis: show the distribution of data for each of the columns you selected (including the **Age, Type_A, Type_B, lor_M** columns), and show plots for TARGET vs. Age, TARGET vs. Type_A, TARGET vs. Type_B, TARGET vs. lor_M, as well as the other columns that you selected. It is up to you to decide which types of plots to use for these tasks. Include text describing your EDA findings.
- Include some text describing your preliminary conclusions about whether any of the other columns you've included in your subset *(i.e., aside from the TARGET indicator)* could be helpful in predicting whether a specific customer is likely to purchase an additional product.
- For both the **Type_A** and **Type_B** columns, create a set of dummy variables. This is necessary because your downstream processing in Project 4 using **scikit-learn** requires that categorical data values be converted to binary indicator variables. See the pandas **get_dummies()** method for one possible approach to doing this.

Phase II: **Build Predictive Models (40 Points)**

- Start with the data (including the dummy variables) in the pandas DataFrame that you constructed in Phase I.
- Use **scikit-learn** to construct appropriate regression model(s) and use the information provided by those models to determine which of the predictor columns that you selected (including **Age, Type_A, Type_B, lor_M**) most accurately predicts whether or not a customer is likely to purchase an additional product. How you go about doing this with **scikit-learn** is up to you as a practitioner of data analytics.
- Clearly state your conclusions along with any recommendations for further analysis.

*****HINT***** : If you understand the process used in the DataSchool videos on Machine Learning with scikit-learn to predict iris species from four predictor variables, you should be able to apply what you've learned to complete this Project.

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Save all of your work for this project within **a single Jupyter Notebook** and submit it via the Project 4 page within Canvas. Be sure to save your Notebook using the following nomenclature : **first initial_last name_Project4**" (e.g., J_Smith_Project4).). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***