DAV 6150 Module 7 Assignment

Logistic Regression

***** You may work in small groups of no more than four (4) people for this Assignment *****

Like many industries, the insurance industry is always interested in broadening its relationships with existing customers. To that end, insurance companies will often attempt to sell additional products to their existing customers. For example, if you have a homeowner's policy with a particular insurance company, they will likely try to also sell you an auto insurance policy, or perhaps a water damage supplemental policy to your homeowner's policy, etc. You've been tasked by a large insurance company with the development of a model that can predict whether or not a given existing customer is likely to purchase an additional insurance product from the company. The insurance 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 using is sourced from a Kaggle contribution:

- https://www.kaggle.com/rluyck/insurance-company?select=Customer_data.csv

The data set is comprised of more than 14,000 observations of 1 response/dependent variable (which indicates whether or not the new insurance product was purchased) and 14 explanatory/independent variables. The insurance 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. A data dictionary for the dataset is provided below.

Attribute	Description
ID	Unique customer identifier
TARGET	Indicator of buying the new product (0 = no, 1= yes)
Loyalty	Loyalty level, from low to high (0 to 3), 99 = unclassified
Age	Age in years
City	Unique code per city
Age_p	Age of partner in years
LOR	Length of relationship in years
LOR_m	Length of relationship in months
Prod_A	Bought Product A (0=no, 1=yes)
Type_A	Type of product A
Turnover_A	Amount of money spent on Product A
Prod_B	Bought Product B (0=no, 1=yes)
Type_B	Type of product B
Turnover_B	Amount of money spent on Product B
Contract	Type of contract

Your task for this Assignment is to construct and compare/contrast a series of **binary logistic regression models** (after completing the necessary EDA and data prep work) that predict whether or not a given insurance company customer is likely to purchase an additional insurance product. The response variable you will be modeling is the data set's "**TARGET**" attribute, which indicates whether or not a given insurance company customer purchased an additional insurance product. It is up to you as the data science practitioner to determine which features should be included in these models. Your work should include EDA, data

preparation (including transforms as needed), feature selection, and a thorough evaluation of model performance metrics. Get started on the Assignment as follows:

1) Load the provided **M7_Data.csv** file to your DAV 6150 Github Repository.
2) Then, using a Jupyter Notebook, read the data set from your Github repository and load it into a Pandas dataframe.
3) Perform EDA work as necessary.
4) Perform any required data preparation work, including any feature engineering adjustments you deem necessary for your work.
5) Apply your knowledge of feature selection and/or dimensionality reduction techniques to identify explanatory variables for inclusion within your models. You may select the features manually via the application of domain knowledge, use forward or backward selection, or use a different feature selection method (e.g., decision trees, etc.). It is up to you as the data science practitioner to decide upon the most appropriate feature selection and/or dimensionality reduction techniques to be used with the data set.
6) After splitting the data into training and testing subsets, use the training subset to construct **at least three different binomial logistic regression models** using different combinations of explanatory variables (or the same variables if they have been transformed via different transformation methods).
7) After training your various models, decide how you will select the "best" regression model from those you have constructed. For example, are you willing to select a model with slightly lower performance if it is easier to interpret or less complicated to implement? What metrics will you use to compare/contrast your models? Evaluate the performance of your models via cross validation using the training data set. Then apply your preferred model to the testing subset and assess how well it performs on that previously unseen data.

Your deliverable for this Assignment is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: Summarize the problem + explain the steps you plan to take to address the problem
2) **Exploratory Data Analysis (20 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis, including any preliminary predictive inferences. This section should include any Python code used for the EDA.
3) **Data Preparation (10 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering techniques you have applied to the data set. This section should include any Python code used for Data Preparation.
4) **Prepped Data Review (5 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.
5) **Regression Modeling (40 Points)**: Explain + present your regression modeling work, including your feature selection work + interpretation of the coefficients your models are generating. Do they make

sense intuitively? If so, why? If not, why not? Comment on the magnitude and direction of the coefficients + whether they are similar from model to model.

6) **Select Models (15 Points)**: Explain your model selection criteria. Identify your preferred model. Compare / contrast its performance with that of your other models. Discuss why you've selected that specific model as your preferred model. Apply your preferred model to the testing subset and discuss your results. Did your preferred model perform as well as expected?

7) **Conclusions (5 Points)**

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.

Upload your Jupyter Notebook within the provided M7 Assignment Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_M7_assn**" (e.g., J_Smith_M7_assn_). ***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.***